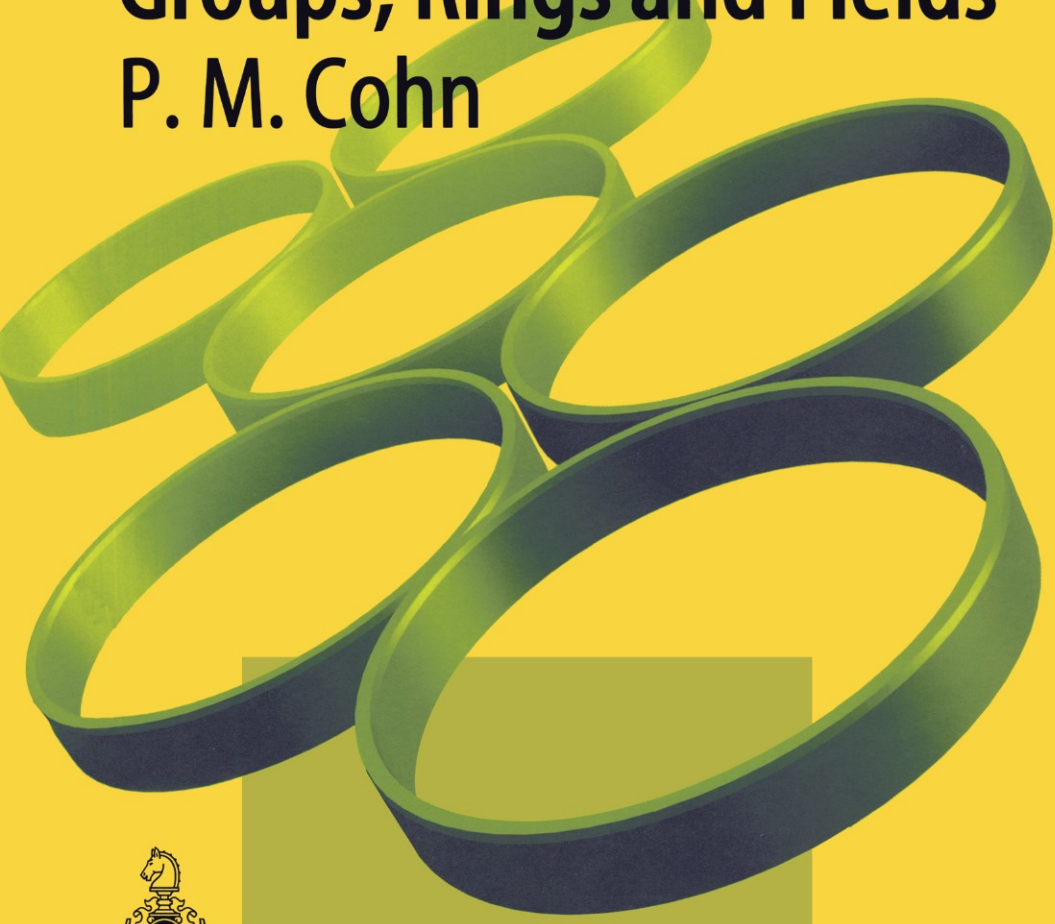
Basic Algebra

Groups, Rings and Fields

P. M. Cohn



Springer

Basic Algebra

P.M. Cohn

Basic Algebra

Groups, Rings and Fields

Springer

P.M. Cohn, MA, PhD, FRS
Department of Mathematics, University College London,
Gower Street, London WC1E 6BT, UK

British Library Cataloguing in Publication Data
Cohn, P. M. (Paul Moritz)
Basic algebra: groups, rings and fields
I. Algebra 2. Rings (Algebra) 3. Algebraic fields
I. Title
512
ISBN 978-1-4471-1060-6

Library of Congress Cataloging-in-Publication Data
Cohn, P.M. (Paul Moritz)
Basic algebra: groups, rings, and fields/P.M. Cohn.
p. cm.
Includes bibliographical references and indexes.
ISBN 978-1-4471-1060-6 ISBN 978-0-85729-428-9 (eBook)
DOI 10.1007/978-0-85729-428-9
1. Algebra. I. Title.

QA154.3.C64 2002
512–dc21 2002070686

springeronline.com

Originally published by Springer-Verlag London Berlin Heidelberg in 2003
Softcover reprint of the hardcover 1st edition 2003

2nd printing 2005

Typesetting by BC Typesetting, Bristol BS31 1NZ
12/3830-54321 Printed on acid-free paper SPIN 11329916

Contents

Preface ix
Conventions on Terminology xi

1. Sets
   - 1.1 Finite, Countable and Uncountable Sets 1
   - 1.2 Zorn's Lemma and Well-ordered Sets 8
   - 1.3 Graphs 15

2. Groups
   - 2.1 Definition and Basic Properties 25
   - 2.2 Permutation Groups 32
   - 2.3 The Isomorphism Theorems 34
   - 2.4 Soluble and Nilpotent Groups 37
   - 2.5 Commutators 42
   - 2.6 The Frattini Subgroup and the Fitting Subgroup 46

3. Lattices and Categories
   - 3.1 Definitions; Modular and Distributive Lattices 51
   - 3.2 Chain Conditions 60
   - 3.3 Categories 65
   - 3.4 Boolean Algebras 70

4. Rings and Modules
   - 4.1 The Definitions Recalled 79
   - 4.2 The Category of Modules over a Ring 84
   - 4.3 Semisimple Modules 91
   - 4.4 Matrix Rings 96
   - 4.5 Direct Products of Rings 101
   - 4.6 Free Modules 105
   - 4.7 Projective and Injective Modules 110
   - 4.8 The Tensor Product of Modules 117
   - 4.9 Duality of Finite Abelian Groups 125

5. Algebras
5.1 Algebras; Definition and Examples 131
5.2 The Wedderburn Structure Theorems 137
5.3 The Radical 141
5.4 The Tensor Product of Algebras 146
5.5 The Regular Representation; Norm and Trace 153
5.6 Möbius Functions 157

6. Multilinear Algebra
6.1 Graded Algebras 165
6.2 Free Algebras and Tensor Algebras 168
6.3 The Hilbert Series of a Graded Ring or Module 173
6.4 The Exterior Algebra on a Module 179

7. Field Theory
7.1 Fields and their Extensions 189
7.2 Splitting Fields 195
7.3 The Algebraic Closure of a Field 200
7.4 Separability 203
7.5 Automorphisms of Field Extensions 206
7.6 The Fundamental Theorem of Galois Theory 211
7.7 Roots of Unity 217
7.8 Finite Fields 223
7.9 Primitive Elements; Norm and Trace 227
7.10 Galois Theory of Equations 232
7.11 The Solution of Equations by Radicals 238

8. Quadratic Forms and Ordered Fields
8.1 Inner Product Spaces 249
8.2 Orthogonal Sums and Diagonalization 252
8.3 The Orthogonal Group of a Space 256
8.4 The Clifford Algebra and the Spinor Norm 259
8.5 Witt's Cancellation Theorem and the Witt Group of a Field 268
8.6 Ordered Fields 272
8.7 The Field of Real Numbers 275
8.8 Formally Real Fields 279
8.9 The Witt Ring of a Field 291
8.10 The Symplectic Group 298
8.11 Quadratic Forms in Characteristic Two 301

9. Valuation Theory
9.1 Divisibility and Valuations 307
9.2 Absolute Values 312
9.3 The p-adic Numbers 322
9.4 Integral Elements 331
9.5 Extension of Valuations 336

10. Commutative Rings
10.1 Operations on Ideals 347
10.2 Prime Ideals and Factorization 349
10.3 Localization 354
10.4 Noetherian Rings 361
10.5 Dedekind Domains 362
10.6 Modules over Dedekind Domains 371
10.7 Algebraic Equations 376
10.8 The Primary Decomposition 380
10.9 Dimension 386
10.10 The Hilbert Nullstellensatz 391

11. Infinite Field Extensions
11.1 Abstract Dependence Relations 397
11.2 Algebraic Dependence 402
11.3 Simple Transcendental Extensions 405
11.4 Separable and p-radical Extensions 409
11.5 Derivations 414
11.6 Linearly Disjoint Extensions 418
11.7 Composites of Fields 427
11.8 Infinite Algebraic Extensions 431
11.9 Galois Descent 437
11.10 Kummer Extensions 441

Bibliography 449
List of Notations 453
Author Index 457
Subject Index 459

Preface

Much of the second and third year undergraduate course in mathematics (as well as some graduate work) was covered by Volumes 2 and 3 of my book on algebra, now out of print.[1] So I was very pleased when Springer Verlag offered to bring out a new version of these volumes. The present book is based on both these volumes, complemented by the definitions and basic facts on groups and rings. Thus the volume is addressed to students who have some knowledge of linear algebra and who have met groups and fields, though all the essential facts are recalled here. My overall aim has been to present as many of the important results in algebra as would conveniently fit into one volume. It is my hope to collect the remaining parts of Volumes 2 and 3 into a second book, more oriented towards applications.[2]

Apart from chapters on groups (Chapter 2), rings and modules (Chapters 4, 5 and 6) and fields (Chapters 7 and 11), a number of concepts are treated that are less central but nevertheless have many uses. Chapter 1, on set theory, deals with countable and well-ordered sets, as well as Zorn's lemma and a brief section on graphs. Chapter 3 introduces lattices and categories, both concepts that form an important part of the language of modern algebra. The general theory of quadratic forms has many links with ordered fields, which are developed in Chapter 8. Chapters 9 and 10 are devoted to valuation theory and commutative rings, a subject that has gained in importance through its use in algebraic geometry.

On a first encounter some readers may find the style of this book somewhat concise, but they should bear in mind that mathematical texts are best read with paper and pencil, to work out the full consequences of what is being said and to check examples. The matter has been well put by Einstein, who said: "Everything should be explained as far as possible but no further." There are numerous exercises throughout, with occasional hints (but no solutions), and some historical remarks.

My thanks are due to the staff of Springer Verlag for the efficient way they have produced this volume.

University College London
June 2002

P.M. Cohn

1 *Algebra*, Vol. 2 (2nd edn, 1989) and Vol. 3 (2nd edn, 1991), Wiley and Sons.

2 *Further Algebra and Applications*, Springer Verlag, London (2003). Referred to in the text as FA.

Conventions on Terminology

We assume that our readers are acquainted with the notion of a set (and even with groups and rings, though their definitions will be recalled in Chapters 2, 4). They will have seen notations such as $x \in S$ (x is a member of S), $S' \subseteq S$ or $S \supseteq S'$ (S' is a subset of S) and $T \subset S$ or $S \supset T$ (T is a proper subset of S) and $\varnothing$ for the empty set. For any propositions P, Q we write '$P \Rightarrow Q$' or '$Q \Leftarrow P$' to indicate that P implies Q, and '$P \Leftrightarrow Q$' to mean '$P \Rightarrow Q$ and $Q \Rightarrow P$', i.e. that P is equivalent to Q.

A property (of members of a set S) is said to hold for *almost all* members of S if it holds for all but a finite number of members of S. If T is a subset of S, its complement in S will be denoted by $S\backslash T$. This notation is also used occasionally for the left coset space (see Section 2.1); the risk of confusion is small.

We can list the elements of a set S by indexing them, e.g. if S is finite, with n elements, we can write $S = \{x_1, x_2, \ldots, x_n\}$; we also write $|S| = n$. More generally, any set can be indexed by a suitable indexing set: $S = \{x_\lambda\}_{\lambda \in I}$, where I is the indexing set. A set in this form is often called a family indexed by I; it is in effect prescribing a mapping from I to S. This mapping is generally not assumed to be injective, thus x_λ may equal x_μ even if $\lambda \neq \mu$.

All mappings between sets are as a rule written on the right, so that fg means: first f, then g. If $f : S \to T$, i.e. f is a mapping from S to T and S' is a subset of S, then the restriction of f to S' is denoted by $f|S'$. A mapping $f : S \to T$ is called *injective* or *one–one* if different members of S have different images, *surjective* or *onto* if every member of T is an image of some member of S, and *bijective* if it is both injective and surjective. Mappings are often arranged as diagrams (see Section 4.2); a diagram is *commutative* if the different ways of going from one point to another along the arrows give the same result.

Frequently a two-index expression $f(i, j)$ is equal to 1 if $i = j$ and 0 otherwise. This is indicated by using the *Kronecker symbol* δ_{ij}; thus $f(i, j) = \delta_{ij}$.

A set S is *partially ordered*, often just called *ordered*, if there is a binary relation $\leq$, called a *partial ordering*, defined on S with the properties:

O.1 $x \leq x$ for all $x \in S$ (reflexive),
O.2 $x \leq y, y \leq z \Rightarrow x \leq z$ for all $x, y, z \in S$ (transitive),
O.3 $x \leq y, y \leq x \Rightarrow x = y$ for all $x, y \in S$ (antisymmetric).

If only **O.1** and **O.2** hold, we speak of a *preordering*.

The ordering is *total* if any two elements are comparable, i.e. $x \leq y$ or $y \leq x$ for any $x, y \in S$. If '$\leq$' is a partial ordering on a set S, we shall write '$x < y$' (x is *strictly*

less than y) to mean '$x \leq y$ and $x \neq y$', and we write $x \geq y$, $x > y$ for $y \leq x$, $y < x$ respectively. As is easily verified, the opposite ordering '$\geq$' again satisfies **O.1–O.3** and so is again a partial ordering. Thus any general statement about ordered sets has a dual, which is obtained by interpreting the original statement for the oppositely ordered set. This principle can often be used to shorten proofs.

A binary relation $\sim$ on a set S is called an *equivalence relation* if it is reflexive, transitive and *symmetric,* i.e. $x \sim y \Rightarrow y \sim x$, for all $x, y \in S$. For example, equality is an equivalence relation. Given an equivalence on S, we can list all members equivalent to a given one together in a class, and in this way obtain a *partition* of S into a collection of disjoint subsets, the *equivalence classes* or *blocks.* The set of equivalence classes is denoted by $S/_{\sim}$ and is called the *quotient set* of S by the equivalence $\sim$.

Given sets S, T, their *Cartesian* or *direct product,* denoted by $S \times T$, is the set of pairs (x, y), where $x \in S$, $y \in T$. If S, T are any ordered sets, their direct product can again be ordered by writing $(x, y) \leq (x', y')$ to mean: $x \leq x'$ or $x = x'$ and $y \leq y'$. This is easily verified to be an ordering, called the *lexicographic ordering*; it is a total ordering whenever both S and T are totally ordered.

References to the bibliography are by the name of the author and the date. In each section all the results are numbered consecutively, e.g. in Section 4.7 we have Theorem 4.7.1, Lemma 4.7.2, Proposition 4.7.3. We shall also use iff as an abbreviation for 'if and only if' and ■ indicates the end (or absence) of a proof. Many exercises are provided with hints, and the harder ones are starred.

1 Sets

Much of algebra can be done using only very little set theory; all that is needed is a means of comparing infinite sets, and the axiom of choice in the form of Zorn's lemma. These topics occupy Sections 1.1 and 1.2. They are followed in Section 1.3 by an introduction to graph theory. This is an extensive theory with many applications in algebra and elsewhere; all we shall do here is to present a few basic results, some of which will be used later, which convey the flavour of the topic.

1.1 Finite, Countable and Uncountable Sets

Most of our readers will have met sets before; a *set* for us is a collection of objects, its members or *elements*. These elements may themselves be sets; of course one has to be careful to avoid situations like Russell's paradox: 'the set Ω of all sets that are not members of themselves'; this quickly leads to a contradiction when one asks if $\Omega \in \Omega$. There are several ways of resolving this paradox, but they will not concern us here; all that is needed is some care in forming 'large' sets.

Given two sets, we may wish to compare them for size, i.e. the number of elements in each. We can use the natural numbers to count the members, but this may not be necessary. When Man Friday wanted to tell Robinson Crusoe that he had seen a boat with 17 men in it, he did this by exhibiting another 17-element set, and he could do this without being able to count up to 17. Even for a fully numerate person it may be easier to compare two sets rather than to count each; e.g. in a full lecture room a brief glance may suffice to convince us that there are as many people as seats. This suggests that it may be easier to determine when two sets have the same 'number of elements' than to find that number. Let us call two sets *equipotent* if there is a bijection (i.e. a one–one correspondence) between them. This relation of equipotence is an equivalence relation on any given collection of sets; here we avoid talking about the collection of *all* sets, as that would bring us dangerously close to the paradox mentioned above.

A set S is said to be *finite*, of *cardinal* n, if S is equipotent to the set $\{1, 2, \ldots, n\}$ consisting of the natural numbers from 1 to n. By convention the empty set, having no elements, is reckoned among the finite sets; its cardinal is 0 and it is denoted by $\varnothing$.

It is clear that two finite sets are equipotent if they have the same cardinal, and this may be regarded as the basis of counting. It is also true that sets of different finite cardinalities are not equipotent. This may seem intuitively obvious; we shall assume it here and defer to FA its derivation from the axioms for the natural numbers. More generally, we shall assume that for any natural numbers m, n, if there is an injective mapping from $\{1, 2, \ldots, m\}$ to $\{1, 2, \ldots, n\}$, then $m \leq n$. Let us abbreviate $\{1, 2, \ldots, n\}$ by $[n]$, for any $n \in \mathbf{N}$. It follows that if there is a bijection between $[m]$ and $[n]$, then $m \leq n$ and $n \leq m$, hence $m = n$. Thus for any finite set, the natural number which indicates its cardinal is uniquely determined. The contrapositive form of the above assertion states that if $m > n$, then there can be no injective mapping from $[m]$ to $[n]$. A more illuminating way of expressing this observation is Dirichlet's celebrated

Box Principle (Schubfachprinzip). *If $n + 1$ objects are distributed over n boxes, then some box must contain more than one of the objects.*

Although intuitively obvious, this principle is of great use in number theory and elsewhere.

Having given a formal definition of finite sets, we now define a set to be *infinite* if it is not finite. Until relatively recent times the notion of 'infinity' was surrounded by a good deal of mystery and uncertainty, even in mathematics. Thus towards the middle of the 19th century, Bernard Bolzano propounded as a paradox the fact that (in modern terms) an infinite set might be equipotent to a proper subset of itself. A closer study reveals the fact that every infinite set has this property, and this has even been taken as the basis of a definition of infinite sets; it certainly no longer seems a paradox. The work of Georg Cantor, Richard Dedekind and others from 1870 onwards has dispelled most of the uncertainties, and though mysteries remain, they will not hamper us in the relatively straightforward use we shall make of the theory.

In order to extend the notion of counting to infinite sets, we associate with every set X, finite or not, an object $|X|$ called its *cardinal* or cardinal *number*, defined in such a way that two sets have the same cardinal iff they are equipotent. Such a definition is possible because, as we have seen, equipotence is an equivalence relation on any collection of sets.

A non-empty finite set has as its cardinal a natural number; the empty set has cardinal 0. All other sets are infinite; their cardinals are said to be *transfinite* or *infinite*. In particular, the set $\mathbf{N}$ of all natural numbers is infinite; its cardinal is denoted by $\aleph_0$. The letter aleph, $\aleph$, the first of the Hebrew alphabet, is customarily used for infinite cardinal numbers. A set of cardinal $\aleph_0$ is also said to be *countable* (or *enumerable*); thus A is countable iff there is a bijection from $\mathbf{N}$ to A. If a set A is countable, it can be written in the form

$$A = \{a_1, a_2, a_3, \ldots\}, \tag{1.1.1}$$

where the a_i are distinct. Such a representation of A is called an *enumeration* of A, and a proof that a set is countable will often consist in giving an enumeration. Sometimes the term 'enumeration' is used for a set written as in (1.1.1) even if the a_i

are not all distinct; in that case we can always produce a strict enumeration by going through the sequence and omitting all repetitions. The set so obtained is finite or countable.

Many sets formed from countable sets are again countable, as our first result shows:

Theorem 1.1.1. *Any subset and any quotient of a countable set is countable or finite. If A and B are countable sets, then the union $A \cup B$ and Cartesian product $A \times B$ are again countable; more generally, the Cartesian product of any finite number of countable sets is countable. Further, a countable union of countable sets is countable and the collection of all finite subsets of a countable set is countable.*

We recall that a *quotient set* of A is the set of all blocks, i.e. equivalence classes, of some equivalence on A.

Proof. Any countable set A may be taken in the form (1.1.1); if A' is a subset, we go through the sequence $a_1, a_2, \ldots$ of elements of A and omit all terms not in A' to obtain an enumeration of A'. If A'' is a quotient set, and $x \mapsto \bar{x}$ is the natural mapping from A to A'', then $\{\bar{a}_1, \bar{a}_2, \ldots\}$ is an enumeration of A'', possibly with repetitions; hence A'' is countable (or finite).

Next let A be given by (1.1.1) and let $B = \{b_1, b_2, \ldots\}$; then $A \cup B$ may be enumerated as $\{a_1, b_1, a_2, b_2, \ldots\}$, where repetitions (which will occur if $A \cap B \neq \varnothing$) may be discarded. Similarly we can enumerate $A \times B$ as $\{(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_1, b_3), (a_2, b_2), (a_3, b_1), \ldots\}$ by writing $A \times B$ as a square table and going along the finite diagonals. Now the result for a product of r countable sets follows by induction on r. If we have a countable family $\{A_n\}$ of countable sets, say $A_n = \{a_{ni}\}$, then we can enumerate the union $\cup A = \{a_{ni} | n, i \in \mathbf{N}\}$ by writing the elements a_{ni} as a matrix and going again along the diagonals.

Finally let A be any countable set and denote by A_r for $r = 1, 2, \ldots$ the set of all r-element subsets of A. Clearly A_r is countable, for it may be mapped into the Cartesian power A^r by the rule

$$\{a_{i_1}, \ldots, a_{i_r}\} \mapsto (a_{j_1}, \ldots, a_{j_r}),$$

where $j_1, \ldots, j_r$ is the sequence $i_1, \ldots, i_r$ arranged in ascending order. This provides a bijection of A_r with a subset of A^r, and it follows that A_r is countable. Now the earlier proof shows that the union $\cup A_r$ is countable, and adding $\varnothing$ as a further member we still have a countable set. ∎

With the help of this result many sets can be proved to be countable which do not at first sight appear to be so. Thus the set $\mathbf{Z}$ of all integers can be written as a union of $\mathbf{N} = \{1, 2, \ldots\}$ and $\mathbf{N}' = \{0, -1, -2, \ldots\}$; both $\mathbf{N}$ and $\mathbf{N}'$ are countable hence so is $\mathbf{Z}$. The set $\mathbf{Q}_+$ of all positive rational numbers is countable, as the image of $\mathbf{N}^2$ under the mapping $(a, b) \mapsto ab^{-1}$. Now $\mathbf{Q}$ itself can be written as the union of the set of all positive rational numbers, the negative rational numbers and 0; therefore $\mathbf{Q}$ is countable. The set of all algebraic numbers (see Section 7.1 below) is countable:

for a given degree n, the set of all monic equations of degree n over $\mathbf{Q}$ is equipotent to $\mathbf{Q}^n$, if we map

$$(a_1, \ldots, a_n) \mapsto x^n + a_1 x^{n-1} + \ldots + a_n = 0.$$

Each equation has at most n complex roots, so the set S_n of all roots of equations of degree n is countable, and now the set of all algebraic numbers is $S_1 \cup S_2 \cup \ldots$, which is again countable.

At this point a newcomer might be forgiven for thinking that perhaps every infinite set is countable. If that were so, there would of course be no need for an elaborate theory of cardinal numbers. In fact the existence of uncountable sets is one of the key results of Cantor's theory, and we shall soon meet examples of such sets.

Our next task is to extend the natural order on $\mathbf{N}$ to cardinal numbers. If α, β are any cardinals, let A, B be sets such that $|A| = \alpha$, $|B| = \beta$. We shall write $\alpha \leq \beta$ whenever there is an injective mapping from A to B. Whether such a mapping exists clearly depends only on α, β and not on A, B themselves, so the notation is justified. Further, $\alpha \leq \alpha$ holds for all α, because the identity mapping on A is injective, and since the composition of two injections is an injection, it follows that $\alpha \leq \beta$, $\beta \leq \gamma$ implies $\alpha \leq \gamma$. Thus we have a preordering; this will in fact turn out to be a total ordering, but for the moment we content ourselves with proving that it is an ordering, i.e. that '$\leq$' is antisymmetric. In terms of sets we must establish

Theorem 1.1.2 (Schröder–Bernstein theorem). *Let A, B be any sets and $f : A \to B$, $g : B \to A$ be any injective mappings. Then there is a bijection $h : A \to B$.*

Proof. By alternating applications of f and g we produce an infinite sequence of successive images starting from $a \in A : a, af, afg, afgf, \ldots$. Further, each element $a \in A$ is the image of at most one element of B under g, which may be written ag^{-1}, and each $b \in B$ is the image of at most one element bf^{-1} of A under f, so from $a \in A$ we obtain a sequence of inverse images which may or may not break off: $ag^{-1}, ag^{-1}f^{-1}, \ldots$. If we trace a given element $a \in A$ as far back as possible we find one of three cases: (i) there is a first 'ancestor' in A, i.e. $a_0 \in A \backslash Bg$, such that $a = a_0(fg)^n$ for some $n \geq 0$; (ii) there is a first ancestor in B, i.e. $b_0 \in B \backslash Af$, such that $a = b_0(gf)^n g$ for some $n \geq 0$; (iii) the sequence of inverse images continues indefinitely.

Each element of A comes under one of these headings, and likewise each element of B. Thus A is partitioned into three subsets A_1, A_2, A_3; similarly B is partitioned into $B_1 = A_1 f, B_2 = A_2 g^{-1}$ and $B_3 = A_3 f = A_3 g^{-1}$. It is clear that the restriction of f to A_1 is a bijection between A_1 and B_1, for each element of B_1 comes from one element of A_1. For the same reason the restriction of g to B_2 provides a bijection between B_2 and A_2, and we can use either f restricted to A_3 or g restricted to B_3 to obtain a bijection between A_3 and B_3. Thus we have found a bijection between A_i and B_i $(i = 1, 2, 3)$ and putting these together we obtain the desired bijection between A and B. ∎

This proof is essentially due to Gyula König (in 1906).

The sum and product of cardinals may be defined as follows. Let α, β be any cardinals, say $\alpha = |A|, \beta = |B|$, and assume that $A \cap B = \varnothing$. Then it is easily seen that $|A \cup B|$ depends only on α, β, not on A, B and we may define

$$\alpha + \beta = |A \cup B|.$$

Similarly we put

$$\alpha\beta = |A \times B|.$$

It is easy to verify that these operations satisfy the commutative and associative laws, and a distributive law, as in the case of the natural numbers. Moreover, for finite cardinals these operations agree with the usual operations of addition and multiplication. On the other hand, the cancellation law does not hold, thus we may have $\alpha + \beta = \alpha' + \beta$ or $\alpha\beta = \alpha'\beta$ for $\alpha \neq \alpha'$, and there is nothing corresponding to subtraction or division. In fact, it can be shown that if $\alpha, \beta \neq 0$ and at least one of α, β is infinite, then

$$\alpha + \beta = \alpha\beta = \max\{\alpha, \beta\}. \tag{1.1.2}$$

For any cardinals α, β we define β^α as $|B^A|$, where A, B are sets such that $|A| = \alpha$, $|B| = \beta$ and B^A denotes the set of all mappings from A to B. It is again clear that β^α is independent of the choice of A, B, and we note that for finite cardinals, β^α has its usual meaning: if A has m elements and B has n elements, then there is a choice of n elements to which to map each element of A, and these choices are independent, so there are $n.n \ldots n$ (m factors) $= n^m$ choices. Of course this interpretation applies only to finite sets.

If B is a 1-element set, then so is B^A, for any set A: each element of A is mapped to the unique element of B, and this applies even if A is empty, for a mapping $A \to B$ is defined as soon as we have specified the images of the elements of A; so when $A = \varnothing$, nothing needs to be done. When B is empty, then so is B^A, unless also $A = \varnothing$, for there is nowhere for the elements of A to map to. Hence we have

$$1^\alpha = 1, 0^\alpha = \begin{cases} 0 & \text{if } \alpha \neq 0, \\ 1 & \text{if } \alpha = 0. \end{cases} \tag{1.1.3}$$

Let us now assume that B has more than one element. Then we necessarily have

$$|B^A| \geq |A|. \tag{1.1.4}$$

For let b, b' be distinct elements of B; we can map A to B^A by the rule $a \mapsto \delta_a$, where

$$x\delta_a = \begin{cases} b & \text{if } x = a, \\ b' & \text{if } x \neq a. \end{cases}$$

This mapping is injective because for $a \neq a'$, δ_a differs from $\delta_{a'}$ at a. It is a remarkable fact that the inequality (1.1.4) is always strict. As usual we write $\alpha < \beta$ or $\beta > \alpha$ to mean '$\alpha \leq \beta$ and $\alpha \neq \beta$'.

Theorem 1.1.3. *For any cardinals α, β, if $\beta > 1$, then $\alpha < \beta^\alpha$. In particular,*

$$\alpha < 2^\alpha \tag{1.1.5}$$

for any cardinal α.

Proof. We have just seen that $\alpha \leq \beta^\alpha$ and it only remains to show that equality cannot hold. Taking sets A, B such that $|A| = \alpha$, $|B| = \beta$, we shall show that there is no surjective mapping from A to B^A; it then follows that these sets are not equipotent. Thus let $f : A \to B^A$ be given; in detail, f associates with each $a \in A$ a mapping from A to B, which may be denoted by f_a. We must show that f is not surjective, i.e. we must find $g : A \to B$ such that $g \neq f_a$ for all $a \in A$. This may be done very simply by constructing a mapping g to differ from f_a at a. By hypothesis, B has at least two elements, say b, b', where $b \neq b'$. We put

$$ag = \begin{cases} b' & \text{if } af_a = b, \\ b & \text{otherwise.} \end{cases}$$

Then g is well-defined and for each $a \in A$, $g \neq f_a$ because $ag \neq af_a$. ■

If in this theorem we take A to be countable and B a 2-element set, simply denoted by 2, then 2^A is again infinite, but uncountable. Moreover, we can in this way obtain arbitrarily large cardinals by starting from any infinite cardinal α and forming in succession $2^\alpha, 2^{2^\alpha}, \ldots$.

Theorem 1.1.3 again illustrates the dangers of operating with the 'set of all sets'. If we could form the union of all sets, U say, then U would contain 2^U as a subset, and it would follow that $|2^U| \leq |U|$, in contradiction to Theorem 1.1.3. This paradox was discussed by Cesare Burali-Forti and others in the closing years of the 19th century, and it provided the impetus for much of the axiomatic development that followed. Any axiomatic system now in use is designed to avoid the possibility of such paradoxes. For our purpose it is sufficient to note that we can avoid the paradoxes by not admitting constructions involving 'all sets' without further qualification.

We conclude this section with some applications of Theorem 1.1.3. Given any set A, we denote by $\mathscr{P}(A)$ the set whose members are all the subsets of A; e.g. $\mathscr{P}(\varnothing) = \{\varnothing\}$, $\mathscr{P}(\{x\}) = \{\varnothing, \{x\}\}$. This set $\mathscr{P}(A)$ is often called the *power set* of A; it is equipotent with 2^A. To obtain a bijection we associate with each subset C of A its *characteristic function* $\chi_C \in 2^A$; taking $2 = \{0, 1\}$, we have

$$\chi_C(x) = \begin{cases} 1 & \text{if } x \in C, \\ 0 & \text{if } x \notin C. \end{cases}$$

It is easily seen that the mapping $C \mapsto \chi_C$ provides a bijection between $\mathscr{P}(A)$ and 2^A. The inverse mapping is obtained by associating with each $f \in 2^A$ the inverse image of 1: $1f^{-1} = \{x \in A | xf = 1\}$. Now Theorem 1.1.3 shows the truth of

Corollary 1.1.4. *No set is equipotent with its power set. More precisely, given any set A, there is no surjection from A to* $\mathscr{P}(A)$. ■

As a further application we determine the cardinal of the set **R** of all real numbers. This cardinal is usually denoted by $\mathfrak{c}$ and is called the *cardinal* (or *power*) of the *continuum*.

Proposition 1.1.5. $\mathfrak{c} = 2^{\aleph_0}$.

Proof. We can replace **R** by the open interval $(0, 1) = \{x \in \mathbf{R} | 0 < x < 1\}$, for there is a bijection, e.g.

$$x \mapsto \frac{1}{2} + \frac{x}{2(1+x^2)^{1/2}}.$$

If we express each number in the binary scale: $a = 0.a_1a_2\ldots(a_i = 0 \text{ or } 1)$, then $a \mapsto f_a$, where $f_a(n) = a_n$, is a mapping $(0, 1) \to 2^{\mathbf{N}}$ which is injective, for distinct real numbers have distinct binary expansions. Indeed, some have more than one, e.g. $0.0111\ldots = 0.1000\ldots$, but we can achieve uniqueness by excluding representations in which only finitely many digits are 0. It follows that $\mathfrak{c} \leq 2^{\aleph_0}$. On the other hand, there is an injective mapping from $2^{\mathbf{N}}$ to $(0, 1)$, obtained by mapping f_a, defined as before, to $0.a_1a_2\ldots$ in the decimal scale; thus the image consists of the real numbers between 0 and 1 whose decimal expansion contains only 0's and 1's. This shows that $2^{\aleph_0} \leq \mathfrak{c}$, and the desired equality follows. ■

It was conjectured by Cantor that $\mathfrak{c}$ is the least cardinal greater than $\aleph_0$; this is known as *Cantor's continuum hypothesis* (CH). In 1939 Kurt Gödel showed that it is consistent with the usual axioms of set theory; thus if the usual system of axioms (which we have not given explicitly) is consistent, then it remains consistent when CH is added. In 1963 Paul J. Cohen showed CH to be independent of the usual axioms of set theory. Thus if the negation of CH is added to the axioms of set theory (assumed consistent), we again get a consistent system. This means that within the usual axiom system of set theory CH is undecidable.

Exercises

1. Show that the set of all intervals in **R** with rational endpoints is countable.
2. Let A be an infinite set, A' be a finite subset and B be its complement in A. By picking a countable subset of B, show that $|A| = |B|$ without assuming Equation (1.1.2).
3. Let A be an uncountable set, A' be a countable subset and B be its complement in A. Show that $|A| = |B|$ without assuming Equation (1.1.2).
4. Fill in the details of the following proof that the interval $(0, 1)$ is uncountable. If the real numbers in binary form (as in the proof of Proposition 1.1.5) could be enumerated as $a^{(1)}, a^{(2)}, \ldots$, we can find a number not included in the enumeration by putting $a = 0.b_1b_2\ldots$, where $b_n = 0$ or 1 according as $a^{(n)}$ has 1 or 0 in

the n-th place. (This is Cantor's diagonal argument: b_n is chosen so as to differ from the diagonal term $a_n^{(n)}$.)

5. Show that the set of all real functions on the interval $[0, 1]$ has cardinal greater than the cardinal of the continuum. What about the subset of all continuous functions?
6. Show that for any cardinals α, β, γ if $\gamma \neq 0$ and $\alpha \leq \beta$, then $\alpha\gamma \leq \beta\gamma$.
7. Show that $\alpha^\gamma\beta^\gamma = (\alpha\beta)^\gamma$, $\alpha^\beta\alpha^\gamma = \alpha^{\beta+\gamma}$, $(\alpha^\beta)^\gamma = \alpha^{\beta\gamma}$.
8. Let $f : \mathbf{R} \to \mathbf{Q}$ be such that $x \leq y$ implies $xf \leq yf$. Show that there is an interval in which f is constant.

1.2 Zorn's Lemma and Well-ordered Sets

In Section 1.1 we have already defined the relation $\alpha \leq \beta$ for cardinals and we have shown in Theorem 1.1.2 that it is a partial ordering. Two elements x, y in a partially ordered set S are said to be *comparable* if $x \leq y$ or $y \leq x$. A subset of S in which any two elements are comparable is called a *chain* or said to be *totally ordered.* A subset in which no two elements are comparable is called an *anti-chain.* For example, the set of natural numbers $\mathbf{N}$ is totally ordered for the usual ordering by magnitude and partially ordered with respect to divisibility: $a|b$ iff $b = ac$ for some $c \in \mathbf{N}$. For the divisibility ordering on $\mathbf{N}$ the set of all prime numbers is an anti-chain.

In any partially ordered set S an element c is a *greatest* element if $x \leq c$ for all $x \in S$, while c is *maximal* if $c < x$ for no $x \in S$. Thus a greatest element is maximal, but the converse need not hold. A greatest element, if it exists, is clearly unique, unlike a maximal element. *Least* and *minimal* elements are defined dually; e.g. $\mathbf{N}$ with its usual ordering has a least element but no greatest element, while $\mathbf{Q}$ has neither a least nor a greatest element.

An *upper bound* of a subset X of S is an element $b \in S$ such that $x \leq b$ for all $x \in X$; here b may or may not belong to X. *Lower bounds* are defined dually, and a subset of S is *bounded* if it has both an upper and a lower bound.

We now take up the question of the comparability of cardinals left open in the last section, i.e. whether the ordering of cardinals is in fact total. In terms of sets the question is whether, given two sets A, B, we can find an injective mapping from one of them to the other. In intuitive terms one might try to answer this question by choosing an element from each of A and B, say a_1, b_1, and pairing them off, then choosing another pair of elements $a_2 \in A$, $b_2 \in B$ and pairing them off, and so on. For sets that are at most countable this solves the problem, but we have seen that there are uncountable sets, and here the procedure adopted is rather more problematic. One way to overcome the difficulty is to introduce the concept of a well-ordering:

An ordered set A is said to be *well-ordered* if every non-empty subset of A has a least element.

A well-ordered set is always totally ordered, as we see by applying the definition to 2-element subsets. It is also clear from the definition that any subset of a well-ordered set is again well-ordered. A countable set may be well-ordered simply by

enumerating its elements, e.g. the natural order of the positive integers is a well-ordering, but there are many other well-orderings which do not put the countability into evidence, e.g. $\{2, 3, 4, \ldots, 1\}$, where the order intended is that in which the numbers are written, or $\{1, 3, 5, \ldots, 2, 4, 6, \ldots\}$ or even $\{1, 2, 4, 6, \ldots, 3, 9, 15, 21, \ldots, 5, 25, 35, \ldots, 7, 49, \ldots\}$. By contrast, the negative numbers in their natural order form a set which is not well-ordered, although we can well-order it, e.g. by writing it in the opposite order.

For well-ordered sets it is possible to prove the comparability in a strong form. Let us call a subset A' of an ordered set A a *lower segment* if for any $u \in A'$, $v \in A$, $v \leq u$ implies $v \in A'$. This definition can be used for any ordered set, not necessarily well-ordered, or even totally ordered. In particular, for any $a \in A$, the set $|a) = \{x \in A | x < a\}$ is a lower segment in A, called a *principal* lower segment. In a well-ordered set A every lower segment not the whole of A is principal, for if A' is a proper lower segment of A and a is the first element of $A \backslash A'$, then $A' = |a)$.

Two ordered sets A, B are said to be *order-isomorphic* or of the same *order-type* if there is a bijection between them which preserves the ordering, $f : A \to B$ such that $x \leq y \Leftrightarrow xf \leq yf$.

Lemma 1.2.1. *A well-ordered set cannot be order-isomorphic to one of its proper lower segments.*

Proof. Let A be well-ordered, $|a)$ be a proper lower segment and suppose that $f : A \to |a)$ is an order-isomorphism. Then clearly $af < a$; if a_0 is the least element of A such that $a_0 f < a_0$, then by applying f and remembering that f preserves the order, we find that $a_0 ff < a_0 f$; so we have found an earlier element with the same property, namely $a_0 f$. This contradiction shows that f cannot exist. ∎

We now show that any two well-ordered sets can be compared.

Theorem 1.2.2. *Let A, B be two well-ordered sets. Then one of them is order-isomorphic to a lower segment of the other.*

Proof. Let us call a pair of elements $a \in A$ and $b \in B$ *matched* if the corresponding lower segments $|a)$ and $|b)$ are order-isomorphic. Two distinct lower segments of A cannot be order-isomorphic, for one of them will be a lower segment of the other and this would contradict Lemma 1.2.1. It follows that any element of B can be matched against at most one element of A and vice versa. Let A' be the set of elements of A that can be matched against elements of B, and B' be the set of elements of B matched against elements of A. Then A' and B' are order-isomorphic, as we see by using the correspondence provided by the matching. Moreover, A' is a lower segment of A, for if $a \in A'$ and $a_1 < a$, let a be matched to $b \in B$; then a_1 is matched to the element of $|b)$ which corresponds to it under the isomorphism between $|a)$ and $|b)$. Similarly B' is a lower segment of B. If $A' \neq A$, then $A' = |a')$ for some $a' \in A$; likewise, if $B' \neq B$, then $B = |b')$ for some $b' \in B$, and by construction there is an order-isomorphism between $|a')$ and $|b')$, so that a' and b' are matched. But this contradicts the fact that $a' \notin |a')$, so we conclude that either $A' = A$ or $B' = B$ (or both), as was claimed. ∎

The problem of comparing cardinals is thus reduced to the problem of well-ordering sets. If every set could be well-ordered, Theorem 1.2.2 would tell us that any two cardinals are comparable. Now it was proved by Ernst Zermelo in 1904 that every set can be well-ordered, but he had to make an assumption which was somewhat less intuitive than the other axioms used. This is the following:

Axiom of Choice. *Given a family of non-empty sets* $\{A_i\}_{i\in I}$, *there exists a function which associates with each set* A_i *a member of* A_i.

At first sight this is an innocent-sounding assumption, which acquires its force from the fact that it applies to collections with arbitrary indexing set; only for finite families $\{A_1, \ldots, A_n\}$ is no axiom needed. The axiom may be illustrated by the following example due to Bertrand Russell. A certain millionaire has infinitely many pairs of shoes and infinitely many pairs of socks. He wants to pick one shoe from each pair: this causes no problems; he simply picks the left shoe each time. But when he wants to pick a sock from each pair, he needs the axiom of choice.

In some respects the axiom of choice occupies a position analogous to the parallel axiom in geometry (although set theory without the axiom of choice is not as interesting as non-Euclidean geometry). Like the continuum hypothesis it has been proved consistent with and independent of the other axioms of set theory (by K. Gödel in 1939 and P. J. Cohen in 1963, respectively). It is of interest to note that the proof of Theorem 1.1.2 did not use the axiom of choice.

A logical step at this point would be to prove (as Zermelo did) that every set can be well-ordered, using the axiom of choice. To prove the well-ordering theorem we shall introduce another axiom, first proved by Kazimierz Kuratowski in 1922 and rediscovered by Max Zorn in 1935, known as *Zorn's lemma*. Although equivalent to the axiom of choice, it seems more appropriate in the present context, for it is Zorn's lemma rather than the axiom of choice that is used in algebra; we shall meet many examples later on.

Zorn's Lemma. *Let A be a partially ordered set. If every chain in A has an upper bound, then A has a maximal element.*

A partially ordered set is called *inductive* if every chain in it has an upper bound. In particular, such a set must be non-empty, as we see by taking the upper bound of the empty chain. In this terminology Zorn's lemma states that every partially ordered set which is inductive has a maximal element.

This statement sounds plausible, but any attempt at a direct proof soon encounters the situation typical of the axiom of choice. The actual derivation of Zorn's lemma from the axiom of choice can be found in most books on set theory. For an excellent account we refer to Kaplansky (1972). Below, in Theorem 1.2.3, we shall prove the well-ordering theorem (W) on the basis of Zorn's lemma (Z), and it is easy to derive the axiom of choice (C) from the well-ordering theorem: if $\{X_i\}$ is any family of non-empty sets, well-order $\cup X_i$ and assign to each X_i the element of it which comes first in the well-ordering. Thus C $\Rightarrow$ Z, Z $\Rightarrow$ W, W $\Rightarrow$ C; so the three assertions, C, Z, W are all equivalent.

Later on we shall meet many situations where the hypotheses of Zorn's lemma are satisfied; for the moment we shall give an illustration where the hypotheses do not hold. Let A be an infinite set and let $\mathscr{F}$ be the collection of all its finite subsets, partially ordered by inclusion. It is clear that $\mathscr{F}$ has no maximal element, and by Zorn's lemma this means that $\mathscr{F}$ must contain chains that have no upper bound in $\mathscr{F}$; such chains are of course easily found. In verifying the hypotheses of Zorn's lemma it is important to test arbitrary chains and not merely ascending sequences, as is shown by examples (see Exercise 3).

Theorem 1.2.3. *Every set can be well-ordered.*

Proof. The idea of the proof is to consider well-orderings of parts of the given set, make these well-orderings into a partially ordered set and show it to be inductive, so that Zorn's lemma can be applied.

Given a set A, let $\mathscr{W}$ be the collection of all subsets of A that can be well-ordered; if a subset can be well-ordered in more than one way we list all the versions separately. For example, any finite subset of A can be well-ordered (usually in more than one way). This shows that $\mathscr{W}$ is not empty; even if $A = \varnothing$, $\mathscr{W}$ contains the set $\varnothing$ as a member. We order $\mathscr{W}$ by writing $X \leq Y$ for $X, Y \in \mathscr{W}$ whenever X is a subset of Y and the inclusion mapping from X to Y is an order-isomorphism of X with a lower segment of Y; in particular the ordering of X is then the same as that induced by Y. It is clear that this defines a partial ordering on $\mathscr{W}$ and we have to show that $\mathscr{W}$ is inductive. Let $\{X_\lambda\}$ be a chain in $\mathscr{W}$, where λ runs over an indexing set (not necessarily countable); thus for any λ, μ either X_λ is a lower segment of X_μ or X_μ is a lower segment of X_λ. To get an upper bound for this chain we put $X = \cup X_\lambda$ and define an ordering on X as follows: let $x, y \in X$ and choose an index λ such that $x, y \in X_\lambda$. If $x \leq y$ in the ordering of X_λ, then the same is true in the ordering of X_μ for any μ such that $x, y \in X_\mu$; for of X_λ, X_μ, one is a lower segment of the other, with the same ordering. Thus we may without ambiguity put $x \leq y$ in X if this holds in some X_λ, and the relation so defined on X is easily seen to be a well-ordering, with each X_λ as a lower segment. Hence X is an upper bound of the given chain in $\mathscr{W}$, and this shows $\mathscr{W}$ to be inductive.

We can now apply Zorn's lemma and obtain a maximal element X' in $\mathscr{W}$. We claim that $X' = A$; for if not, then there exists $z \in A \backslash X'$. We form $X'' = X' \cup \{z\}$ into a well-ordered set by taking the given order on X' and letting z follow all of X'. Then X'' is a member of $\mathscr{W}$ which is strictly greater than X', contradicting the maximality of X'. Hence $X' = A$ and this is the desired well-ordering of A. ∎

This result allows us to conclude that any two cardinals can be compared; thus the relation '$\leq$' is a total ordering of cardinals. But we can say rather more than this. Theorems 2.2 and 2.3 suggest a classification of well-ordered sets according to their order-type. Thus with every well-ordered set A we associate a symbol α, called the *ordinal number* or *order-type*, or simply *ordinal*, such that two well-ordered sets have the same ordinal precisely when they are order-isomorphic. Further, we can define a relation $\alpha \leq \beta$ between ordinals, whenever a set of type α is order-isomorphic to a lower segment of a set of type β. This is a well-defined relation

on any set of ordinals, clearly transitive and by Lemma 1.2.1 antisymmetric, i.e. an ordering, which is total by Theorem 1.2.2. In fact it is a well-ordering (see Exercise 7).

With each ordinal number α a cardinal number $|\alpha|$ may be associated, namely the cardinal of a well-ordered set of ordinal α. This is an order-preserving mapping from ordinals to cardinals, but not injective: to each finite cardinal there corresponds just one ordinal of the same type, but an infinite cardinal always corresponds to many different ordinals. However, we obtain a well-defined mapping by assigning to each cardinal the *least* ordinal which corresponds to it. For example, a countable set, **N** say, may be well-ordered as $1, 2, 3, \ldots$; this order-type is denoted by ω and is the least countable ordinal. Another ordering, not isomorphic to the first, is $2, 3, 4, \ldots, 1$; it is denoted by $\omega + 1$. Similarly, $n+1, n+2, \ldots, 1, 2, \ldots, n$ has ordinal $\omega + n$. The type of $1, 3, 5, \ldots, 2, 4, 6, \ldots$ is written $\omega + \omega$ and generally, given ordinals α, β, we define $\alpha + \beta$ as the type of a well-ordered set of type α followed by one of type β. It is easily checked that such an arrangement gives rise to a well-ordered set whose type depends only on α and β. We observe that the addition of ordinal numbers is still associative, but no longer commutative: $1 + \omega = \omega \neq \omega + 1$.

We shall write 2ω for $\omega + \omega$ and generally $n\omega$ for $\omega + \omega + \ldots + \omega$ to n terms. The limit of the sequence $\omega, 2\omega, 3\omega, \ldots$, i.e. the first ordinal following all of them, is written ω^2. We shall not pursue this topic further except to mention that every ordinal number α can be written in just one way as

$$\alpha = a_1\omega^{\alpha_1} + a_2\omega^{\alpha_2} + \ldots + a_r\omega^{\alpha_r},$$

where $r, a_1, \ldots, a_r$ are natural numbers and $\alpha_1, \ldots, \alpha_r$ is a decreasing sequence of ordinal numbers (for a proof see e.g. Sierpiński (1956)).

There is a particular situation allowing Zorn's lemma to be applied which frequently occurs in algebra. Let S be a set and P be a property of certain subsets of S; by a P-set we shall understand a subset with the property P. A property P of subsets of S is said to be of *finite character* if any subset T of S is a P-set precisely when all finite subsets of T are P-sets. For example, if S is a partially ordered set, then being totally ordered is a property of finite character: a subset T of S is totally ordered iff every 2-element subset of T is totally ordered. On the other hand, being well-ordered is not a property of finite character in ordered sets, because every finite subset of a totally ordered set is well-ordered, but the set itself need not be well-ordered.

For a property of finite character there is always a maximal subset with this property:

Proposition 1.2.4. *Let S be a set and P be a property of subsets of S. If P is a property of finite character, then the collection of all P-sets in S has a maximal member.*

Proof. The result will follow by Zorn's lemma if we can show that the set $\mathscr{F}$ of all P-sets in S is inductive. Let $\{T_\alpha\}$ be a chain of P-sets and write $T = \cup T_\alpha$. If T fails to have property P, then there is a finite subset $\{x_1, \ldots, x_r\}$ of T which does not have P. Let $x_i \in T_{\alpha_i}$; since the T_α form a chain, there is a largest among the sets $T_{\alpha_1}, \ldots, T_{\alpha_r}$, say T'. But then $T' \supseteq \{x_1, \ldots, x_r\}$, so T' does not have P, which is a contradiction.

This shows $\mathscr{F}$ to be inductive; by Zorn's lemma it has a maximal member which is the desired maximal P-set. ■

We assume that all our readers will have met proofs by induction, where a theorem involving a natural number n, say $T(n)$, is proved first for $n = 1$ (or for $n = 0$) and then $T(n+1)$ is proved assuming $T(n)$. It would be equally acceptable to prove $T(n+1)$ assuming $T(\nu)$ for all $\nu \leq n$. A slight variant is a *recursive definition*, where a concept $C(n)$ involving an integer n is defined for $n = 0$ or 1, and then $C(n+1)$ is defined in terms of $C(\nu)$ for $\nu \leq n$. Both notions can be justified by a proof based on Peano's axioms for the natural numbers (see FA Chapter 1).

For well-ordered sets there is a form of induction, known as the *principle of transfinite induction*. This is embodied in the remark (which practically reproduces the definition) that any non-empty subset of a well-ordered set has a least element. Let us pause briefly to examine how a transfinite induction proof looks in practice. One can distinguish three kinds of ordinals: an ordinal number β may have an immediate predecessor α, so that $\beta = \alpha + 1$, or it may have no immediate predecessor. In that case it is either the first ordinal 1, or it is the first ordinal after an infinite set of ordinals, in which case it is called a *limit ordinal*. For example, the first limit ordinal is ω. We note that sometimes the first ordinal is taken to be 0.

Now the principle of transfinite induction may be stated as follows:

Theorem 1.2.5. *Let A be a well-ordered set; if its ordinal is τ, the set may be indexed by the ordinals $< \tau$: $A = \{a_\alpha\}_{\alpha<\tau}$. Suppose that X is a subset of A satisfying the conditions:*

(i) $a_1 \in X$;
(ii) *if* $a_\alpha \in X$, *then* $a_{\alpha+1} \in X$;
(iii) *if* $a_\alpha \in X$ *for all* $\alpha < \lambda$, *where* $\lambda < \tau$ *is a limit ordinal, then* $a_\lambda \in X$.

Then $X = A$.

Proof. Suppose X is a proper subset of A and let a_β be the least element of $A \backslash X$. Then $\beta > 1$ by (i); if β is a non-limit ordinal, it has an immediate predecessor α say. Now $a_\alpha \in X$ and $\beta = \alpha + 1$, so $a_\beta \in X$ by (ii), a contradiction. So β must be a limit ordinal and by definition, $a_\alpha \in X$ for all $\alpha < \beta$. Hence by (iii), $a_\beta \in X$, again a contradiction. It follows that $X = A$, as claimed. ■

This analysis allows us to give an explicit description of well-ordered sets:

Corollary 1.2.6. *Any well-ordered set consists of a well-ordered set of countable sequences, possibly followed by a finite sequence.*

Proof. Let A be a well-ordered set and consider the set L of limit ordinals of A, together with the first element. This is a well-ordered set, and each $\lambda \in L$ which does not come last in L is the first of a countable sequence. If L has no last element, then A consists of a family of countable sequences indexed by L. If L has a last element, then this is the first of a countable or finite sequence. Thus in either case A has the required form. ■

We saw earlier that the smallest infinite cardinal is denoted by $\aleph_0$; this notation, introduced by Cantor, is part of the aleph notation for infinite cardinals. It is defined as a function from ordinals to cardinals by transfinite recursion, as follows:

The first infinite cardinal is $\aleph_0$. If $\beta = \alpha + 1$, then $\aleph_\beta$ is the least cardinal greater than $\aleph_\alpha$ and for a limit ordinal λ, $\aleph_\lambda = \cup_{\alpha<\lambda}\aleph_\alpha$. For example, $\aleph_1$ is the least uncountable cardinal and the continuum hypothesis may be expressed as $\aleph_1 = 2^{\aleph_0}$.

We conclude this section with a proof of a special case of a formula mentioned earlier, Equation (1.1.2).

Proposition 1.2.7. *For any infinite cardinal α and any $n \geq 1$, $n\alpha = \alpha$; moreover,*

$$\aleph_0\alpha = \alpha. \tag{1.2.1}$$

Proof. We shall need the associative law of multiplication of cardinals: $(\alpha\beta)\gamma = \alpha(\beta\gamma)$; it is easily proved by observing that each side may be regarded as the cardinal number of the product set $A \times B \times C$, where A, B, C are sets of cardinals α, β, γ respectively.

In the case where $\alpha = \aleph_0$, (1.2.1) states that $\mathbf{N}^2$ is equipotent with $\mathbf{N}$ and this was proved in Theorem 1.1.1. Secondly, if α is of the form $\aleph_0\beta$, for some cardinal β, then by the associative law,

$$\aleph_0\alpha = \aleph_0(\aleph_0\beta) = \aleph_0^2\beta = \aleph_0\beta = \alpha,$$

which proves (1.2.1) in this case. We complete the proof by showing that every infinite cardinal is of the form $\aleph_0\beta$. This amounts to showing that every infinite set A is equipotent with a set of the form $\mathbf{N} \times B$, for a suitable set B.

Let A be an infinite set. By Theorem 1.2.3 A can be well-ordered, and by Corollary 1.2.6, A consists of a well-ordered set of countable sequences, possibly followed by a finite sequence. Since A is infinite, at least one infinite sequence occurs, and we may rearrange A by taking the finite sequence from the end and putting it in front of the first sequence. The set A now consists entirely of countable sequences, i.e. well-ordered sequences of type ω. If they are indexed by a set B, it follows that A is equipotent with $\mathbf{N} \times B$, and this is what we had to show. The rest is clear. ■

Exercises

1. Show that the axiom of choice is equivalent to the following axiom: every surjective mapping has a left inverse.
2. Let Φ be a partial ordering relation on a set A. Show that there is a total order Φ' on A such that $\Phi \subseteq \Phi'$.
3. Let A be an uncountable set and $\mathscr{F}$ be the collection of all its countable subsets. Show that every countable ascending sequence of members of $\mathscr{F}$ has an upper bound in $\mathscr{F}$, but that $\mathscr{F}$ has no maximal element.
4. Show that any totally ordered set X has a well-ordered subset Y (where the ordering of Y is that induced by X) with the property: for each $x \in X$ there exists $y \in Y$ such that $x \leq y$ (i.e. Y is a *cofinal* subset of X).

5. Determine all order-automorphisms (i.e. order-preserving permutations) of **Z**.
6. Find a well-ordering of the set **Z** of all integers. Find well-orderings of **N** of type 2ω, $2\omega + 1$, $\omega^2 + 1$.
7. Show that the set of all lower segments of a well-ordered set is well-ordered by inclusion. Deduce that any set of ordinals is well-ordered by $\leq$.
8. Check that the addition of cardinals is well-defined and satisfies the associative law.
9. Ordinal multiplication may be defined by taking, for any ordinals α, β, sets A, B of type α, β respectively and denoting by $\alpha\beta$ the ordinal of the product $A \times B$, ordered lexicographically. Show that this multiplication is well-defined, and that it agrees with the following recursive definition: (i) $1\beta = \beta$; (ii) $(\alpha + 1)\beta = \alpha\beta + \beta$; (iii) if λ is a limit ordinal, then $\lambda\beta = \sup\{\gamma\beta | \gamma < \lambda\}$.
10. Show that with the definition of Exercise 9, for any ordinals α, β, γ, $(\alpha + \beta)\gamma = \alpha\gamma + \beta\gamma$, $(\alpha\beta)\gamma = \alpha(\beta\gamma)$, but that in general, $\alpha(\beta + \gamma) \neq \alpha\beta + \alpha\gamma$. (Hint. For the inequality take β, γ finite and α infinite.)
11. From the definition in Exercise 9 show that for any ordinals α, β, γ, if $\alpha < \beta, \gamma > 0$, then $\alpha\gamma < \beta\gamma$. Deduce that $\gamma \neq 0$ and $\alpha\gamma = \beta\gamma$ imply $\alpha = \beta$; give examples to show that $\gamma\alpha = \gamma\beta$ does not imply $\alpha = \beta$.
12. Show that if β is a limit ordinal, then so is $\alpha + \beta$, for any ordinal α. Is $\beta + \alpha$ necessarily a limit ordinal? If $\alpha > 0$ and β is a limit ordinal, show that $\alpha\beta$ is a limit ordinal.

1.3 Graphs

Many problems both in mathematics and elsewhere can best be solved diagrammatically; the diagrams involved are more or less subtle reformulations of the problem, and the efforts to solve such problems have given rise to the theory of graphs. We can do no more here than present the beginnings of the theory, but it seems appropriate to do so since the methods are often algebraic and graphs are increasingly being used in other parts of mathematics as well as in algebra itself.

A *graph* Γ consists of a pair of sets $V(\neq \varnothing)$, E. The members of V are the *points* or *vertices* of Γ while the members of E are its *edges*. With each edge of E two points are associated, its *endpoints*. Here the endpoints of an edge need not be distinct; if they coincide, the edge is a *loop*. Two edges may have the same pair of (distinct) endpoints, giving a *multiple edge*. Two vertices are *adjacent* if they are joined by an edge. A point is *isolated* if it is not an endpoint of an edge. Some simple examples are illustrated in Figure 1.1, where the first three represent *simple* graphs, i.e. graphs without loops or multiple edges.

Given a graph $\Gamma = \{V, E\}$, if V' is a subset of V and E' is a subset of E such that any edge in E' has its endpoints in V', then $\{V', E'\}$ is again a graph, called a *subgraph* of Γ. A subgraph Γ' is said to be *induced* if for any vertices p, q in Γ' all edges between p and q belong to Γ'.

For any set S, the graph with S as vertex set and an edge between each pair of distinct vertices is called the *complete graph* on S and denoted by $C(S)$. Every simple graph $\Gamma = \{V, E\}$ is clearly a subgraph of $C(V)$; the graph with vertex set V and

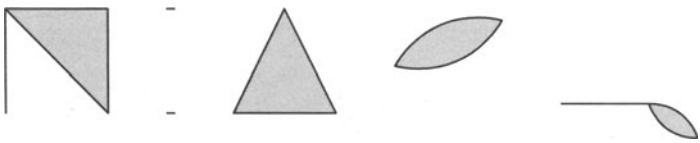


Figure 1.1

the set of edges in $C(V)$ that are not in Γ is written Γ' and called the *complementary graph* of Γ.

Example 1. In a group of six people it is always possible to find either three people who know each other or three people who are all strangers to each other.

In order to prove this statement we represent the six individuals by points and join any two points representing acquaintances by an edge. We thus obtain a graph Γ, the 'acquaintanceship graph' of our group and we have to show that either Γ contains three edges forming a triangle, or its complement Γ' does so. Take a vertex p_1; it is adjacent to each of the five other points in just one of Γ, Γ'. Hence it must be adjacent to three points in one of these graphs, say in Γ it is adjacent to p_2, p_3, p_4. If two of p_2, p_3, p_4 are adjacent in Γ, then these two vertices together with p_1 form a triangle in Γ; otherwise p_2, p_3, p_4 form a triangle in Γ'.

Example 2 (The Königsberg bridge problem). A famous problem concerns the seven bridges in Königsberg crossing the river Pregel, which are situated as shown in Figure 1.2.

The problem was to cross in the course of a single walk each bridge exactly once. It is not hard to convince oneself that this is impossible; this was first proved by

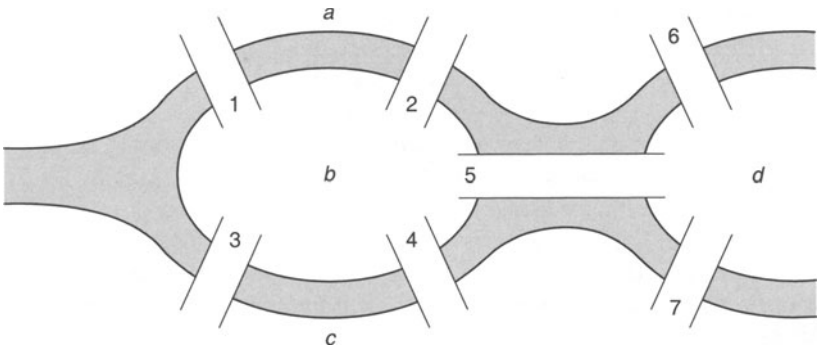


Figure 1.2

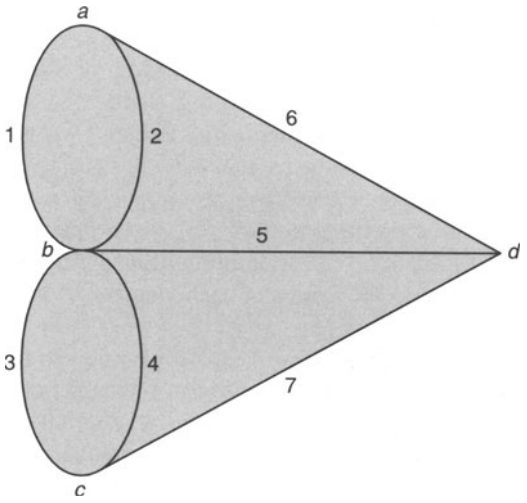


Figure 1.3

Leonhard Euler in 1736. The first step is to represent the problem by a graph in which the areas are points and the bridges become edges (Figure 1.3).

For each vertex we define its *valency* as the number of edges ending in it (counting loops twice). A walk across all the bridges becomes a path which includes each edge just once. If this path begins and ends at the same point, each point has even valency; otherwise there are just two points of odd valency, the beginning and end point of our path. But in the above graph all four points have odd valency, so there cannot be such a path.

Formally we define a *path* of *length* n from p to q in a graph as a set of edges $e_1, \ldots, e_n$ such that e_i has endpoints p_{i-1}, p_i and $p_0 = p$, $p_n = q$. A path from p to p is called a *cycle* (or also *closed*) and a graph without cycles of positive length is said to be *acyclic*.

We note that every finite partially ordered set may be considered as a graph by drawing an edge from p to q if q *covers* p, i.e. $p < q$ but $p < x < q$ for no x. In fact we thus obtain a *directed* graph or *digraph*, i.e. a graph in which the endpoints for each edge form an ordered pair, the *initial* vertex and the *final* vertex; the edges in a digraph are also called *arrows* and a finite digraph is called a *quiver*. In a digraph only paths are allowed which go in the direction of each arrow.

It is clear that the digraph derived from an ordered set is acyclic; conversely, given any acyclic graph Γ, we can define a partially ordered set on the vertex set of Γ by fixing a direction in Γ and writing $p \leq q$ whenever there is a path from p to q. Thus finite partially ordered sets may be identified with directed acyclic graphs. Our first result, though nominally about ordered sets, is really graph-theoretic in nature. In a finite ordered set S the maximum number of elements in an anti-chain will be called the *width* of S.

Theorem 1.3.1 (Dilworth's theorem). *Let S be a finite partially ordered set. Then the minimum number of disjoint chains into which S can be decomposed is the width of S.*

Proof. We shall use induction on $|S|$. Let m be the width of S and take an anti-chain A of m elements; any chain meets A in at most one element, so we cannot have fewer than m chains. Let S^+ be the set of elements of S that are $\geq$ some element of A and S^- be the set of elements $\leq$ some element of A. Then clearly $S^+ \cup S^- = S$, $S^+ \cap S^- = A$ and the elements of A are both minimal elements of S^+ and maximal elements of S^-. We assume first that both S^+, S^- are proper subsets of S. Then by induction we can write each of S^+, S^- as a union of m chains, with an element of A at one end. Hence by joining the chains at each element of A we obtain m chains for S.

There remains the case where S^+ or S^- equals S for any choice of A; this means that any anti-chain of m elements consists entirely of maximal or entirely of minimal elements of S. Choose a maximal element u and a minimal element $v \leq u$; then $S\backslash\{u, v\}$ has width $< m$ and by induction it can be written as a union of at most $m - 1$ chains. Together with $\{u, v\}$ this gives a decomposition of S into m chains, as required. ∎

Dilworth's theorem can be used to prove Philip Hall's theorem on distinct representatives.

Theorem 1.3.2 (P. Hall's theorem). *Let $T_1, \ldots, T_m$ be subsets of a finite set S. Then there is a family of distinct elements $a_1, \ldots, a_m$ in S such that $a_i \in T_i$, provided that the union of any k of the subsets T_i contains at least k elements, for $k = 1, \ldots, m$.*

Proof. Let $S = \{a_1, \ldots, a_n\}$ and define a partial ordering on the set W consisting of all the T_i and a_j by writing $x \leq y$ precisely when $x = y$ or $y = T_i$ and $x = a_j \in T_i$. Let us take any anti-chain in W:

$$\{T_1, \ldots, T_r, a_1, \ldots, a_s\}. \tag{1.3.1}$$

By hypothesis $T_1 \cup \ldots \cup T_r$ contains at least r elements, which must be distinct from $a_1, \ldots, a_s$ because (1.3.1) is an anti-chain. It follows that $r + s \leq n$, so every anti-chain in W has at most n elements; in fact this bound is reached by the anti-chain $\{a_1, \ldots, a_n\}$. By Theorem 1.3.1 we can decompose W into n disjoint chains. Each T_i is in just one of the chains, and each contains an a_j; thus $a_j \in T_i$ and we have a system of distinct representatives. ∎

Remark 1. Let $A_1, \ldots, A_m$ and $B_1, \ldots, B_n$ be two families of disjoint subsets of a set S. If any k of the A_i between them meet at least k of the B_j then we can find $a_i \in A_i$ such that the distinct a_i belong to distinct B's. For example, in a finite group G with a subgroup H we can take the A_i to be the left cosets and the B_j to be the right cosets of H in G (see Section 2.1). We thus obtain a set of common representatives for the left and right cosets of H (this was Hall's original purpose in proving this theorem in 1935).

Remark 2. The assertion of Theorem 1.3.2 has been formulated more picturesquely by Paul Halmos as the 'marriage theorem': in a group of m bachelors and n spinsters, if any k men between them are acquainted with at least k of the women, then each man can be married off to an acquaintance.

A graph is said to be *connected* if any two of its points can be joined by a path; if moreover, there is a unique path joining any two points, the graph is called a *tree*. Clearly a tree can also be characterized as a connected graph which is acyclic. It is an important fact that any connected graph contains a tree, called a *spanning tree*, which includes all vertices of the graph:

Theorem 1.3.3. *Let Γ be a connected graph. Then there is a subgraph Γ_0 of Γ which is a spanning tree. Moreover, for any finite tree (V, E) we have*

$$|V| = |E| + 1. \tag{1.3.2}$$

Proof. We observe that a graph Γ_1 is a tree iff any induced subgraph on finitely many vertices is a tree, for if Γ_1 contains a cycle, this already appears in a finite subgraph, and if Γ_1 fails to be connected, then the same is true of an induced 2-vertex subgraph. It follows by Proposition 1.2.4 that Γ contains a maximal subgraph Γ_0 which is a tree, i.e. a maximal subtree of Γ. We claim that Γ_0 contains all the vertices of Γ. For otherwise there is a vertex p adjacent to a vertex in Γ_0 but not itself in Γ_0. Since Γ is connected, there is an edge e from p to $q \in \Gamma_0$. Now the graph obtained by adjoining p and the edge e to Γ_0 is still a tree, but it contains Γ_0 as a proper subtree, contradicting the maximality of the latter. Thus Γ_0 contains all vertices of Γ and it is the required tree.

Now let $\Gamma = \{V, E\}$ be a finite tree; in Γ we can find a vertex p_0 which is adjacent to only one other vertex. To find p_0 we start from any vertex of Γ along a path and continue as far as possible without traversing an edge more than once. As long as all the vertices we reach have valency greater than 1 we can continue, and we never pass a vertex twice because Γ is a tree. Since Γ is finite, this process must come to a halt, and it can only do so when we reach a vertex of valency 1; this is the required vertex p_0. If we omit p_0 and the single edge ending at p_0 from Γ we obtain a tree $\Gamma' = \{V', E'\}$ with fewer vertices than Γ. By induction we have $|V'| = |E'| + 1$, and since $|V| = |V'| + 1$, $|E| = |E'| + 1$, (1.3.2) follows because the result holds for the trivial graph with one vertex and no edges. ∎

Our final result in this section, proved by Frank P. Ramsey in 1930, is a far-reaching generalization of an earlier example involving the acquaintanceship graph (p. 16). For brevity let us call a set or a subset consisting of r elements an r-set, respectively an r-subset, and write $\mathscr{P}_r(S)$ for the set of all r-subsets of S; for example, every r-set has exactly one r-subset and one 0-subset. The earlier example showed that if the collection of all 2-subsets of a 6-set S is partitioned in any way into two disjoint sets A_1, A_2, then either S contains a 3-subset all of whose 2-subsets lie in A_1 or it contains a 3-subset all of whose 2-subsets lie in A_2. In terms of graphs, if

in a complete graph on six vertices each edge is painted either blue or red, then there is a complete subgraph on three vertices in which all edges have the same colour. More generally, given integers $p, q \geq 2$, there exists an integer N such that if in a complete graph on N vertices the edges are painted blue or red, then there is either a complete blue subgraph on p vertices or a complete red subgraph on q vertices. This is just the special case $r = 2$ of the following theorem.

Theorem 1.3.4 (Ramsey's theorem). *Let p_1, p_2, r be integers such that $0 \leq r \leq p_i$ $(i = 1, 2)$. Then there exists an integer $N(p_1, p_2, r)$ with the following property: if in any set S of at least $N(p_1, p_2, r)$ elements the family of r-subsets is partitioned into two disjoint sets A_1 and A_2, then for $i = 1$ or 2, S contains a p_i-subset all of whose r-subsets lie in A_i.*

In symbols we have $\mathscr{P}_r(S) = A_1 \cup A_2, A_1 \cap A_2 = \varnothing$, and the assertion is that there is a p_i-subset T of S such that $\mathscr{P}_r(T) \subseteq A_i$ for $i = 1$ or 2.

Proof. We shall use induction on r, and for a given r, on p_1 and p_2. For $r = 0$ the result holds trivially: there is only one 0-subset of any set and this lies in A_1 or A_2. We may therefore assume that $r > 0$.

Firstly we note that $N(p_1, r, r) = p_1$. For if S has at least p_1 $(\geq r)$ elements, and the family of r-subsets of S has been partitioned into disjoint sets A_1 and A_2, then either $A_2 \neq \varnothing$ and so S contains an r-subset whose unique r-subset lies in A_2, or $A_2 = \varnothing$ and any p_1-subset of S has all its r-subsets in A_1. This shows that $N(p_1, r, r) \leq p_1$ and it is easily seen that the inequality here cannot be strict. Thus $N(p_1, r, r) = p_1$ and a similar argument shows that $N(r, p_2, r) = p_2$.

We now put $q_1 = N(p_1 - 1, p_2, r), q_2 = N(p_1, p_2 - 1, r)$ and claim that $N(p_1, p_2, r) \leq N(q_1, q_2, r - 1) + 1$. For let S be a set with at least $N(q_1, q_2, r - 1) + 1$ elements and let A_1, A_2 be a given partition of the family of r-subsets of S. Fix $c_0 \in S$, write $S' = S \setminus \{c_0\}$ and define a partition $\{B_1, B_2\}$ of the family of $(r - 1)$-subsets of S' by taking an $(r - 1)$-subset T to B_i whenever $T \cup \{c_0\} \in A_i$. Since S' has at least $N(q_1, q_2, r - 1)$ elements, we can apply the induction hypothesis. Two cases can arise: (i) S' has a q_1-subset U all of whose $(r - 1)$-subsets are in B_1. If U contains a q_2-subset all of whose r-subsets lie in A_2, the conclusion follows. Otherwise, by the definition of q_1, U has a $(p_1 - 1)$-subset whose r-subsets are all in A_1, and adjoining c_0 we get a p_1-subset all of whose r-subsets are in A_1. The second case (ii) is that S' has a q_2-subset all of whose $(r - 1)$-subsets are in B_2; this can be dealt with by symmetry, so the conclusion holds for S. ∎

The least value of $N(p_1, p_2, r)$ is sometimes called the *Ramsey number*. Thus the example given earlier shows that $N(3, 3, 2) \leq 6$, and it is easily checked that equality holds here. For $r = 1$ we have $N(p_1, p_2, 1) = p_1 + p_2 - 1$ and the theorem states that when a set of at least $p_1 + p_2 - 1$ elements is partitioned into two disjoint subsets A_1, A_2, then A_i has at least p_i elements for either $i = 1$ or 2.

Exercises

1. Construct all simple graphs with four edges and no isolated points (there are 11, up to isomorphism). Construct all trees with five edges (there are six).
2. Let S be a partially ordered set in which each chain and each anti-chain is finite. Define $l(a)$ for $a \in S$ as the minimum of the lengths of maximal chains below a, and show that $A_n = \{a \in S \mid l(a) = n\}$ is an anti-chain. By considering the maximal elements of S, write S as a union of a finite number of the A_n and hence show that S must be finite.
3. Prove Dilworth's theorem in the infinite case: every partially ordered set of finite width m can be written as the disjoint union of m chains.
4. Prove Dénes König's lemma: every infinite connected graph in which all points have finite valency has an infinite path. (Hint. Choose $p_1, p_2, \ldots, p_n$ so that there is a path along these points and p_n can be connected to infinitely many points by paths not passing through p_{n-1}.) Show that the result need not hold if there is at least one point of infinite valency.
5. Let $\Gamma = \{V, E\}$ be a finite graph, not necessarily connected but for which each connected component is a tree (such Γ is sometimes called a 'forest'). Show that Γ has $|V| - |E|$ connected components. (Hint. Use induction on the number of edges.)
6. Show that $N(p, q, 0) = \max\{p, q\}$, $N(p, q, 1) = p + q - 1$, $N(p, q, 2) \leq \binom{p+q-2}{p-1}$.
7. Show that when $p, q \geq 2$, then $N(p, q, 2) \leq N(p-1, q, 2) + N(p, q-1, 2)$. Find $N(p, q, 2)$ for $p \leq 4$, $q \leq 5$.
8. Show that $N(r, m, r) = m$.
9. Show that of five points in a plane, no three of which are collinear, there are four that form vertices of a convex quadrilateral.
10. Given n points in a plane of which no three are collinear, show that if all the quadrilaterals formed from these points are convex, then there is a convex n-gon with these points as vertices.
11. Show that for any $n \geq 1$, if at least $N(n, 5, 4)$ points in the plane are given, no three of which are collinear, then there are n points in the given set forming a convex n-gon. (Erdős–Szekeres. For $n = 3$ take $N = 3$; otherwise use Theorem 1.3.4 and Exercises 9 and 10.)
12. Show that there is a number $N = N(p_1, \ldots, p_t, r)$ such that if for any set S of at least N elements, the family of all subsets is partitioned into disjoint sets $A_1, \ldots, A_t$ then for some $i = 1, \ldots, t$, S contains a p_i-subset all of whose r-subsets lie in A_i.
13. Show that if the r-subsets of an infinite set S are partitioned into disjoint sets $A_1, \ldots, A_t$, then there is an infinite subset of S whose r-subsets all lie in A_i, for some i. (Hint. Adapt the proof of Theorem 1.3.4.)
14. Using Exercise 13, show that any infinite partially ordered set contains either an infinite chain or an infinite anti-chain.
15. Show that for any graph either it or its complement is connected.

Further Exercises for Chapter 1

1. Show that the union of a finite family of finite sets is finite.
2. Show that if α, β are cardinals of which at least one is infinite, then $\alpha + \beta = \max\{\alpha, \beta\}$.
3. Show that a set is finite iff every collection of subsets has either a maximal or a minimal member.
4. (Sierpiński) A collection T of sets will (for this exercise) be called a *tower* if $\varnothing \in T$ and $X, Y \in T \Rightarrow X \cup Y \in T$. Show that a set A is finite iff A belongs to every tower T such that $\{x\} \in T$ for all $x \in A$.
5. Prove that Zorn's lemma is equivalent to Hausdorff's maximal principle: given a collection A of sets, any chain in A is contained in a maximal chain.
6. (Tarski) Show that a set is finite iff it can be ordered so that both the ordering and its opposite are well-orderings.
7. Define an *upper segment* in an ordered set as a lower segment in the opposite ordering. Verify that the upper segments are precisely the complements of the lower segments. If a set A is well-ordered, and order-isomorphic to every non-empty upper segment of itself, what can be said about A? Show that A cannot be of type 2ω.
8. Let A be a countable set. Show that the set of all equivalences on A with finitely many equivalence classes is uncountable, but the subset of those equivalences in which only one class is infinite is countable.
9. Deduce Zorn's lemma from the well-ordering theorem.
10. Let S be a set of n elements and $\mathscr{C} = \{C_i\}$ be a collection of subsets of S such that $C_i \neq C_j$ for $i \neq j$ and $C_i \cap C_j \neq \varnothing$. Show that $\mathscr{C}$ has at most 2^{n-1} members.
11. Let A be a well-ordered set. Show that any interval of the form $\{x \in A | a \leq x < b\}$ is order-isomorphic to a lower segment of A.
12. Let S be a set of n elements. Show that $\mathscr{P}(S)$ has $n!$ maximal chains and that each r-subset of S is contained in exactly $r!(n-r)!$ maximal chains. Show further that if $\mathscr{P}(S)$ has an anti-chain containing ν_r r-subsets, then

$$\sum_{r=0}^{n} \nu_r \binom{n}{r}^{-1} \leq 1.$$

Deduce Sperner's lemma: the maximal length of any anti-chain in $\mathscr{P}(S)$ is the number of $[n/2]$-subsets in S (D. Lubell).

13. Let $\Gamma = (V, E)$ be a finite connected graph. Show that if T is a spanning tree of Γ, then T has $|V| - 1$ edges. Verify that for each edge e not in T there is a cycle C_e in $T \cup \{e\}$ and any two cycles C_e, $C_{e'}$ have only the edges of T in common. Deduce that the total number of such cycles is $|E| - |V| + 1$.
14. Let Γ be a graph and Δ be a subgraph. The graph Δ^c whose edges are the edges of Γ not in Δ and whose vertices are the vertices of Γ that are either not in Δ or incident with an edge of Γ not in Δ is called the *complement* of Δ in Γ. Show that Δ^{cc} is obtained from Δ by deleting the vertices of Δ that are incident with an edge of Γ not in Δ but not incident with any edge in Δ.
15. Show that in any finite graph the number of points of odd valency is even.

16. A finite graph is said to be *Eulerian* if it is connected and has a closed path (i.e. cycle) including each edge once. Show that a finite connected graph is Eulerian iff each vertex has even valency. Find a condition for a graph to have a path, not necessarily closed, that includes each edge just once.
17. For any finite graph $\Gamma = (V, E)$ define p_0 as the number of its connected components and p_1 as the least number of edges which need to be removed to make each component into a tree. Show that $|V| - |E| = p_0 - p_1$. Show also that if $v(x)$ is the valency of $x \in V$, then

$$\sum_{x \in V} [v(x) - 2] = 2(p_1 - p_0).$$

18. Let S be a finite set and $\mathscr{F}$ be a family of subsets of S whose union is S itself. The *intersection graph* of $\mathscr{F}$ is defined as the simple graph with $\mathscr{F}$ as vertex set, and where $X, Y \in \mathscr{F}$ are joined by an edge whenever $X \neq Y$ and $X \cap Y \neq \varnothing$. Show that every finite simple graph is the intersection graph of an appropriate family of sets.
19. Let Γ be a finite graph and for each vertex p define $d(p)$ as the length of the longest simple path (i.e. without loops) starting at p and not passing again through p. Show that for a tree, $d(p)$ assumes its least value either at a single vertex or at several adjacent vertices.
20. With every finite graph Γ one associates its *adjacency matrix* A, a square matrix whose rows and columns are indexed by the vertices of Γ and whose (i, j)-entry is the number of edges from i to j. Show that A is symmetric and that Γ is connected iff there is no permutation matrix P such that $P^{-1}AP$ is the diagonal sum of two matrices. Interpret the entries of A^n, for $n = 2, 3, \ldots$.
21. Define the *incidence matrix* of a finite graph $\Gamma = (V, E)$ as the matrix B whose rows are indexed by the vertices and columns indexed by the edges of Γ and whose (i, j)-entry is 1 if the i-th vertex is an endpoint of the j-th edge and 0 otherwise. Show that $BB^T - A$ (where T indicates the transpose) is a diagonal matrix and interpret its entries.

2 Groups

Our readers will have met groups before, so we shall be fairly brief in recalling the fundamentals, which occupy Sections 2.1–2.3. The remainder of this chapter deals with some notions of importance in elucidating the structure of groups, such as solubility, nilpotence (Section 2.4) and commutator subgroups (Section 2.5). In Section 2.6 we describe the constructions of Frattini and Fitting, which have their counterpart in rings in the form of the radical.

2.1 Definition and Basic Properties

We recall that a *group* is a set G on which a binary operation is defined, with values in G, denoted by $x.y$ or simply xy, the *product*, such that

G.1 $(xy)z = x(yz)$ for all $x, y, z \in G$ (associative law).
G.2 There exists an element $e \in G$ such that $xe = ex = x$ for all $x \in G$.
G.3 For each $x \in G$ there exists $x' \in G$ such that $xx' = x'x = e$.

The element e in **G.2** is uniquely determined by the equations in **G.2**; it is called the *neutral element* or *unit element*, and is usually denoted by e or 1. The element x' in **G.3** is uniquely determined by x; it is called the *inverse* of x and is denoted by x^{-1}. If the commutative law holds in G: $xy = yx$ for all $x, y \in G$, the group is said to be *abelian*. For an abelian group the additive notation is sometimes used for the operation; it is then called the *sum* and denoted by $x + y$, the neutral element is written 0 (and called the *zero element*), and the inverse of x is written $-x$.

As a consequence of **G.1**, the value of a repeated product $x_1x_2 \dots x_n$ is independent of bracketing, as long as the order of the factors is preserved. This follows easily by induction on n: for $n = 3$ it is just **G.1**; when $n > 3$, suppose that the product is bracketed in two different ways. We must show that

$$(x_1 \dots x_{i-1})(x_i \dots x_n) = (x_1 \dots x_{j-1})(x_j \dots x_n), \tag{2.1.1}$$

and we may suppose that $1 < i < j \leq n$. Moreover the portions inside the brackets are independent of the bracketing, by the induction hypothesis. Writing $u = x_1 \dots x_{i-1}, v = x_i \dots x_{j-1}, w = x_j \dots x_n$, we can rewrite (2.1.1) as $u(vw) = (uv)w$, and this is seen to follow from **G.1**. Thus $x_1x_2 \dots x_n$ is well-defined. In particular,

taking all the x_i equal to x, we obtain $xx \ldots x$ (n factors), which is denoted by x^n. This can be defined even for negative n by putting $x^{-n} = (x^{-1})^n$. We note the rules for any group element x and any $m, n \in \mathbf{Z}$:

$$(x^m)(x^n) = (x^n)(x^m) = x^{m+n}, \quad (x^m)^n = x^{mn}, \tag{2.1.2}$$

whose verification is straightforward. Similarly, in additive notation, $x + x + \ldots + x$ with n terms is written as nx. A group G is called *elementary abelian* if for some prime p, $px = 0$ for all $x \in G$. An example is the abelian group consisting of four elements e, a, b, c with $a^2 = b^2 = c^2 = e, ab = c$; it is known as the *Klein 4-group*.

Probably the most familiar example is the additive group of integers, with neutral element 0 and $-x$ as the inverse of x; another example is the multiplicative group of non-zero rational numbers with neutral element 1 and x^{-1} as the inverse of x. An even simpler example is the *trivial* group, consisting of a single element e with multiplication $ee = e$; the trivial group will usually be denoted by 1.

The groups mentioned so far are all abelian; to obtain a non-abelian group, recall that for any set S, by a *permutation* of S one understands a bijective mapping from S to itself. It is clear that for any set S the set $\Sigma(S)$ of all permutations of S is a group under composition of mappings, known as the *symmetric group* on S; the cardinal of S is called the *degree* of $\Sigma(S)$. It is not hard to see that any symmetric group of degree greater than two is non-abelian.

Let G be a group; a *subgroup* of G is a subset H of G which is a group relative to the operations of G. Thus H is a subgroup iff $1 \in H$ and for any $x, y \in H$, $xy, x^{-1} \in H$. An element c satisfying $c^n = 1$ for some $n \geq 1$ is said to be of *finite order*, and the least such n is called the *order* of c. Clearly in a finite group every element is of finite order, but there are also infinite groups with this property; they are usually called *torsion groups*. In an abelian group the set of all elements of finite order is a subgroup, called the *torsion subgroup*.

It is easily verified that the intersection of two subgroups is again a subgroup. More generally, this holds for any set of subgroups. Given a subset X of a group, the set consisting of 1 and all finite products of members of X and their inverses, briefly *group words* in X, forms a subgroup, denoted by $\mathrm{gp}\{X\}$ and called the subgroup *generated* by X. Clearly any subgroup containing X also contains $\mathrm{gp}\{X\}$; it follows that $\mathrm{gp}\{X\}$ is equal to the intersection of all subgroups containing X. If G is a group generated by a set X, then any equation of the form $f = g$, where f, g are group words in X, is called a *relation*. Let G be generated by X and let R be a set of relations such that every relation in the elements of X holding in G is a consequence of relations in R; then we write $G = \mathrm{gp}\{X|R\}$ and call this a *presentation* of G by X as *generating set* with R as set of *defining relations*. In a relation of the form $f = 1$, f is also called a *relator*. For example, the quaternions i, j generate a group of order 8, the *quaternion group*, which has the presentation $\mathrm{gp}\{a, b|a^4 = e, a^2 = b^2, b^{-1}ab = a^{-1}\}$. The group of motions of a regular n-gon, which consists of n rotations and n reflexions, has the presentation $\mathrm{gp}\{a, b|a^n = b^2 = e, b^{-1}ab = a^{-1}\}$; it is called the *dihedral group* of order $2n$ and is denoted by $\mathbf{D}_n$.

Let G, H be any groups. A mapping $f : \mathrm{G} \to \mathrm{H}$ is called a *homomorphism* if $(xy)f = xf.yf$ for all $x, y \in G$; it follows that $1_G f = 1_H$, where $1_G, 1_H$ denote the unit elements of G, H respectively. The set mapped to $1 : \{x \in G|xf = 1\}$ is a sub-

group called the *kernel* of f. When $H = G$, we call f an *endomorphism* of G. If $f : G \to H$ is bijective, it is called an *isomorphism* and G is said to be *isomorphic* to H; clearly f^{-1} is then an isomorphism from H to G. An isomorphism of G with itself is called an *automorphism*. For example, in any group G, the mapping $x \mapsto c^{-1}xc (c \in G)$ is an automorphism, the *inner automorphism* or *conjugation by c*.

A group generated by a single element is said to be *cyclic*. For example, the integers $\mathbf{Z}$ form a cyclic group under addition, with 1 as generator. Any cyclic group is a homomorphic image of $\mathbf{Z}$; the kernel is a subgroup of $\mathbf{Z}$, and hence of the form $m\mathbf{Z}$, for some $m \in \mathbf{Z}$, $m \geq 0$. Thus every cyclic group is either isomorphic to $\mathbf{Z}$ or of the form $\mathbf{Z}/m = \{x | mx = 0\}$ $(m \in \mathbf{N})$. In multiplicative notation this is the group of m-th roots of 1; so up to isomorphism there is one infinite cyclic group and one cyclic group of order m, for each positive integer m. It will usually be denoted by $\mathbf{C}_m$ or $\mathbf{Z}/m$.

For any group G, the subgroups of G that are maximal among the proper subgroups of G are simply called the *maximal subgroups*. For finitely generated (non-trivial) groups their existence is guaranteed by Zorn's lemma:

Proposition 2.1.1. *Let G be a finitely generated non-trivial group and H be a proper subgroup. Then there is a maximal subgroup containing H.*

Proof. Let X be a finite generating set of G and consider the set $\mathscr{P}$ of all proper subgroups containing H. Given any chain in $\mathscr{P}$, its union A is clearly a subgroup of G; if $A = G$, then $A \supseteq X$ and since X is finite, all the members of X belong to some member B of this chain; but this would mean that $B = G$, a contradiction. This shows that A must be proper, hence $\mathscr{P}$ is inductive and by Zorn's lemma it has a maximal member, which is the desired subgroup. ■

Given groups G, H, if $f : G \to H$ is a general homomorphism, the image Gf of G is easily seen to be a subgroup of H, the *homomorphic image* of G under f. The inverse image $1f^{-1}$ of the unit element of H, the *kernel* of f, written $\ker f$, is a *normal subgroup* of G, i.e. it is a subgroup such that $x^{-1}(\ker f)x = \ker f$ for all $x \in G$. This follows because if $af = 1$, then $(x^{-1}ax)f = (xf)^{-1}1(xf) = 1$ for all $x \in G$.

A group G with two subgroups H, K is said to be the *direct product* of H and K if every element $g \in G$ can be uniquely expressed as $g = xy = yx$, where $x \in H$, $y \in K$. It follows that H, K are both normal in G, $HK = G$ and $H \cap K = 1$. Conversely, any two subgroups H, K of G satisfying these conditions give rise to a direct product representation of G.

Let G be any group and H be a subgroup; any set of the form $Ha = \{ha | h \in H\}$ is called a *right coset* of H in G. If two cosets Ha, Hb have a common element $ua = vb$ (where $u, v \in H$), say, then for any $h \in H$, $ha = hu^{-1}vb$, $hb = hv^{-1}ua$ and it follows that $Ha = Hb$. Thus the right cosets of H in G form a partition of G. A *left coset* of H is similarly defined as $aH = \{ah | h \in \mathrm{H}\}$. We shall denote the set of right cosets of H in G by G/H and the set of left cosets by $H\backslash G$. In general left cosets define a different partition of G, but their number is the same, since the correspondence

$$Ha \leftrightarrow a^{-1}H \tag{2.1.3}$$

is independent of the choice of a in the coset: if a is replaced by ha, a^{-1} is replaced by $a^{-1}h^{-1}$. Thus (2.1.3) is a bijection between the coset spaces G/H and $H\backslash G$. A complete set of coset representatives is also called a *transversal.*

For a normal subgroup N the right and left cosets coincide, thus $G/N = N\backslash G$; more explicitly, we have

$$aN = Na \quad \text{for all} \quad a \in G; \tag{2.1.4}$$

'conversely, when (2.1.4) holds, the subgroup N is normal in G, as is easily verified. In that case we can define a multiplication on the set of cosets:

$$Na.Nb = Nab \quad (a, b \in G),$$

and it is easily checked that this defines a group on the set of cosets of N. It is denoted by G/N and is called the *quotient group* or *quotient* of G by N. Clearly the mapping $x \mapsto Nx$ defines a homomorphism from G to G/N which is surjective, with kernel N. We shall also write $N \lhd G$ to indicate that N is normal in G.

The following almost trivial remark is often useful:

Proposition 2.1.2 (Modular law). *If H, K, L are any subgroups of a group G, and $H \subseteq L$, then $H(K \cap L) = HK \cap L$.*

Proof. The inclusion $\subseteq$ is clear. Suppose now that $c \in HK \cap L$, say $c = ab$, where $a \in H$, $b \in K$. Then $a, c \in L$, hence $b = a^{-1}c \in L$, so $b \in K \cap L$ and $a \in H$, hence $c = ab \in H(K \cap L)$. ∎

Let G be a finite group; the number of its elements is called its *order* and is denoted by $|G|$. Generally we shall write $|X|$ for the number of elements in any subset X of G. Any subgroup H is again finite, and for any $a \in G$, Ha has as many elements as H, because the map $h \mapsto ha$ is a bijection. Since the right cosets form a partition of G, we have $|G| = n.|\mathrm{H}|$, where n is the number of right cosets; this number is called the *index* of H in G and is written $(G : H)$. This proves

Theorem 2.1.3.(Lagrange's theorem). *If G is any finite group and H is a subgroup of G, then*

$$|G| = (G : H).|H|. \tag{2.1.5}$$

In particular, the order of H divides the order of G. ∎

A group G is said to have *finite exponent* if there exists a natural number m such that $x^m = 1$ for all $x \in G$, and the least such m is called the *exponent* of G. Clearly every finite group has finite exponent; of course a group of finite exponent need not be finite, since it may not be finitely generated. A group is called *locally finite* if every finitely generated subgroup is finite, and the question whether every group of finite exponent is locally finite was raised by William Burnside in 1902 and is known as the *Burnside problem.* This question was answered negatively in 1968 by Sergei I. Adyan and Petr S. Novikov, but a restricted form of the Burnside problem (showing that there is a largest finite r-generator group of exponent e) was answered positively by Efim I. Zelmanov in 1989.

Let G be any group; by a *G-set* one understands a set S such that each $g \in G$ defines a permutation of S, $s \mapsto sg (s \in S, g \in G)$, also called a *group action* or *G-action*, such that $s(gh) = (sg)h, s1 = s$. The G-set is called *transitive* if for any $s, t \in S, t = sg$ for some $g \in G$. It is clear that any G-set can be described as the union of its transitive components sG, also called its *orbits*. For any $s \in S$ the set $G_s = \{g \in G | sg = s\}$ is easily verified to be a subgroup of G, called the *stabilizer* of s under the action of G. For example, if H is any subgroup of G, then the set of right cosets of H in G is a transitive G-set under the operation $Ha \mapsto Hag$; the stabilizer of Ha is $a^{-1}Ha$, in particular $H = H.1$ has the stabilizer H. It turns out that any transitive G-set can be described in this way. Let us call two G-sets S, T *isomorphic* if there is a bijection $\phi : S \to T$ such that $(s\phi)g = (sg)\phi$ for all $s \in S$, $g \in G$.

Theorem 2.1.4. *Given any group G and a subgroup H, the right coset space G/H is a transitive G-set with H as the stabilizer of a point. Moreover, any transitive G-set with H as the stabilizer of a point is isomorphic to G/H.*

Proof. Only the second part still needs proof, so let S be a transitive G-set with H as the stabilizer of $s \in S$. Then any $t \in S$ has the form $t = sg$ for some $g \in G$. We define a mapping $\phi : G/H \to S$ by the rule $Hg \mapsto sg$. It is well-defined, in fact it is a bijection, since $Hx = Hy \Leftrightarrow xy^{-1} \in H \Leftrightarrow sxy^{-1} = s \Leftrightarrow sx = sy$. Now $(Hx)y = Hxy$ and $(sx)y = s(xy)$, and this shows ϕ to be an isomorphism of G-sets. ∎

This result describes the orbits of any G-set. Since the size of G/H is the index $(G : H)$, we obtain the following orbit formula for the size of an orbit of a G-set sG with stabilizer G_s:

ORBIT FORMULA $$|sG| = (G : G_s). \tag{2.1.6}$$

When G is finite, this with (2.1.5) leads to the formula

$$|G| = |sG| . |G_s|. \tag{2.1.7}$$

We also note the formula for the number of orbits in a G-set:

Proposition 2.1.5. *Given a finite group G and a finite G-set S, denote by c_g the number of elements of S fixed by $g \in G$. Then the number of orbits of S is*

$$\omega = \frac{1}{|G|} \sum_{g \in G} c_g. \tag{2.1.8}$$

Proof. Consider the product set $S \times G$ and count the number of pairs (s, g) such that $sg = s$. In an orbit of r points, each point s is fixed by the members of its stabilizer, which has $|G|/r$ elements, by (2.1.7). So the orbit contributes $|G|$ pairs in all, and we therefore have $\Sigma c_g = |G| . \omega$, where ω is the number of orbits. But each $g \in G$ fixes just c_g points, so the number of pairs with $sg = s$ is Σc_g; now (2.1.8) follows on dividing by $|G|$. ∎

In any group G, two elements a, b are said to be *conjugate* if $c^{-1}ac = b$ for some $c \in G$. It is not hard to see that the conjugation mapping $a \mapsto c^{-1}ac$ is a group action by G on itself. The orbits under this action are called the *conjugacy classes* of G. The stabilizer of $g \in G$ is the set of elements commuting with g, called its *centralizer*:

$$Z_G(g) = \{x \in G | xg = gx\};$$

more generally, $Z_G(X)$, for any $X \subseteq G$ is the intersection of all the centralizers $Z_G(x)$, $x \in X$, and any element of $Z_G(X)$ is said to *centralize* X. In particular, $Z(G) = Z_G(G)$ is the set of all elements centralizing the whole of G; it is a normal subgroup, called the *centre* of G.

For any group G and a subset A of G, any $x \in G$ is said to *normalize* A if $A^x = x^{-1}Ax = A$. The set of all elements of G normalizing A is clearly a subgroup of G, the largest subgroup of G in which A is normal. It is called the *normalizer* of A in G and is written

$$N_G(A) = \{x \in G | A^x = A\}.$$

The orbit formula (2.1.6) shows that the number of elements in the conjugacy class containing g is the index in G of the centralizer of g. By Theorem 2.1.3 it follows that in a finite group G the number of elements in each conjugacy class divides the order of G. This allows us to obtain some information about certain groups. A group whose order is a positive power of a prime number p is called a *p-group*. Now we have

Theorem 2.1.6. *Any p-group has a non-trivial centre.*

Proof. Let G be a p-group and denote its conjugacy classes by $C_1, \ldots, C_r$. The unit element forms a conjugacy class, C_1 say, and each central element forms a separate conjugacy class, so if the centre of G is trivial, $C_1 = \{1\}$, while the number of elements in any other class is a power of p. By enumerating all elements of G we obtain the *class equation*

$$|G| = |C_1| + \ldots + |C_r|.$$

Here each $|C_i|$ for $i > 1$ is a positive power of p, as well as $|G|$, while $|C_1| = 1$, a contradiction, which shows the centre of G to be non-trivial. ■

Theorem 2.1.6 shows that every non-trivial p-group G has a central subgroup P of order p; clearly P is normal in G and so we can form G/P. Here the same argument applies, leading to a normal subgroup of G of order p^2. By induction on $|G|$ we obtain

Corollary 2.1.7. *Any p-group G has normal subgroups of all orders dividing* $|G|$. ■

A non-empty set S with a binary operation with values in S, which is associative, is called a *semigroup*; if S also has a neutral element, it is called a *monoid*. It is clear how

the notions of generating set, defining relations and presentation can also be defined for a monoid. A monoid S is said to have *right cancellation*, if

$$ac = bc \Rightarrow a = b \text{ for all } a, b, c \in S.$$

If $ca = cb \Rightarrow a = b$, S has *left cancellation*. Clearly every group has both left and right cancellation. Moreover a finite monoid S with left (or right) cancellation is a group. For the mapping $x \mapsto ax$ is an injective mapping of S into itself, and since S is finite, it is bijective (by the Box Principle, Section 1.1), thus $ax = 1$ has a solution a' for any $a \in S$. Hence $a'x = 1$ also has a solution, a'', say and $a = a(a'a'') = (aa')a'' = a''$; thus $aa' = a'a = 1$ and so S is indeed a group.

For each monoid S there is a 'universal group' which may be described as follows.

Proposition 2.1.8. *Let S be a monoid. Then there is a group $Q(S)$ with a homomorphism $\lambda : S \to Q(S)$ such that any homomorphism of S into a group G can be factored by λ, i.e. given a homomorphism $f : S \to G$, there exists $f' : Q(S) \to G$ such that $f = \lambda f'$; moreover, f' is uniquely determined by f.*

If S is commutative, then λ is injective if and only if S has cancellation.

Proof. To form $Q(S)$ we take a presentation of S as monoid and interpret it as presentation of a group. In other words, for each element x of S we introduce an inverse x^{-1} with the relations $x^{-1}x = xx^{-1} = 1$. Clearly $Q(S)$ is a group and by interpreting the elements of S as elements of $Q(S)$ we obtain the homomorphism λ. Now let $f : S \to G$ be a homomorphism to a group. For any $x \in S$ we define xf' as xf, while $(x^{-1})f' = (xf)^{-1}$; this provides a homomorphism from $Q(S)$ to G; clearly $\lambda f' = f$, and f' is unique, since it is prescribed on a generating set of $Q(S)$.

Suppose now that S is commutative. If λ is injective, S can be embedded in the group $Q(S)$ and so must have cancellation. Conversely, if S has cancellation, we define an equivalence relation on the set of pairs S^2 by putting

$$(a, b) \sim (a', b') \Leftrightarrow ab' = ba'.$$

This relation is clearly reflexive and symmetric; to verify transitivity, suppose that $(a, b) \sim (a', b'), (a', b') \sim (a'', b'')$; then $ab' = ba', a'b'' = b'a''$, hence $ab'b'' = ba'b'' = bb'a''$ and by cancellation, $ab'' = ba''$, i.e. $(a, b) \sim (a'', b'')$ as claimed. We denote the equivalence class of (a, b) by a/b and define multiplication by

$$a/b.a'/b' = aa'/bb'.$$

It is easily verified that this product depends only on the equivalence classes and not on the representatives chosen. The group laws can also be checked without difficulty; for example, the inverse of a/b is b/a, since $a/b.b/a = ab/ab = 1$. This group may be obtained by taking the elements $a/1$ $(a \in S)$ and their inverses $1/a$; thus it is nothing other than $Q(S)$, and S is embedded in $Q(S)$, for if $a/1 = a'/1$, then $a = a.1 = a'.1 = a'$. ∎

Exercises

1. Verify that the neutral element in a group is unique and that each element has a unique inverse.
2. Let G be a set with an element e and a binary operation $f(x, y)$ such that (i) $f(f(x, y), f(z, y)) = f(x, z)$, (ii) $f(x, x) = e$, (iii) $f(x, e) = x$, (iv) $f(f(x, y), f(e, y)) = x$. Show that G is a group with respect to the operation $xy = f(x, f(e, y))$.
3. Show that for any group G, $(x^{-1})^{-1} = x$, $(xy)^{-1} = y^{-1}x^{-1}$.
4. Verify the rules (2.1.2) for any group element x.
5. Show that any group G has a non-trivial proper subgroup unless G is trivial or of prime order.
6. Show that every group G can be defined as a G-set by the action $x \mapsto xg$. However the rule $x \mapsto gx$ does not define a G-action unless G is abelian. How can this rule (of left multiplication) be modified to produce a G-action for general G?
7. Let G be a group and S, T be any G-sets. Define a G-action on the Cartesian product $S \times T$ and compare the stabilizers of points $s \in S$, $t \in T$ and $(s, t) \in S \times T$.
8. (Poincaré's theorem) Let G be a group and H, K be subgroups of index r, s respectively in G. Show that $H \cap K$ has index at most rs in G. (Hint. Take transitive G-sets with H, K as stabilizers and consider their direct product.)
9. Let G be a group and S be a G-set. Define a G-action on $S \times S$ and compare the stabilizer of a point $(s, s) \in S \times S$ with that of $s \in S$.
10. Let G be a group and H be a subgroup of index r in G. Show that for any $g \in G$, $H \cap g^{-1}Hg$ has index at most $r(r-1)$ in G.
11. Let G be a group and H, K be subgroups. Show that $HK = \{hk | h \in H, k \in K\}$ need not be a subgroup, but that it is one whenever $HK = KH$.
12. Let G be a group. Show that any subgroup of index 2 is normal in G.

2.2 Permutation Groups

We have already met the symmetric group $\Sigma(X)$, consisting of all permutations of a set X. Any subgroup of $\Sigma(X)$ will be called a *permutation group* on X. This is a useful way of representing groups, which is of importance since it applies to all groups:

Theorem 2.2.1 (Cayley's theorem). *Every group is isomorphic to a group of permutations on a set.*

Proof. Let G be any group, and for each $g \in G$, define a mapping $\phi_g : x \mapsto xg$ of G into itself. This is easily seen to be a permutation, and $\phi_{gh} = \phi_g \phi_h$ by the associative law: $x(gh) = (xg)h$; moreover $\phi_g = 1$ iff $g = 1$, so that G is isomorphic to the group of permutations on G by right multiplication. ■

In particular, every finite group, of order n say, is isomorphic to a subgroup of Sym_n, the symmetric group on n symbols. These symbols are usually taken to be $1, 2, \ldots, n$; the number of permutations is $n!$, so that $|\mathrm{Sym}_n| = n!$. To write down

a permutation σ explicitly, one writes $1, 2, \ldots, n$ in one row and under i writes $i\sigma$. Thus for example

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 1 & 3 \end{pmatrix}. \tag{2.2.1}$$

However, a shorter notation, the *cycle notation* is often used, where we write the numbers $1, 1\sigma, 1\sigma^2, \ldots$ in a bracket, continuing until we get back to 1, then writing down the next number not yet written and again applying powers of σ and continuing in this way until all numbers from 1 to n and their images are recorded. Thus in this notation Equation (2.2.1) reads $[(1\ 3\ 4)(2)]^{-1} = (1\ 4\ 3)(2)$; of course any fixed symbols (like 2 in this case) may be omitted.

Another method of writing permutations is as *transpositions*, i.e. cycles of length 2. Every permutation can be written as a product of transpositions. If our permutation is $\sigma : i \mapsto i\sigma$, suppose first that $1\sigma = 1$; then σ is a permutation of $2, \ldots, n$ and by induction on n this can be expressed as a product of transpositions. If $1\sigma \neq 1$, then $k\sigma = 1$ for some $k \neq 1$; now consider the permutation of $2, \ldots, n$ which maps i to $i\sigma$ except for k, which is mapped to 1σ. This is again a product of transpositions and its product with $(1\ 1\sigma)$ is σ.

The number of transpositions needed to represent a given permutation can vary according to the method chosen, e.g. $(1\ 2\ 3) = (1\ 2)(1\ 3) = (1\ 3)(1\ 2)(2\ 3)(1\ 3)$. However, for a given permutation σ the number is either always odd or always even. This follows by applying σ to the variables $x_1, x_2, \ldots, x_n$ in the expression

$$\Delta(x_1, \ldots, x_n) = \prod_{i>j} (x_i - x_j). \tag{2.2.2}$$

We note that each transposition changes its sign, so the number of transpositions in any representation of σ has the same parity (even or odd). Thus we can with each permutation σ associate a sign $\operatorname{sgn}(\sigma)$, and it is clear that $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau)$; this result can be stated as follows:

Theorem 2.2.2 *Any permutation can be written as a product of transpositions. If σ is expressed as a product of r transpositions, define $\operatorname{sgn}(\sigma) = (-1)^r$. Then although r may vary for different representations of σ as a product of transpositions, $\operatorname{sgn}(\sigma)$ is independent of this representation and we have*

$$\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau). \qquad \blacksquare \tag{2.2.3}$$

From (2.2.3) we see that for any symmetric group Sym_n the mapping $\sigma \mapsto \operatorname{sgn}(\sigma)$ is a homomorphism from Sym_n to the cyclic group of order 2, whose kernel is the subgroup of even permutations. It is called the *alternating group* of *degree n* and is denoted by Alt_n.

Just as Sym_n is generated by transpositions, Alt_n can be generated by 3-cycles. In fact, a generating set for Sym_n is given by $\{(1\ 2), (1\ 3), \ldots, (1\ n)\}$, as is easily verified (see Exercise 1), while for Alt_n we have

Theorem 2.2.3. *The alternating group of degree n is generated by* $(1\ 2\ 3), (1\ 2\ 4), \ldots, (1\ 2\ n)$.

Proof. The above remark shows that Alt_n is generated by the set $\{(1\ i)(1\ j)\}$ for all $i \neq j,\ i, j = 2, \ldots, n$. Now $(1\ i)(1\ j) = (1\ i\ j) = (1\ 2\ j)(1\ 2\ i)(1\ 2\ j)^{-1}$, so the result follows. ■

By Theorem 2.1.4, Alt_n has order $n!/2$. It is easily checked that Alt_n is trivial for $n = 1$ or 2, while Alt_3 is cyclic of order 3. Alt_4 has order 12 and is easily seen to have a subgroup of order 4: $K_4 = \{1, (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$, the *Klein 4-group*. This subgroup is normal in Alt_4, indeed it is normal in Sym_4. However for $n > 4$, Alt_n has no normal subgroups other than itself and the trivial group, i.e. it is *simple* (see M. Hall (1959); also Section 7.11 below for a proof that Alt_5 is simple).

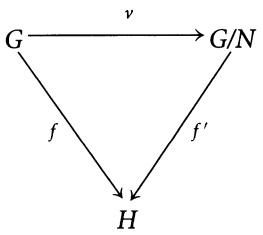
Exercises

1. Show that $\{(1\ 2), (1\ 3), \ldots, (1\ n)\}$ is a generating set of Sym_n. Show also that Sym_n has the generating set $\{(1\ 2), (2\ 3), \ldots, (n-1\ n)\}$. By examining the conjugates of $(1\ 2)$, show that G has a generating set consisting of $(1\ 2)$ and one other permutation.
2. Show that an n-cycle $(1\ 2 \ldots n)$ can be expressed as a product of $n-1$ transpositions, but no fewer.
3. Verify that the Klein 4-group K_4 is a normal subgroup of Sym_4 and show that Alt_4 has no normal subgroups other than itself, K_4 and 1.
4. Let G be an infinite group with a subgroup H of finite index. Show that G has a normal subgroup of finite index contained in H.
5. A permutation group is said to be *regular* if it is transitive (i.e. the set permuted has a single orbit) with trivial stabilizers. Show that every finite group can be represented as a regular permutation group. (Hint. Use Cayley's theorem.)

2.3 The Isomorphism Theorems

We have seen that for any group G and a normal subgroup N, the mapping $g \mapsto gN$ is a homomorphism from G onto G/N, the *natural homomorphism*, with kernel N. As a consequence we have the *factor theorem*:

Theorem 2.3.1. *Given a group homomorphism $f : G \to H$ and a normal subgroup N of G such that $N \subseteq \ker f$, there exists a unique mapping $f' : G/N \to H$ such that $f = vf'$, where $v : G \to G/N$ is the natural homomorphism, i.e. the triangle*



commutes. Moreover, f' is a homomorphism which is injective if and only if $N = \ker f$ and surjective if and only if f is surjective.

Proof. The mapping f', if it exists at all, must satisfy

$$(xN)f' = xf \quad (x \in G); \tag{2.3.1}$$

it follows that there is at most one such mapping. To show that it exists, we observe that xf is independent of the choice of x in its coset: if $xN = x'N$, then $x'x^{-1} \in N \subseteq \ker f$, whence $xf = x'f$. Thus (2.3.1) is well-defined, and clearly it is a homomorphism. The rest follows easily, since the cosets mapped to 1 by f are precisely the ones in $\ker f$. ■

This result can be applied to obtain an analysis of group homomorphisms. Given $f : G \to H$, by taking $N = \ker f$, we obtain an isomorphism $G/\ker f \to \mathrm{im} f$. This is the content of the *first isomorphism theorem* for groups:

Theorem 2.3.2. *Any group homomorphism $f : G \to H$ admits a factorization $f = \alpha f_1 \beta$, where $\alpha : G \to G/\ker f$ is the natural homomorphism, $\beta : \mathrm{im} f \to H$ is the inclusion map and $f_1 : G/\ker f \to \mathrm{im} f$ is the induced map, an isomorphism.* ■

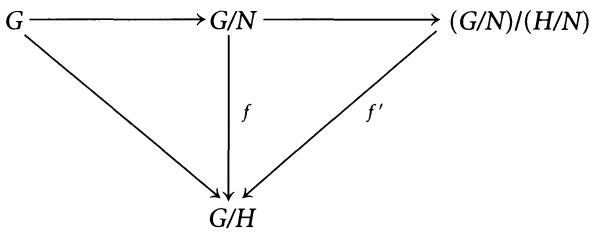
To state the *second isomorphism theorem*, also called the *parallelogram law*, we recall that for any subgroups H, K of G, HK is a subgroup whenever at least one of H, K is normal in G (so that $HK = KH$).

Theorem 2.3.3. *Let G be a group with subgroups H, N, where N is normal in G. Then $H \cap N$ is normal in H and*

$$H/(H \cap N) \cong HN/N. \tag{2.3.2}$$

Proof. Let $f : G \to G/N$ be the natural homomorphism and $\phi = f|H$ be its restriction to H. Then $\ker \phi = H \cap N$, $\mathrm{im}\, \phi = HN$, hence (2.3.2) follows by Theorem 2.3.2. ■

We shall use this result to compare the subgroup structure of a group with that of its quotient group. Let G be any group and N be a normal subgroup; the natural homomorphism $v : G \to G/N$ maps any subgroup H of G to the subgroup HN/N of G/N. In particular, any subgroup contained in N is mapped to the trivial subgroup of G/N, but if we restrict H to contain N, we obtain a bijection in this way, and it is



easily checked that H/N is normal in G/N iff H is normal in G. Thus we obtain the above commutative diagram.

The mapping f is defined by Theorem 2.3.1, because $N \subseteq H$, and since $\ker f = H/N$, we obtain f'. This mapping is a bijection, hence an isomorphism, and subgroups correspond under this isomorphism. The result is known as the *third isomorphism theorem:*

Theorem 2.3.4. *Let G be any group with a normal subgroup N. Then there is a natural bijection between the subgroups of G/N and the subgroups of G containing N, with H/N corresponding to H. Moreover, when H is normal in G, we have an isomorphism*

$$(G/N)/(H/N) \cong G/H. \qquad \blacksquare \qquad (2.3.3)$$

These results can be applied to obtain an analysis of groups, in particular, to derive the Jordan–Hölder theorem, which we shall state here without proof (see e.g. M. Hall (1959) and Exercise 7 below). For a given group G, a chain of subgroups

$$G = G_0 \supseteq G_1 \supseteq \ldots \supseteq G_n = 1, \qquad (2.3.4)$$

is called a *normal chain* if G_i is normal in G_{i-1} for $i = 1, \ldots, n$, while G_{i-1}/G_i is called a *factor* of G. By a *refinement* of (2.3.4) we understand a normal chain obtained from (2.3.4) by inserting further terms. This is always possible as long as some factor of (2.3.4) is not simple or trivial; when this is no longer possible, i.e. all factors are simple (and repetitions are omitted), (2.3.4) is called a *composition series.* Now we have

Theorem 2.3.5 (Jordan–Hölder theorem). *Every finite group G has a composition series; any two composition series of G have the same number of terms and the resulting factors are the same in both series, except for the order in which they occur.* ■

If we take a chain as in (2.3.4), where each G_i is normal in G (and not merely in G_{i-1}), we obtain what is called an *invariant chain*; a maximal invariant chain is called a *chief series* and the factors in a chief series are called the *chief factors* of G. Any two chief series of a finite group are again isomorphic and each chief factor is a direct power of a simple group; in particular in a soluble group each chief factor is an elementary abelian group (see M. Hall (1959) Chapter 8).

Camille Jordan proved in 1870 that factors in the two series could be paired off so that members of each pair had the same order, while Otto Hölder in 1889 established their isomorphism.

We also recall the Sylow theorems, proved by Ludvig Sylow in 1872 (see e.g. M. Hall (1959)); by a *Sylow p-subgroup* one understands a subgroup of order a power of p and of index prime to p. Thus a Sylow p-subgroup of G is a maximal p-subgroup in G.

Theorem 2.3.6. *Let G be a finite group and p be a prime dividing the order of G. Then G has a Sylow p-subgroup and any p-subgroup is contained in a Sylow p-subgroup. Further, the number of Sylow p-subgroups is congruent to $1 (\mathrm{mod}\, p)$ and they are all conjugate in G.* ■

Exercises

1. Let G be a group and H be a subgroup. Show that the largest normal subgroup of G contained in H is the intersection of all the conjugates of H.
2. Let G be a finite group and H be a subgroup whose order is not divisible by any prime $p < (G : H)$. Show that H is normal in G.
3. Find all composition series of Sym_n.
4. If G is a simple group with a subgroup of index $n > 1$, show that G can be represented as a permutation group of degree n. Deduce that G is finite and obtain a bound on its order.
5. Let G be a group and X be a subset of G and define the centralizer $Z_G(X)$ and normalizer $N_G(X)$ of X in G as in Section 2.1. Verify that the action $x \mapsto g^{-1}xg$ defines an action of the normalizer on X and hence a homomorphism $N_G(X) \to \Sigma(X)$ and determine its kernel.
6. (Zassenhaus lemma) Let G be a group. Given subgroups H, K and normal subgroups H' of H and K' of K, show that $(H' \cap K)K'$ is normal in $(H \cap K)K'$ and $(H \cap K)K'/(H' \cap K)K' \cong (H \cap K)/(H \cap K')(H' \cap K)$. Deduce that $(H \cap K)K'/(H' \cap K)K' \cong (H \cap K)H'/(H \cap K')H'$.
7. (Schreier refinement theorem) Show that any two normal chains of a group have refinements that are isomorphic. (Hint. Use the Zassenhaus lemma to project each chain on to the other.) Use this result to deduce the Jordan–Hölder theorem.
8. In any group G let H, K be normal subgroups such that $H \cap K = 1$. Show that H and K commute elementwise, i.e. $xy = yx$ for all $x \in H$, $y \in K$.

2.4 Soluble and Nilpotent Groups

In an abelian group every subgroup is normal, hence in a finite abelian group every element generates a normal subgroup, and by taking an element of prime order we obtain a normal subgroup of prime order. Thus every finite abelian group has a normal chain with factors all of prime order, hence a composition series:

$$G = G_0 \supset G_1 \supset \ldots \supset G_r \supset 1. \tag{2.4.1}$$

We have the following property of finitely generated abelian groups:

Theorem 2.4.1 (Basis theorem for abelian groups). *Every finitely generated abelian group can be written as a direct product of cyclic groups of prime power order or infinite cyclic, and the summands are determined up to isomorphism and order.*

Proof. We shall write our group in additive notation, so we have to express it as a direct sum of cyclic groups of infinite or prime power order. We first assume that we have an abelian group A of finite order $n = q_1 \dots q_r$, where the q_i are powers of distinct primes. It is not hard to check that the elements of order dividing q_i form a subgroup A_i say. Since $n/q_1, \dots, n/q_r$ are integers without a common factor, there exist integers $t_1, \dots, t_r$ such that $t_1 n/q_1 + \dots + t_r n/q_r = 1$. Any $c \in A$ can be expressed as $c = c_1 + \dots + c_r$, where $c_i = t_i n/q_i . c$. Clearly $q_i c_i = 0$, and it follows that $A = A_1 + \dots + A_r$; moreover, this sum is direct, because any element of $A_2 + \dots + A_r$ has order prime to q_1, hence $(A_2 + \dots + A_r) \cap A_1 = 0$, and similarly for the other terms.

It remains to show that a finite abelian p-group A is a direct sum of cyclic groups, and the number of terms of a given order is uniquely determined. If A is cyclic, there is nothing to prove. Otherwise let B be a cyclic subgroup of maximal order p^r say, generated by b. Then there is a subgroup C of order p such that $B \cap C = 0$. For if $c \in A \setminus B$ has order p^s mod B, then $s > 0$. We have $p^s c = mb$ and $p^{r-s} mb = p^r c = 0$, hence $p^r | p^{r-s} m$. It follows that $p | m$, say $m = pm'$. Now $c' = p^{s-1} c - m' b$ has order p and $c' \notin B$, by the choice of s, hence $C = \mathrm{gp}\{c'\}$ is the desired subgroup.

We next show that

$$A = B \oplus D, \tag{2.4.2}$$

for some subgroup D of A. Let us put $\overline{A} = A/C$, where C is the subgroup just found. In $\overline{A}$ the image $\overline{B}$ of B has the same order as B and so is cyclic of maximal order in $\overline{A}$. By induction on $|A|$ there is a subgroup $\overline{D}$ of $\overline{A}$ such that $\overline{A} = \overline{B} \oplus \overline{D}$. Hence $A = B + D$, where D is the inverse image of $\overline{D}$ in A. Moreover, $B \cap D \subseteq C$, but $B \cap C = 0$, therefore $B \cap D = 0$, and (2.4.2) holds. Now an induction on $|A|$ yields the desired decomposition. If the number of elements of order dividing p^i is n_i, then there are n_i/n_{i-1} summands of order p^i.

Suppose next that A is a finitely generated abelian group with no elements of finite order (apart from 0). Take a generating set $a_1, \dots, a_n$ and suppose that there is a non-trivial relation $\sum c_j a_j = 0$. Since there are no non-zero elements of finite order, the c_j are relatively prime and so form the first row of an invertible matrix $C = (c_{ij})$. We put $b_i = \sum c_{ij} a_j$; then the b_i clearly again form a generating system, but $b_1 = \sum c_{1j} a_j = 0$, so A is generated by $b_2, \dots, b_n$. By taking a generating set of least cardinal we conclude that A is free abelian, i.e. a direct sum of infinite cyclic groups, and the rank, i.e. the number of free generators, is unique; it is r, where $[A : 2A] = 2^r$.

Finally, let A be any finitely generated abelian group and let A' be its torsion subgroup; then A/A' has no elements of finite order and is therefore free abelian. Let $a_1, \dots, a_r \in A$ be elements mapping to a basis of A/A' and denote the subgroup generated by A''. Then $A = A' \oplus A''$; by the first part A' is a direct sum of finite cyclic groups of prime power orders and by the second part A'' is a direct sum of infinite cyclic groups, so we have reached the required decomposition. ∎

In the proof we saw that every finite abelian group can be written as a direct sum of p-groups for the different primes dividing the order. Conversely, it is easy to verify that a direct sum of a cyclic p-group and a cyclic p'-group, where p, p' are different primes, is again a cyclic group. By collecting up all cyclic prime power groups of highest order, then those of next highest and so on, we obtain

Corollary 2.4.2. *Every finite abelian group can be written as a direct sum of cyclic groups: $A = B_1 \oplus \ldots \oplus B_r$, where the order of B_i divides that of $B_{i+1} (i = 1, \ldots, r-1)$.* ∎

In general, even for finite groups, a composition series does not have the simple form (2.4.1), as is shown by the existence of non-abelian simple groups. One therefore defines a group G to be *soluble* if it has a normal chain (2.4.1) such that each factor G_{i-1}/G_i is abelian. Now it is clear that a finite soluble group has a composition series in which all factors are of prime order, and conversely, if a finite group G has a composition series with all factors of prime order, then G is soluble.

In order to test a group for solubility one can form the largest abelian quotient. Given G and a normal subgroup N, the quotient G/N is abelian iff $xy \equiv yx (\mathrm{mod}\, N)$, i.e. iff

$$(x, y) = x^{-1}y^{-1}xy \in N \text{ for all } x, y \in G. \tag{2.4.3}$$

The element (x, y) defined by (2.4.3) is called the *commutator* of x and y and the subgroup generated by all commutators is called the *commutator subgroup* or *first derived group* of G and is denoted by G'. It is easily seen to be normal in G and in fact G/G' is the largest abelian quotient of G, in the sense that any normal subgroup N of G such that G/N is abelian satisfies $N \supseteq G'$. Thus for any group G we can form the *derived series*

$$G \supseteq G' \supseteq G'' \supseteq \ldots \supseteq G^{(s)}. \tag{2.4.4}$$

This series terminates when we reach a group that is *perfect*, i.e. equal to its derived group. We note

Theorem 2.4.3. *A group is soluble if and only if its derived series ends in 1 after a finite number of steps.*

Proof. Let G be a soluble group with a normal chain (2.4.1) whose factors are abelian. It follows that $G' \subseteq G_1$ and by induction on r we find that $G^{(r)} \subseteq G_r = 1$, hence $G^{(r)} = 1$. Conversely, when $G^{(r)} = 1$ for some r, then (2.4.4) (with $s = r$) is a normal chain with abelian factors, hence G is soluble. ∎

The number of steps in the derived series of a soluble group is called its *derived length.*

Let G be a group. A invariant chain $\{G_n\}$ such that G_{i-1}/G_i lies in the centre of G/G_i for $i = 1, \ldots, r$, is called a *central chain*, and G is called *nilpotent* if it has a central chain starting at G and ending in 1. Clearly every nilpotent group G is

soluble, but the converse is false, e.g. Sym_3 is soluble but not nilpotent (since its centre is trivial). If we put $\gamma_1(G) = G$ and for $i > 1$, define recursively $\gamma_i(G) = (G, \gamma_{i-1}(G)) = \mathrm{gp}\{(x, y)|x \in G,\ y \in \gamma_{i-1}(G)\}$, where $(x, y) = x^{-1}y^{-1}xy$, then we have the central chain

$$G = \gamma_1(G) \supseteq \gamma_2(G) \supseteq \ldots, \tag{2.4.5}$$

which has the property that for any other central chain $G = H_1 \supseteq H_2 \supseteq \ldots$, $H_i \supseteq \gamma_i(G)$. Thus $\{\gamma_i(G)\}$ is the *lower central series.* This shows in particular that G is nilpotent iff $\gamma_{r+1}(G) = 1$ for some $r \geq 0$. The least such r is called the (precise) *class of nilpotence* of G.

There is also an *upper central series* $\{Z_i(G)\}$ defined recursively by setting $Z_0(G) = 1$ and for $i \geq 1$ defining $Z_i(G)$ by the condition that $Z_i(G)/Z_{i-1}(G)$ is the centre of $G/Z_{i-1}(G)$. If G *has the central chain*

$$G = H_1 \supseteq H_2 \supseteq \ldots \supseteq H_{r+1} = 1,$$

then

$$Z_i(G) \supseteq H_{r+1-i}. \tag{2.4.6}$$

For we have $Z_1(G) \supseteq H_r$ and if (2.4.6) holds for some $i \geq 1$, then $Z_{i+1}(G) = \{x \in G|(G, x) \subseteq Z_i(G)\}$ and $(G, H_{r-i}) \subseteq H_{r+1-i} \subseteq Z_i(G)$, so $H_{r-i} \subseteq Z_{i+1}(G)$, and this shows that (2.4.6) holds generally.

In Theorem 2.1.6 we saw that every p-group has a non-trivial centre; in fact the argument leading to Corollary 2.1.7 shows that every p-group has an upper central series with quotients of order p. It follows that the group is nilpotent; thus we obtain

Theorem 2.4.4. *Every p-group is nilpotent.* ■

Given two groups G_1, G_2, we can form a group P which is their direct product by taking their Cartesian product $P = G_1 \times G_2$ and defining the multiplication componentwise: $(x, y)(x', y') = (xx', yy')$. This construction is sometimes called the *external direct product* in contrast to the earlier notion, which is the *internal direct product.* Of course the external direct product $G_1 \times G_2$ may be regarded as the internal direct product of $G_1^* = \{(x, 1)|x \in G_1\}$ and $G_2^* = \{(1, y)|y \in G_2\}$.

It is clear how the notion of direct product can be extended to any finite number of factors. As an illustration we have the following result.

Theorem 2.4.5. *A finite group G is the direct product of its Sylow subgroups if and only if every Sylow subgroup is normal in G.*

Proof. If G is the direct product of its Sylow subgroups, these subgroups must clearly be normal in G. Conversely, suppose that all the Sylow subgroups are normal in G. Then for each prime dividing $|G|$ there is just one Sylow subgroup, the different Sylow subgroups meet in 1 and hence commute elementwise. Any element of G can be written uniquely as a product of elements of prime power orders for different primes, and it follows that there is a direct product representation with the Sylow subgroups as factors. ■

In a direct product representation $G = H \times K$, the elements of H commute with those of K; sometimes we have a slightly more general situation, where there are two subgroups H, K of G such that $G = HK$, $H \cap K = 1$ but only K is normal in G. It is still true that each element of G can be expressed uniquely as a product xy, where $x \in H$, $y \in K$, but now H, K no longer commute elementwise. Thus the product rule now becomes

$$(xy)(x'y') = xx'y^{x'}y' \quad \text{where} \quad x, x' \in H, y, y' \in K, y^{x'} = x'^{-1}yx'. \tag{2.4.7}$$

The group G is then called the (internal) *semidirect product* of H and K, with H acting on K and we write

$$G = H \ltimes K \quad \text{or} \quad G = K \rtimes H. \tag{2.4.8}$$

For example, the symmetric group of degree three is a semidirect product of $\mathbf{C}_2$ by $\mathbf{C}_3$, with $\mathbf{C}_2$ acting on $\mathbf{C}_3$: $\mathrm{Sym}_3 = \mathbf{C}_3 \rtimes \mathbf{C}_2$. Here we must specify the action of one factor on the other; since $\mathbf{C}_3$ has only one automorphism of order 2, there is only one choice of non-trivial action.

Given two groups H, K, with an action of H on K by automorphisms, we can form their semidirect product as follows. Let us denote the action by y^x, where $x \in H$, $y \in K$; now the multiplication on the set $H \times K$ is given by (2.4.7). Of course it has to be checked that the multiplication is associative, there is a neutral element $(1_H, 1_K)$ and each element (x, y) has an inverse, namely $(x^{-1}, (y^{-1})^{x^{-1}})$; this is a routine verification, which may be left to the reader. The resulting group is again denoted as in (2.4.8) and is called the *external semidirect product*, in contradistinction to the *internal* semidirect product defined earlier.

Exercises

1. Show that any homomorphism of groups, $f : G \to H$ maps G' into H', where $'$ indicates the derived group.
2. Show that a group has a composition series with all factors of prime order iff it is finite soluble.
3. Show that a group G is the direct product of its subgroups H, K whenever both H and K are normal in G, $H \cap K = 1$ and $HK = G$.
4. Show that a group G with a normal subgroup N is soluble iff both N and G/N are soluble. Deduce that any direct product of soluble groups is soluble.
5. Let G be a group, p be a prime and S be a Sylow p-subgroup. If S is the only Sylow p-subgroup of G, show that S is normal in G.
6. Show that a direct product of nilpotent groups is nilpotent. Deduce that any group which is the direct product of its Sylow p-subgroups is nilpotent.
7. Show that in a finite nilpotent group G every Sylow subgroup is normal and deduce that G is the direct product of its Sylow subgroups.
8. Show that in any finite group G the inner automorphisms form a permutation group P acting on G, and P is a homomorphic image of G. For any prime p, show that P has as many Sylow p-subgroups as G itself.

9. Verify that the external semidirect product, as defined in the text, satisfies all the group properties.
10. Let A be a group acting by automorphisms on a group G, such that in the semidirect product $(G, A) = G$. Show that if $N \lhd G$ is such that $(A, N) = 1$, then N is contained in the centre of G.
11. Let G be a soluble finite group. Show that any minimal subgroup has prime order and any minimal normal subgroup is elementary abelian, i.e. a direct power of a group of prime order.

2.5 Commutators

In (2.4.3) we met the commutator, an important concept, which is a useful tool for studying group properties. Below we give some basic formulae and a few simple applications.

Given a group G and $x, y \in G$, we recall that the conjugate of x by y is $x^y = y^{-1}xy$ and the commutator of x and y is $(x, y) = x^{-1}y^{-1}xy$. We note the following general formulae:

$$x^y = x(x, y), xy = yx(x, y), \tag{2.5.1}$$

$$(x, y)^{-1} = (y, x) = (x^{-1}, y)^x, \tag{2.5.2}$$

$$(xy, z) = (x, z)^y(y, z) = (x, z)((x, z), y)(y, z), \tag{2.5.3}$$

$$(x, yz) = (x, z)(x, y)^z = (x, z)(x, y)((x, y), z). \tag{2.5.4}$$

The proofs are straightforward; we have e.g. for (2.5.2): $(x^{-1}, y)^x = x^{-1}(xy^{-1}x^{-1}y)x = y^{-1}x^{-1}yx = (y, x)$, and for (2.5.4): $(x, yz) = x^{-1}z^{-1}y^{-1}xyz = x^{-1}z^{-1}xz.z^{-1}x^{-1}y^{-1}xyz = (x, z)(x, y)^z$.

We further note

Lemma 2.5.1. *Let G be any group and $x, y \in G$.*

(i) *If (x, y) commutes with x, then*

$$(x^n, y) = (x, y)^n \quad \text{for all } n \in \mathbf{Z}. \tag{2.5.5}$$

(ii) *If (x, y) commutes with x and y, then*

$$(xy)^n = x^n y^n (y, x)^{n(n-1)/2} \quad \text{for all } n \geq 0. \tag{2.5.6}$$

Proof. (i) By (2.5.3) we have $(x^{n+1}, y) = (x, y)^{x^n}(x^n, y)$; since x commutes with (x, y), we can omit the exponent. Now (2.5.5) follows for $n \geq 0$ by induction; for negative n, say $n = -m$, we have $(x^{-m}, y) = (x, y)^{-m}$ by (2.5.2).

(ii) For $n = 0$ or 1, (2.5.6) reduces to an obvious identity. If it holds for some $n \geq 1$, then on writing $s = n(n-1)/2$, we have

$$\begin{aligned}(xy)^{n+1} = (xy)^n(xy) &= x^n y^n (y,x)^s (xy)\\ &= x^n y^n (xy)(y,x)^s\\ &= x^{n+1} y^n (y^n, x) y (y,x)^s\\ &= x^{n+1} y^n (y,x)^n y (y,x)^s\\ &= x^{n+1} y^{n+1} (y,x)^{s+n},\end{aligned}$$

which is the required formula, since $s + n = n(n-1)/2 + n = (n+1)n/2$. ∎

Higher commutators are defined recursively by putting

$$(x_1, x_2, \ldots, x_n) = ((x_1, \ldots, x_{n-1}), x_n). \tag{2.5.7}$$

Thus a higher commutator within a single pair of brackets is understood to be *left-normed*, i.e. all brackets begin on the left of the first argument. We remark that the identities between commutators correspond to identities in a Lie algebra, but this correspondence will not be pursued further here. For the moment we note an analogue of the Jacobi identity:

Proposition 2.5.2 (Witt identity). *In any group G*

$$(x, y^{-1}, z)^y (y, z^{-1}, x)^z (z, x^{-1}, y)^x = 1. \tag{2.5.8}$$

Proof. We have $(x, y^{-1}, z)^y = y^{-1}(yx^{-1}y^{-1}x)z^{-1}(x^{-1}yxy^{-1})zy = x^{-1}y^{-1}xz^{-1}x^{-1}.yxy^{-1}zy$. This has the form $u^{-1}v$, where $u = xzx^{-1}yx$, $v = yxy^{-1}zy$. If we put $w = zyz^{-1}xz$, then the left-hand side of (2.5.8) can be written $u^{-1}v.v^{-1}w.w^{-1}u$, which reduces to 1. ∎

For any subgroups H, K of a group G we define the commutator subgroup (H, K) as the subgroup generated by all commutators (x, y), where $x \in H$, $y \in K$. Thus the general element of (H, K) is a product of commutators, but need not itself be a commutator. Higher commutator subgroups are again defined recursively as left-normed products by

$$(H_1, \ldots, H_n) = ((H_1, \ldots, H_{n-1}), H_n). \tag{2.5.9}$$

Of course the same groups are obtained by using right-normed products, as can easily be verified by means of (2.5.2). If each H_i is normal in G, then (2.5.9) is the subgroup generated by all $(x_1, \ldots, x_n)$, where $x_i \in H_i$, but in general equality need not hold. We note the following obvious properties of commutator subgroups, whose proof is left to the reader:

Proposition 2.5.3. *If G is any group with subgroups H, K, then*

(i) $(H, K) = (K, H)$;
(ii) (H, K) *is normal in the group generated by H and K;*
(iii) (H, K) *is normal in H if and only if $K \subseteq N_G(H)$, the normalizer of H in G;*

(iv) $(H, K) = 1$ *if and only if* H *and* K *commute elementwise;*
(v) *for any homomorphism* φ *of* G *into another group,* $(H, K)^{\varphi} = (H^{\varphi}, K^{\varphi})$. ■

From Witt's identity we can easily deduce a result known as Philip Hall's three-subgroup lemma:

Lemma 2.5.4. *Given three subgroups* A, B, C *of a group, if* $(A, B, C) = (B, C, A) = 1$, *then* $(C, A, B) = 1$.

Proof. Let $x \in A$, $y \in B$, $z \in C$. By hypothesis, $(x, y^{-1}, z)^y = (y, z^{-1}, x)^z = 1$, hence Witt's identity shows that $(z, x^{-1}, y)^x = 1$, so $(z, x^{-1}, y) = 1$. This means that (z, x^{-1}) centralizes B, i.e. $((C, A), B) = 1$. ■

Nilpotent groups may be characterized in various ways. First we note a property of normalizers of Sylow subgroups:

Proposition 2.5.5. *Let* G *be a finite group,* H *be a subgroup and* P *be a Sylow subgroup of* H. *If* $N_G(H)$ *is the normalizer of* H *in* G *and* P^* *is the normalizer of* P *in* $N_G(H)$, *then*

$$N_G(H) = P^*H. \tag{2.5.10}$$

Proof. Clearly $N_G(H) \supseteq P^*H$. To prove the converse, take $a \in N_G(H)$, so that $H^a = H$. Then P and P^a are Sylow subgroups of $H = H^a$ for the same prime, hence they are conjugate in H, i.e. $P^a = P^b$ for some $b \in H$. Since $a \in N_G(H)$, it follows that $c = ab^{-1} \in P^*$, and so $a = cb \in P^*H$, as claimed. ■

We note two consequences. If H is normal in G, the left-hand side of (2.5.10) becomes G, while the right-hand side is $N_G(P)H$; so we have

Corollary 2.5.6. *If* G *is a finite group,* H *is a normal subgroup and* P *is a Sylow subgroup of* H, *then* $G = N_G(P)H$. ■

This reasoning is often described as the *Frattini argument.*

Secondly we take P to be a Sylow subgroup of G and $H \supseteq N_G(P)$; then the right-hand side of (2.5.10) is H and we obtain

Corollary 2.5.7. *Let* G *be a finite group with a Sylow subgroup* P. *Then any subgroup containing the normalizer of* P *is its own normalizer.* ■

We now have the following characterizations of a nilpotent group:

Theorem 2.5.8. *For any finite group the following conditions are equivalent:*

(a) G *is nilpotent;*
(b) $\gamma_{r+1}(G) = 1$ *for some* $r \geq 0$;
(c) $Z_r(G) = G$ *for some* $r \geq 0$;

(d) *every proper subgroup is distinct from its normalizer;*
(e) *every maximal subgroup is normal in G.*

Moreover, the least r in (b) *and* (c) *are the same and are equal to the precise class of nilpotence of G.*

Here 'maximal' is of course understood to mean 'maximal among the proper subgroups'.

Proof. We have already seen the equivalence of (a), (b), (c) and the assertion about the precise class of nilpotence. Now assume that G is nilpotent with a central series

$$G = H_1 \supset H_2 \supset \ldots \supset H_{r+1} = 1, \tag{2.5.11}$$

and let K be a proper subgroup of G, say $H_i \subseteq K, H_{i-1} \not\subset K$. Then we have

$$(K, H_{i-1}) \subseteq (G, H_{i-1}) \subseteq H_i \subseteq K,$$

hence $H_{i-1} \subseteq N_G(K)$. If we take $x \in H_{i-1} \setminus K$, then $K^x = K$ and it follows that $N_G(K) \supset K$, i.e. (d) holds. Clearly (d) $\Rightarrow$ (e), so assume (e). If P is any Sylow subgroup of G and P is not normal, then $N_G(P) \neq G$, hence there is a maximal subgroup $M \supseteq N_G(P)$ and M is its own normalizer by Corollary 2.5.7, but this contradicts (e). Thus all Sylow subgroups are normal in G, so G is the direct product of its Sylow subgroups and hence is nilpotent. ■

Exercises

1. Fill in the details in the proof of (2.5.1)–(2.5.4).
2. Prove Proposition 2.5.3.
3. Show that a group satisfying the law $x^2 = 1$ is abelian. Deduce that any commutator can be written as a product of squares and find an expression of (x, y) in this form.
4. Let G be a group and K, L, N be subgroups such that $N \triangleleft K$, $K \triangleleft G$, $(L, K) = K$, $KL = G$, $(L, N) = 1$. Show that N lies in the centre of K.
5. Prove the identity $((x, y), z^x)((z, x), y^z)((y, z), x^y) = 1$.
6. Show that if $H_1, \ldots, H_n$ are normal subgroups of a group G, then $(H_1, \ldots, H_n)$ is the subgroup generated by all commutators $(x_1, \ldots, x_n)$, where $x_i \in H_i$ and give an example to show that normality cannot be omitted.
7. (H. Wielandt) Give a direct proof that a finite group is nilpotent iff every maximal subgroup is normal.
8. Let G be a group and $a \in G$. For any $x \in G$ write $x\alpha = x^{-1}ax, x\lambda = (a, x)$. Show that for $n = 1, 2, \ldots, x\alpha^n = a.x\lambda^n$.
9. Show that in a symmetric group every cycle of odd length is a commutator (and not merely a product of commutators).
10*. (D. Ornstein) Show (by induction on r) that $(a, b)^r$ can be written as a product of $(ba)^{-r}(ab)^r$ and $r - 1$ commutators. Deduce that if in a group G, the centre of G has finite index in G, then the derived group G' is finite.

2.6 The Frattini Subgroup and the Fitting Subgroup

Let G be any group and denote by $\Phi(G)$ the intersection of all its maximal (proper) subgroups; of course, if there are no such subgroups, then $\Phi(G) = G$. It is clear that $\Phi(G)$ is a subgroup of G, which is *characteristic*, i.e. invariant under all automorphisms of G. It is called the Φ-*subgroup* or *Frattini subgroup*, after Giovanni Frattini, who introduced it in 1885. Clearly any non-trivial finite group has maximal subgroups, so in that case $\Phi(G) \subset G$. In fact this holds for any finitely generated (non-trivial) group, by Proposition 2.1.1.

There is another characterization of $\Phi(G)$ which is often useful. In any group G, an element c will be called a *non-generator* if any subset of G which together with c generates G is itself a generating set of G.

Proposition 2.6.1. *Let G be any group. Then $\Phi(G)$ consists precisely of all the non-generators of G.*

Proof. Let $c \in G$ be a non-generator. If M is a maximal subgroup, then $c \in M$, for otherwise the subgroup generated by $M \cup \{c\}$ would be larger than M, hence equal to G, and so by definition of c as non-generator, $M = G$, which is false. Thus c lies in every maximal subgroup, and hence in $\Phi(G)$.

Conversely, if $c \in \Phi(G)$ and $X \cup \{c\}$ generates G, we have to show that X generates G. If this is not so, then $\mathrm{gp}\{X\}$ is proper in G, and hence, by Zorn's lemma, there is a subgroup M containing X but not c and maximal with those properties. Since $c \notin M$, $M \neq G$, but any subgroup properly containing M must also contain c and hence be equal to G. Thus M is maximal in G, but $c \in \Phi(G) \subseteq M$, which is a contradiction. Hence X generates G and c is indeed a non-generator. ∎

The property of elements described in Proposition 2.6.1 can be extended to subgroups as follows:

Proposition 2.6.2. *Let G be a group and U be a subgroup. Then $U \subseteq \Phi(G)$ whenever U satisfies the following condition:*

$$\text{For any proper subgroup } H \text{ of } G,\ \mathrm{gp}\{U, H\} \subset G. \tag{2.6.1}$$

When G is finitely generated, this condition is necessary as well as sufficient. In particular, we have $H\Phi(G) \subset G$ for every proper subgroup H of G.

Proof. If (2.6.1) holds, then in particular, for any maximal subgroup M of G, $M \subseteq \mathrm{gp}\{U, M\} \subset G$, hence $M = \mathrm{gp}\{U, M\}$ and so $U \subseteq M$; therefore $U \subseteq \Phi(G)$. Now assume that G is finitely generated; if $U \subseteq \Phi(G)$, and H is a proper subgroup of G, then there is a maximal subgroup M containing H, but also $U \subseteq M$, hence $\mathrm{gp}\{U, H\} \subseteq M \subset G$, and (2.6.1) is satisfied. ∎

We shall also need another property of Φ.

Lemma 2.6.3. *Let G be any group and N be a finitely generated normal subgroup. Then $\Phi(N) \subseteq \Phi(G)$.*

Proof. $\Phi(N)$ is a characteristic subgroup of N, hence normal in G. Now let M be a maximal subgroup of G and suppose that $\Phi(N) \not\subset M$. Then $\Phi(N)M = G$, hence by the modular law, since $\Phi(N) \subseteq N$,

$$N = \Phi(N)M \cap N = \Phi(N)(M \cap N).$$

Thus $M \cap N$ together with $\Phi(N)$ generates N; hence by Proposition 2.6.2, $M \cap N = N$, i.e. $M \supseteq N$ and so $\Phi(N) \subseteq M$, a contradiction. It follows that $\Phi(N) \subseteq M$ for any maximal M, and so we find that $\Phi(N) \subseteq \Phi(G)$. ■

We remark that without the normality of N this result need not hold, as is shown by the case of a simple group (whose Φ-subgroup is trivial) and a cyclic subgroup.

Lemma 2.6.4. *Let G be a finite group and N, K be normal subgroups such that $N \subseteq K \cap \Phi(G)$. If K/N is nilpotent, then so is K.*

Proof. Let P be a Sylow p-subgroup of K. Then PN/N is a Sylow p-subgroup of K/N, because $PN/N \cong P/(P \cap N)$ and this again has p-power order, while its index in K is prime to p. Since K/N is nilpotent, PN/N is normal in K/N, therefore characteristic in K/N and so normal in G/N, hence $PN \lhd G$. Now P is a Sylow p-subgroup of PN; hence by Corollary 2.5.6,

$$G = PN.N_G(P) = N.N_G(P).$$

Now $N \subseteq \Phi(G)$ by hypothesis, so by Proposition 2.6.2, $N_G(P) = G$, and so $P \lhd G$. Thus $P \lhd K$ and this holds for all Sylow subgroups of K, hence by Theorem 2.4.5, K is nilpotent, as claimed. ■

We can now establish a useful criterion for nilpotency in terms of Φ.

Proposition 2.6.5 (Wielandt). *Let G be a finite group and N be a normal subgroup of G. Then N is nilpotent if and only if $N' \subseteq \Phi(G)$.*

Proof. If $N' \subseteq \Phi(G)$, then $N' \lhd G$ and N/N' is abelian, hence nilpotent, and so by Lemma 2.6.4, N is nilpotent. Conversely, if N is nilpotent, then every proper subgroup of N is properly contained in its normalizer, hence by Theorem 2.5.8, every maximal subgroup of N is normal in N, therefore of prime index and so it contains N'. This means that $N' \subseteq \Phi(N)$; by Lemma 2.6.3, $\Phi(N) \subseteq \Phi(G)$, hence $N' \subseteq \Phi(G)$. ■

This result has several useful consequences. In the first place, since $\Phi(G)$ is normal in G and $\Phi' \subseteq \Phi$, we have

Theorem 2.6.6 (Frattini). *The Frattini subgroup of any finite group is nilpotent.* ■

Secondly, taking $N = G$, we obtain Wielandt's characterization of finite nilpotent groups:

Theorem 2.6.7. *A finite group G is nilpotent if and only if $G' \subseteq \Phi(G)$.* ■

The Frattini subgroup may be thought of as a kind of radical, in analogy to the Jacobson radical (see Chapter 5). But whereas the latter is the largest of all the commonly used radicals in a ring, this is not so for the Frattini subgroup. Radicals are not as useful for groups as for rings; nevertheless there is another type, first introduced by Hans Fitting in 1938, which plays a role in group theory. To prove its existence, we shall need another property of nilpotent groups. In the proof we shall use the commutator notation: $(x, y) = x^{-1}y^{-1}xy$, while for subgroups H, K, (H, K) denotes the subgroup generated by all (x, y), where $x \in H$, $y \in K$.

Lemma 2.6.8. *If G is nilpotent and $1 \neq K \triangleleft G$, then K meets the centre of G non-trivially.*

Proof. We shall write $Z_i = Z_i(G)$, $Z = Z_1$. Since G is nilpotent, we have $Z_n = G$ for some n, so there exists a positive integer r such that $K \subseteq Z_r$, $K \not\subseteq Z_{r-1}$. On defining recursively $K_0 = K$, $K_{i+1} = (K_i, G)$, we have $K_i \subseteq Z_{r-i}$, $K_i \not\subseteq Z_{r-i-1}$ by induction on i, hence $K_{r-1} \subseteq Z_1$, $K_{r-1} \not\subseteq Z_0 = 1$. Since K is normal in G, it contains K_1 and generally, $K_i \supseteq K_{i+1}$; therefore $K \cap Z \supseteq K_{r-1} \cap Z_1 = K_{r-1} \neq 1$. ■

For any finite group G we shall define the *Fitting subgroup,* denoted by $F(G)$, as the subgroup generated by all the nilpotent normal subgroups of G.

Theorem 2.6.9. *Let G be a finite group. Then the Fitting subgroup $F(G)$ is a nilpotent normal subgroup of G which contains every nilpotent normal subgroup of G.*

Proof. If H, K are any nilpotent normal subgroups of G, then clearly $HK \triangleleft G$; we claim that HK is nilpotent and here we may assume that $H, K \neq 1$ and use induction on $|G|$. Since H is nilpotent, $Z(H) \neq 1$; let us write $Z = Z(H)$. If $(Z, K) = 1$, then Z is in the centre of HK and we have $HK/Z \cong (H/Z)(ZK/Z)$, a product of nilpotent normal subgroups of HK/Z, hence nilpotent by the induction hypothesis, and so HK is then nilpotent. If $(Z, K) \neq 1$, then $(Z, K) \triangleleft K \cap Z$, hence $T = (Z, K) \cap Z(K) \neq 1$, by Lemma 2.6.8, and we can repeat the argument with N replaced by T. Thus in either case HK is nilpotent.

Now let F be a nilpotent normal subgroup of G of maximal order. Then for any nilpotent normal subgroup H, FH is again nilpotent and normal and $FH \supseteq F$, hence $FH = F$ by the maximality of F, so $H \subseteq F$, as required. ■

Finally we establish a connexion between the Frattini and the Fitting subgroups.

Theorem 2.6.10. *Let G be a finite group, $\Phi(G)$ be its Frattini subgroup and $F(G)$ be its Fitting subgroup. Then*

$$F(G)' \subseteq \Phi(G) \subseteq F(G), \tag{2.6.2}$$

and

$$F(G/\Phi(G)) = F(G)/\Phi(G). \tag{2.6.3}$$

Proof. The inclusion (2.6.2) is clear from Proposition 2.6.5, Theorem 2.6.6 and the definition of F. To prove (2.6.3), let us write $\Phi = \Phi(G)$, $F = F(G)$, $\overline{G} = G/\Phi$, $\overline{F} = F(\overline{G})$, and denote by K the inverse image of $\overline{F}$ in G. Now take a Sylow subgroup $P \neq 1$ of K; its image $\overline{P}$ in $\overline{G}$ is a Sylow subgroup of $\overline{F}$ and since $\overline{F}$ is nilpotent, $\overline{P}$ is normal in $\overline{F}$, hence characteristic and so normal in $\overline{G}$. Therefore $P\Phi \triangleleft G$; we claim that $N_G(P)\Phi = G$. For P is a Sylow subgroup of $L = P\Phi$; hence $G = N_G(P)L$ by Corollary 2.5.6, i.e. $G = N_G(P)\Phi$, therefore $G = N_G(P)$ by Proposition 2.6.2, and so $P \triangleleft G$. Thus each Sylow subgroup of K is normal and it follows that K is nilpotent, so $K \subseteq F$. But the image of F in $\overline{G}$ is certainly nilpotent and so is contained in $\overline{F}$. It follows that $F = K$, in particular, $\overline{F} = F/\Phi$. ∎

The Frattini subgroup can also be used to give an estimate for the number of automorphisms of a finite group G, i.e. the order of $\mathrm{Aut}(G)$:

Theorem 2.6.11 (P. Hall). *Let G be a finite group with a d-element generating set. Then* $|\mathrm{Aut}(G)|$ *divides* $|\mathrm{Aut}(G/\Phi)|.|\Phi(G)|^d$.

Proof. Since $\Phi = \Phi(G)$ is a characteristic subgroup, each automorphism α of G induces an automorphism $\overline{\alpha}$ of G/Φ, and the correspondence $\alpha \mapsto \overline{\alpha}$ clearly is a homomorphism. Denote the kernel by K, thus $\alpha \in K$ iff $x^\alpha x^{-1} \in \Phi$ for all $x \in G$.

Let $\{u_1, \ldots, u_d\}$ be a generating system of G. The number of families $\{u_1 a_1, \ldots, u_d a_d\}$, where $a_i \in \Phi$, is $|\Phi|^d$ and each again generates G. The set U of all these generating systems is permuted by K, and since any automorphism leaving $\{u_1, \ldots, u_d\}$ fixed must be the identity (because this is a generating set), it follows that the stabilizer of any member of U is 1. Hence each orbit has $|K|$ members and if there are r orbits, we have $r|K| = |\Phi|^d$. Now $|\mathrm{Aut}(G)| = |\mathrm{Aut}(G/\Phi(G))|.|K|$, and the result follows, since $|K|$ divides $|\Phi|^d$. ∎

Exercises

1. Show that $\Phi(\mathbf{Q}) = \mathbf{Q}$.
2. Show that if $N \triangleleft G$, then $\Phi(G/N) \geq \Phi(G)N/N$, but that equality need not hold. (Hint. Try $N = \Phi(G)$.)
3. For a finite group G show that if G/Φ is nilpotent, then so is G.
4. Show that for any nilpotent group G (not necessarily finite), $G' \subseteq \Phi(G)$. (Hint. Use Proposition 2.6.2.)
5. Show that a finite group with a single maximal subgroup is cyclic of prime power order.
6. Show that for any p-group G, $\Phi(G) = G'G^p$, where $G^p = \mathrm{gp}\{x^p | x \in \mathrm{G}\}$, while for $p = 2$, $\Phi(G) = G^2$.
7. If P is a finite p-group, then any subset X of P generates P iff its residues mod $\Phi(P)$ span $P/\Phi(P)$, regarded as vector space over the field of p elements. Deduce that all minimal generating sets have the same number of elements and that every element not in $\Phi(P)$ is contained in a minimal generating set.

8. Let G be a finite group and A be an abelian normal subgroup of G. Show that if $A \cap \Phi(G) = 1$, then A has a complement in G, i.e. a subgroup B such that $AB = G$, $A \cap B = 1$.
9. A permutation group G acting on a set S is *faithful* if only the unit element fixes all members of S; it is *primitive* if S cannot be partitioned into sets which are permuted among themselves by G. Show that in a faithful primitive permutation group any nilpotent normal subgroup is abelian.
10. Let G be a finite group and α be an automorphism of G of order prime to $|\Phi(G)|$ and inducing the identity on $G/\Phi(G)$. Show that $\alpha = 1$.
11. Let G be a group of order p^n and $(G : \Phi(G)) = p^d$. Show that $\mathrm{Aut}(G)$ has order dividing $p^{d(n-d)}(p^d - 1)(p^d - p)\ldots(p^d - p^{d-1})$.
12. Show that any non-abelian group of order 8 has a cyclic subgroup of index 2 and hence is either of the form $\mathrm{gp}\{a, b | a^4 = b^2 = (ab)^2 = 1\}$ (dihedral group) or $\mathrm{gp}\{a, b | a^4 = 1, a^2 = b^2, aba = b\}$ (quaternion group).

Further Exercises for Chapter 2

1. Describe the group of symmetries of (i) a square and (ii) a regular hexagon.
2. Describe the group of rotations of a cube.
3. Let G be a finite group and X be a subset. Show that the number of conjugates of X is equal to $(G : N_G(X))$.
4. Let G be a finite group of order r. Find a bound for the order of its automorphism group $\mathrm{Aut}(G)$ in terms of r. Let B be the subgroup of $\mathrm{Aut}(G)$ consisting of all automorphisms which map each conjugacy class of G into itself. Show that B is normal in $\mathrm{Aut}(G)$ and its order contains only prime factors occurring in $|G|$. Is the order necessarily a divisor of $|G|$?
5. Show that the presentation $\mathrm{gp}\{a, b | a^{2n} = 1, a^n = b^2, aba = b\}$ defines a group of order $4n$ (known as a *dicyclic group*), by verifying that each element can be expressed uniquely as $a^r b^s$, where $0 \leq r < 2n$, $0 \leq s < 2$, and that it has the dihedral group $\mathbf{D}_n$ as homomorphic image.
6. Let G be a group with an automorphism α. Show that there is a semidirect product H of G and an infinite cyclic group such that α is induced by an inner automorphism of H.
7. Let G be a finite group and A be an abelian normal subgroup of G. Show that if $A \cap \Phi(G) = 1$, then A has a complement in G.
8. Let G be a finite group and H be a subgroup. A *partial complement* for H is a subgroup $K \neq H$ such that $HK = G$. Show that a normal subgroup of G has a partial complement iff it is not contained in $\Phi(G)$.
9. Show that the Frattini subgroup $\Phi(G)$ of a p-group G is the smallest subgroup H of G such that G/H is elementary abelian. Deduce that a p-group G is elementary abelian iff $\Phi(G) = 1$.

3 Lattices and Categories

The subsets of a set permit operations quite similar to those performed on numbers. If for the moment we denote the union of two subsets A, B by $A + B$ and their intersection by AB, a notation that will not be used later (despite some historical precedents), then we have laws like $AB = BA$, $A(B + C) = AB + AC$, similar to the familiar laws of arithmetic, as well as new laws such as $A + A = A$, $A + BC = (A + B)(A + C)$. The algebra formed in this way is called a *Boolean algebra*, after George Boole who introduced it around the middle of the 19th century, and who made the interesting observation that Boolean algebras could also be used to describe the propositions of logic.

The more general notion of a lattice was first used towards the end of the 19th century by Richard Dedekind to study the relations between ideals in rings of numbers. The lattice concept helps to unify a number of disparate ideas and this chapter deals with the basic properties of lattices in Sections 3.1 and 3.2 and Boolean algebras in Section 3.4. The latter have recently found applications in switching theory, but we shall not enter on this aspect. We shall concentrate on applications to algebra, and describe such important ideas as chain conditions in the general setting of partially ordered sets.

Section 3.3 introduces categories, which provide a succinct way of describing many aspects of groups and are also of use in other parts of algebra.

3.1 Definitions; Modular and Distributive Lattices

The definitions relating to partially ordered sets (see Conventions on Terminology, p. xi) form the foundation for much that follows. Whereas sets of numbers, like **N** or **Q**, are totally ordered, this property is not shared by most partially ordered sets. Examples are the set $\mathscr{P}(S)$ of all subsets of a set S (with more than one element) relative to inclusion, or the set **N** of natural numbers relative to divisibility. In both these cases any two members have a least upper bound or *supremum*, briefly sup; given x, y, there exists z such that (i) $x \leq z$, $y \leq z$ and (ii) for any z', $x \leq z'$, $y \leq z' \Rightarrow z \leq z'$. Dually, any two members have a greatest lower bound or *infimum*, briefly inf, defined similarly. Thus if X and Y are subsets of S, their sup is $X \cup Y$ and their inf is $X \cap Y$. If m, n are natural numbers, their sup (relative to divisibility) is their least common multiple while their inf is their highest common factor.

A partially ordered set in which any two elements have a supremum and an infimum is called a *lattice*. The sup of x, y is written $x \vee y$ and is called the *join* of x and y; their inf is written $x \wedge y$ and is called the *meet* of x and y. It is clear that the lattice concept is self-dual, so that the dual of a lattice, obtained by reversing the ordering, is again a lattice. As examples of lattices we have the set **N** of natural numbers under divisibility, as well as $\mathscr{P}(S)$, the set of all subsets of S under inclusion; further, every totally ordered set is trivially a lattice, with $x \wedge y = x$, $x \vee y = y$ if $x \leq y$.

Partially ordered sets are often represented as directed graphs: the elements of the set form the vertices, and edges are drawn so that $a < b$ holds precisely when b is higher than a and there is a descending path from b to a. Such diagrams are mainly used for finite sets; for **N** or **Z** the structure can be hinted at in a partial diagram, but the case of **Q** or **R** would be more difficult. Some examples are given in Figures 3.1–3.3, where Figures 3.1 and 3.2 are lattices, but Figure 3.3 is not.

A lattice may be regarded as a set with two binary operators, $\vee$ and $\wedge$. These operators satisfy a number of laws, reminiscent of the laws governing addition and multiplication of numbers, and these laws can be used to give an alternative definition of lattices.

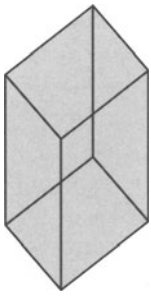


Figure 3.1

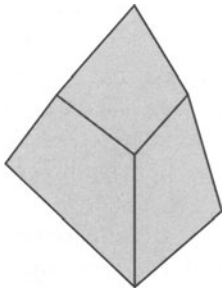


Figure 3.2

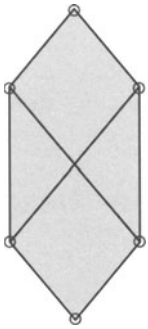


Figure 3.3

Proposition 3.1.1. *Let L be a lattice. Then for any $a, b, c \in L$,*

$$a \vee (b \vee c) = (a \vee b) \vee c, \quad a \wedge (b \wedge c) = (a \wedge b) \wedge c, \quad \text{(associative law)} \tag{3.1.1}$$

$$a \vee b = b \vee a, \quad a \wedge b = b \wedge a, \quad \text{(commutative law)} \tag{3.1.2}$$

$$a \wedge (a \vee b) = a, \quad a \vee (a \wedge b) = a, \quad \text{(absorptive law)} \tag{3.1.3}$$

$$a \vee a = a, \quad a \wedge a = a. \quad \text{(idempotent law)} \tag{3.1.4}$$

Conversely, if L is a set with two binary operators $\vee$, $\wedge$ satisfying (3.1.1)–(3.1.3)*, then* (3.1.4) *also holds and a partial ordering may be defined on L by the rule*

$$a \leq b \quad \textit{if and only if} \quad a \vee b = b. \tag{3.1.5}$$

Relative to this ordering L is a lattice such that the join of a, b is $a \vee b$ and the meet is $a \wedge b$.

Proof. By the definition of $a \vee b$ as sup we see that the unique sup of a, b can be written as either $a \vee b$ or $b \vee a$, and a similar remark applies to $a \wedge b$, so (3.1.2) follows. Likewise the sup of a, b, c may be written as $a \vee (b \vee c)$ or $(a \vee b) \vee c$, hence (3.1.1) holds. Now (3.1.3) follows because $a \wedge b \leq a \leq a \vee b$, and (3.1.4) is a trivial consequence of the definition, but we observe that it also follows from (3.1.3): if in the first Equation (3.1.3) we replace b by $a \wedge a$ we get $a = a \wedge [a \vee (a \wedge a)] = a \wedge a$, by the second Equation (3.1.3). This proves the first idempotent law; the second follows by duality.

Now let L be a set with two operators $\vee$, $\wedge$ satisfying (3.1.1)–(3.1.3); then (3.1.4) also holds, as we have just seen, and moreover,

$$a \vee b = b \Leftrightarrow a \wedge b = a. \tag{3.1.6}$$

For if $a \vee b = b$, then by (3.1.3), $a \wedge b = a \wedge (a \vee b) = a$, and the converse follows by duality (and the commutative law (3.1.2)). If we define the relation '$\leq$' by (3.1.5), we have a partial ordering: if $a \leq b$, $b \leq c$, then $a \vee b = b$, $b \vee c = c$; hence by (3.1.1), $c = b \vee c = (a \vee b) \vee c = a \vee (b \vee c) = a \vee c$, so $a \leq c$. Further, $a \leq a$ by (3.1.4), and if $a \leq b$, $b \leq a$, then $b = a \vee b = b \vee a = a$, by (3.1.2).

By the definition of $a \leq b$ and (3.1.2), $a \vee b$ is an upper bound of a, b. If c is another upper bound, then $a \leq b$, $b \leq c$, hence $c = a \vee c = b \vee c$, and so $c = a \vee (b \vee c) = (a \vee b) \vee c$, i.e. $a \vee b \leq c$, which shows $a \vee b$ to be the least upper bound, i.e. the sup. Now by duality and (3.1.6) it follows that the inf of a, b is $a \wedge b$. ■

Any subset of an ordered set S is again ordered, but it need not be a lattice, even if S is one. Guided by Proposition 3.1.1, we define a *sublattice* of a lattice L as a subset M which admits the operators $\vee$, $\wedge$ of L, i.e. given $a, b \in M$, we have $a \vee b$, $a \wedge b \in M$. It is clear that M is again a lattice. In terms of the partial ordering the definition may be expressed as follows: M is a sublattice of L if for any $a, b \in M$, their sup and inf (taken in L) again lie in M. We note that it may very well be possible for a subset of L to be a lattice without being a sublattice of L. For example, let G be a group, $\mathscr{P}(G)$ *be the set of all subsets of* G and Lat(G) be the set of all subgroups of G. As we shall soon see, Lat(G) is again a lattice with respect to the ordering by inclusion, as is $\mathscr{P}(G)$, but Lat(G) is not usually a sublattice of $\mathscr{P}(G)$, because the union $H \cup K$ of two subgroups need not be a subgroup.

It is clear from the definition that in a lattice L any intersection of sublattices is again a sublattice. Thus we can define the sublattice *generated by* a subset X of L as the intersection of all sublattices containing X. As in the case of subgroups, this sublattice can be obtained by repeated application of the lattice operations to the elements of X.

As for groups we define a *homomorphism* of lattices as a mapping $f : L \to L'$ between lattices L, L' such that for all $a, b \in L$,

$$(a \vee b)f = af \vee bf, \quad (a \wedge b)f = af \wedge bf. \tag{3.1.7}$$

It is clear that a lattice-homomorphism preserves the ordering: $a \leq b$ implies $af \leq bf$, but not every order-preserving mapping between lattices is a lattice-homomorphism. For example, the mapping $\alpha_c : x \mapsto x \vee c$ (for a fixed element c) in any lattice is order-preserving:

$$a \leq b \Rightarrow a \vee c \leq b \vee c, \tag{3.1.8}$$

and although α_c satisfies the first Equation (3.1.7), it does not generally satisfy the second (this is in fact the distributive law, to be discussed later). However, an order-preserving bijection with an order-preserving inverse between lattices is always a lattice-isomorphism, because the sets are then order-isomorphic and the lattice operations can be defined in terms of the ordering (as in (3.1.5)).

In any lattice each finite (non-empty) subset has a sup and an inf, as an easy induction shows. Explicitly the sup and inf of $a_1, \ldots, a_n$ are given by

$$a_1 \vee \ldots \vee a_n \quad \text{and} \quad a_1 \wedge \ldots \wedge a_n$$

respectively. Here we may omit brackets, by associativity, omit repetitions, by the idempotent law, and the order of the factors is immaterial, by commutativity.

The notions of sup and inf can also be defined for infinite subsets, but in a general lattice they may not exist. A lattice L in which every subset has a sup and an inf is said

to be *complete*. In particular such a lattice L has a greatest element ($\sup L$ or $\inf \varnothing$), denoted by 1, and a least element ($\inf L$ or $\sup \varnothing$), denoted by 0. For example, every finite lattice is complete and so is $\mathscr{P}(S)$, for any set S. The following criterion for completeness is often useful:

Proposition 3.1.2. *If L is a partially ordered set such that every subset has an inf, then L is a complete lattice.*

Proof. Given $X \subseteq L$, let Y be the set of all upper bounds of X in L and set $y = \inf Y$. Any element of X is a lower bound of Y, hence $x \leq y$ for all $x \in X$. If also $x \leq z$ for all $x \in X$, then $z \in Y$, by the definition of Y, and so $y \leq z$, therefore $y = \sup X$. ■

For example, the set of all subgroups of a group G is partially ordered by inclusion and if $\{H_\lambda\}$ is a family of subgroups, then their intersection $\cap H_\lambda$ is again a subgroup; thus every subset of our set has an inf. Applying Proposition 3.1.2, we see that it is a complete lattice, which we denote by $\mathrm{Lat}(G)$. The inf of a family $\{H_\lambda\}$ is their intersection, while the sup is the least subgroup containing all the H_λ, i.e. the subgroup generated by all the H_λ. We see that although the inf in $\mathrm{Lat}(G)$ is the same as in $\mathscr{P}(G)$, the sup in general is not.

Many of the lattices we shall meet in the sequel satisfy the modular law already encountered in Section 2.1:

$$a \vee (b \wedge c) = (a \vee b) \wedge c \quad \text{for all} \quad a, b, c \in L \quad \text{such that} \quad a \leq c. \tag{3.1.9}$$

A lattice satisfying (3.1.9) is said to be *modular* (Dedekind's name was "Dualgruppe vom Modultypus"). For example the set $\mathscr{N}(G)$ of all normal subgroups of a group G is a modular lattice under the operations $H \cap K$, HK, as is easily verified. By contrast, the lattice of *all* subgroups of a group will not in general be modular, and $H \vee K$ need not equal HK (see Exercise 10). The modular law holds more generally for submodules of a module (see Chapter 4), which accounts for the name.

To see why modular lattices are more tractable, let us return to a general lattice L for a moment. With $a, b \in L$ such that $a \leq b$ we can associate the interval

$$[a, b] = \{x \in L | a \leq x \leq b\}.$$

Such an interval need not be a chain, but it is always a sublattice of L, with least element a and greatest element b. More generally, for any $a, b \in L$ we can form the intervals $I = [a \wedge b, a]$ and $J = [b, a \vee b]$. Let us define a mapping $\alpha : I \to J$ by

$$\alpha : x \mapsto x \vee b,$$

and a mapping $\beta : J \to I$ by

$$\beta : y \mapsto y \wedge a.$$

For any $x \in I$ we have $x\alpha\beta = (x \vee b) \wedge a$. Here $x \leq a$, hence if L is modular, then $x\alpha\beta = (x \vee b) \wedge a = x \vee (b \wedge a) = x$. Thus $\alpha\beta = 1$ and dually, $\beta\alpha = 1$, i.e. in a modular lattice the mappings α, β are mutually inverse. Since α and β are both order-preserving, it follows that I and J are isomorphic as lattices. But in concrete cases the explicit form of the mappings α, β often tells us more than this.

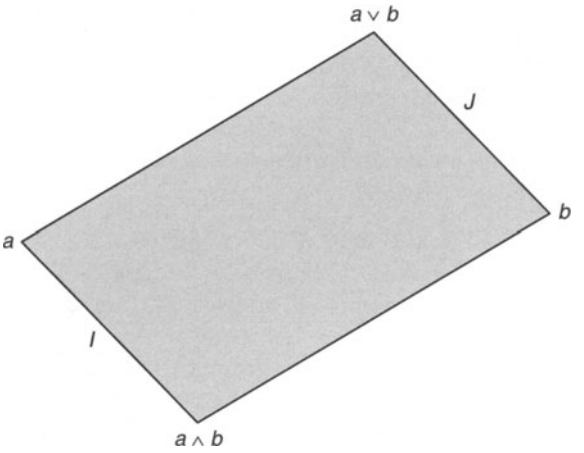


Figure 3.4

Let us call two intervals I, J related as in Figure 3.4 *perspective*, and two intervals related by a series of perspectivities *projective*. The second isomorphism theorem of group theory (Theorem 2.3.3) shows that in the lattice $\mathcal{N}(G)$ of all normal subgroups of G, perspective intervals define isomorphic quotients. Hence the same is true of projective intervals and the usual proof of the Schreier refinement theorem, using the Zassenhaus lemma (see Exercise 6 of Section 2.3) shows that any two normal chains have refinements in which corresponding intervals are projective, and hence define isomorphic factor groups (however, this theorem does not operate within $\mathcal{N}(G)$).

To obtain a criterion for modularity, let us call two elements x, y in an interval $[a, b]$ *complementary* and call each a *relative complement* (in $[a, b]$) of the other if $x \wedge y = a$, $x \vee y = b$.

Proposition 3.1.3. *A lattice L is modular if and only if, for each interval I of L, any two elements of I which are comparable and have a common complement in I are equal.*

Proof. In any lattice L, given $a, b, c \in L$, if $a \leq c$, then $a \vee (b \wedge c) \leq a \vee b$ and $a \vee (b \wedge c) \leq c$, hence

$$a \vee (b \wedge c) \leq (a \vee b) \wedge c. \tag{3.1.10}$$

Therefore L is non-modular iff the inequality (3.1.10) is strict for at least one triple (a, b, c) such that $a \leq c$. When $a = c$, the two sides of (3.1.10) are equal by the absorptive law (3.1.3), so we may assume that $a < c$. Suppose first that strict inequality holds in (3.1.10). Put $a' = a \vee (b \wedge c)$, $c' = (a \vee b) \wedge c$; then by (3.1.10),

$$a \leq a' < c' \leq c, \tag{3.1.11}$$

and $b \wedge c' = b \wedge (a \vee b) \wedge c = b \wedge c$, $a' \vee b = a \vee (b \wedge c) \vee b = a \vee b$. Moreover, $c' \leq a \vee b$, hence $b \vee c' \leq a \vee b \leq b \vee c'$ by (3.1.10); therefore $b \vee c' = a \vee b$, and

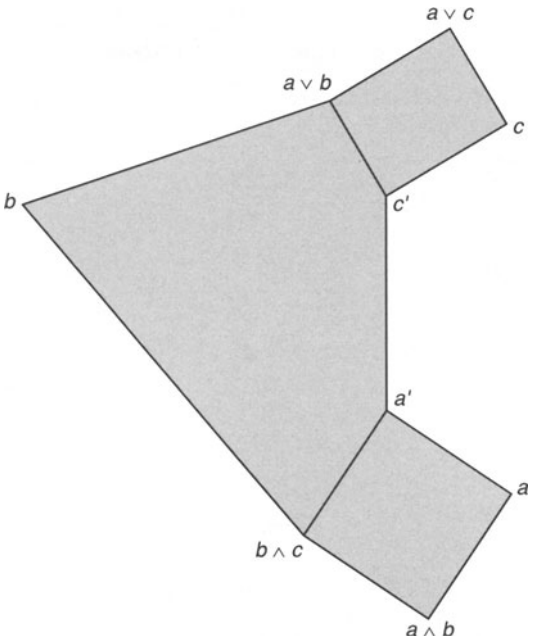


Figure 3.5

dually, $a' \wedge b = b \wedge c$. This shows that a' and c' have the common complement b in $[b \wedge c, a \vee b]$, and by (3.1.11) they are comparable but distinct (see Figure 3.5). Conversely, if a', c' are distinct elements which are comparable and have a common complement in $[u, v]$, say $a' \wedge b = c' \wedge b = u$, $a' \vee b = c' \vee b = v$ and $u \leq a' < c' \leq v$, then

$$a' \vee (b \wedge c') = a' < c' = (a' \vee b) \wedge c',$$

hence L is not modular. ■

The property characterizing modularity involves only five elements, namely the endpoints of the interval, an element and its two complements. Thus we have

Corollary 3.1.4. *A lattice is modular if and only if it does not contain a sublattice isomorphic to the pentagon lattice of Figure 3.6.* ■

For example, the lattice of Figure 3.1 is modular, but that of Figure 3.2 is not.

If S is any set, then the lattice of subsets of S, $\mathscr{P}(S)$, is modular, but not every modular lattice can be represented as a lattice of subsets. For example, the Klein 4-group $\mathrm{gp}\{a, b | a^2 = b^2 = (ab)^2 = 1\}$ has the subgroup lattice shown in Figure 3.7 (the 'diamond' lattice), but it cannot be represented as a lattice of subsets, as the reader can discover by a little experimentation.

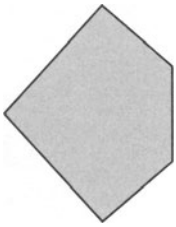


Figure 3.6

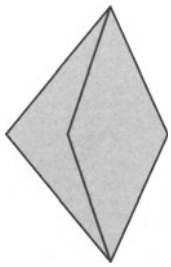


Figure 3.7

It turns out that the lattices $\mathscr{P}(S)$ satisfy a further law not holding in the diamond lattice, namely the distributive law. There are two dual forms of this law, which however are equivalent; moreover, they imply the modular law. Before proving this fact we note that in any lattice we have $a \vee b \geq a \wedge c, b \wedge c$ and $c \geq a \wedge c, b \wedge c$ for any a, b, c; hence

$$(a \vee b) \wedge c \geq (a \wedge c) \vee (b \wedge c). \tag{3.1.12}$$

Now the distributive law is expressed by equality in (3.1.12). More precisely we have

Proposition 3.1.5. *In any lattice L the following conditions are equivalent:*

(α) $(a \vee b) \wedge c = (a \wedge c) \vee (b \wedge c)$ *for all* $a, b, c \in L$;
(α^*) $(a \wedge b) \vee c = (a \vee c) \wedge (b \vee c)$ *for all* $a, b, c \in L$;
(β) $(a \vee b) \wedge c \leq a \vee (b \wedge c)$ *for all* $a, b, c \in L$.

Proof. If (α) holds, then

$$(a \vee b) \wedge c = (a \wedge c) \vee (b \wedge c) \leq a \vee (b \wedge c),$$

i.e. (β). Conversely, assume (β): $(a \vee b) \wedge c \leq a \vee (b \wedge c)$. Applying $\wedge c$ to both sides and using (β) again, we obtain

$$(a \vee b) \wedge c \leq [(b \wedge c) \vee a] \wedge c \leq (b \wedge c) \vee (a \wedge c).$$

The reverse inequality holds by (3.1.12), hence we obtain (α). Thus (α) $\Leftrightarrow$ (β), and since (α^*) is the dual of (α) and (β) is self-dual, we also have (α^*) $\Leftrightarrow$ (β). ■

A lattice satisfying these three equivalent conditions is said to be *distributive*; specifically, either (α) or (α^*) is called the *distributive law*. From (β) it is clear that every distributive lattice is modular.

There is a criterion analogous to Proposition 3.1.3: a lattice is distributive iff relative complements in any interval are unique. We shall only need the necessity of this condition:

Proposition 3.1.6. *In any distributive lattice, relative complements in each interval are unique.*

Proof. Let $a \wedge b = a' \wedge b = u$, $a \vee b = a' \vee b = v$; then

$$a = a \wedge v = a \wedge (a' \vee b) = (a \wedge a') \vee (a \wedge b) = a \wedge a';$$

hence $a \leq a'$. By symmetry $a' \leq a$ and hence $a' = a$. ■

If we single out the five elements involved we obtain the following alternative formulation:

A lattice is distributive if and only if it does not contain a sublattice isomorphic to the pentagon lattice in Figure 3.6 or the diamond lattice in Figure 3.7.

For a proof of this criterion see Birkhoff (1967) or Cohn (1981) (see also Exercise 12 below).

Exercises

1. Show that in any lattice, if $a \leq a'$, $b \leq b'$, then $a \wedge b \leq a' \wedge b'$, $a \vee b \leq a' \vee b'$.
2. Find the smallest partially ordered set in which any two elements have an upper bound and a lower bound but which is not a lattice.
3. Find all lattices on at most five elements. Which of them are anti-isomorphic with themselves? Which are modular, or distributive?
4. Show that the least element 0 in a lattice (if it exists) is characterized by $0 \wedge x = 0$, $0 \vee x = x$, and give a corresponding characterization of the greatest element.
5. Let L be a modular lattice. Show that if $a, b, c \in L$ satisfy $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$, then the sublattice generated by a, b, c is distributive.
6. Show that in any modular lattice the sublattice generated by any two chains is distributive.
7. Let L be a system with two binary operators $\vee$, $\wedge$ and a particular element 1 in L satisfying (i) $a \wedge a = a$, (ii) $a \vee 1 = 1 \vee a = 1$, (iii) $a \wedge 1 = 1 \wedge a = a$, (iv) $a \wedge (b \vee c) = (a \wedge b) \vee (a \wedge c)$, $(b \vee c) \wedge a = (b \wedge a) \vee (c \wedge a)$ for all $a, b, c \in L$. Show that L is a distributive lattice with greatest element 1.
8. Show that in a topological space the closed sets form a complete distributive lattice.
9. (O. Ore) Show that, for any group G, Lat(G) is distributive iff G is locally cyclic (i.e. every finitely generated subgroup is cyclic).

10. Let $G = \mathrm{Sym}_4$, $H = \{1, (1\ 2)\}$, $K = \{1, (1\ 2\ 3\ 4)\}$; show that G is generated by H and K, and deduce that $H \vee K \neq HK$, where $H \vee K$ is the join of H and K in $\mathrm{Lat}(G)$. Show that $\mathrm{Lat}(G)$ is not modular.
11. Show, by examining the lengths of maximal chains in $\mathrm{Lat}(\mathrm{Alt}_4)$, that this lattice is not modular.
12. Show that in any lattice L,

$$(x \wedge y) \vee (y \wedge z) \vee (z \wedge x) \leq (x \vee y) \wedge (y \vee z) \wedge (z \vee x) \text{ for all } x, y, z \in L, \quad \text{(i)}$$

with equality iff L is distributive. Denote the two sides of (i) by u, v respectively and put $x' = (x \wedge v) \vee u$, $y' = (y \wedge v) \vee u$, $z' = (z \wedge v) \vee u$. If L is modular but not distributive, choose $x, y, z \in L$ such that the inequality (i) is strict and verify that u, v, x', y', z' form a diamond lattice. Deduce that a modular lattice is distributive iff it does not contain a diamond lattice as sublattice.

3.2 Chain Conditions

Although most of the lattices we shall meet are infinite, many of them satisfy finiteness conditions; these take several equivalent forms. We state them for any partially ordered set:

Proposition 3.2.1. *In any partially ordered set S the following conditions are equivalent:*

(a) *(Ascending chain condition) Every ascending chain becomes stationary: if*

$$a_1 \leq a_2 \leq \ldots, \tag{3.2.1}$$

then there exists n_0 such that $a_m = a_n$ for all $m, n \geq n_0$.

(b) *Every strictly ascending chain terminates: if*

$$a_1 < a_2 < \ldots, \tag{3.2.2}$$

then the chain has only finitely many terms.

(c) *(Maximum condition) Every non-empty subset of S has a maximal element.*

Proof. (a) $\Rightarrow$ (b) follows because any chain (3.2.2) can become stationary only by terminating. To prove (b) $\Rightarrow$ (c), let M be a non-empty subset of S. Pick $a_1 \in M$; if a_1 is not maximal in M, we can find $a_2 \in M$ such that $a_2 > a_1$, and generally, for each $a_n \in M$, either a_n is maximal or there exists $a_{n+1} \in M$ such that $a_{n+1} > a_n$. Thus we obtain a chain (3.2.2) which must terminate, by (b), and the last element is maximal in M.

(c) $\Rightarrow$ (a). Given (3.2.1), let a_n be maximal in the set $\{a_1, a_2, \ldots\}$; then $a_n \geq a_m$ for all m, hence $a_n = a_{n+1} = \ldots$, so (3.2.1) becomes stationary. ∎

We note that the Axiom of Choice was used in the deduction (b) $\Rightarrow$ (c); it can be shown that this is indispensable. Thus without the Axiom of Choice the maximum condition is stronger than the ascending chain condition, but in the presence of the Axiom of Choice they are equivalent (see Hodges [1974]).

There is a useful induction principle holding in sets with maximum condition:

Proposition 3.2.2 (Noetherian induction). *Let S be a partially ordered set with maximum condition. If X is a subset of S which contains any element a of S whenever it contains all elements $x \in S$ such that $x > a$, then $X = S$.*

Proof. Consider the complement X' of X in S. If $X' \neq \varnothing$, let c be a maximal element of X'. Then any $x > c$ must be in X, hence by hypothesis, $c \in X$, which contradicts the fact that $c \in X'$. Therefore X' is empty and $X = S$, as claimed. ∎

By duality we obtain from Proposition 3.2.1 the equivalence of the minimum condition and (two forms of) the descending chain condition, and as in Proposition 3.2.2 we obtain an induction principle for ordered sets with minimum condition. For well-ordered sets, i.e. totally ordered sets with minimum condition, this is just the principle of transfinite induction (see Section 1.2).

Every finite ordered set clearly satisfies both maximum and minimum conditions; the converse is false, as any infinite set shows whose elements are all incomparable. Even for modular lattices the converse need not hold (see Exercise 2), but it does hold for distributive lattices, as we shall see in Section 3.4. For the moment we shall describe the modular lattices satisfying both chain conditions. Given a chain C between certain points p, q, any chain from p to q which includes C is called a *refinement* of C; clearly C is a maximal chain from p to q iff it has no proper refinements.

Proposition 3.2.3. *In a partially ordered set S with both chain conditions, every chain is finite and can be refined by inserting further terms to yield a maximal chain between the given endpoints.*

Proof. By the minimum condition, every chain in S has a minimal element, necessarily unique. Given a chain C in S, let a_1 be its least element, and generally define a_ν as the least element of $C \setminus \{a_1, \ldots, a_{\nu-1}\}$. Then

$$a_1 < a_2 < \ldots, \tag{3.2.3}$$

and by the maximum condition this chain terminates. If the last term is a_n, it follows that $C = \{a_1, \ldots, a_n\}$, hence C is finite. Next, given a chain (3.2.3), let b_1 be minimal in S such that $a_1 < b_1 \leq a_2$. If $b_1 < a_2$, we choose $b_2 \in S$ such that b_2 is minimal subject to $b_1 < b_2 \leq a_2$. Continuing in this way, we obtain a chain

$$a_1 < b_1 < b_2 < \ldots \leq a_2,$$

which cannot be refined further. By the maximum condition it must terminate, which can only happen when $b_k = a_2$ for some k. We now have a maximal chain from a_1 to a_2; by induction on the number of terms in (3.2.3) we can find a maximal chain from a_2 to a_n, and together with the part found this provides a maximal chain from a_1 to a_n. ∎

Let us define the *length* of a chain as the number of its links, thus

$$a = a_0 < a_1 < \ldots < a_n = b \tag{3.2.4}$$

has length n. In general there is no reason why two maximal chains between the same endpoints should have the same length. For example, in the pentagon lattice of Figure 3.6 there are two maximal chains in the lattice of lengths 2 and 3. But if we have two chains between the points a and b in a modular lattice, then the proof of the Schreier refinement theorem (see Exercise 7 of Section 2.3), applied to lattices, shows that they have refinements whose links can be paired off in such a way that corresponding links are projective. In particular, if both chains are maximal, it follows that they both have the same length. Thus we have

Proposition 3.2.4. *Let L be a modular lattice and $a \leq b$ in L such that there is a maximal chain from a to b, of length n. Then every chain between a and b has length at most n, and the length equals n if and only if the chain is maximal.* ■

In view of this result we can define the *length* of a lattice as the supremum of the lengths of its chains. Modular lattices of finite length can be defined as follows:

Corollary 3.2.5. *A modular lattice has finite length if and only if it satisfies both the ascending and the descending chain condition.*

Proof. Clearly any modular lattice of finite length satisfies both chain conditions. Conversely, if a modular lattice L satisfies both chain conditions, take a maximal element c in L. Then $x < c$ or $x = c$ for any $x \in L$, hence c is in fact the greatest element 1 in L, and dually, L has a least element 0. By Proposition 3.2.4, there is a maximal chain from 0 to 1; this is finite, of length n say, and no chain can be longer. ■

A modular lattice has finite length if there is an element c such that all chains below c are finite, and likewise all chains above c. But sometimes we shall want the corresponding assertion when only one of the chain conditions holds.

Proposition 3.2.6. *Let L be a modular lattice and $c \in L$. Denote by $L_c, {}_cL$ the sublattices of elements $\leq c$ and $\geq c$ respectively. If both L_c and ${}_cL$ satisfy the maximum condition, then so does L.*

Proof. Let

$$a_1 \leq a_2 \leq \ldots \tag{3.2.5}$$

be an ascending chain in L. Then $a_1 \wedge c \leq a_2 \wedge c \leq \ldots$, and by hypothesis this becomes stationary. Likewise for $a_1 \vee c \leq a_2 \vee c \leq \ldots$; thus we may choose n_0 such that for all $m, n \geq n_0$,

$$a_m \wedge c = a_n \wedge c = u, \text{ say}, \quad a_m \vee c = a_n \vee c = v, \text{ say}.$$

Hence a_m, a_n have a common complement c in $[u, v]$, and since $a_m \leq a_n$ for $m \leq n$, it follows by Proposition 3.1.3 that $a_m = a_n$ for $m, n \geq n_0$, so (3.2.5) becomes stationary, as claimed. ■

In lattices with a chain condition there is a decomposition lemma that is frequently used. It is convenient to formulate it more generally for ordered monoids. By a *partially ordered monoid* we understand a monoid which is partially ordered as a set and such that $x \leq x', y \leq y'$ implies $xy \leq x'y'$. We shall further require the condition

$$xy \leq x, xy \leq y. \tag{3.2.6}$$

Examples are lattices with greatest element 1 and with $x \wedge y$ as the operation, or **N** with the ordering opposite to the usual one. An element c in such a monoid is called *irreducible* if $c \neq 1$ and c cannot be written as a product of two elements that are $> c$.

Lemma 3.2.7 (Decomposition Lemma). *Let M be a partially ordered monoid satisfying the maximum condition and* (3.2.6). *Then every element can be written as a product of irreducible elements.*

Proof. Denote by I the subset of elements of S which cannot be expressed as a product of a finite number of irreducible elements; we have to show that I is empty. If $I \neq \emptyset$, take a maximal element c in I; then c cannot be irreducible and $c \neq 1$ because 1 is the product of the empty family. Hence $c = ab$ for some $a > c, b > c$. By the maximality of c, a and b are products of irreducible elements, say $a = a_1 \dots a_r$, $b = b_1 \dots b_s$ and it follows that $c = a_1 \dots a_r b_1 \dots b_s$. This contradicts the choice of c and it shows that $I = \emptyset$. Thus every element of S is a product of irreducible elements. ∎

Example 1. Let L be a lattice with maximum condition. Then L has a greatest element 1 and it satisfies the hypothesis of the lemma with respect to the operation $x \wedge y$. In this case the irreducible elements are called *meet-irreducible* and the lemma tells us that in L every element c can be written in the form

$$c = a_1 \wedge \dots \wedge a_r, \text{ where each } a_i \text{ is meet-irreducible.}$$

Example 2. Dually, if the operation is $x \vee y$, the irreducible elements are called *join-irreducible.* By the above lemma, or rather, its dual, in a lattice with minimum condition every element c can be written as

$$c = b_1 \vee \dots \vee b_s, \text{ where each } b_i \text{ is join-irreducible.}$$

Example 3. In **N** the minimum condition holds for the usual ordering and $xy \geq x, y$; the irreducible elements in this case are the prime numbers, so we can apply the lemma to deduce that every natural number can be written as a product of prime numbers. In Section 10.8 we shall apply Lemma 3.2.7 in a similar situation which generalizes this case.

In a partially ordered set with minimum condition there is a relation between anti-chains and lower segments which is sometimes useful.

Proposition 3.2.8. *Let S be a partially ordered set with minimum condition. Then there is a natural bijection between lower segments and anti-chains: To each lower segment L*

there corresponds the anti-chain L^0 consisting of all minimal elements of the complement L' of L; to each anti-chain A there corresponds the complement of the upper segment $[A|$ generated by A, $A^0 = [A|'$.

Proof. From the definitions it is clear that for any lower segment L we have $L \subseteq L^{00}$. If $x \notin L$, then $x \in L'$, hence by the minimum condition there is a minimal element a of L' such that $a \leq x$, so $x \in [L^0|$ and therefore $x \notin L^{00}$; this proves that $L^{00} = L$. Next take an anti-chain A; again it is clear that $A \subseteq A^{00}$. To prove equality here we note that $A^0 = [A|'$; hence for any $x \in A^{00}$, x is a minimal element of $A^{0\prime} = [A|$. By definition this means that $x \geq a$ for some $a \in A$, but by the minimality of x, $x = a$, so $x \in A$, as claimed. ■

Our final result does not strictly deal with a chain condition, but it is a construction arising from an ascending chain of subgroups.

Proposition 3.2.9. *Let $\{G_n\}$ be a sequence of groups such that G_n is a subgroup of G_{n+1}. Then the union $\cup G_n$ is a group with each G_n as a subgroup.*

Proof. On the union $G = \cup G_n$ we can define a group structure in just one way. namely, if $x, y \in G$, then $x \in G_r, y \in G_s$ where $r \leq s$ say. Hence $x, y \in G_s$ and the product xy is defined in G_s. Moreover, for any $n \geq s$, G_s is a subgroup of G_n, so the product xy is the same in G_n as in G_s. In this way we define a multiplication on G and it is easily checked that G is a group with respect to this multiplication, with each G_n as a subgroup. ■

We observe that this result holds more generally if instead of an ascending sequence of groups we have a 'directed system' $\{G_\lambda\}$ of groups, i.e. for each pair λ, μ there exists ν such that $G_\lambda \cup G_\mu \subseteq G_\nu$.

Exercises

1. Show that a partially ordered set in which each chain has at most m and each anti-chain has at most n elements, has at most mn elements.
2. Give an example of an infinite modular lattice of length 2.
3. Let G be a finitely generated group. Show that the union of any countable strictly ascending sequence of subgroups without last term is a proper subgroup.
4. Give an example of a non-trivial abelian group without maximal proper subgroups.
5. Show that a lattice in which every lower segment is principal (i.e. generated by a single element) satisfies the maximum condition.
6. Prove without the Axiom of Choice that in a group with maximum condition all subgroups are finitely generated.

3.3 Categories

Many readers will have met categories, but in view of their importance we recall their definitions in some detail and describe some of their simpler properties. A *category* $\mathscr{A}$ consists of a class of *objects* and a class of *morphisms* or *maps*. With each morphism α two objects are associated, its *source* and *target*; if they are X, Y respectively, we write $\alpha : X \to Y$ or $X \xrightarrow{\alpha} Y$ and say: α *goes from* X *to* Y. The collection of all morphisms from X *to* Y is written $\mathrm{Hom}_{\mathscr{A}}(X, Y)$ or simply $\mathscr{A}(X, Y)$. Further, certain pairs of morphisms can be combined to yield another morphism. Given $\alpha : X \to Y$, $\beta : Y \to Z$, so that the target of α is the source of β, we can compose α and β to a morphism from X to Z, denoted by $\alpha\beta$. These objects and morphisms are subject to the following rules:

C.1 $\mathscr{A}(X, Y)$ is a set and $\mathscr{A}(X, Y) \cap \mathscr{A}(X', Y') = \varnothing$ unless $X = X'$ and $Y = Y'$.

C.2 If $\alpha : X \to Y$, $\beta : Y \to Z$, $\gamma : Z \to T$, so that $(\alpha\beta)\gamma$ and $\alpha(\beta\gamma)$ are both defined, then $(\alpha\beta)\gamma = \alpha(\beta\gamma)$.

C.3 For each object X there exists a morphism $1_X : X \to X$ such that for each $\alpha : X \to Y$, we have $1_X\alpha = \alpha 1_Y = \alpha$.

It is easily seen that 1_X, for each object X, is uniquely determined by these properties; it is called the *identity morphism* for X. Given $\alpha : X \to Y$, a morphism $\alpha' : Y \to X$ such that $\alpha\alpha' = 1_X$, $\alpha'\alpha = 1_Y$ is called an *inverse* of α; such an α' may not exist, but if it does, it is uniquely determined by α; it will denoted by α^{-1}. Any morphism with an inverse is called an *isomorphism*, and two objects are *isomorphic* if there is an isomorphism between them.

An obvious example is the category whose objects are all sets, with mappings between them as morphisms; this category is usually denoted by Ens (for 'ensemble', French for 'set'). More precisely, the morphisms are triples (α, X, Y), where the source and target are named, to distinguish (α, X, Y) from (α, X, Y'), where Y' is a subset of Y containing the image of X under α. Similarly we have Gp, the category of groups and homomorphisms, and Top, the category of topological spaces and continuous mappings. In the next chapter we shall meet Rg, the category of rings and homomorphisms, and for each ring R, the category Mod_R whose objects are all right R-modules, while the morphisms are all R-homomorphisms; corresponding definitions apply for the category ${}_R\mathrm{Mod}$ of left R-modules.

As we saw in Section 1.1, we cannot speak of the 'set of all sets' without rapidly reaching contradictions; the simplest way out of this dilemma is to refer to the *class* of all sets, and to keep a distinction between classes and sets. A set may be thought of as a 'small' class; in this sense a category is said to be *small* if the class of its objects is a set. The categories listed above, Ens, Gp, Top, Rg, Mod_R, ${}_R\mathrm{Mod}$, are not small. But any group (more generally, any monoid) can be regarded as a category with a single object, with multiplication everywhere defined; this provides an example of a small category.

Given a category $\mathscr{A}$, a *subcategory* $\mathscr{B}$ is a collection of objects and morphisms of $\mathscr{A}$ which forms a category with respect to the composition in $\mathscr{A}$. Thus $\mathscr{B}(X, Y) \subseteq \mathscr{A}(X, Y)$ for any $\mathscr{B}$-objects X, Y; if equality always holds here, $\mathscr{B}$ is called a *full* subcategory of $\mathscr{A}$. Thus a full subcategory is determined once we have specified

the objects. For example, the category Ab of abelian groups is a full subcategory of Gp.

From every category $\mathscr{A}$ we obtain another category $\mathscr{A}^O$, called its *opposite*, by reversing all the arrows. Thus $\mathscr{A}^O$ has the same objects as $\mathscr{A}$ but for each $\mathscr{A}$-morphism $\alpha : X \to Y$ there is a morphism $\alpha^o : Y \to X$ in $\mathscr{A}^O$, with multiplication $(\alpha\beta)^O = \beta^o\alpha^o$, whenever both sides are defined.

A *functor* F from one category $\mathscr{A}$ to another, $\mathscr{B}$, is a function which assigns to each $\mathscr{A}$-object X a $\mathscr{B}$-object X^F and to each $\mathscr{A}$-morphism $\alpha : X \to X'$ a $\mathscr{B}$-morphism $\alpha^F : X^F \to X'^F$ such that

F.1 If $\alpha\beta$ is defined in $\mathscr{A}$, then $\alpha^F.\beta^F$ is defined in $\mathscr{B}$ and $\alpha^F.\beta^F = (\alpha\beta)^F$.
F.2 $1_X^F = 1_{X^F}$, for each $\mathscr{A}$-object X.

Thus a functor may be described succinctly as a homomorphism of categories. More precisely, the functor defined above is called *covariant*; by a *contravariant* functor one understands a functor G from $\mathscr{A}$ to $\mathscr{B}$ which assigns to each $\mathscr{A}$-object X a $\mathscr{B}$-object X^G and to each $\mathscr{A}$-morphism $\alpha : X \to X'$ a $\mathscr{B}$-morphism $\alpha^G : X'^G \to X^G$ (note the reversed order) such that **F.2** holds, while **F.1** is replaced by

F.1^O. If $\alpha\beta$ is defined in $\mathscr{A}$, then $\beta^G.\alpha^G$ is defined in $\mathscr{B}$ and equals $(\alpha\beta)^G$.

Thus a contravariant functor from $\mathscr{A}$ to $\mathscr{B}$ may be described as an *anti*-homomorphism from $\mathscr{A}$ to $\mathscr{B}$, or also as a homomorphism from $\mathscr{A}^O$ to $\mathscr{B}$ (or from $\mathscr{A}$ to $\mathscr{B}^O$).

As an example of a functor we have the derived group G' of a group G, that is, the subgroup generated by all commutators $(x, y)(x, y \in G)$. For every group homomorphism $f : G \to H$ there is a homomorphism $f' : G' \to H'$, obtained by restriction from f, and it is clear that $(fg)' = f'g'$, $1' = 1$. On the other hand, the centre of a group cannot be regarded as a functor; if the centre of G is $Z(G)$, then a homomorphism $G \to H$ need not map $Z(G)$ into $Z(H)$, as we see by taking G to be an abelian subgroup of H, not contained in $Z(H)$, for a non-abelian group H.

All the categories mentioned above are *concrete*, in the sense that there is a function F to Ens, such that the induced mapping of hom-sets $\mathscr{A}(X, Y) \to \mathrm{Ens}(X^F, Y^F)$ is injective. This functor F, associating with each group, ring etc. its underlying set, is called the *forgetful* functor: it 'forgets' the group (resp. ring etc.) structure.

Frequently one wants to compare two functors. Given two functors S, T from one category $\mathscr{A}$ to another (possibly the same) $\mathscr{B}$, we define a *natural transformation* from S to T as a family of $\mathscr{B}$-morphisms $\varphi_X : X^S \to X^T$ for each $\mathscr{A}$-object X, such that for any $\mathscr{A}$-morphism $f : X \to Y$ and $S, T : \mathscr{A} \to \mathscr{B}$, we have $f^S\varphi_Y = \varphi_X f^T$:

$$\begin{array}{ccc} X^S & \xrightarrow{f^S} & Y^S \\ {\scriptstyle \varphi_X}\downarrow & & \downarrow{\scriptstyle \varphi_Y} \\ X^T & \xrightarrow{f^T} & Y^T \end{array}$$

A natural transformation with an inverse which is again a natural transformation is called a *natural isomorphism*. Here S and T were assumed covariant throughout, but the same definition applies mutatis mutandis when both S and T are contravariant.

Examples. If V is a finite-dimensional vector space over a field k, and $V^* = \mathrm{Hom}(V, k)$ is the dual space, then $\dim V = \dim V^*$ and so the spaces V, V^* are isomorphic, but the isomorphism is not natural (it depends on the choice of bases in V, V^*). On the other hand, there is a natural isomorphism between V and its bidual V^{**}. Let us write $\langle x, \alpha\rangle$ for the value of $\alpha \in V^*$ at $x \in V$ and (f, α) for the value of $f \in V^{**}$ at $\alpha \in V^*$. Then a natural transformation from V to V^{**} is given by

$$x \mapsto \bar{x}, \quad \text{where } \bar{x} \in V^{**} \text{ is defined by } (\bar{x}, \alpha) = \langle x, \alpha\rangle. \tag{3.3.1}$$

We observe that we cannot expect to find a natural transformation from V to V^* because the correspondence $V \mapsto V^*$ is a contravariant functor. Now it can be shown that for finite-dimensional spaces the correspondence (3.3.1) is a natural isomorphism (see Section 4.6 and Section 4.9).

As another example, consider, for any group G, the quotient $G^{ab} = G/G'$, also called the *abelianization* of G. It is easy to verify that for any homomorphism $f : G \to A$ from G to an abelian group A there exists a homomorphism $f' : G^{ab} \to A$ such that f is equal to the natural map $G \to G^{ab}$ followed by f'. This is expressed by saying that G^{ab} is the *universal abelian homomorphic image* of G and it has the *universal mapping property*.

Two categories $\mathscr{A}$, $\mathscr{B}$ are said to be *isomorphic* if there is a functor $T : \mathscr{A} \to \mathscr{B}$ with an inverse, i.e. a functor $S : \mathscr{B} \to \mathscr{A}$, such that $ST = 1$, $TS = 1$. For example, the category of abelian groups is isomorphic to the category of $\mathbf{Z}$-modules; it is well known that every abelian group may be considered as a $\mathbf{Z}$-module and vice versa. Nevertheless the notion of isomorphism between categories is rather restrictive; it leaves out of account the fact that isomorphic objects in a category are for many purposes interchangeable. For this reason the following notion of equivalence is more useful:

Two categories $\mathscr{A}$, $\mathscr{B}$ are said to be *equivalent* if there are two covariant functors $T : \mathscr{A} \to \mathscr{B}$, $S : \mathscr{B} \to \mathscr{A}$ such that TS is naturally isomorphic to the identity functor on $\mathscr{A}$, and similarly ST is naturally isomorphic to the identity on $\mathscr{B}$. When this holds for contravariant functors, $\mathscr{A}$ and $\mathscr{B}$ are called *dual* or *anti-equivalent*. For example, the reversal operator $\mathrm{op} : \mathscr{A} \to \mathscr{A}^O$ is a duality (which happens to be its own inverse).

Any functor $T : \mathscr{A} \to \mathscr{B}$ defines for each pair of $\mathscr{A}$-objects X, Y a mapping

$$\mathscr{A}(X, Y) \to \mathscr{B}(X^T, Y^T). \tag{3.3.2}$$

T is called *faithful* if (3.3.2) is injective, *full* if (3.3.2) is surjective and *dense* if each $\mathscr{B}$-object is isomorphic to one of the form X^T, for some $\mathscr{A}$-object X. For an equivalence functor T, (3.3.2) is a bijection, so in this case T is full and faithful, and clearly it is also dense. Conversely, suppose that T is full, faithful and dense. Then (3.3.2) is

an isomorphism; moreover, for each $\mathscr{B}$-object Z we can by density find an $\mathscr{A}$-object Z^S such that $Z^{ST} \cong Z$, and now we can use the isomorphism (3.3.2) to transfer any map between $\mathscr{B}$-objects to a map between the corresponding $\mathscr{A}$-objects. Thus we obtain

Proposition 3.3.1. *A functor is an equivalence if and only if it is full, faithful and dense.* ■

To illustrate the result, let k be a field and consider Vec_k, the category of all finite-dimensional vector spaces over k with linear mappings as morphisms. In Vec_k we have the subcategory Col_k consisting of all column vectors over k, i.e. all spaces $k^n (n \geq 0)$. Let us choose, for each vector space V of dimension n, an isomorphism $\theta_V : V \to k^n$. Define $T : \mathrm{Col}_k \to \mathrm{Vec}_k$ as the inclusion functor and $S : \mathrm{Vec}_k \to \mathrm{Col}_k$ as follows: $V^S = V\theta_V = k^n$, where $n = \dim V$, and given $f : U \to V$, we put $f^S = \theta_U^{-1} f \theta_V$. This definition ensures that θ is a natural transformation from Vec_k to Col_k. We may without loss of generality take θ_V to be the identity when $V = k^n$. In that case we have $TS = 1$, while ST is naturally isomorphic to 1, via θ. Thus Vec_k is equivalent to Col_k. We observe that Col_k is a small category, so Vec_k is equivalent to a small category, though not itself small. We also note that we cannot choose a smaller category than Col_k, for it has only one object of any given isomorphism type. A category with this property is said to be *skeletal*, and for any category, a skeletal subcategory equivalent to it is called a *skeleton*. It is clear that any category has a skeleton, for we can always choose a subcategory by taking one copy from each isomorphism class of objects. However, the skeleton need not be small; e.g. Ens does not have a small skeleton, since there are sets of arbitrary size, but a small skeleton exists for the subcategory of finite sets.

In any category $\mathscr{A}$ an *initial object* is an $\mathscr{A}$-object I such that there is just one morphism from I to any $\mathscr{A}$-object; thus $\mathscr{A}(I, X)$ always has just one element and in particular, the only map $I \to I$ is the identity on I. A category may have more than one initial object, but they are all isomorphic, for if I, I' are both initial, then there exist unique morphisms $\alpha : I \to I'$, $\beta : I' \to I$; hence $\alpha\beta : I \to I$ must be the identity on I, and likewise $\beta\alpha$ is the identity on I'; therefore α is an isomorphism. Now a *final object* in $\mathscr{A}$ is defined as an initial object in the opposite category $\mathscr{A}^{\mathrm{O}}$. What we have proved can be stated as follows:

Proposition 3.3.2. *In any category any two initial (or final) objects are isomorphic, by a unique isomorphism.* ■

For example, given any group G, we can form the category (G, Ab) whose objects are homomorphisms from G to an abelian group, $\lambda_A : G \to A$, while the morphisms are homomorphisms between abelian groups, $f : A \to B$ such that $\lambda_B = \lambda_A f$. An initial object in this category is the natural homomorphism from G to its abelianization: $G \to G^{ab}$.

Categories with initial objects often arise in the following way. Let $\mathscr{A}$ be any category and P be any $\mathscr{A}$-object. We form a new category $(P, \mathscr{A})$, called the

comma category determined by P, whose objects are the morphisms $\varphi_X : P \to X$ to an $\mathscr{A}$-object X, while the morphisms are $\mathscr{A}$-morphisms $f : X \to Y$ such that $\varphi_Y = \varphi_X f$. It can easily be verified that this new category has $1 : P \to P$ as initial object. For example, let $U : \mathrm{Gp} \to \mathrm{Ens}$ be the forgetful functor associating with each group its underlying set and for any set X form the comma category (X, Gp^U). Its objects are maps from X to G^U, the set underlying the group G, and its morphisms are commutative triangles arising from homomorphisms $f : G \to H$. This category has an initial object, consisting of a group F_X and a map ν from X to F_X^U such that any map from X to a group can be factored by ν. This group F may be described as the *universal group* on X; it is better known as the *free group* on X (see FA Chapter 3).

As a further illustration of a universal mapping property we have the factor theorem (Theorem 2.3.1), which may be stated as follows:

Theorem 3.3.3. *For any group G and a normal subgroup N of G there exists a group G/N (the quotient of G by N) with a homomorphism $\nu : G \to G/N$ (the natural homomorphism) which is universal for homomorphisms from G whose kernel contains N.* ■

Categories were introduced in the 1940s by Eilenberg and Mac Lane [1945] to state topological results concisely in general form.

Exercises

1. Show that the set $\mathbf{N}_0$ consisting of all the natural numbers and 0 is a skeleton for the category of finite sets and mappings.
2. Show that Ab is a full subcategory of Gp, and that the category of monoids is a subcategory of the category of semigroups which is not full.
3. If $\mathscr{A}$ is any category and I is a small category, show that there is a category $\mathrm{Fun}(I, \mathscr{A})$ whose objects are functors from I to $\mathscr{A}$ and the morphisms are natural transformations.
4. Let I be a small category. Show that the functor from I^{O} to $\mathrm{Fun}(I, \mathrm{Ens})$ which maps X to $I(X, -)$ is full and faithful. Is it dense?
5. Let I be a small category and for any I-objects X, Y write $X \leq Y$ iff $I(X, Y) \neq \varnothing$. Show that '$\leq$' is a preordering on the object set of I. Verify that conversely, any preordered set can be made into a small category by introducing a morphism $\alpha : x \to y$ whenever $x \leq y$. Show that any skeleton of the resulting category is an ordered set.
6. Show that there is no natural transformation from a covariant funtor to a contravariant functor.
7. Find a category with a skeleton which is not small.

3.4 Boolean Algebras

We have seen that in any distributive lattice complements in an interval, when they exist, are unique. Thus if L is a distributive lattice with 0 and 1 in which every element has a complement, we can regard the process of associating with each element x its complement x' as a *unary* operator, i.e. an operator with one argument. A complemented distributive lattice with greatest and least element is called a *Boolean algebra* (after George Boole). Thus a Boolean algebra is a set with two binary operators $\vee$, $\wedge$, a unary operator $'$ and constants 0, 1, satisfying the laws (3.1.1)–(3.1.4) and the distributive law, as well as the equations

$$x \wedge x' = 0, \quad x \vee x' = 1. \tag{3.4.1}$$

Clearly $0' = 1$, $1' = 0$, and since x' is the unique complement of x,

$$x'' = x. \tag{3.4.2}$$

We note a consequence, known as *De Morgan's laws* (after Augustus De Morgan):

$$(x \wedge y)' = x' \vee y', \quad (x \vee y)' = x' \wedge y'. \tag{3.4.3}$$

For we have $(x' \vee y') \wedge (x \wedge y) = [x' \wedge (x \wedge y)] \vee [y' \wedge (x \wedge y)] = 0$, and $(x' \vee y') \vee (x \wedge y) = [(x' \vee y') \vee x] \wedge [(x' \vee y') \vee y] = 1$. Hence $x' \vee y'$ is the complement of $x \wedge y$ and we obtain the first Equation (3.4.3); the second follows by duality.

We also note that

$$x \leq y \Leftrightarrow x \wedge y' = 0 \Leftrightarrow x' \vee y = 1. \tag{3.4.4}$$

For if $x \leq y$, then $x \wedge y' \leq y \wedge y' = 0$, hence $x \wedge y' = 0$. Conversely, if $x \wedge y' = 0$, then $x \vee y = (x \vee y) \wedge 1 = (x \vee y) \wedge (y' \vee y) = (x \wedge y') \vee y = y$, and so $x \leq y$. This proves the first equivalence in (3.4.4); the second follows by applying (3.4.3).

Example 1. The set of all subsets $\mathscr{P}(X)$ of any set X is a Boolean algebra if we put $0 = \varnothing$, $1 = X$ and for $Y \in \mathscr{P}(X)$ take Y' to be the complement of Y in X. More generally, any system of subsets of X closed under finite unions and complements is a Boolean algebra; the closure under finite intersections follows by (3.4.3) and 0, 1 are present as the empty union and its complement. This is merely a subalgebra of $\mathscr{P}(X)$, also called a *field of sets* in X.

Let X be any infinite set. Then the subalgebra of $\mathscr{P}(X)$ generated by all the finite subsets of X consists of all the finite subsets of X and their complements, called the *cofinite* subsets of X.

We remark that a sublattice of $\mathscr{P}(X)$ need not be a Boolean algebra, because it may not be closed under complements.

Example 2. In any interval $I = [a, b]$ of a distributive lattice L, the set of elements of I which have a complement in I forms a Boolean algebra.

Example 3. The 2-element lattice is a Boolean algebra. If the elements are a, b and $a \vee b = b$, say, then $a < b$ and we get a Boolean algebra by putting $a = 0$, $b = 1$, $0' = 1$, $1' = 0$. This lattice will be denoted by **2**. The 1-element lattice is also a Boolean algebra, called the *trivial* algebra; it is usually excluded from consideration. Thus **2** is the smallest non-trivial Boolean algebra.

Example 4. The set of all propositions in logic forms a Boolean algebra, taking $\wedge$, $\vee$ to be conjunction ('and') and disjunction ('or') respectively and x' to be the negation of x ('not x'), and for 1, 0 propositions T, F known to be true and false respectively (e.g. $T =$ 'someone is looking at this page now' and $F = T'$). In classical logic each proposition is either true or false, once specific values are assigned to the variables, i.e. each proposition A has a truth-value $f(A)$, and it is easily verified that f is a homomorphism of the Boolean algebra of all propositions into **2**.

It is clear that the duality holding in general lattices extends to Boolean algebras. Thus from any law holding in Boolean algebras we obtain another law by interchanging $\vee$, $\wedge$ and 0, 1.

Given Boolean algebras A and B, by a *dual homomorphism* one understands a mapping $f : A \to B$ such that

$$x'f = (xf)', \quad (x \vee y)f = xf \wedge yf, \quad (x \wedge y)f = xf \vee yf. \tag{3.4.5}$$

Dual isomorphisms etc. are defined similarly. From De Morgan's laws (3.4.3) we see that every Boolean algebra admits a dual automorphism, the *natural duality*, defined by the complementation mapping $x \mapsto x'$. In a non-trivial Boolean algebra this duality has no fixed point, and its square is the identity. This shows that every finite non-trivial Boolean algebra has an even number of elements.

The Boolean algebra **2** possesses a remarkable property, its functional completeness: we shall find that every function of n variables on **2** with values in **2** can be expressed as a Boolean polynomial. Here a *Boolean polynomial* in $x_1, \ldots, x_n$ is defined by the following rules:

1. Each x_i is a Boolean polynomial.
2. If u is a Boolean polynomial, then so is u'.
3. If u, v are Boolean polynomials, then so is $u \wedge v$.

Of course $u \vee v$ can be expressed as $(u' \wedge v')'$. For example $(x \wedge y') \vee (z'' \wedge x)$ is a Boolean polynomial. Two Boolean polynomials are said to be *equal* if we can pass from one to the other by applying the laws of Boolean algebras, viz. (3.1.1)–(3.1.4) and (3.4.1)–(3.4.3). It is clear that equal Boolean polynomials are equal as functions on any Boolean algebra; below we shall prove a converse: two Boolean polynomials which define the same function on a given non-trivial Boolean algebra must be equal as polynomials. Clearly it will be enough to prove this for the smallest non-trivial algebra **2**. The result is proved by finding a normal form for polynomials, which is shown to be unique; more precisely, we shall find two normal forms dual to each other.

To describe these normal forms, let us define a *minterm* in $x_1, \ldots, x_n$ as

$$X_1 \wedge \ldots \wedge X_n, \text{ where } X_i \text{ is } x_i \text{ or } x_i' \quad (i = 1, \ldots, n). \tag{3.4.6}$$

Now a Boolean polynomial f is said to be in *disjunctive normal form* if it has the form

$$f = f_1 \vee \ldots \vee f_r, \tag{3.4.7}$$

where each f_j is a minterm in $x_1, \ldots, x_n$. For example, the disjunctive normal form for $x' \wedge y$ is $x' \wedge y$, while for $x \vee y'$ it is $(x \wedge y') \vee (x' \wedge y') \vee (x \wedge y)$. Dually we define a *maxterm* in $x_1, \ldots, x_n$ as

$$X_1 \vee \ldots \vee X_n, \text{ where } X_i \text{ is } x_i \text{ or } x_i' \quad (i = 1, \ldots, n) \tag{3.4.8}$$

and the *conjunctive normal form* for f is

$$f = f_1 \wedge \ldots \wedge f_r, \tag{3.4.9}$$

where each f_j is a maxterm in $x_1, \ldots, x_n$. For example, the conjunctive normal form for $x \wedge y'$ is $(x \vee y') \wedge (x' \vee y') \wedge (x \vee y)$ while for $x' \vee y$ it is $x' \vee y$. With these definitions we have

Lemma 3.4.1. *Any Boolean polynomial in $x_1, \ldots, x_n$ is equal to a polynomial in disjunctive normal form, and likewise to a polynomial in conjunctive normal form.*

Proof. Let p be any Boolean polynomial in $x_1, \ldots, x_n$. By the distributive law we can write p as a disjunction (3.4.7), where each f_j is a conjunction of some of the $x_1, \ldots, x_n, x_1', \ldots, x_n'$. By the idempotent law we can omit repetitions and if both x_i and x_i' occur in f_j, then f_j is 0 and so may be omitted, while x_i'' can be replaced by x_i. In order to obtain a disjunction of minterms we order the variables in f_j (by the commutative law), say in ascending order; if neither x_i nor x_i' occurs, then since $f_j = (f_j \wedge x_i) \vee (f_j \wedge x_i')$, we can replace f_j by $f_j \wedge x_i$ and $f_j \wedge x_i'$. In this way we obtain an expression (3.4.7) for p, where each f_j is a minterm in $x_1, \ldots, x_n$, so p has been expressed in disjunctive normal form. The result for conjunctive normal forms follows by duality. ■

Let us consider a minterm (3.4.6). It takes the value 1 for just one set of values in $\mathbf{2}^n$, namely $a = (a_1, \ldots, a_n)$, where

$$a_i = \begin{cases} 1 & \text{if } X_i = x_i, \\ 0 & \text{if } X_i = x_i'. \end{cases} \tag{3.4.10}$$

From (3.4.10) we see that for each n-tuple $a \in \mathbf{2}^n$ there is just one minterm taking the value 1 at a and 0 elsewhere; this minterm will be denoted by ε_a. Thus ε_a is the characteristic function of the 1-point subset $\{a\}$ of $\mathbf{2}^n$. More generally, if S is any subset of $\mathbf{2}^n$, then its characteristic function can be written as a Boolean polynomial:

$$\chi_S = \bigvee_{a \in S} \varepsilon_a. \tag{3.4.11}$$

Now any function $f : \mathbf{2}^n \to \mathbf{2}$ is the characteristic function of a subset of $\mathbf{2}^n$, namely $1f^{-1} = \{a \in \mathbf{2}^n | f(a) = 1\}$. Hence (3.4.11) can be used to express f as a Boolean polynomial; moreover this expression is unique, since f is completely determined by $1f^{-1}$. Thus every function on $\mathbf{2}^n$ can be represented by exactly one Boolean polynomial in disjunctive normal form. A dual argument applies to conjunctive normal forms. Let us sum up our conclusions:

Theorem 3.4.2. *Each Boolean polynomial is equal to a unique expression in disjunctive normal form* (3.4.7) *and to a unique expression in conjunctive normal form* (3.4.9) *and it defines a function on each Boolean algebra. Moreover, on the Boolean algebra* **2**, *every function of n variables is given by a Boolean polynomial, but on any Boolean algebra of more than two elements there are functions not represented by Boolean polynomials.*

Proof. We have seen that every Boolean polynomial can be put in disjunctive normal form (Lemma 3.4.1) and that every function on **2** has a unique expression in disjunctive normal form. This shows that the disjunctive normal form is unique (since it is unique on **2**). Now the number of mappings from $\mathbf{2}^n$ to **2** is 2^{2^n} so this must be the number of expressions in disjunctive normal form (this is also easily verified directly). But on a Boolean algebra with b elements there are b^{b^n} functions of n arguments, so for $b > 2$ not all of them can be written as Boolean polynomials. ■

The fact that on **2** every function is a Boolean polynomial is expressed by saying that **2** is *functionally complete.*

As an example consider the function given by Table 3.1.

Table 3.1

000	001	010	100	011	101	110	111
0	1	0	0	0	1	0	1

The corresponding polynomial in disjunctive normal form is

$$(x \wedge y' \wedge z) \vee (x' \wedge y' \wedge z) \vee (x \wedge y \wedge z).$$

We note that this may be simplified to give $(y' \wedge z) \vee (x \wedge y \wedge z)$ or also $(x' \wedge y' \wedge z) \vee (x \wedge z)$. Thus uniqueness holds only for the full disjunctive normal form, where each variable occurs in each minterm.

We also note the interpretation of Theorem 3.4.2 in terms of the propositional calculus (Example 5 above). It states that every assignment of truth values to n propositions can be realized by a propositional function of n variables.

An important help in understanding finite Boolean algebras is a representation theorem which shows them to have the form $\mathcal{P}(X)$ for finite sets X. We shall establish this result in the slightly more general context of distributive lattices. Let P be a partially ordered set and denote by P^* the set of its lower segments. Each element c of P defines a *principal lower segment* $|c] = \{x \in P | x \leq c\}$; by identifying c with $|c]$ we

may regard P as a subset of P^*. It is clear that P^* is a lattice with union and intersection as operations; thus it is a sublattice of $\mathcal{P}(P)$ and hence is distributive. In this way every partially ordered set can be embedded in a distributive lattice. It is a remarkable fact that every finite distributive lattice is of this form.

Theorem 3.4.3. *Let L be a distributive lattice of finite length. Then there is a finite partially ordered set P, unique up to order-isomorphism, such that $L \cong P^*$. If L and P correspond in this way, then the length of L equals the number of elements of P.*

Proof. Denote by L^* the set of all join-irreducible elements of L, partially ordered by inclusion. By Lemma 3.2.7, each $c \in L$ can be represented by the set of all join-irreducible elements below it, and the sets of join-irreducible elements occurring in this way are just the lower segments of L^*; thus L is order-isomorphic to the set L^{**} of these lower segments, as claimed.

To show that P is uniquely determined by P^* we shall verify that $P^{**} \cong P$. Consider $\alpha \in P^*$; by definition, α is a lower segment in P. If $a_1, \ldots, a_r$ are the different maximal elements of α, then $x \in \alpha$ iff $x \leq a_1$ or $\ldots$ or $x \leq a_r$. Hence

$$\alpha = |a_1] \vee \ldots \vee |a_r],$$

and it follows that α is join-irreducible in P^* iff it is principal. Thus P^{**}, the set of join-irreducible elements of P^*, is just the set of principal lower segments of P. But the latter is order-isomorphic to P, as we saw; therefore $P^{**} \cong P$, as claimed.

Finally, if L and P correspond and P has n elements, then we can form a maximal chain in L by picking a minimal element $a_1 \in P$, next a minimal element a_2 in $P\backslash\{a_1\}$ and so on; therefore each maximal chain in L has length n. ■

It is not hard to see that this correspondence between ordered sets and lattices is a contravariant functor in each direction, providing a duality, i.e. an anti-equivalence, between the category of finite partially ordered sets and order-homomorphisms and finite distributive lattices and lattice-homomorphisms. Here the natural isomorphism from P to P^{**} takes the following form: With each $x \in P$ we associate an element $\bar{x} \in P^{**}$ defined as a mapping $P^* \to \mathbf{2}$ by

$$\bar{x}(\alpha) = x\alpha \quad \text{for all} \quad \alpha \in P^*.$$

By definition of P^*, if $x \leq y$ in P, then $x\alpha \leq y\alpha$ for all $\alpha \in P^*$, hence $\bar{x}(\alpha) \leq \bar{y}(\alpha)$, and so $\bar{x} \leq \bar{y}$. This shows that the natural mapping $P \to P^{**}$ is an order-isomorphism. In a similar way it can be shown that the natural mapping $L \to L^{**}$ is a lattice-isomorphism.

In any pair L, P that correspond, P is generally simpler than L. For example, the free distributive lattice on three generators has length 6 and consists of 18 elements. The corresponding partially ordered set is the three-cornered crown shown in Figure 3.8.

Other examples of partially ordered sets and their corresponding lattices are given in Figure 3.9.

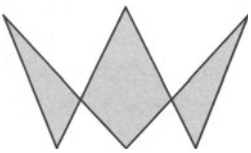


Figure 3.8

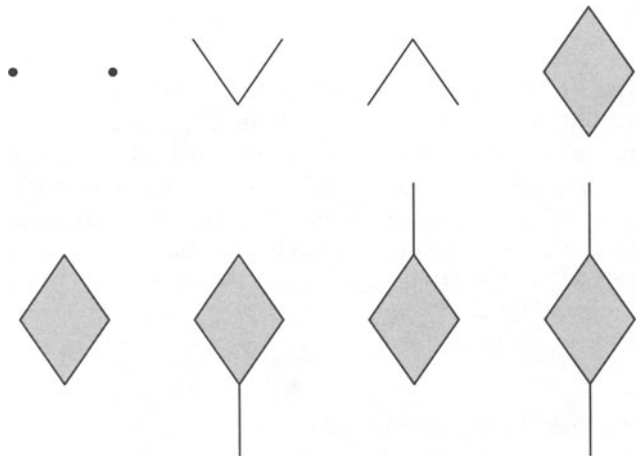


Figure 3.9

Let L be a distributive lattice of finite length n. By Theorem 3.4.3, the set L^* of its join-irreducible elements has n elements and L is isomorphic to a sublattice of 2^{L^*}. In particular, it follows that L is finite:

Corollary 3.4.4. *Any distributive lattice L of finite length n is a sublattice of $\mathscr{P}(P)$, where P, the set of join-irreducible elements of L, has n elements. In particular, $|L| \leq 2^n$.* ■

The bound can be attained, since for a totally unordered set P of n elements $\mathbf{2}^P$, the set of all order-homomorphisms from P to $\mathbf{2}$, has exactly n elements.

Every Boolean algebra is a distributive lattice and it is natural to ask which partially ordered sets correspond to finite Boolean algebras in the duality of Theorem 3.4.3. This is answered by the next result. To state it let us define an *atom* in a Boolean algebra as an element a such that $a > 0$ but no element x satisfies $a > x > 0$. It is clear that any atom is join-irreducible; conversely, if $p \in L$ is join-irreducible, then $p \neq 0$ and since $p = p \wedge 1 = p \wedge (a \vee a') = (p \wedge a) \vee (p \wedge a')$ for any $a \in L$, we have either $p = p \wedge a$, i.e. $p \leq a$, or $p = p \wedge a'$, i.e. $p \leq a'$ and so $p \wedge a = 0$. Therefore a cannot satisfy $0 < a < p$, and it follows that p is an atom.

Theorem 3.4.5. *Let L be a finite distributive lattice and $P = L^*$ be the associated partially ordered set. Then L is complemented (and hence a Boolean algebra) if and only if P is totally unordered. Hence any finite Boolean algebra, of length n, has 2^n elements and is isomorphic to $\mathscr{P}(A)$, where A, the set of its atoms, has n elements.*

Proof. If P is totally unordered (i.e. an anti-chain), every subset of P is a lower segment, hence $P^* \cong \mathscr{P}(P)$, and this is the Boolean algebra of all subsets of P, with 2^n elements. Conversely, if L is a Boolean algebra, then the join-irreducible elements are just its atoms, as we have just seen, and it is clear that the atoms of L form a totally unordered set. ∎

Theorem 3.4.5 (or rather, the Boolean algebra case of Theorem 3.4.3) has been generalized to arbitrary Boolean algebras by Marshall H. Stone, whose representation theorem establishes a duality between the category of all Boolean algebras and the category of all Boolean spaces, i.e. totally disconnected Hausdorff spaces. The space corresponding to a Boolean algebra B is the set of all homomorphisms $B \to \mathbf{2}$, as subset of $\mathbf{2}^B$ with the product topology, while the algebra corresponding to a space T is the set of all continuous mappings $T \to \mathbf{2}$, as subset of the product $\mathbf{2}^T$.

Exercises

1. Show that in any Boolean algebra, $(x \vee y) \wedge (x' \vee z) = (x' \wedge y) \vee (x \wedge z)$.
2. Show that in (3.4.5) the first two formulae imply the third.
3. Prove the general distributive law: $(\bigvee_I a_i) \wedge (\bigvee_J b_j) = \bigvee_{I \times J} (a_i \wedge b_j)$, for any finite sets I, J.
4. Show that any Boolean algebra with ascending chain condition is finite.
5. Express the following functions in conjunctive normal form for x, y, z:

   (i) $(x \wedge y' \wedge z) \vee (x \wedge y \wedge z') \vee (x \wedge y' \wedge z') \vee (x' \wedge y)$;
   (ii) $(x \wedge y \wedge z') \vee (x' \wedge y' \wedge z) \vee (x \wedge y' \wedge z') \vee (x' \wedge y' \wedge z') \vee (x' \wedge y \wedge z)$;

   and for x, y, z, u:

   (iii) $(x' \wedge y \wedge z) \vee (y' \wedge z) \vee (x' \wedge z \wedge u) \vee (z \wedge u')$.

6. Verify that for any finite distributive lattice L, the natural mapping $L \to L^{**}$ is a lattice isomorphism.
7. Show that a finite partially ordered set P is order-isomorphic to P^* precisely when it is totally ordered.
8. Show that two Boolean algebras with the same number of elements are isomorphic.
9. In a Boolean algebra let $a \leq x \leq b$; show that $(x' \vee a) \wedge b$ is a relative complement of x in $[a, b]$. Verify that for an element x with a complement x' in a modular lattice this remains true, but not in general lattices.
10. Let B be a Boolean algebra. Show that for $a, b \in B$ there is a homomorphism $B \to \mathbf{2}$ which maps a to 1 and b to 0, unless $a \leq b$. Deduce that B can be embedded in $\mathbf{2}^I$, for some set I.

Further Exercises for Chapter 3

1. Let $\mathscr{C}$ be a family of subsets of a set S and assume that $\mathscr{C}$ is closed under arbitrary intersections. Show that $\mathscr{C}$ is a lattice: what are the conditions on S for $\mathscr{C}$ to be a sublattice of $\mathscr{P}(S)$?
2. Let L be a complete lattice and f an order-homomorphism of L into itself. Define $A = \{x \in L | x \leq xf\}$; show that if $a = \sup A$, then af is an upper bound for A, and deduce that $af = a$. Show that the fixed points of f again form a complete lattice with respect to the order induced by L. Is this lattice necessarily a sublattice of L?
3. Let A, B be partially ordered sets such that A is order-isomorphic to a lower segment of B and B is order-isomorphic to an upper segment of A. Show that there is a bijection $f : A \to B$ such that for no pair $x < y \in A$ is $xf \geq yf$. Deduce the Schröder–Bernstein theorem (Theorem 1.1.2). (Hint. If $g : A \to B$, $h : B \to A$ are the given order-homomorphisms and $\mathscr{S}(A)$ is the set of all lower segments of A, define $x\theta = ((xg)'h)'$ for $x \in \mathscr{S}(A)$, where $'$ denotes the complement. Verify that θ is an order-homomorphism of a complete lattice and use Exercise 2.)
4. Let A, B be totally ordered sets. Show that if A is order-isomorphic to a lower segment of B and B is order-isomorphic to an upper segment of A, then A, B are order-isomorphic.
5. Show that the finite unions of half-closed intervals $[a, b) = \{x \in \mathbf{R} | a \leq x < b\}$, including $(-\infty, b)$ and $[a, \infty)$, form a field of sets in $\mathbf{R}$.
6. Show that in any modular lattice of finite length, $[a \wedge b, a]$ and $[b, a \vee b]$ have the same length. Deduce that if a is the join of n atoms, then $[0, a]$ has length at most n. Under what conditions does equality hold?
7. A partially ordered set is said to be *graded* if there is a function $f(x)$ with integer values such that $x < y \Rightarrow f(x) < f(y)$ and if y covers x (i.e. $x < y$ and there is no element between x and y) then $f(y) = f(x) + 1$. In any partially ordered set S with least element 0, define the *height function* $h(x)$ as the sup of lengths of chains from 0 to x. Show that all maximal chains between the same endpoints have the same length iff S is graded by h. Moreover, in this case all bounded chains are finite.
8. Show that in a modular lattice, if the height h is defined as in Exercise 7, then $h(x \vee y) + h(x \wedge y) = h(x) + h(y)$. Conversely, show that in a lattice of finite length this condition implies modularity.
9. A lattice L is said to be *lower semimodular* if whenever a covers b and $c (b \neq c)$, then b and c both cover $b \wedge c$. Show that in a lower semimodular lattice the height function h satisfies $h(x \vee y) + h(x \wedge y) \geq h(x) + h(y)$.

   Show that the lattice of subgroups of a finite p-group is lower semimodular. (Hint. Use the fact that in finite p-group every maximal subgroup has index p and observe that $(B : B \cap C) = (BC : C)$ for any subgroups B, C.)
10. Let V be a 3-dimensional space over $\mathbf{F}_2$, the field of 2 elements and X_i $(i = 1, 2, 3)$ be the set of all vectors with i-component $\neq 0$. Show that the lattice generated by X_1, X_2, X_3 has 18 elements (this is the free distributive lattice on 3 free generators, see FA, Chapter 1).

11. Show that for a complemented modular lattice the maximum condition is equivalent to the minimum condition. Does this remain true if modularity is not assumed?
12. A subset of a lattice L is called an *ideal* if it is a lower segment and is closed under $\vee$; it is *principal* if it has the form $|a] = \{x \in L | x \leq a\}$. Show that the ideals in L form a lattice and the principal ideals form a sublattice isomorphic to L. Show that if L satisfies the ascending chain condition, then every ideal is principal.
13. Let X be an infinite set and $\mathscr{F}$ be the collection of all its finite subsets. Show that $\mathscr{F}$ is an ideal in $\mathscr{P}(X)$ and verify that $\mathscr{P}(X)/\mathscr{F}$ has no atoms and is not complete, as a lattice.
14*. (Yoneda's lemma) Let $F : \mathscr{A} \to$ Ens be a functor and for any $p \in X^F$ define a natural transformation $p^* : h^X = \mathscr{A}(X, -) \to F$ by the rule; if $\alpha \in \mathscr{A}(X, Y)$, then $\alpha \mapsto p\alpha^F$ maps $Yh^X (= \mathscr{A}(X, Y))$ to Y^F. Verify that this is indeed a natural transformation and prove that the resulting mapping $X^F \to \mathrm{Nat}(h^X, F)$ to the set of natural transformations is an isomorphism. (Hint. Define the inverse $\tau \mapsto (1_X)\tau \in X^F$.)
15. (E. V. Huntingdon) Let A be a non-empty finite set with a binary operator $\vee$ and a unary operator $'$ and define $a \wedge b = (a' \vee b')'$. Show that if (i) $a \vee b = b \vee a$, (ii) $a \vee (b \vee c) = (a \vee b) \vee c$, (iii) $(a \wedge b) \vee (a \wedge b') = a$, and (iv) for some $e \in A$ and all $a \in A$, $a \vee e = e$, then A is a Boolean algebra.
16. (M. Sheffer) Let A be a set with a binary operation $a|b$ and define $a' = a|a$, $a \vee b = (a|b)'$, $a \wedge b = a'|b'$. Show that if this operation satisfies (i) $(b|a)|(b'|a) = a$, (ii) $[(c'|a)|(b'|a)]' = a|(b|c)$, and (iii) for some $e \in A$ and all $a \in A$, $e|a = a'$, then A is a Boolean algebra with respect to these operations.
17. Show that any two countably infinite Boolean algebras without atoms are isomorphic.
18. Show that for any partially ordered set P, the set P^* of its lower segments is a distributive lattice and there is a natural order-homomorphism $P \to P^{**}$.
19. Let B be a Boolean algebra. Show that $B \cong [0, a] \times [0, a']$ for any $a \in B$. Deduce that $\mathbf{2}$ is the only indecomposable Boolean algebra, and that any finite Boolean algebra has the form $\mathbf{2}^n$.
20. Let P be a finite partially ordered set and $L = P^*$ be the corresponding distributive lattice. Show that P and L have isomorphic automorphism groups.
21. (W. H. Gottschalk) For any Boolean polynomial $f(x) = f(x_1, \ldots, x_n)$ show that the operations $\alpha : f(x) \mapsto f(x)'$ and $\beta : f(x) \mapsto f(x')$ are automorphisms. Show that the group generated by α, β is the Klein 4-group.
22. A Boolean algebra is called *atomic* if every element > 0 contains an atom. Show that a complete atomic Boolean algebra is isomorphic to $\mathscr{P}(A)$, where A is the set of its atoms; in particular, every finite Boolean algebra is of this form (cf. Theorem 3.4.5).
23. Show that any generating set of the Boolean algebra $\mathbf{2}^r$ has cardinal at least $\log_2 r$, and that this estimate is best possible.

4 Rings and Modules

Linear algebra deals with fields and vector spaces; here we are concerned with the generalizations to rings and modules over them. Whereas a vector space over a given field is determined up to isomorphism by its dimension, there is much greater variety for modules. Another way of regarding modules is as abelian groups, written additively, with operators. This means that much of general group theory applies, and after recalling the isomorphism theorems, proved for groups in Section 2.3, we treat a number of special situations. Semisimple modules (Section 4.3) come closest to vector spaces; they are direct sums of simple modules, but we have to bear in mind that over a given ring there may be more than one type of simple module. In the free modules (Section 4.6) we have another generalization of vector spaces. The homological treatment of module theory requires the notions of projective and injective module (Section 4.7), and they can usefully be introduced here, as they will occur again in Chapter 10, although their main use will be in FA. Other important notions introduced here are those of matrix ring (Section 4.4) and tensor product (Section 4.8).

4.1 The Definitions Recalled

We recall that a *ring* is a set R with two binary operations with values in R, addition: $x + y$, and multiplication: xy, such that R is an abelian group under addition with neutral element 0, the *zero element*, a monoid under multiplication with neutral element 1, called the *unit element* or *one*, and these operations are linked by the *distributive laws*:

$$x(y + z) = xy + xz, \quad (x + y)z = xz + yz \quad \text{for all } x, y, z \in R.$$

If $1 = 0$, then for any $x \in R$, $x = x.1 = x.0 = 0$, so the ring consists of a single element. This is the *trivial* ring; usually our rings are assumed to be non-trivial. If the commutative law holds: $xy = yx$ for all $x, y \in R$, the ring is said to be *commutative*, but we shall not generally make this assumption.

The best-known example of a ring is the ring of integers $\mathbf{Z}$, which is in fact commutative. The integers also have the property that a product is zero iff at least one of the factors is zero. For any element c of a ring R either (i) $c = 0$, or (ii) $c \neq 0$ and $ca = 0$ or $ac = 0$ for some $a \neq 0$ in R, or (iii) ac and ca never vanish for $a \neq 0$.

If (ii) holds, c is called a *zerodivisor*; if (iii) holds, c is a *non-zerodivisor*. A non-trivial ring without zerodivisors is called an *integral domain*. Some authors use this term only for commutative rings, but we shall not make this restriction. Thus a ring R is an integral domain iff $R\backslash\{0\}$ contains 1 and is closed under multiplication; we shall usually write $R\backslash\{0\} = R^\times$. In an integral domain R the additive order of 1 is called the *characteristic* of R. If $n.1 \neq 0$ for all $n > 0$, R is said to have *characteristic zero*. The alternative is that 1 has finite order, which is easily seen to be a prime; thus every integral domain has characteristic zero or a prime number.

An integral domain R for which $R^\times$ is a group under multiplication is called a *division ring*, or when R is commutative, a *field*. Instead of 'division ring' the term *skew field* will often be used. Generally an element $a \in R$ is called a *unit* or *invertible* if it has an inverse, i.e. there exists $a^{-1} \in R$ such that $aa^{-1} = a^{-1}a = 1$. Clearly any $a \in R$ can only have one inverse, for if a', a'' are two inverses of a, then $a' = a'aa'' = a''$. A *non-unit* is an element which is not a unit; however, an exception will be made for the 'matrix units' e_{ij} $(i, j = 1, \ldots, n)$ to be introduced in Section 4.4, which are of course non-units when $n > 1$.

Two rings R, S are said to be *isomorphic*, if there is a mapping $f : R \to S$ which is a bijection preserving all the ring operations, i.e.

$$(x + y)f = xf + yf, \tag{4.1.1}$$

$$(xy)f = xf.yf, \tag{4.1.2}$$

$$1f = 1, \tag{4.1.3}$$

for all $x, y \in R$. More generally, a mapping $f : R \to S$ satisfying (4.1.1)–(4.1.3) is called a *homomorphism*; such a mapping need not be a bijection. We note that for any such homomorphism the image $\operatorname{im} f$ is a *subring* of S, i.e. a subset of S admitting the operations of S, while the *kernel* of f, $\ker f = 0f^{-1}$, is an *ideal* of R, i.e. a subgroup of the additive group of R such that $R(\ker f) \subseteq \ker f$, $(\ker f)R \subseteq \ker f$. It is worth noting that for an isomorphism the condition (4.1.3) holds automatically, since the 1 of S is uniquely determined as the solution of $xa = ax = a$ for all $a \in S$; however, for a homomorphism it has to be postulated separately. As in the case of groups we define an *endomorphism* of a ring as a homomorphism into itself and an *automorphism* as an isomorphism with itself.

In a commutative ring A the ideal generated by a single element c is denoted by (c) and is called a *principal ideal*. If A is a commutative integral domain in which every ideal is principal, A is called a *principal ideal domain*, often abbreviated as PID.

In any ring R the set $\mathscr{B}(R)$ of all central idempotents of R, i.e. elements e in the centre of R such that $e^2 = e$, is a Boolean algebra under the operations

$$x \wedge y = xy, \quad x \vee y = x + y - xy, \quad x' = 1 - x. \tag{4.1.4}$$

Shortly we shall see that every Boolean algebra may be obtained as the algebra of idempotents of some commutative ring R; in fact R may be chosen to consist entirely of idempotents. Let us define a *Boolean ring* as a ring satisfying the identity

$$x^2 = x. \tag{4.1.5}$$

Such a ring is necessarily commutative. For we have $2x = (2x)^2 = 4x^2 = 4x$, hence $2x = 0$ and so

$$x = -x. \tag{4.1.6}$$

Next we have $x + y = (x + y)^2 = x^2 + xy + yx + y^2$; hence $xy + yx = 0$, so by (4.1.6), $xy = yx$. Thus every Boolean ring is commutative and of characteristic 2.

Proposition 4.1.1. (i) *Given any ring R, the set $\mathscr{B}(R)$ of its central idempotents forms a Boolean algebra relative to the operations* (4.1.4).
(ii) *On any Boolean algebra B define the two operations*

$$x + y = (x \wedge y') \vee (x' \wedge y), \quad xy = x \wedge y. \tag{4.1.7}$$

Then B forms a Boolean ring $\mathscr{R}(B)$ relative to these operations.

Moreover, $\mathscr{B}(\mathscr{R}(B)) \cong B$ for any Boolean algebra B and $\mathscr{R}(\mathscr{B}(R)) \cong R$ for any Boolean ring R.

The first operation (4.1.7) is called the *symmetric difference* of x and y; in the case where $B = \mathscr{P}(X)$, $Y + Z$ represents the subset of X consisting of all elements that are either in Y or in Z but not in both (this is also called 'addition mod 2').

Proof. The verification that $\mathscr{B}(R)$ is a Boolean algebra under the operations (4.1.4) and that $\mathscr{R}(B)$ is a ring under the operations (4.1.7) is routine and may be left to the reader.

Now let B be a Boolean algebra and put $B_1 = \mathscr{B}(\mathscr{R}(B))$. Since $\mathscr{R}(B)$ is a Boolean ring, B_1 and B are the same set, and both are Boolean algebras with the same definition of $\wedge$. In B_1 we have $x' = 1 - x$; from (4.1.7) and the fact that $\mathscr{R}(B)$ is a Boolean ring we see that $1 - x = (1 \wedge x') \vee (0 \wedge x) = x'$. So x' is the same on B and B_1; likewise for $x \wedge y$, hence this also holds for $x \vee y = (x' \wedge y')'$.

Next take a Boolean ring R and put $R_1 = \mathscr{R}(\mathscr{B}(R))$. Again R and R_1 are the same set; both are Boolean rings with the same definition of product, and if $+_1$ denotes the sum in R_1, then we have $x +_1 y = x(1 - y) + (1 - x)y = x + y$; hence $R_1 \cong R$, as required. ∎

It is clear that Boolean algebra homomorphisms correspond precisely to ring homomorphisms; we have here an example of a category equivalence: $\mathscr{R}$ and $\mathscr{B}$ are mutually inverse functors defining an equivalence between the categories of Boolean rings and Boolean algebras.

Just as a group can act by permutations on a set, so a ring can act by linear transformations on an additive group. This gives rise to the notion of a module, which may also be regarded as a generalization of a vector space. Thus let R be any ring. By a *right* R-module one understands an abelian group M, written additively, with a mapping from $M \times R$ into M, $(x, a) \mapsto xa$, such that for all $x, y \in M$, $a, b \in R$,

$$(x + y)a = xa + ya, \tag{4.1.8}$$

$$x(a + b) = xa + xb, \tag{4.1.9}$$

$$x(ab) = (xa)b, \tag{4.1.10}$$

$$x1 = x. \tag{4.1.11}$$

Sometimes we have, instead of (4.1.10), the rule $x(ab) = (xb)a$; in such cases it is expedient to write the operator on the left, putting ax instead of xa, so this law now takes the form

$$(ab)x = a(bx). \tag{4.1.12}$$

The resulting structure is called a *left* R-module. If we define the *opposite ring* of R as the ring R^O whose additive group is the same as that of R but with the multiplication

$$x.y = yx \quad (x, y \in R)$$

then we can say: a left R-module is a right R^O-module and a right R-module is a left R^O-module. In other words, changing the side from which a ring operates on a module corresponds to a reversal of the order of multiplication in the ring. This means that instead of using an R^O-module we can take the coefficients on the other side so as to get an R-module. Of course when R is commutative then $R^O = R$; in this case the difference between left and right R-modules is purely notational. We shall sometimes write ${}_RM$ resp. M_R to indicate that M is a left resp. right R-module. In general we shall speak of an 'R-module' when the side is not specified, and a 'module' when we do not wish to specify the ring. A module is called *cyclic* if it can be generated by a single element.

As an example we may take the ring itself. Any ring R is a right R-module, with the action of R being defined by right multiplication. The distributive laws ensure that (4.1.8), (4.1.9) hold while (4.1.10) follows by the associative law and (4.1.11) is the property of the unit element. In fact R as right (or left) R-module is cyclic, generated by 1. Any submodule, i.e. subgroup admitting the right multiplication by elements of R, is called a *right ideal* of R. Similarly, left multiplication defines R as a left R-module, with left ideals as submodules; the ideals defined earlier are just subsets of R that are both left ideals and right ideals, also called *two-sided ideals.*

Another example of importance in the sequel is obtained by taking any abelian group A, written additively, and considering the set $E = \mathrm{End}(A)$ of all its endomorphisms. If a, b are endomorphisms acting on the right, and $x, y \in A$, then (4.1.8) just expresses their defining property. We can now define an addition on E by (4.1.9) and a multiplication by (4.1.10), and can verify that E forms a ring for these operations, with the identity mapping as unit element. This verification is straightforward and so may be left to the reader. In this way we obtain a ring structure on E, called the *endomorphism ring* of A, and A is seen to be a right E-module.

Given two right R-modules M, N over a ring R, an R-homomorphism $f : M \to N$, also called an *R-linear map,* is defined as a map f satisfying

$$f(x+y) = fx + fy,\ f(xa) = (fx)a \quad \text{for all } x, y \in M, a \in R.$$

Here we have written f on the left, on the opposite side from the ring coefficients. We shall usually follow this practice; thus a homomorphism of left modules will

be written on the right. This convention is chiefly of use for endomorphisms, which again form a ring, but it is not always possible to adhere to it rigidly (e.g. in the case of bimodules). We shall frequently use left modules, so as to be able to put maps on the right.

If M, N are left R-modules and f, g are homomorphisms from M to N, then so is the map $f + g$ defined by

$$x(f + g) = xf + xg;$$

in this way the set $\mathrm{Hom}_R(M, N)$ of all R-homomorphisms from M to N becomes an abelian group. When R is commutative, this is actually an R-module, taking rf to be defined by

$$x.(rf) = rx.f = r(xf).$$

For general rings this is no longer so; we shall soon meet the appropriate generalization. The set $\mathrm{Hom}_R(M, M)$ of all endomorphisms of M is denoted by $\mathrm{End}_R(M)$. We recall that it is a ring, using the composition of endomorphisms as multiplication.

Let R, S be rings and let M be an abelian group which is both a left R-module and a right S-module, such that

$$(rx)s = r(xs) \quad \text{for all } x \in M, r \in R, s \in S. \tag{4.1.13}$$

Then M is said to be an *(R, S)-bimodule*, and we indicate this by writing ${}_RM_S$. In the case $S = R$ we speak of an R-bimodule. This is then a left and right R-module M such that $(ax)b = a(xb)$ for $x \in M$ and $a, b \in R$. For example, the ring R itself is an R-bimodule, as the associative law shows. Further, when R is commutative, any left R-module may also be considered as a right R-module and hence an R-bimodule.

We remark that any left R-module may be defined as (R,S)-bimodule by means of a homomorphism $f : S \to \mathrm{End}_R(M)$. Likewise a right S-module may be defined as an (R, S)-bimodule by means of an anti-homomorphism $g : R \to \mathrm{End}_S(M)$; here we need an *anti*-homomorphism because R acts on M from the left.

Given an (R,S)-bimodule M and an (R,T)-bimodule N, we can define $\mathrm{Hom}_R(M, N)$ in a natural way as (S,T)-bimodule by the rules

$$x(sf) = (xs)f, x(ft) = (xf)t, \quad \text{where } f : M \to N, x \in M, s \in S, t \in T.$$

The bimodule property follows because

$$x[(sf)t] = [x(sf)]t = ((xs)f)t, \quad x[s(ft)] = (xs)(ft) = ((xs)f)t.$$

This rule: $({}_RM_S, {}_RN_T) \mapsto ({}_S\mathrm{Hom}_R(M, N)_T)$ is easily remembered if it is borne in mind that Hom is contravariant in the first and covariant in the second argument, so the order in S is reversed while that in T is preserved. In particular, when R is commutative, any R-modules may be regarded as R-bimodules and $\mathrm{Hom}_R(M, N)$ is again an R-bimodule.

Exercises

1. Verify that End(A), for any abelian group A, is a ring for the definitions given in the text. Examine what goes wrong if A is not abelian.
2. Show that for any $m \in \mathbf{Z}$ the set of multiples of m is an ideal (m) and that there is a homomorphism from $\mathbf{Z}$ to the ring of integers mod m with (m) as kernel. Show that every ideal of $\mathbf{Z}$ is of this form.
3. Show that the order of 1 in an integral domain, if finite, is necessarily a prime.
4. Show that in any Boolean algebra $x + y = (x \vee y) + (x \wedge y)$, where addition has been defined by (4.1.7).
5. Show that a Boolean ring R is simple iff $R \cong \mathbf{2}$. Deduce that an ideal I in a Boolean ring is a maximal ideal iff $R/I \cong \mathbf{2}$.
6. Verify that for any ring R, $\mathscr{B}(R)$ is a Boolean algebra and for any Boolean algebra B, $\mathscr{R}(B)$ is a Boolean ring.
7. Let M, N be left R-modules and $H = \mathrm{Hom}(M, N)$ be the abelian group of all homomorphisms (not necessarily R-linear) from M to N, qua abelian groups. Verify that $\mathrm{Hom}_R(M, N)$ is a subgroup of H. Taking $N = M$, show that the R-module action on M defines a ring homomorphism $\varphi : R \to \mathrm{End}(M)$ and show that $\mathrm{End}_R(M)$ is the centralizer in End(M) of the image of φ.
8. Let M, N be right R-modules and $H = \mathrm{Hom}(M, N)$ as in Exercise 7. Show that H can be regarded (i) as a right R-module, by defining $x(fr) = (xf)r$; (ii) as a left R-module, by defining $x(rf) = (xr)f$, where $x \in M$, $r \in R$, $f \in H$. Verify that H is an R-bimodule, and show that this argument breaks down if we replace $\mathrm{Hom}(M, N)$ by $\mathrm{Hom}_R(M, N)$.
9. Let R be any ring. Show that $R^O \cong R$ precisely when R has an anti-automorphism. Show also that $R^{OO} = R$; when is $R^O = R$?
10. Let M be a left R-module. Show that for each $u \in M$, the annihilator of u in R, $\mathrm{Ann}(u) = \{x \in R | xu = 0\}$ is a left ideal and that $\mathrm{Ann}(M) = \{x \in R | xM = 0\}$ is an ideal in R.

4.2 The Category of Modules over a Ring

Let R be a ring which will be fixed in what follows but may be quite arbitrary. Given R-modules M, N and an R-homomorphism

$$f : M \to N, \tag{4.2.1}$$

it is clear that the kernel $\ker f = \{x \in M | xf = 0\}$ is a submodule of M and the image $\mathrm{im}\, f = \{y \in N | y = xf \text{ for some } x \in M\}$ is a submodule of N. Let us also recall the notion of an exact sequence. A sequence

$$\ldots \longrightarrow M_{r-1} \xrightarrow{f_{r-1}} M_r \xrightarrow{f_r} M_{r+1} \xrightarrow{f_{r+1}} \ldots$$

of R-modules and homomorphisms is said to be *exact at* M_r if $\mathrm{im}\, f_{r-1} = \ker f_r$; if the sequence is exact at each module it is called an *exact sequence*. Thus the exactness of $0 \to M \to N$ means that the homomorphism $M \to N$ is injective, while surjectivity

is expressed by the exactness of $M \to N \to 0$. If we are given a 3-term exact sequence

$$0 \longrightarrow M' \xrightarrow{f} M \xrightarrow{g} M'' \longrightarrow 0, \tag{4.2.2}$$

usually called a *short* exact sequence, this means that M'' is a quotient of M by a submodule isomorphic to M'. The sequence (4.2.2) is said to *split* or be *split exact* if $\operatorname{im} f = \ker g$ is a direct summand in M.

Returning to the situation (4.2.1), we define the *cokernel* of f as $\operatorname{coker} f = N/\operatorname{im} f$ and the *coimage* as $\operatorname{coim} f = M/\ker f$. Their relations are summarized in the following diagram.

$$\begin{array}{ccccccccc} 0 & \to & \ker f & \to & M & \xrightarrow{f} & N & \to & \operatorname{coker} f \to 0 \\ & & & & \alpha \downarrow & & \uparrow \beta & & \\ & & & & \operatorname{coim} f & \xrightarrow{f'} & \operatorname{im} f & & \end{array}$$

Here $f = \alpha f' \beta$, so either way of going around the square gives the same result; this is expressed briefly by saying that the diagram is *commutative*. In the horizontal sequence the image of the incoming arrow equals the kernel of the outgoing arrow, i.e. the sequence is *exact*. The existence of this factorization $f = \alpha f' \beta$, where α is surjective, β is injective and f' is an isomorphism, is just an instance of the first isomorphism theorem for modules. For reference we restate this and the other isomorphism theorems, which are simpler for modules than for groups, because all submodules are normal subgroups:

First Isomorphism Theorem. *Given a homomorphism of modules, $f : M \to N$, there is a factorization $f = \alpha f_1 \beta$, where $\alpha : M \to M/\ker f$ is the natural homomorphism, $\beta : \operatorname{im} f \to N$ is the inclusion mapping and $f_1 : M/\ker f \to \operatorname{im} f$ is an isomorphism.*

Factor Theorem. *Given a homomorphism of modules $f : M \to N$ and a submodule M' of M with the natural mapping $v : M \to M/M'$ such that $M' \subseteq \ker f$, there exists a unique mapping $f' : M/M' \to N$ such that $f = vf'$; f' is injective precisely if $\ker f = M'$.*

Second Isomorphism Theorem (Parallelogram law). *For any submodules A, B of a module M we have the isomorphism*

$$(A + B)/B \cong A/(A \cap B).$$

Third Isomorphism Theorem. *Given a module M and a submodule M', there is a natural (order-preserving) bijection between the set of submodules of M containing M' and the set of submodules of M/M', $N \leftrightarrow N/M'$, with an isomorphism of corresponding quotients:*

$$(M/M')/(N/M') \cong M/N.$$

Let $\phi : R \to S$ be a homomorphism of rings. Given a left S-module M, we can define a left R-module ${}^{\phi}M$ by taking the additive group of M with the R-action defined by

$$a.x = (a\phi)x, \quad x \in M, a \in R.$$

It is clear that we obtain a left R-module in this way, said to be obtained from the S-module M by *pullback* along ϕ. It is also clear that a subgroup of M is an R-submodule of ${}^{\phi}M$ iff it is an S-submodule of M.

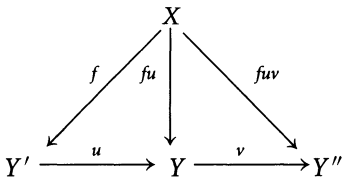
Suppose now that $\phi : R \to S$ is a surjective homomorphism and U is a left R-module such that $aU = 0$ for all $a \in R$ such that $a\phi = 0$. Then we can define an S-module structure on U by the rule

$$c.x = ax, \text{ where } x \in M, c \in S \text{ and } a \in R \text{ is such that } a\phi = c. \tag{4.2.3}$$

If also $b\phi = c$, then $(a - b)\phi = 0$, hence $(a - b)x = 0$ and so $ax = bx$, which shows that the definition (4.2.3) is unambiguous. It is easily checked that this defines an S-module structure on U; if it is denoted by V, then $U \cong {}^{\phi}V$, thus the R-module structure on U can be obtained by pullback along f. We shall not use a special symbol for this construction.

For any ring R we have a category Mod_R whose objects are the right R-modules while the morphisms are the R-linear maps. Similarly we can form the category ${}_R\mathrm{Mod}$ of left R-modules and the category ${}_S\mathrm{Mod}_R$ of (S, R)-bimodules.

Given a left R-module X and a sequence of left R-modules and homomorphisms $Y' \xrightarrow{u} Y \xrightarrow{v} Y''$, we have for each homomorphism $f : X \to Y'$ a commutative diagram



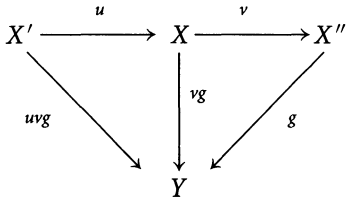
and hence a sequence of homomorphisms of abelian groups

$$\mathrm{Hom}_R(X, Y') \xrightarrow{u_*} \mathrm{Hom}_R(X, Y) \xrightarrow{v_*} \mathrm{Hom}_R(X, Y''),$$

where u_* maps f to fu and v_* maps g to gv. Similarly $(uv)_*$ maps f to $f(uv) = (fu)v$, so that $(uv)_* = u_*v_*$. If $Y' = Y$ and $u = 1_Y$ then clearly $1_* = 1$ and this shows $\mathrm{Hom}_R(X, -) : Y \mapsto \mathrm{Hom}_R(X, Y)$ to be a covariant functor from ${}_R\mathrm{Mod}$ to the category Ab of abelian groups.

If in $\mathrm{Hom}_R(X, Y)$ we keep Y fixed and vary X we obtain a contravariant functor in the same way. The contravariance is expressed by the reversal of the arrows in passing from $X' \xrightarrow{u} X \xrightarrow{v} X''$ to

$$\mathrm{Hom}_R(X'', Y) \xrightarrow{v_*} \mathrm{Hom}_R(X, Y) \xrightarrow{u_*} \mathrm{Hom}_R(X', Y).$$



Thus $\mathrm{Hom}_R(-, -)$ is a functor of two arguments, also called a *bifunctor.*

For any family of left R-modules (M_i) $(i \in \mathrm{I})$ their *direct product* $P = \prod_I M_i$ is defined as the Cartesian product of the M_i, consisting of all families (x_i) $(x_i \in M_i)$, in which the operations are defined componentwise: $(x_i) + (y_i) = (x_i + y_i)$, $r(x_i) = (rx_i)$ (or for right modules, $(x_i)r = (x_ir)$). If $M_i = M$ for all i, we have a *direct power* of M, written M^I. The subset S of P consisting of all families (x_i) with at most finitely many non-zero terms is again a module, called the *direct sum* of the M_i and written $S = \oplus_I M_i$ or $S = \coprod_I M_i$. For each $i \in I$ we have the canonical surjection $\pi_i : P \to M_i$ which picks out the i-th coordinate, and the canonical injection $\mu_i : M_i \to S$ which maps $x \in M_i$ to the family (x_j), where $x_j = \delta_{ij}x$. When $M_i = M$ for all $i \in I$, we have a direct sum of copies of M, written ${}^IM$. For a finite index set I the direct sum and direct product coincide as modules; nevertheless it is often convenient to distinguish between them; their significance will become clearer with the homological development of module theory in FA.

A module M is called the *sum* of a family of submodules (M_i), $M = \sum_I M_i$, if every element of M can be written as $\sum x_i (x_i \in M_i)$, where almost all the x_i vanish, so that the sum is well-defined. If each element of M can be so expressed in only one way, we write $M = \oplus_I M_i$ and call it the *direct sum.* This is sometimes called the *internal* direct sum, to distinguish it from the *external* direct sum or *coproduct* $\coprod_I M_i$. It is clear that $\coprod_I M_i$ is in fact the internal direct sum of the submodules $\mathrm{im}\ \mu_i$, where the μ_i are the canonical injections.

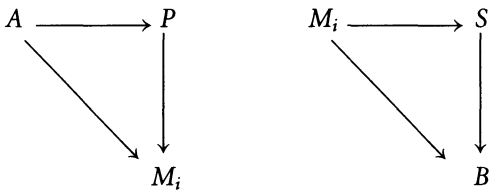
We note the universal properties of $P = \prod_I M_i$ and $S = \coprod_I M_i$ which can also be used to define them.

(i) Given a family (M_i) $(i \in I)$ of R-modules, there exists an R-module P with homomorphisms $\pi_i : P \to M_i$ such that for any family of homomorphisms $f_i : A \to M_i$ from an R-module A there is a unique homomorphism $f : A \to P$ such that $f_i = f\pi_i$ for all $i \in I$.

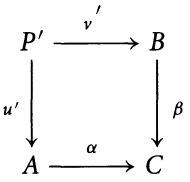
(ii) Given a family (M_i) $(i \in I)$ of R-modules, there exists an R-module S with homomorphisms $\mu_i : M_i \to S$ such that for any family of homomorphisms $g_i : M_i \to B$ to an R-module B there exists a unique homomorphism $g : S \to B$ such that $g_i = \mu_i g$ for all $i \in I$.

We remark that these universal properties can be expressed by the equations

$$\mathrm{Hom}_R\Big(A, \prod M_i\Big) \cong \prod \mathrm{Hom}_R(A, M_i), \quad \mathrm{Hom}_R\Big(\coprod M_i, B\Big) \cong \prod \mathrm{Hom}_R(M_i, B). \tag{4.2.4}$$

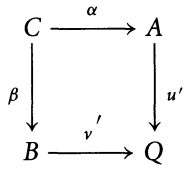


The direct product has a generalization which is often useful. Let A, B, C be R-modules and $\alpha : A \to C$, $\beta : B \to C$ be homomorphisms and in the direct product $A \times B$ with projection maps u, v consider $P' = \ker(u\alpha - v\beta)$ with the restrictions u', v' of u, v respectively. This gives rise to a commutative square which is universal for all such squares (completing α and β)



where $u'\alpha = v'\beta$; moreover, given any R-module M with homomorphisms f to A and g to B such that $f\alpha = g\beta$, there is a unique homomorphism $\phi : M \to P'$ such that $f = \phi u'$, $g = \phi v'$. In other words, (P', u', v') is universal for a module with maps to A and B giving a commutative square. P' (with the maps u', v') is called the *pullback* of α and β. Like all universal constructions it is unique up to isomorphism. Borrowing from the terminology for rings, we may describe the pullback of α, β as a *least common left multiple* of α, β.

Dually, a *pushout* of two homomorphisms $\alpha : C \to A$, $\beta : C \to B$ can be described as a *least common right multiple* of α, β. Explicitly we form the direct sum $A \oplus B$ with injections u, v from A resp. B and for each $x \in C$ identify the images of $x\alpha u$ and $x\beta v$, so as to obtain again a commutative square



Sometimes the pullback is called the 'fibre product' and the pushout the 'fibre sum'. We note that the direct product of A and B is a special case of the pullback, obtained by taking $C = 0$, and similarly the direct sum of A and B is a special case of the pushout when $C = 0$. The pushout has an analogue in groups: if A, B are groups and C is

a subgroup of A and B, with inclusion homomorphisms, then their 'pushout' is the 'free product of A and B amalgamating C'.

We record a fact about pullbacks and pushouts which is often useful.

Proposition 4.2.1. *In the above pullback diagram,* $\ker v' \cong \ker \alpha$, *hence if* α *is injective, then so is* v'. *Dually, in the pushout diagram,* $\operatorname{coker} v' \cong \operatorname{coker} \alpha$, *so if* α *is surjective, then so is* v'.

Proof. This is an easy verification, which may be left to the reader. ∎

The results of Section 3.2 can be applied to modules by observing that, for any R-module M, the set of all submodules forms a lattice, $\mathrm{Lat}_R(M)$ say, under the partial ordering by inclusion, with intersection as meet and sum (not union) as join. As in the case of groups, modules satisfy the modular law, so $\mathrm{Lat}_R(M)$ is a modular lattice; in fact the proofs of the Jordan–Hölder and Schreier refinement theorems are based on this fact.

An R-module M is called *Noetherian* (after Emmy Noether) if $\mathrm{Lat}_R(M)$ satisfies the maximum condition, and *Artinian* (after Emil Artin) if $\mathrm{Lat}_R(M)$ satisfies the minimum condition. If we apply this terminology to R itself, regarded as a right R-module, we obtain the notion of a *right Noetherian* or *right Artinian* ring, as a ring with maximum or minimum conditions respectively on right ideals. Left Noetherian and Artinian rings are defined similarly, using left ideals.

The maximum condition for modules can be stated in another way:

Proposition 4.2.2. *An* R*-module is Noetherian if and only if all its submodules are finitely generated.*

Proof. ($\Rightarrow$) Let N be a submodule of M and choose $x_1, x_2, \ldots \in N$ such that on writing $N_i = x_1R + \ldots + x_iR$, we have $x_{i+1} \notin N_i$. Then

$$0 = N_0 \subset N_1 \subset \ldots; \tag{4.2.5}$$

by the ascending chain condition this must break off, at N_r say, and this means that $N = N_r$, so N is finitely generated.

($\Leftarrow$) Given an ascending chain (4.2.5) of submodules in M, write $N = \sum N_i$ and choose a finite generating set $x_1, \ldots, x_r$ of N. If $x_i \in N_{\nu_i}$ and $\nu = \max\{\nu_1, \ldots, \nu_r\}$, then $x_i \in N_\nu$ for $i = 1, \ldots, r$, hence $N_\nu = N$ and (4.2.5) breaks off; this shows M to be Noetherian. ∎

If a ring R is right Noetherian (or Artinian), then any cyclic right R-module, being of the form $R/\mathfrak{a}$ for some right ideal $\mathfrak{a}$ of R, is again Noetherian (or Artinian), because the submodules of $R/\mathfrak{a}$ are in natural (order-preserving) bijection with the right ideals of R containing $\mathfrak{a}$, by the third isomorphism theorem. But we can say more than this:

Theorem 4.2.3. *Let R be a right Noetherian (resp. Artinian) ring. Then any finitely generated right R-module is again Noetherian (resp. Artinian).*

Proof. We have just seen that the result holds for cyclic modules, and we shall now use induction on the number of generators of M: let M be generated by n elements and denote by M' the submodule generated by n-1 of these. Then $M'' = M/M'$ is cyclic and we have the exact sequence

$$0 \to M' \to M \to M'' \to 0.$$

Assume that R is right Noetherian; then $\mathrm{Lat}(M')$ satisfies the maximum condition, by the induction hypothesis, and $\mathrm{Lat}(M'')$ satisfies the maximum condition because M'' is cyclic. Now the lattice of submodules of M containing M' corresponds to $\mathrm{Lat}(M'')$ (by the third isomorphism theorem) and so satisfies the maximum condition; hence by Proposition 3.2.6, $\mathrm{Lat}(M)$ satisfies the maximum condition and this shows M to be Noetherian. The Artinian case is proved similarly. ■

An integral domain in which every right ideal is principal (i.e. generated as right ideal by a single element) will be called a *principal right ideal domain*; such a ring allows the following conclusion from Theorem 4.2.3:

Corollary 4.2.4. *Over a principal right ideal domain every finitely generated right module is Noetherian.* ■

Let R be any ring and M be a left R-module; an element $m \in M$ is called a *torsion element* if $rm = 0$ for some non-zerodivisor r in R. This definition is usually applied when R is an integral domain. We note that if R is a Noetherian domain, then the torsion elements of any module M form a submodule tM, called the *torsion submodule* of M. When $tM = 0$, the module M is called *torsion-free*, while M is called a *torsion module* when $tM = M$.

A finitely generated module over any ring has an important maximality property, which in the case of groups we have already met in Proposition 2.1.1, with a similar proof, whose details here will therefore be left to the reader:

Proposition 4.2.5. *Let R be any ring and M be a finitely generated R-module. Then every proper submodule of M is contained in a maximal proper submodule.* ■

This result may be applied to R itself. As right R-module, R is generated by the single element 1 and the submodules are right ideals. By a *maximal* right ideal is meant a right ideal maximal in the set of all proper right ideals. Similarly for left ideals and for two-sided ideals, regarding R as R-bimodule. Thus we have

Theorem 4.2.6 (Krull's theorem). *Let R be any ring. Then any proper right ideal is contained in a maximal right ideal. In particular, every non-trivial ring has maximal right ideals. A corresponding result holds for left ideals, and for two-sided ideals.* ■

This result makes essential use of the existence of a unit element in rings; in the absence of 1 it need not hold (see Exercise 9).

Exercises

1. Let R be a ring and $\mathfrak{a}$ be an ideal in R. Show that every $(R/\mathfrak{a})$-module can be defined as an R-module by pullback along the natural homomorphism $R \to R/\mathfrak{a}$, and conversely, every right R-module M such that $M\mathfrak{a} = 0$ can be defined as an $(R/\mathfrak{a})$-module.
2. Let R be a ring, $\mathfrak{a}$ be an ideal in R and $\mathrm{Mod}_{(R,\mathfrak{a})}$ be the category of right R-modules M such that $M\mathfrak{a} = 0$. Show that $\mathrm{Mod}_{(R,\mathfrak{a})}$ is a full subcategory of Mod_R which is equivalent to $\mathrm{Mod}_{R/\mathfrak{a}}$.
3. Let M be a module. Given an endomorphism f of M such that $f^2 = f$, show that $M = \ker f \oplus \operatorname{im} f$; show that this holds even if only $\operatorname{im} f^2 = \operatorname{im} f$, $\ker f^2 = \ker f$.
4. Show that $A \xrightarrow{f} B \xrightarrow{g} C$ is exact iff the compositions $\operatorname{im} f \to B \to \operatorname{coim} g$ and $\ker g \to B \to \operatorname{coker} f$ are both zero.
5. Show that the pushout of two homomorphisms $\alpha : C \to A$, $\beta : C \to B$, one of which is the zero map, is $A \oplus B$.
6. Show that a module (over any ring) is finitely generated iff it cannot be expressed as a union of a well-ordered chain of proper submodules without the last term. Deduce another proof of Propositions 4.2.2 and 4.2.5. Does the result still hold if instead of chains we take countable ascending sequences of proper submodules?
7. Show that every quotient of a finitely generated module is finitely generated, but not necessarily every submodule.
8. Let R be a Noetherian domain. Show that in any R-module the set of torsion elements forms a submodule.
9. Show that the 'non-unital ring' (i.e. ring without 1) defined on the additive group of rational numbers by the multiplication $xy = 0$ has no maximal ideals.

4.3 Semisimple Modules

For any ring R, an R-module M is said to be *simple* if $M \neq 0$ and M has no submodules other than 0 and M. For example, over a field the only simple module is a one-dimensional vector space. When $R = \mathbf{Z}$, the $\mathbf{Z}$-modules are just abelian groups, and the simple $\mathbf{Z}$-modules are the cyclic groups of prime order $\mathbf{C}_p$.

A module which can be expressed as a direct sum of simple modules is said to be *semisimple*. For example, $\mathbf{C}_6$ is semisimple, since it can be written$\mathbf{C}_6 = \mathbf{C}_3 \oplus \mathbf{C}_2$, but $\mathbf{C}_4$ is not semisimple. To show that a module M is semisimple we need to find a family of simple submodules S_λ such that the sum $\sum S_\lambda$ is direct and equal to M. Here it is not necessary for the family to be finite, in fact we have the following condition for the sum to be finite:

Proposition 4.3.1. *Let M be a semisimple module, say*

$$M = \oplus_I S_i, \tag{4.3.1}$$

where each S_i is simple. Then the number of summands in (4.3.1) *is finite if and only if M is finitely generated.*

Proof. If the index set I is finite, then since each S_i, being simple, is cyclic, M is finitely generated. Conversely, let M be finitely generated, by $u_1, \ldots, u_r$ say. For each u_j we can find finitely many terms S_i whose sum contains u_j, hence all the u_j are contained in a finite subfamily of the S_i, and this family generates M, so that I must be finite. ■

Below we shall describe semisimple modules in some alternative ways which are often used, but first we single out a property needed for the proof.

Lemma 4.3.2. *Let M be a module of the form $M = \sum_I S_i$, where each S_i is simple (but the sum need not be direct). If N is any submodule of M, then there is a subset J of I such that*

$$M = N \oplus (\oplus_J S_i).$$

In particular, $M = \oplus_L S_i$ for some $L \subseteq I$.

Proof. Let $\mathscr{F}$ be the family of all subsets J of I such that the sum $N + \sum_J S_i$ is direct. The sum fails to be direct iff there is a finite subset J_0 of J and $x_j \in S_j$ $(j \in J_0)$, $u \in N$ such that $x_j \neq 0$ and $u + \sum_{J_0} x_j = 0$. This shows $\mathscr{F}$ to be of finite character and hence inductive. By Zorn's lemma there is a maximal subset J in $\mathscr{F}$; with this set J the sum $P = N + \sum_J S_j$ is direct and we wish to show that $P = M$. Now for any S_k, $P \cap S_k$ is either S_k or 0; if it is 0, then the sum $N + \sum_J S_j + S_k$ is still direct, contradicting the maximality of J. Hence $P \cap S_k = S_k$ for all k, so $P \supseteq \sum S_k = M$, and hence $P = M$, as claimed. The final assertion is the case $N = 0$. ■

Let us recall that a submodule N of a module M is said to be *complemented* in M if there is a submodule N' such that $M = N \oplus N'$. Here N' is called a *complement* of N in M; generally it will not be unique, but it is unique up to isomorphism, because $N' \cong M/N$, by the second isomorphism theorem. We remark that when $M = N \oplus N'$, then for any submodule P containing N we have $P = M \cap P = (N + N') \cap P = N + (N' \cap P)$, by the modular law, and $N \cap (N' \cap P) = 0$, hence $P = N \oplus (N' \cap P)$; this proves

Lemma 4.3.3. *Let M be any module and N be a submodule. If N is complemented in M, then N is also complemented in any submodule containing it.* ■

We now have the following characterization of semisimple modules:

Theorem 4.3.4. *Let R be any ring. For any R-module M the following conditions are equivalent:*

(a) *M is a sum of simple modules,*
(b) *M is semisimple,*
(c) *every submodule of M is complemented.*

Proof. (a) $\Rightarrow$ (b) follows by Lemma 4.3.2 and clearly (b) $\Rightarrow$ (a), so (a) and (b) are equivalent. To prove (b) $\Rightarrow$ (c), let $M = \sum_I S_i$, where each S_i is simple. For any submodule N of M we have, by Lemma 4.3.2, a subset J of I such that $M = N \oplus (\oplus_J S_j)$; hence $(\oplus_J S_j)$ is a complement of N in M.

(c) $\Rightarrow$ (a). Let S be the sum of all the simple submodules of M; we have to show that $S = M$. If this were not so, then $M = S \oplus T$, where $T \neq 0$. Let N be a non-zero cyclic submodule of T; by Proposition 4.2.5 there exists a maximal proper submodule N' of N. Now N' has a complement in M, so by Lemma 4.3.3, N' also has a complement in N, say $N = N' \oplus P$. Since $P \cong N/N'$, P is simple and it is contained in T; thus $P \subseteq S \cap T$, which is a contradiction. This shows that $T = 0$ and so $S = M$, as required. ∎

Corollary 4.3.5. *Let M be a semisimple module, say $M = \oplus_I S_i$, where S_i is simple. Then any submodule N of M has a complement of the form $\oplus_K S_i$ for some subset K of I, and*

$$N \cong \oplus_{K'} S_i, \quad \text{where } K' = I \backslash K. \tag{4.3.2}$$

In particular, if N is simple, then $N \cong S_i$ for some $i \in I$ and $\oplus_{j \neq i} S_j$ is a complement for N.

Proof. By Lemma 4.3.2 we have $M = N \oplus (\oplus_K S_i)$ for some $K \subseteq I$; thus $N' = (\oplus_K S_i)$ is a complement of N. Since we also have $M = (\oplus_{K'} S_i) \oplus (\oplus_K S_i)$, it follows that $N \cong M/N' \cong \oplus_{K'} S_i$ and (4.3.2) follows. When N is simple, K' reduces to a single element and so $N \cong S_i$ in this case, and there is a complement as stated. ∎

As an example to show that the isomorphism (4.3.2) cannot always be strengthened to equality, take a simple module P and in P^2 define the submodules $P_1 = \{(x, 0) | x \in P\}$, $P_2 = \{(0, x) | x \in P\}$, $P_3 = \{(x, x) | x \in P\}$. Clearly $P^2 = P_1 \oplus P_2 = P_1 \oplus P_3 = P_2 \oplus P_3$, and $P_1 \cong P_2 \cong P_3$, but these submodules are all different. This example also illustrates the point that to prove the directness of a sum $\sum P_i$ of more than two terms it is not enough to have $P_i \cap P_j = 0$ for $i \neq j$; we must have $P_i \cap (\sum_{j \neq i} P_j) = 0$ or at least $P_i \cap (\sum_1^{i-1} P_j) = 0$.

The form of (4.3.2), together with the fact that every quotient of M is isomorphic to a submodule, shows the truth of

Corollary 4.3.6. *If M is semisimple, then so is every submodule and every quotient.* ∎

We observe that all of (a)–(c) in Theorem 4.3.4 are lattice conditions but they are

not equivalent in all lattices. Of course (a) $\Leftrightarrow$ (b) in any lattice and (a) $\Leftrightarrow$ (c) in any modular lattice of finite length, but in general neither of (a), (c) implies the other. Thus (c) but not (a) is self-dual, and in the lattice of ideals of $\mathbf{Z}$ the dual of (a) but not of (c) holds. To find an example satisfying (c) but not (a), let B be the Boolean algebra of all subsets of an infinite set S, and $\mathfrak{F}$ be the ideal of all finite subsets. Then $B/\mathfrak{F}$ is again a Boolean algebra, hence a complemented distributive lattice, but it has no atoms, so (c) holds, but not (a).

Condition (c) of Theorem 4.3.4 may also be expressed by saying that any exact sequence $0 \to N \to M \to M/N \to 0$ splits; in other words, every short exact sequence with middle term M splits. This is equivalent to either of the following conditions on (4.2.2):

(a) There is an R-homomorphism $f' : M \to M'$ such that $ff' = 1_{M'}$; this mapping f' is called a *right inverse* or *retraction* for f.
(b) There is an R-homomorphism $g' : M'' \to M$ such that $g'g = 1_{M''}$; the mapping g' is called a *left inverse* or a *section* for g.

Clearly (a), (b) hold when $\operatorname{im} f$ is complemented; conversely, in case (a) we have $M = \operatorname{im} f \oplus \ker f'$ and in case (b) $M = \ker g \oplus \operatorname{im} g'$, as is easily verified.

Over a field (even skew) the simple modules are just the one-dimensional vector spaces; since every vector space can be written as a sum of one-dimensional spaces, it follows that over a field every module (= vector space) is semisimple. In particular, this proves the existence of a basis for any vector space, even infinite-dimensional. For if $V = \oplus_I S_i$ and u_i is a generator of S_i, then $\{u_i\}_{i \in I}$ is a basis of V, as is easily checked.

For an arbitrary ring R the theory of semisimple modules is quite similar to the theory of vector spaces over a field. The main difference is that there may be more than one type of simple module. We shall say that two simple R-modules have the same *type* if they are isomorphic. A semisimple R-module is called *isotypic* if it can be written as a sum of simple modules all of the same type. In any R-module M the sum of all simple submodules is called the *socle* of M; thus the semisimple modules are those that coincide with their socle. The sum of all simple submodules of a given isomorphism type α is called the *α-socle* of M or a *type component* in M.

Let M be any R-module; a submodule N of M is said to be *fully invariant* in M if it admits (i.e. is mapped into itself by) all R-endomorphisms of M. We note that for an R-module M, the set $S = \operatorname{End}_R(M)$ is just the centralizer in $\operatorname{End}(M)$ of the image of R defining the R-action on M, so a subgroup N of M is a fully invariant submodule iff it admits both R and S. In a semisimple module the fully invariant submodules are easily described: they are the sums of type components.

Theorem 4.3.7. *Let R be a ring and M be an R-module. Then:*

(i) *For any type α the α-socle is an isotypic submodule of M containing all simple submodules of type α, and the socle is the direct sum of the α-socles, for different α.*
(ii) *Any sum of type components in M is fully invariant in M; when M is semisimple, the converse holds: any fully invariant submodule is a sum of type components.*

(iii) *If M is semisimple and S is the centralizer in* $\mathrm{End}(M)$ *of the R-action, then M is also semisimple as S-module.*

Proof. (i) Denote the α-socle of M by H_α. By Theorem 4.3.4, H_α is semisimple, and it contains all simple submodules of type α, by definition. By Corollary 4.3.5, any submodule of H_α is a sum of modules of type α, hence $H_\alpha \cap (\sum_{\beta\neq\alpha} H_\beta) = 0$, so the sum $\sum H_\alpha$ is direct.

(ii) Let H_α again be the α-socle of M, say $H_\alpha = \sum P_i$, where the P_i are all the simple submodules of type α. For any R-endomorphism f of M we have $H_\alpha f = \sum P_i f$; each $P_i f$ is a homomorphic image of P_i, hence either 0 or simple of type α, so in any case $P_i f \subseteq H_\alpha$. Hence $H_\alpha f \subseteq H_\alpha$, i.e. H_a is fully invariant. It follows that any sum of type components is fully invariant.

Now let N be a fully invariant submodule of a semisimple module M. To prove that N is a sum of type components of M we need only show: if P is a simple submodule of N and Q is a simple submodule of M such that $Q \cong P$, then $Q \subseteq N$. This is clear if $Q = P$; otherwise $P \cap Q = 0$, because the intersection must be a proper submodule of both. Hence $P + Q$ is direct and $M = P \oplus Q \oplus U$, for some U, by Theorem 4.3.4(c). Let $f : P \to Q$ be the given isomorphism and define a mapping $\theta : M \to M$ by

$$\theta : x + y + z \mapsto xf + yf^{-1} + z, \quad \text{where } x \in P, y \in Q, z \in U.$$

Clearly θ is an R-endomorphism, hence $N\theta \subseteq N$ and $P \subseteq N$, therefore $Q = P\theta \subseteq N$. Thus N contains with P the whole α-socle of M and so is a sum of type components of M.

(iii) Let us write M as left R-, right S-module. We have $M = \sum xS$, where x runs over all elements of all simple R-submodules, for every element of M is a sum of such elements. Now the result will follow if we show that xS is simple or 0. Let $y \in xS$, $y \neq 0$, say $y = xs$. Then for any $a \in R$, the mapping $\lambda_a : ax \mapsto axs = ay$ is a homomorphism $Rx \to Ry$, which is surjective and $Ry \neq 0$. Since Rx is simple, λ_a must be an isomorphism, so there is also an R-homomorphism which maps x to y. As in (ii) we can find an R-endomorphism of M which maps y to x. Hence yS contains x, therefore $yS = xS$ and so xS is indeed simple. ∎

Corollary 4.3.8. *Let M be a semisimple R-module and express M as the sum of its α-socles: $M = \oplus_\alpha H_\alpha$. Then*

$$\mathrm{End}_R(M) = \prod_\alpha \mathrm{End}_R(H_\alpha). \tag{4.3.3}$$

Proof. Any family (f_α), where f_α is an endomorphism of H_α, defines an endomorphism f of M and it is clear that the correspondence $(f_\alpha) \mapsto f$ is an injective homomorphism from the right- to the left-hand side of (4.3.3). In the other direction consider $f \in \mathrm{End}_R(M)$; since each H_α is fully invariant, it is mapped to itself by f. If the restriction $f|H_\alpha$ is denoted by f_α, then f is just the endomorphism defined by the family (f_α). Thus the correspondence $(f_\alpha) \mapsto f$ is surjective and (4.3.3) is established. ∎

In any semisimple module M the simple components in a direct decomposition are determined up to isomorphism; on the other hand, the decomposition $M = \oplus_\alpha H_\alpha$ into α-socles is unique (not merely up to isomorphism). The next result is a special case of the Krull–Schmidt theorem (see FA Chapter 4, also Corollary 4.6.5 below):

Proposition 4.3.9. *Let M be a finitely generated semisimple R-module. Then any two direct decompositions of M into simple submodules have the same number of terms, and for suitable numbering corresponding terms are isomorphic.*

Proof. Let $M = P_1 \oplus \ldots \oplus P_n = Q_1 \oplus \ldots \oplus Q_m$ be two direct decompositions into simple submodules, where the number of terms is finite, by Proposition 4.3.1. By Corollary 4.3.5, $Q_1 \cong P_j$ for some j, say $j = 1$ (by renumbering the P's), and moreover, $P_2 \oplus \ldots \oplus P_n \cong Q_2 \oplus \ldots \oplus Q_m$. Now the result follows by induction on n. ∎

Exercises

1. Show in detail how Theorem 4.3.4 may be used to prove the existence of a basis in any vector space over a field.
2. Show that the socle of any module is semisimple.
3. Let R be any ring. Show that when R is considered as a right R-module, the fully invariant submodules are just the two-sided ideals.
4. For any homomorphism between R-modules, $f : M \to N$, define its *graph* as the subset of $M \coprod N$ given by $F = \{(x, xf) | x \in M\}$. Verify that F is a submodule and that $F \cap N = 0$; when is $F \cap M = 0$?
5. A module M is said to be *distributive* if the lattice of its submodules, $\mathrm{Lat}(M)$, is distributive. Show that a semisimple module is distributive iff any two distinct simple submodules are non-isomorphic. (Hint. Examine $\mathrm{Lat}(P^2)$, where P is simple.)
6. (J.-E. Roos) By examining the graph of a homomorphism $M \to N$ (see Exercise 4), show that if $M \coprod N$ is distributive, then $\mathrm{Hom}_R(M, N) = 0$.
7. Show that a module M fails to be distributive iff M has distinct submodules U, V such that $U/(U \cap V) \cong V/(U \cap V)$. (Hint. For the necessity take a submodule with two relative complements.)
8. Show that if M is distributive and $M = U_1 \oplus \ldots \oplus U_r$, then $\mathrm{End}_R(M) = \prod \mathrm{End}_R(U_i)$.

4.4 Matrix Rings

We first encounter matrix rings in the description of linear mappings between vector spaces; in particular, any matrix ring over a field may be described as the endomorphism ring of a vector space. In order to gain a better understanding of general matrix rings we shall consider direct sums of modules and their endomorphisms.

We shall find that an endomorphism of a direct sum can be expressed as a ring of matrices whose (i, j)-entry is a homomorphism from the i-th to the j-th summand.

Let us consider, for any ring R, a left R-module M which is expressed as a direct sum of certain submodules

$$M = U_1 \oplus \ldots \oplus U_n. \tag{4.4.1}$$

Let $\pi_i : M \to U_i$ be the canonical projections and $\mu_i : U_i \to M$ be the canonical injections $(i = 1, \ldots, n)$; thus $(x_1, \ldots, x_n)\pi_i = x_i$, $x\mu_i = (0, \ldots, 0, x, 0, \ldots, 0)$, with x in the i-th place. It is clear that

$$\mu_i\pi_j = \delta_{ij}, \tag{4.4.2}$$

$$\sum \pi_i\mu_i = 1. \tag{4.4.3}$$

With each endomorphism $f : M \to M$ we can associate the matrix (f_{ij}), where $f_{ij} : U_i \to U_j$ is defined by $f_{ij} = \mu_i f \pi_j$. Similarly, any family (α_{ij}) of homomorphisms $\alpha_{ij} : U_i \to U_j$ gives rise to an endomorphism $\alpha : M \to M$ defined by $\alpha = \sum \pi_i \alpha_{ij} \mu_j$. These two processes are mutually inverse: If $\alpha = \sum \pi_i \alpha_{ij} \mu_j$, then $\mu_r \alpha \pi_s = \sum \mu_r \pi_i \alpha_{ij} \mu_j \pi_s = \alpha_{rs}$ by (4.4.2), and if $f_{ij} = \mu_i f \pi_j$, then $\sum \pi_i f_{ij} \mu_j = \sum \pi_i \mu_i f_{ij} \pi_j \mu_j = f$ by (4.4.3). Moreover, the families (f_{ij}) are added and multiplied 'matrix fashion': $(f + g)_{ij} = \mu_i(f + g)\pi_j = \mu_i f \pi_j + \mu_i g \pi_j = f_{ij} + g_{ij}$ and $(fg)_{ik} = \mu_i fg \pi_k = \sum \mu_i f \pi_j \mu_j g \pi_k = \sum f_{ij} g_{jk}$. Thus we obtain

Theorem 4.4.1. *Let R be any ring. If M is a left R-module, expressed as a direct sum: $M = U_1 \oplus \ldots \oplus U_n$ with projections $\pi_i : M \to U_i$ and injections $\mu_i : U_i \to M$, then the elements of $\mathrm{End}_R(M)$ can be written as matrices (f_{ij}), where $f_{ij} : U_i \to U_j$, with the usual addition and multiplication of matrices.* ∎

In particular, when all the summands are isomorphic, then $M \cong U^n$, and we obtain

Corollary 4.4.2. *Let R be a ring, U be a left R-module and $S = \mathrm{End}_R(U)$. Then for any $n \geq 1$,*

$$\mathrm{End}_R(U^n) \cong \mathfrak{M}_n(S). \tag{4.4.4}$$ ∎

Here U^n means the direct sum of n copies of U. It is most convenient to write its elements as rows; then it is clear how the matrix ring $\mathfrak{M}_n(S)$, written more briefly as S_n, operates from the right on these rows. Generally, if V is an (R, S)-bimodule, R acting on the left and S on the right, we write ${}^mV^n$ for the set of all $m \times n$ matrices with entries in V. This is acted on from the left by R_m and from the right by S_n, and it is easily verified to be an (R_m, S_n)-bimodule. In the special case $m = 1$ we write ${}^1V^n$ simply as V^n, as remarked earlier. Similarly we write ${}^mV$ for the set ${}^mV^1$ of column vectors. The same notation ${}^mR^n$ is used in the case of a ring R; when $m = n$, we have $\mathfrak{M}_n(R) = R_n$. For example, R itself is an R-bimodule, as we saw, and ${}^mR^n$ becomes an (R_m, R_n)-bimodule in this way.

Let us return to the case considered in Corollary 4.4.2. Writing $e_{ij} = \pi_i \mu_j$, we obtain from (4.4.2), (4.4.3) the equations

$$e_{ij}e_{rs} = \delta_{jr}e_{is}, \quad \sum e_{ii} = 1. \tag{4.4.5}$$

It follows that the e_{ij} are just the familiar matrix units in $(\mathrm{End}_R(U))_n$; thus e_{ij} is the matrix with 1 in the (i, j)-position and 0 elsewhere. The isomorphism (4.4.4) is given explicitly by

$$\alpha \leftrightarrow (\alpha_{ij}), \quad \text{where } \alpha_{ij} = \mu_i \alpha \pi_j, \quad \alpha = \sum \pi_i \alpha_{ij} \mu_j.$$

It is worth remarking that matrix rings can be defined in terms of their matrix units.

Proposition 4.4.3. *Let R be any ring with n^2 elements e_{ij} satisfying* (4.4.5). *Then $R \cong C_n$, where C is the centralizer in R of the e_{ij}.*

Proof. Given $a \in R$, we define $a_{ij} = \sum_\nu e_{\nu i} a e_{j\nu}$. Since $e_{j\nu}e_{rs} = \delta_{\nu r}e_{js}$, we have $a_{ij}e_{rs} = \sum_\nu e_{\nu i} a e_{j\nu} e_{rs} = e_{ri} a e_{js}$ and $e_{rs}a_{ij} = \sum_\nu e_{rs}e_{\nu i} a e_{j\nu} = e_{ri} a e_{js}$, hence $a_{ij} \in C$. Moreover, $\sum a_{ij}e_{ij} = \sum e_{\nu i} a e_{j\nu} e_{ij} = \sum e_{ii} a e_{jj} = a$. Conversely, given a matrix (a_{ij}) over C, we define $a = \sum a_{ij}e_{ij}$ and then find that $\sum e_{\nu i} a e_{j\nu} = \sum e_{\nu i} a_{rs} e_{rs} e_{j\nu} = \sum a_{rs} e_{\nu i} e_{rs} e_{j\nu} = a_{ij} \sum e_{\nu\nu} = a_{ij}$. Thus we have indeed a bijection $a \leftrightarrow (a_{ij})$ and this is easily seen to be an isomorphism. ∎

It is clear that any ring homomorphism $R \to S$ induces a homomorphism of matrix rings $R_n \to S_n$; we simply map (a_{ij}) to $(a_{ij}f)$. We also have a converse:

Corollary 4.4.4. *Let $f : R \to S$ be a ring homomorphism. If R is a full matrix ring, say $R = K_n$, then S has the form $S = L_n$ for some ring L, and there is a homomorphism $\varphi : K \to L$ which induces f.*

Note particularly that f does not have to be surjective, for the conclusion to hold.

Proof. Let e_{ij} be the matrix units in R and write $e'_{ij} = e_{ij}f$. The e'_{ij} again satisfy equations (4.4.5), hence S is a matrix ring; moreover, the centralizer of the e_{ij}, K, maps to the centralizer of the e'_{ij} in S, L say, and this mapping $\varphi : K \to L$ induces f; the details may be left to the reader. ∎

Let A, B be rings and M be an (A, B)-bimodule. Then we can form a ring which is most easily expressed in matrix form

$$R = \begin{pmatrix} A & M \\ 0 & B \end{pmatrix}.$$

Thus the elements of R are 2×2 matrices $\begin{pmatrix} a & m \\ 0 & b \end{pmatrix}$, $a \in A$, $b \in B$, $m \in M$, with the usual matrix operations

$$\begin{pmatrix} a & m \\ 0 & b \end{pmatrix} + \begin{pmatrix} a' & m' \\ 0 & b' \end{pmatrix} = \begin{pmatrix} a+a' & m+m' \\ 0 & b+b' \end{pmatrix},$$

$$\begin{pmatrix} a & m \\ 0 & b \end{pmatrix}\begin{pmatrix} a' & m' \\ 0 & b' \end{pmatrix} = \begin{pmatrix} aa' & am'+mb' \\ 0 & bb' \end{pmatrix}.$$

The ring structures on A and B and the bimodule structure on M ensure that R is again a ring. It will be called the *triangular matrix ring* or the *tiled ring* formed from A, B and M. This construction is often useful in forming asymmetric counter-examples (see Exercise 8). We note the ideal structure of R:

Proposition 4.4.5. *Let A, B be rings, M be an (A, B)-bimodule and let R be the triangular matrix ring formed from A, B, M. Given any left A-submodule N of $A \oplus M$ and any left ideal $\mathfrak{b}$ of B satisfying $M\mathfrak{b} \subseteq N$, the set $N \oplus \mathfrak{b}$ is a left ideal of R and every left ideal of R is of this form.*

Proof. Let I be a left ideal of R and consider a typical element $\begin{pmatrix} x & y \\ 0 & z \end{pmatrix}$. Then I contains the products

$$\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x & y \\ 0 & z \end{pmatrix} = \begin{pmatrix} ax & ay \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & m \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x & y \\ 0 & z \end{pmatrix} = \begin{pmatrix} 0 & mz \\ 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0 \\ 0 & b \end{pmatrix}\begin{pmatrix} x & y \\ 0 & z \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & bz \end{pmatrix}. \tag{4.4.6}$$

Taking $a = 1$, $b = 1$, we see that I contains each row of $\begin{pmatrix} x & y \\ 0 & z \end{pmatrix}$; thus we have $I = N \oplus \mathfrak{b}$, where N denotes the set of first rows and $\mathfrak{b}$ the set of second rows. Equations (4.4.6) further show that $AN \subseteq N$, $B\mathfrak{b} \subseteq \mathfrak{b}$ and $M\mathfrak{b} \subseteq N$; hence I is of the required form. Conversely, it is clear from (4.4.6) that any set of this form is a left ideal of R. ■

By the symmetry of the construction a similar result holds for right ideals. We now return to the situation of Corollary 4.4.2. There we considered U^n as a left R-module and found an endomorphism ring of the form S_n. But we can also consider U^n as a left R_n-module, and then its endomorphism ring turns out to be S. In this case we shall write ${}^nU$ and visualize its elements as column vectors; here we shall moreover find a correspondence between submodules. If M is an R-module, we shall write $\mathrm{Lat}_R(M)$ for the lattice of all R-submodules of M.

Theorem 4.4.6. *Let M be a left R-module and write $S = \mathrm{End}_R(M)$. Then ${}^nM$ may be regarded as a left R_n-module in a natural way, and*

$$\mathrm{End}_{R_n}({}^nM) \cong S. \tag{4.4.7}$$

Moreover, there is a lattice-isomorphism

$$\mathrm{Lat}_R(M) \cong \mathrm{Lat}_{R_n}({}^nM), \tag{4.4.8}$$

and (R, S)*-subbimodules correspond to* (R_n, S)*-subbimodules under this isomorphism.*

Proof. In determining the endomorphism ring of ${}^nM$ we may replace R by its image in $\mathrm{End}(M)$; thus S is the centralizer of R in $\mathrm{End}(M)$. Let e_{ij} be the matrix units in $\mathrm{End}({}^nM)$; together with R they generate a ring isomorphic to R_n in $\mathrm{End}({}^nM)$, and $\mathrm{End}_{R_n}({}^nM)$ is the centralizer of R and the e_{ij}. By Corollary 4.4.2, the centralizer of R is just S_n, so the required endomorphism ring is the centralizer of the e_{ij} in S_n, i.e. S itself; this proves (4.4.7).

Now any R-submodule N of M corresponds to an R_n-submodule ${}^nN$ of ${}^nM$, and the correspondence

$$N \mapsto {}^nN \tag{4.4.9}$$

is clearly order-preserving. To establish (4.4.8) we need only show that (4.4.9) has an inverse which is also order-preserving. Consider the projection on the first factor

$$\pi_1 : {}^nM \mapsto M. \tag{4.4.10}$$

With any R_n-submodule of ${}^nM$ this associates an R-submodule of M, and we claim that this is the required inverse. If we take an R-submodule N of M and apply first (4.4.9) and then (4.4.10), we evidently get back to N. Now let P be any R_n-submodule of ${}^nM$; each projection π_i maps P to an R-submodule N_i of M, but all these submodules agree: $N_i = N_1$, because P admits R_n. This becomes clear if we think of the elements of P as column vectors, acted on by R_n from the left. It follows that $P = {}^nN_1$, and since the S-action on M and on ${}^nM$ is unaffected by R, it follows that S-submodules correspond under (4.4.8); therefore (R, S)-bimodules correspond to (R_n, S)-bimodules. ∎

We can go beyond Theorem 4.4.6 and obtain an equivalence between the categories of left R-modules and left R_n-modules. For any $M \in {}_R\mathrm{Mod}$ consider the functor

$$M \mapsto {}^nM = \mathrm{Hom}_R(R^n, M), \tag{4.4.11}$$

and for $P \in {}_{R_n}\mathrm{Mod}$,

$$P \mapsto P^\blacklozenge = \mathrm{Hom}_{R_n}({}^nR, P) = e_{11}P. \tag{4.4.12}$$

The functorial property is clear and we have $({}^nM)^\blacklozenge \cong M$ and ${}^n(P^\blacklozenge) \cong P$. This shows that the functors (4.4.11) and (4.4.12) provide an equivalence between the categories ${}_R\mathrm{Mod}$ and ${}_{R_n}\mathrm{Mod}$, for any ring R and any $n \geq 1$.

Two rings R, S with the property that ${}_R\mathrm{Mod}$ and ${}_S\mathrm{Mod}$ are equivalent categories are said to be *Morita-equivalent*; it can be shown that this happens iff Mod_R is equivalent to Mod_S. What we have shown above can be expressed by saying that any ring R is Morita-equivalent to the full matrix ring R_n, for any $n \geq 1$. In general R may be Morita-equivalent to rings not of the form R_n (e.g. when R is itself a matrix

ring), but there is an important case where this is the only possibility; this will be discussed in Section 5.1. A general discussion of Morita-equivalence will be reserved for FA.

Exercises

1. Verify that in a full matrix ring R_n the centralizer of the matrix units e_{ij} is precisely R. Show also that $e_{11}R_ne_{11} \cong R$.
2. Let R be a ring with two elements u, v such that $u^2 = v^2 = 0$, $uv + vu = 1$. Show that $R \cong S_2$, where S is the centralizer of u, v.
3. Show that there is a natural bijection between the set of two-sided ideals of R and the set of (R, R_n)-subbimodules of R^n, and likewise between the set of two-sided ideals of R_n and (R, R_n)-subbimodules of R^n. Deduce that there is a natural bijection between the set of all ideals of R and those of R_n.
4. Find the centre of the ring of all upper triangular $n \times n$ matrices over the real numbers (i.e. matrices (a_{ij}) with $a_{ij} = 0$ for $i > j$).
5. Show that if K is a skew field, then K_n is a simple ring (see Section 5.2).
6. Let K be a skew field and $n \geq 1$. Show that any family of (non-zero) orthogonal idempotents in K_n has at most n members. (Hint. Consider direct sums in K^n.)
7. Let $A \subseteq B$ be rings and $R = \begin{pmatrix} A & B \\ 0 & B \end{pmatrix}$, where B is regarded as (A, B)-bimodule. Describe the left and the right ideals of R. (Hint. Take first the case where B is a field.)
8. Show that the ring $\begin{pmatrix} \mathbf{R} & \mathbf{R} \\ 0 & \mathbf{Q} \end{pmatrix}$, where the real numbers $\mathbf{R}$ are regarded as $(\mathbf{R}, \mathbf{Q})$-bimodule, is left Noetherian and Artinian but neither right Noetherian nor right Artinian.

4.5 Direct Products of Rings

Let (R_i), $i \in I$, be any family of rings. Its *direct product* is the ring obtained by forming the Cartesian product

$$R = \prod R_i \tag{4.5.1}$$

and defining all operations componentwise, thus $(x_i) + (y_i) = (x_i + y_i)$, $(x_i)(y_i) = (x_iy_i)$, $1 = (1_i)$, where 1_i is the one of R_i. It is easily seen that R is again a ring; the natural projections $\pi_i : R \to R_i$ are homomorphisms, while the injections $\mu_i : R_i \to R$ preserve addition and multiplication, but not the 1, and so are not homomorphisms. For the same reason the images of the R_i in R are not subrings, although they are rings in their own right and in fact they are ideals in R.

We shall mainly be interested in finite direct products; they may be described as direct sums of ideals:

Theorem 4.5.1. *Let $R_1, \ldots, R_t$ be any rings and denote their direct product by R, with natural projections $\pi_i : R \to R_i$ and injections $\mu_i : R_i \to R$. Then $\mathfrak{a}_i = im\ \mu_i$ is an ideal in R and*

$$R = \mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_t. \tag{4.5.2}$$

Moreover,

$$\mathfrak{a}_i\mathfrak{a}_j = 0 \text{ if } i \neq j, \quad \mathfrak{a}_i\mathfrak{a}_i \subseteq \mathfrak{a}_i. \tag{4.5.3}$$

Conversely, any ring R of the form (4.5.2), where each $\mathfrak{a}_i$ is an ideal in R_i, may be expressed as a direct product of rings $R_1, \ldots, R_t$, where R_i is isomorphic to $\mathfrak{a}_i$, qua ring.

On account of (4.5.2), R is sometimes called the 'direct sum' of the R_i, but this is rather misleading when rings with 1 are considered, as is the case here.

Proof. We have seen that a finite direct product of abelian groups is isomorphic to their direct sum. Hence in the finite case, (4.5.2) follows from (4.5.1). Now whenever (4.5.2) holds, then $\mathfrak{a}_i\mathfrak{a}_j \subseteq \mathfrak{a}_i \cap \mathfrak{a}_j$ because the $\mathfrak{a}_i$ are ideals in R and since $\mathfrak{a}_i \cap \mathfrak{a}_j = 0$ for $i \neq j$, (4.5.3) follows. Conversely, any ring of the form (4.5.2) is a product of the $\mathfrak{a}_i$ as abelian groups, and (4.5.3) ensures that each $\mathfrak{a}_i$ is an ideal in R and the multiplication is performed componentwise. Moreover, the natural homomorphism from R with kernel $\sum_{j\neq i} \mathfrak{a}_j$ has image $\mathfrak{a}_i$ which is therefore a ring R_i say, and so $R \cong \prod R_i$. ∎

Let R be a ring with ideals $\mathfrak{c}_1, \ldots, \mathfrak{c}_t$ and write $R_i = R/\mathfrak{c}_i$. Then we have the natural homomorphisms $f_i : R \to R_i$ which can be combined to give a homomorphism

$$f : R \to \prod R_i. \tag{4.5.4}$$

Explicitly we have $f : x \mapsto (xf_1, \ldots, xf_t)$. We ask: when is (4.5.4) an isomorphism? This is answered by the next result, which generalizes the Chinese remainder theorem. In a ring R, two additive subgroups U, V (e.g. ideals) will be called *comaximal* if $U + V = R$; for left (or right) ideals this holds precisely when $1 \in U + V$.

Theorem 4.5.2. *Let R be a ring with ideals $\mathfrak{c}_1, \ldots, \mathfrak{c}_t$. Put $R_i = R/\mathfrak{c}_i$ and let $f : R \to \prod R_i$ be the homomorphism (4.5.4). Then*

(i) *f is injective if and only if $\cap \mathfrak{c}_i = 0$,*
(ii) *f is surjective if and only if the $\mathfrak{c}_i$ are pairwise comaximal.*

Proof. (i) Clearly $\ker f = \cap \ker f_i = \cap \mathfrak{c}_i$, so (i) follows. (ii) If f is surjective, take $x \in R$ mapping to $(1, 0, \ldots, 0)$; then $x \equiv 1 \pmod{\mathfrak{c}_1}$, $x \equiv 0 \pmod{\mathfrak{c}_2}$, so $1 = 1 - x + x \in \mathfrak{c}_1 + \mathfrak{c}_2$, hence $\mathfrak{c}_1$ and $\mathfrak{c}_2$ are comaximal; similarly for other pairs $\mathfrak{c}_i$, $\mathfrak{c}_j$.

Conversely, assume that the $\mathfrak{c}_i$ are pairwise comaximal and put

$$\mathfrak{a}_i = \bigcap_{j\neq i} \mathfrak{c}_j.$$

Since $\mathfrak{c}_1 + \mathfrak{c}_i = R$ for $i > 1$, there exist $x_i \in \mathfrak{c}_1, y_i \in \mathfrak{c}_i$ such that $x_i + y_i = 1$, and therefore $y_2 \ldots y_t = (1 - x_2) \ldots (1 - x_t) \equiv 1 \pmod{\mathfrak{c}_1}$ and $y_2 \ldots y_t \in \mathfrak{a}_1$, so $\mathfrak{c}_1 + \mathfrak{a}_1 = R$. It follows that there exists $e_1 \in R$ which maps to $(1, 0, \ldots, 0)$ under f. By symmetry there exist $e_i \in R$ mapping to the element with i-th component 1 and the rest 0. Now for any $a_1, \ldots, a_t \in R$, $\sum a_i e_i \mapsto (a_1 f_1, \ldots, a_t f_t)$, where $f_i : R \to R/\mathfrak{c}_i$ is the natural homomorphism. Hence f is surjective, as claimed. ∎

In the decomposition (4.5.2) the unit element e_i of $\mathfrak{a}_i$ is *idempotent*: $e_i^2 = e_i$, *central*: $e_i x = x e_i$ for all $x \in R$, and distinct e_i are *orthogonal*: $e_i e_j = 0$ for $i \neq j$. Moreover, their sum is 1:

$$e_1 + \ldots + e_t = 1. \tag{4.5.5}$$

Thus a direct decomposition (4.5.2) of R yields a decomposition of 1 as a sum of pairwise orthogonal central idempotents. Conversely, let R be a ring in which 1 admits a decomposition (4.5.5) as a sum of pairwise orthogonal central idempotents. Then $\mathfrak{a}_i = e_i R$ $(i = 1, \ldots, t)$ are ideals of R for which (4.5.2) and (4.5.3) hold, so R is then a direct product of the $\mathfrak{a}_i$, by Theorem 4.5.1. This proves

Proposition 4.5.3. *For any ring R the representations* (4.5.2) *of R as a finite direct product correspond to the decompositions of 1 as a sum of pairwise orthogonal central idempotents.* ∎

A central idempotent e is said to be *centrally primitive* if $e \neq 0$ and e cannot be written in the form $e = u + v$, where u, v are central idempotents and $uv = 0$, $u, v \neq 0$. From Proposition 4.5.3 it is clear that a non-zero ring R is indecomposable (into a direct product of several factors) iff 1 is centrally primitive in R.

Any decomposition (4.5.5) with a maximal number of terms has centrally primitive summands. Such maximal decompositions exist in any Noetherian or Artinian ring (but need not exist in general rings). For let R be Noetherian say, and choose a direct decomposition $R = \mathfrak{a}_1 \oplus \mathfrak{a}_1'$ in which $\mathfrak{a}_1'$ is a (proper) maximal ideal. Then $\mathfrak{a}_1$ will be directly indecomposable. Next write $\mathfrak{a}_1' = \mathfrak{a}_2 \oplus \mathfrak{a}_2'$, where $\mathfrak{a}_2'$ is maximal among ideals strictly contained in $\mathfrak{a}_1'$. Continuing in this way, we obtain a sequence of decompositions

$$R = \mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_t \oplus \mathfrak{a}_t';$$

since R is Noetherian, the ascending chain of ideals $\mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_t$ $(t = 1, 2, \ldots)$ must break off and we conclude that $\mathfrak{a}_t' = 0$ at some stage. By the maximality of $\mathfrak{a}_i'$, $\mathfrak{a}_i$ is directly indecomposable for each i, so we have achieved the desired decomposition of R. In an Artinian ring we proceed similarly, choosing $\mathfrak{a}_i$ minimal $\neq 0$ at each stage. Then the $\mathfrak{a}_t'$ form a descending chain which must again break off. So in either case we obtain a decomposition of R as a direct product of indecomposable rings. Such a decomposition must be unique, for if

$$R = \mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_t = \mathfrak{b}_1 \oplus \ldots \oplus \mathfrak{b}_s,$$

where all the $\mathfrak{a}_i$, $\mathfrak{b}_j$ are indecomposable as rings, then $\mathfrak{a}_1 = \mathfrak{a}_1\mathfrak{b}_1 \oplus \ldots \oplus \mathfrak{a}_1\mathfrak{b}_s$ and by indecomposability $\mathfrak{a}_1\mathfrak{b}_j \neq 0$ for just one subscript j, say $j = 1$. Hence $\mathfrak{a}_1 = \mathfrak{a}_1\mathfrak{b}_1$ and similarly $\mathfrak{a}_1\mathfrak{b}_1 = \mathfrak{b}_1$, so $\mathfrak{a}_1 = \mathfrak{b}_1$. Now $\mathfrak{a}_2 \oplus \ldots \oplus \mathfrak{a}_t = \mathfrak{b}_2 \oplus \ldots \oplus \mathfrak{b}_s$, for each side is the annihilator of $\mathfrak{a}_1$ in R. By induction on t we have $s = t$ and for suitable numbering, $\mathfrak{a}_i = \mathfrak{b}_i$. Thus we have

Proposition 4.5.4. *Any Artinian or Noetherian ring can be written in just one way as a direct product of a finite number of indecomposable rings; the factors are uniquely determined as the ideals generated by the centrally primitive central idempotents.* ■

In rings without finiteness conditions the direct product decomposition (4.5.1) may be replaced by a representation of R as a sheaf of simpler rings over a certain topological space, the *decomposition space* of R. This is defined as the Boolean space of central idempotents in R (see Pierce [1967]).

Exercises

1. In any ring R define the characteristic as the additive order of 1, if finite. For a ring R with characteristic n show that $nR = 0$, and R can be written as a direct product of rings whose characteristics are powers of distinct primes.
2. Show that if e is a non-zero idempotent in a ring R. then for any $x \in R$, $e + ex(1 - e)$ is another non-zero idempotent. Deduce that e is central iff it commutes with every idempotent in R.
3. Show that if R is a *reduced* ring (i.e. a ring with no non-zero nilpotent elements), then every idempotent in R is central.
4. Show that any two decompositions of 1 into sums of pairwise orthogonal central idempotents have a common refinement (obtained by decomposing the terms further, resp. taking the products of the idempotents from the two decompositions). Deduce that a decomposition with a maximal number of terms is unique.
5. A ring is said to be *strongly regular* if for each $a \in R$, the equation $xa^2 = a$ has a solution. Show that R is strongly regular iff every principal (left or) right ideal is generated by a central idempotent. (Hint. Show that R is reduced and compute $(axa - a)^2$.)
6. A ring is called *completely primary* if each of its elements is either invertible or nilpotent. Show that every commutative Artinian ring can be written as a finite direct product of completely primary rings. (Hint. If $c \in R$ is neither invertible nor nilpotent, find $a \in R$ and $n \in \mathbf{N}$ such that ac^n is an idempotent $\neq 0, 1$.)
7. Let R be a commutative ring and $\mathfrak{c}_1, \ldots, \mathfrak{c}_t$ be any ideals such that $\cap\, \mathfrak{c}_i = 0$ and the $\sqrt{\mathfrak{c}_i} = \{x \in R | x^n \in \mathfrak{c}_i \text{ for some } n\}$ are distinct maximal ideals. Show that $R \cong \prod R/\mathfrak{c}_i$.
8. Let k be a field and $\alpha_1, \ldots, \alpha_n$ be distinct elements of k. Show that in the polynomial ring $k[x]$ the ideals $(x - \alpha_i)$ are pairwise comaximal. Hence find, for any $\beta_1, \ldots, \beta_n \in k$, a polynomial f in $k[x]$ such that $f \equiv \beta_i \pmod{(x - \alpha_i)}$. Deduce the Lagrange interpolation formula in the form $f = \sum \beta_i e_i$, where e_i is

the unique polynomial of degree less than n determined by $e_i(x)(x - \alpha_i) = c_i f(x)$, the constant c_i being chosen so that $e_i(\alpha_i) = 1$. Deduce that the e_i are pairwise orthogonal idempotents whose sum is 1.

4.6 Free Modules

The theory of vector spaces over a field derives its relative simplicity from the fact that every vector space has a basis. As we have seen in Section 4.3, this holds even for vector spaces over skew fields, whether finite-dimensional or not. But it fails for general rings, and this leads to the notion of a free module.

We recall that in any left R-module M, a family of elements $x_1, \ldots, x_n$ is called *linearly dependent* if for some $\alpha_i \in R$, not all zero, the relation $\sum \alpha_i x_i = 0$ holds. More generally, an infinite family is said to be linearly dependent if some finite subfamily is linearly dependent; explicitly this means that for some family (α_λ) of elements of R, almost all but not all zero, the relation $\sum \alpha_\lambda x_\lambda = 0$ holds. In the contrary case the family is *linearly independent*. A linearly independent generating set is called a *basis* and a module is called *free* if it has a basis. It is clear that any basis of a free module is a *minimal generating set*, i.e. a generating set such that no proper subset generates the whole module.

For any set I there is a free left R-module F_I with a basis of cardinal $|I|$, unique up to isomorphism, obtained by taking the direct sum of $|I|$ copies of R. Explicitly F_I consists of all families of elements of R indexed by I, with almost all components zero, with componentwise addition and multiplication by elements of R. If the family with μ-component $\delta_{\lambda\mu}$ $(\lambda, \mu \in I)$ is denoted by u_λ, then $(\alpha_\lambda) \in F_I$ can be uniquely written as $\sum \alpha_\lambda u_\lambda$ and this shows that the u_λ form a basis of F_I. Any other free R-module F' with a basis (v_λ) of the same cardinal is isomorphic to F_I via the correspondence $\sum \alpha_\lambda u_\lambda \leftrightarrow \sum \alpha_\lambda v_\lambda$.

Free modules can also be characterized by their universal property:

Theorem 4.6.1. *Let R be any ring. Then for any set I there is a left R-module F_I with a mapping $\varphi : I \to F_I$ which is universal for mappings into left R-modules, i.e. given any mapping $f : I \to M$ into a left R-module, there is a unique homomorphism $f' : F_I \to M$ such that $f = \varphi f'$. The module F_I is in fact free on the image of I under φ.*

Proof. Let us define F_I as above, as the free left R-module on a basis (u_λ) $(\lambda \in I)$. Given any mapping $f : I \to M$, if there is a homomorphism $f' : F_I \to M$ of the required form, then clearly $u_\lambda f' = \lambda f$; hence we must have

$$\Big(\sum \alpha_\lambda u_\lambda\Big) f' = \sum \alpha_\lambda (\lambda f). \tag{4.6.1}$$

This shows that at most one such homomorphism can exist. Now it is easily checked that Equation (4.6.1) indeed defines a homomorphism f' with the desired properties. ∎

As a useful consequence we have

Corollary 4.6.2. *Any short exact sequence*

$$0 \longrightarrow M' \xrightarrow{\alpha} M \xrightarrow{\beta} M'' \longrightarrow 0, \tag{4.6.2}$$

in which M'' is free, is split exact.

Proof. To show that (4.6.2) splits it is enough to find a left inverse of β. Let I be a basis of M''; since β is surjective, we can for each $\lambda \in I$ find $x_\lambda \in M$ mapped to λ by β. The mapping $\lambda \mapsto x_\lambda$ extends to a homomorphism $\beta' : M'' \to M$, by Theorem 4.6.1, and $\beta'\beta = 1$, by the definition of β'. Hence the sequence (4.6.2) splits, as required. ■

The importance of free modules stems from the fact that every module can be written as a homomorphic image of a free module.

Theorem 4.6.3. *Let R be any ring and M be an R-module. Then there is a free R-module F with a submodule G such that $M \cong F/G$.*

Proof. Let M be a left R-module say, with generating family (a_λ), $(\lambda \in I)$. Take a free left R-module F with basis u_λ $(\lambda \in I)$; the mapping $\lambda \mapsto a_\lambda$ extends to a homomorphism $f : \sum \alpha_\lambda u_\lambda \mapsto \sum \alpha_\lambda a_\lambda$, which is surjective, because the a_λ generate M; hence $G = \ker f$ is a submodule of F such that $M \cong F/G$. ■

It is clear that a module M is finitely generated iff it can be written as F/G, where F is free on a finite set. If $M \cong F/G$, where F is free and F, G are both finitely generated, we say that M is *finitely presented*; if we merely know that G is finitely generated, M is said to be *finitely related*.

We now turn to the question of comparing the cardinals of different bases of a free module. In the case of vector spaces this cardinal is uniquely determined and is just the dimension of the space. We shall find that for most of the rings encountered here the cardinal is again unique, but this does not hold universally. We therefore consider briefly under what circumstances exceptions can occur.

A free R-module F with a basis of cardinal γ is said to have *rank* γ; if all its bases have the same rank, F is said to have *unique rank*. In the first place we note that any free module which is not finitely generated always has unique rank. This result actually holds in a somewhat more general form. For any subset X of a left R-module M we shall denote by RX the submodule of M generated by X.

Proposition 4.6.4. *Given any ring R, let M be a left R-module with a minimal generating set X. If X is infinite, of cardinal α, then any generating set of M has cardinal at least α. In particular, M is not finitely generated and any two minimal generating sets have the same cardinal.*

Proof. Let Y be any generating set of M, of cardinal β. Every $y \in Y$ is a linear combination of a finite number of elements from X, hence there is a finite subset X_y of X such that $y \in RX_y$. We assert that

$$X = \bigcup_{y \in Y} X_y. \tag{4.6.3}$$

For clearly $\cup X_y \subseteq X$, and $R(\cup X_y)$ is a submodule containing Y and hence equal to M. Thus $\cup X_y$ is a generating set, equal to X, by the minimality of the latter. If Y were finite, (4.6.3) would express X as a finite union of finite sets, which contradicts the fact that X is infinite. Hence Y must be infinite and from (4.6.3) and Proposition 1.2.7,

$$\alpha = |X| \leq \sum |X_y| \leq \aleph_0 \beta = \beta.$$

This shows that Y has cardinal at least α i.e. $\alpha \leq \beta$. If Y is also minimal, then by interchanging the roles of X and Y we see that $\beta \leq \alpha$, hence $\beta = \alpha$. ■

With the help of this result we can establish the general form of Proposition 4.3.9.

Corollary 4.6.5. *Any two decompositions of a semisimple module into simple summands have the same number of terms and for suitable indexing corresponding terms are isomorphic.*

Proof. Suppose first that M is an isotypic module, say $M = \oplus_I S_i$. Choose $u_i \neq 0$ in S_i; then the family $\{u_i\}$ generates M and it is clearly a minimal generating set. If $M = \oplus_J T_j$ is another such decomposition, then $|I| = |J|$ by Proposition 4.6.4 when I is infinite, while the finite case is covered by Proposition 4.3.9. This proves the result for isotypic modules, since $S_i \cong T_j$ in this case. In the general case we can write M as a direct sum of type components: $M = \oplus H_\alpha$; given any expression of M as a direct sum of simple modules, if we group all terms of a given type α together, we find the α-socle H_α. Hence in any decomposition the number of terms of type α is uniquely determined as the cardinal of a minimal generating set of H_α. ■

A ring R is said to have the *invariant basis property* or *invariant basis number* (IBN) if every free left R-module has unique rank. By Proposition 4.6.4 we need only consider finite rank; for a free module F of unique rank r we shall write $\mathrm{rk}(F) = r$. Most rings commonly encountered have IBN and exceptions are usually reckoned among the pathology of rings (see Exercise 2). A trivial example is given by the trivial ring: here $1 = 0$, so the only module is 0 and $0^m \cong 0^n$ for all m, n. This case will usually be excluded in what follows.

Occasionally a stronger condition than IBN is needed. A ring R is said to be *weakly finite* if for each $n \geq 1$, every generating set of n elements of R^n is linearly independent. Given any ring R, suppose that R^n has a generating set of m elements, for some m, n. Then we have a surjective homomorphism $\beta : R^m \to R^n$, giving rise to an exact sequence

$$0 \longrightarrow H \xrightarrow{\alpha} R^m \xrightarrow{\beta} R^n \longrightarrow 0 \tag{4.6.4}$$

which splits, by Corollary 4.6.2. Thus we have

$$R^m \cong H \oplus R^n. \tag{4.6.5}$$

In terms of matrices, let B be the $m \times n$ matrix describing the mapping β in (4.6.4), and let A be the $n \times m$ matrix describing the splitting homomorphism, so that $AB = I_n$. We can now restate the conditions for IBN and weak finiteness:

Proposition 4.6.6. (i) *For any ring R, the following conditions are equivalent:*

(a) *R has invariant basis number;*
(b) *for any m, n, $R^m \cong R^n \Rightarrow m = n$;*
(c) *for any m, n, if $A \in {}^nR^m, B \in {}^mR^n$ and $AB = I$, $BA = I$, then $m = n$.*

(ii) *For any ring R, the following conditions are equivalent:*

(a) *R is weakly finite;*
(b) *for any n, $R^n \cong H \oplus R^n \Rightarrow H = 0$;*
(c) *for any n, if $A, B \in R_n$ and $AB = I$, then $BA = I$.*

Moreover, any non-trivial weakly finite ring has IBN.

Proof. (i) (b) is practically a restatement of the definition (a). The isomorphism $R^m \cong R^n$ is described by matrices as in (c), so the condition for IBN is that $m = n$.

(ii) To say that an n-element generating set of R^n is linearly independent just amounts to the assertion that $H = 0$, and the condition $H = 0$ holds precisely when $\beta\beta' = 1$, i.e. $BA = I$.

Finally, if IBN fails and $R \neq 0$, then we have $R^m \cong R^n$ for some $m \neq n$, say $m > n$. Then $R^n \cong R^{m-n} \oplus R^n$ and by (ii) (b) it follows that R is not weakly finite. ■

It is clear that the trivial ring is weakly finite, but does not have IBN. For examples of rings with IBN that are not weakly finite, see Cohn [1966]. We remark that a ring R with the property $ab = 1 \Rightarrow ba = 1$ is also called 'weakly 1-finite'; other names are 'directly finite' 'von Neumann-finite' or 'inverse symmetric'.

The next result assures us that many of the rings we shall meet are in fact weakly finite (and hence also have IBN).

Theorem 4.6.7. (i) *If a ring is weakly finite or has IBN, then the same is true of the opposite ring.*

(ii) *Any commutative ring is weakly finite.*
(iii) *Any Artinian or Noetherian ring is weakly finite.*
(iv) *If R is weakly finite, the same holds of any subring.*
(v) *Given a ring homomorphism $f : R \to S$, if S has IBN, so does R.*

Proof. (i) follows by the symmetry of conditions (c) in Proposition 4.6.6. To prove (ii) we note that when $AB = I$ over a commutative ring, then $(\det A)(\det B) = 1$, hence A is invertible and $AB = BA = I$.

(iii) Suppose that R fails to be weakly finite, so for some $n \geq 1$, $R^n \cong H \oplus R^n$ with $H \neq 0$. Hence R^n is isomorphic to a proper submodule of itself. By repetition this leads to an infinite descending chain of submodules in R^n, which cannot exist if R is left Artinian, by Theorem 4.2.3. Similarly, if $R^n \cong H \oplus R^n$, we can represent R^n as a proper homomorphic image of itself and when we repeat the process, the kernels of these homomorphisms form an infinite ascending chain of submodules of R^n, which cannot exist when R is left Noetherian.

To establish (iv) we use (ii) (c) of Proposition 4.6.6, which is clearly inherited by subrings. Finally, to prove (v), suppose that R does not have IBN and let A, B be a pair of rectangular (non-square) matrices over R such that $AB = I$, $BA = I$. Applying f, we obtain a pair of non-square matrices Af, Bf satisfying the same equations, hence IBN also fails for S, and (v) is proved. ∎

We conclude this section by noting that any ring other than a field has non-free modules. It is convenient to include several equivalent conditions. In any ring R, if $ab = 1$, we call a a *left inverse* of b and b a *right inverse* of a.

Theorem 4.6.8. *Let R be a non-trivial ring. Any element of R is a unit provided it has a unique left inverse. Moreover, the following conditions on R are equivalent:*

(a) *every left R-module is free;*
(b) *every cyclic left R-module is free;*
(c) *R is simple as left R-module;*
(d) *every non-zero element of R has a left inverse;*
(e) *R is a skew field;*
(a^0)–(e^0) *the left–right analogues of* (a)–(e).

Proof. If $ab = 1$, then $(a + ba - 1)b = ab + bab - b = 1 + b - b = 1$, so by the uniqueness of left inverses, $ba = 1$, and therefore b is a unit.

(a) ⇒ (b) is clear. To prove (b) ⇒ (c), let $\mathfrak{a}$ be a maximal left ideal of R. We have an exact sequence

$$0 \to \mathfrak{a} \to R \to F \to 0,$$

where F is cyclic and hence free, thus $F \cong R/\mathfrak{a}$. Since $\mathfrak{a}$ was maximal, F is simple, and as cyclic free module, $F \cong R$, therefore R is simple as left R-module.

(c) ⇒ (d). Take $c \neq 0$ in R. Then $Rc = R$, hence $bc = 1$ for some $b \in R$ and so c has a left inverse.

(d) ⇒ (e). Any $c \neq 0$ in R has a left inverse b, say, and since $b \neq 0$, it has a left inverse a. Now $ab = bc = 1$, hence $a = a.bc = ab.c = c$, so $a = c$ and $bc = cb = 1$. This shows R to be a skew field. Now (e) ⇒ (a) because over a skew field every module is free, as we saw in Section 4.3 and the equivalence to (a^0)–(e^0) follows by the symmetry of (e). ∎

Exercises

1. Verify that the trivial ring is weakly finite but does not have IBN.
2. Let V be an infinite-dimensional vector space over a field k and put $R = \mathrm{End}_k(V)$. Show that $R^2 \cong R$ as right R-modules and deduce that R does not have IBN.
3. Show that in any non-trivial ring R without IBN there exist positive integers h, k such that $R^m \cong R^n$ iff either $m = n$ or $m, n \geq h$ and $m \equiv n \pmod{k}$. (Hint. Take $(h, h+k)$ to be the first pair (m, n) in the lexicographic ordering such that $m \neq n$ and $R^m \cong R^n$.)
4. Let R be a ring without IBN. Show that there is an integer h such that every finitely generated R-module can be generated by h elements.
5. Show that for any ring R the following conditions are equivalent: (a) for every $n \in \mathbf{N}$ there is a finitely generated R-module which cannot be generated by n elements; (b) for all m, n, $R^n \cong H \oplus R^m$ implies $n \geq m$; (c) for any $A \in {}^mR^n$, $B \in {}^nR^m$, if $AB = I$, then $n \geq m$.
6. A ring with the equivalent properties of Exercise 5 is said to possess UGN (unbounded generating number). Show that any non-trivial weakly finite ring has UGN, and any ring with UGN has IBN. (It can be shown that a ring has UGN iff some non-zero homomorphic image is weakly finite, see e.g. Cohn (1985) p. 8 or Cohn (1995) p. 190).
7. (M. Akgül) Let $V_{m,n}$ be the universal non-IBN ring of type m, n: generated by an $m \times n$ matrix A and an $n \times m$ matrix B, with defining relations $AB = I$, $BA = I$. Show that for any $r \in \mathbf{N}$ there is a homomorphism $V_{m,m+rk} \to V_{m,m+k}$.
8. Let M be a semisimple module and $\theta : N \to M$ be a non-zero homomorphism of a simple submodule N of M into M. Show that θ can be extended to an automorphism of M. (Hint. Reduce to the case of a given type and apply Corollary 4.6.5.) Show that the conclusion may fail if N is not simple, even when θ is injective.

4.7 Projective and Injective Modules

Projective modules form a generalization of free modules which is important in homological algebra. At first sight it looks a little unsatisfactory to replace the (seemingly) well-known notion of 'free module' by that of 'projective module', but the latter has the advantage of being defined categorically. By contrast, we cannot define a free module without using a basis, even though the basis is not an invariant of the module.

We begin with a property of functors. Let $F : \mathscr{A} \to \mathscr{B}$ be a functor between categories of modules. We shall normally assume that F is additive, i.e. for any maps α, α' in $\mathscr{A}$ between the same modules, so that $\alpha + \alpha'$ is defined, we have

$$(\alpha + \alpha')^F = \alpha^F + \alpha'^F.$$

In other words, the mapping $\mathscr{A}(X, Y) \to \mathscr{B}(X^F, Y^F)$ is a group homomorphism. For example, for any $A \in {}_R\mathrm{Mod}$, the hom functors $h^A : X \mapsto \mathrm{Hom}_R(A, X)$ and

$h_A : X \mapsto \mathrm{Hom}_R(X, A)$ are additive functors from ${}_R\mathrm{Mod}$ to Ab, the category of abelian groups. On the other hand, the functor $A \mapsto \mathrm{Hom}_R(\mathrm{Hom}_R(A, R), A)$ is not additive.

Consider any sequence

$$A \xrightarrow{\lambda} B \xrightarrow{\mu} C. \tag{4.7.1}$$

Any functor takes zero maps to zero maps and so, if $\lambda\mu = 0$, then $\lambda^F.\mu^F = 0$. We shall be particularly interested in functors that preserve exact sequences. A functor F is said to be *exact* if it transforms an exact sequence of the form (4.7.1) into the exact sequence

$$A^F \xrightarrow{\lambda^F} B^F \xrightarrow{\mu^F} C^F. \tag{4.7.2}$$

Exact functors are rare and we usually have to be satisfied with less. We define a functor F to be *left exact* if the exactness of

$$0 \xrightarrow{\lambda} B \xrightarrow{\mu} C \tag{4.7.3}$$

implies that the induced sequence

$$0 \to A^F \xrightarrow{\lambda^F} B^F \xrightarrow{\mu^F} C^F \tag{4.7.4}$$

is exact. Similarly, a *right exact* functor preserves exactness when 0 is at the other end. Equivalently, F is left resp. right exact iff it transforms an exact sequence

$$0 \to A \to B \to C \to 0$$

into a sequence $0 \to A^F \to B^F \to C^F \to 0$ which is exact except possibly at C^F resp. A^F.

This shows that F is exact iff it is left and right exact. Similarly a contravariant functor $G : \mathscr{A} \to \mathscr{B}$ is *left exact* if the corresponding covariant functor $\mathrm{op}.G : \mathscr{A}^0 \to \mathscr{B}$ is left exact; a *right exact* contravariant functor G is defined similarly by the right exactness of op.G.

To check that F is left exact we need only verify that it preserves kernels. For, given (4.7.3), we have $A \cong \ker \mu$ and (4.7.4) will be exact at A^F and B^F precisely when $A^F \cong \ker \mu^F$. Similarly F is right exact iff it preserves cokernels. Further, a contravariant functor is left exact iff it maps cokernels to kernels and right exact iff it maps kernels to cokernels.

Theorem 4.7.1. *For any module category* ${}_R\mathrm{Mod}$ *the functor* $\mathrm{Hom}_R(-, -)$ *is left exact in each argument.*

Proof. Consider the functor $h^M = \mathrm{Hom}_R(M, -)$. For any $\mu : B \to C$ the kernel of the induced mapping $\mu^{h^M} : B^{h^M} \to C^{h^M}$ is the set of homomorphisms $f : M \to B$ annihilated by μ, i.e. the maps that factor uniquely through $\ker \mu = A$. Thus $\ker(\mu^{h^M}) = (\ker \mu)^{h^M}$, as claimed. Similarly, for $h_N = \mathrm{Hom}_R(-, N)$ and any $\lambda : A \to$B we have

$$\ker(\lambda^{h_N}) = (\mathrm{coker}\ \lambda)^{h_N}.$$

■

The lack of exactness expressed in Theorem 4.7.1 lends importance to the following

Definition. A module P is called *projective* if the covariant functor $h^P = \mathrm{Hom}_R(P, -)$ is exact; a module I is called *injective* if the contravariant functor $h_I = \mathrm{Hom}_R(-, I)$ is exact.

For checking exactness the following lemma will be useful.

Lemma 4.7.2. *Given a family of modules and homomorphisms*

$$A_i \xrightarrow{\mu_i} B_i \xrightarrow{\beta_i} C_i, \tag{4.7.5}$$

the following conditions are equivalent:

(a) *all the sequences* (4.7.5) *are exact;*
(b) *the sequence*

$$\prod A_i \longrightarrow \prod B_i \longrightarrow \prod C_i \tag{4.7.6}$$

is exact;
(c) *the sequence*

$$\coprod A_i \longrightarrow \coprod B_i \longrightarrow \coprod C_i \tag{4.7.7}$$

is exact.

Proof. Take $x \in \prod B_i$, say $x = (x_i)$, and write $\alpha = \prod \alpha_i$, $\beta = \prod \beta_i$. We have '$x\beta = 0$' $\Leftrightarrow$ '$x_i\beta_i = 0$ for all i', and '$x = y\alpha$ for some $y \in \prod A_i$' $\Leftrightarrow$ '$x_i = y_i$ for some $y_i \in A_i$'. Hence if (4.7.5) is exact for all i, then (4.7.6) is also exact and conversely. For (c) the only difference is that $x = (x_i)$ has only finitely many non-zero components, but the argument is the same. ■

Proposition 4.7.3. *Let (M_i) be a family of modules over any ring R. Then*

(i) $\prod M_i$ *is injective if and only if each M_i is injective,*
(ii) $\coprod M_i$ *is projective if and only if each M_i is projective.*

In particular, for each finite family the direct sum $M_1 \oplus \ldots \oplus M_n$ is injective or projective if and only if each M_i is.

Proof. (i) $\prod M_i$ is injective $\Leftrightarrow$ $\mathrm{Hom}(-, \prod M_i)$ is exact $\Leftrightarrow$ $\prod \mathrm{Hom}(-, M_i)$ is exact, by (4.2.4) $\Leftrightarrow$ $\mathrm{Hom}(-, M_i)$ is exact for each i, by Lemma 4.7.2 $\Leftrightarrow$ M_i is injective for each i. This proves (i); (ii) follows in the same way. ■

This proposition no longer holds when products and coproducts are interchanged (see Further Exercises 27 and 28).

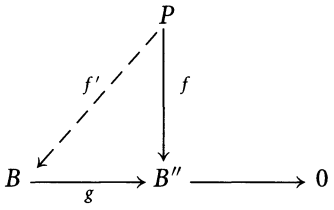
We can now give a number of alternative descriptions of projective modules.

Theorem 4.7.4. *Let R be any ring. For any R-module P the following conditions are equivalent:*

(a) *P is projective,*
(b) *every short exact sequence*

$$0 \to A \xrightarrow{\lambda} B \xrightarrow{\mu} P \to 0 \tag{4.7.8}$$

with P in third position splits,
(c) *P is a direct summand of a free module,*
(d) *every homomorphism from P to a quotient of a module B can be lifted to B; thus the diagram below can be completed by a map $P \to B$ to give a commutative triangle.*



Proof. (a) $\Rightarrow$ (b). Given a short exact sequence (4.7.8), we find by (a) that the sequence of abelian groups

$$0 \to \mathrm{Hom}_R(P, A) \to \mathrm{Hom}_R(P, B) \to \mathrm{Hom}_R(P, P) \to 0$$

is exact. Now $1_P \in \mathrm{Hom}_R(P, P)$ and by exactness there exists $\beta \in \mathrm{Hom}_R(P, B)$ such that $\beta\mu = 1_P$, hence (4.7.8) splits.

(b) $\Rightarrow$ (c). By Theorem 4.6.3 we can write P as a quotient of a free module:

$$0 \to G \to F \to P \to 0.$$

This sequence splits, by (b), so P is a direct summand of the free module F.

(c) $\Rightarrow$ (d). By (c) we can write $F = P \oplus P'$, where F is free. Given a diagram as shown, we obtain a homomorphism $h : F \to B''$ by combining the projection $F \to P$ with the map $f : P \to B''$. Let (u_i) be a basis of F; since g is surjective, we can choose $a_i \in B$ such that $a_i g = u_i h$. Since F is free, there is a homomorphism $h' : F \to B$ which maps u_i to a_i, hence $u_i h' g = a_i g = u_i h$ and so $h = h'g$. It follows that the map $f' : P \to B$ obtained by combining the injection $P \to F$ with h' satisfies $f = f'g$, so it satisfies the conditions of (d).

(d) $\Rightarrow$ (a). Given a short exact sequence, apply $h^P = \mathrm{Hom}_R(P, -)$:

$$0 \to \mathrm{Hom}_R(P, B') \to \mathrm{Hom}_R(P, B) \to \mathrm{Hom}_R(P, B'') \to 0. \tag{4.7.9}$$

By left exactness this can fail to be exact only at $\mathrm{Hom}_R(P, B'')$. But by (d) every map $P \to B''$ lifts to a map $P \to B$ and this means that (4.7.9) is exact at $\mathrm{Hom}_R(P, B'')$ as well. ∎

Although a projective module does not in general have a basis, it has a 'projective coordinate system' with similar properties. We recall that in a free left R-module F with basis (u_i) every element x can be written as

$$x = \sum a_i u_i \quad (a_i \in R). \tag{4.7.10}$$

Here a_i is uniquely determined by x as the value of the i-th projection from F to R. If we denote this projection mapping by α_i and write (α_i, x) for its value at x, so that $(\alpha_i, x) = a_i$, then (4.7.10) becomes

$$x = \sum (\alpha_i, x) u_i. \tag{4.7.11}$$

We shall find that such a representation (4.7.11) still holds for projective modules; in fact it characterizes them, although of course the coefficients are no longer unique, as they were in a free module.

Proposition 4.7.5 (Dual basis lemma). *Let P be a left R-module (for any ring R), and (u_i) be a generating system for P. Then P is projective if and only if there exist $\alpha_i \in \mathrm{Hom}_R(P, R)$ such that for any $x \in P$, (α_i, x) vanishes for almost all i and*

$$x = \sum (\alpha_i, x) u_i \quad \text{for all } x \in P. \tag{4.7.12}$$

Proof. Suppose that P is projective; take a free module F on a basis (v_i) in bijective correspondence with (u_i) and define a mapping $\alpha : F \to P$ by $\alpha : v_i \mapsto u_i$. Since the u_i generate P, we get a surjective mapping and writing $\ker \alpha = N$ we obtain an exact sequence

$$0 \to N \to F \to P \to 0.$$

This sequence splits because P is projective; so there exists $\lambda : P \to F$ such that $x\lambda\alpha = x$ for all $x \in P$. Let $\pi_i : F \to R$ be the projection on the i-th factor and put $\alpha_i = \lambda\pi_i : P \to R$. Then $(\alpha_i, x) = 0$ for almost all i because this is true for π_i. Hence if $x \in P$, then by (4.7.11) applied to F, we have

$$\begin{aligned} x = x\lambda\alpha &= \Big[\sum (\pi_i, x\lambda) v_i\Big]\alpha \\ &= \sum (\alpha_i, x) u_i, \end{aligned}$$

and so (4.7.12) holds. Conversely, given (4.7.12), a homomorphism $f : P \to B''$ and a surjection $g : B \to B''$ as in Theorem 4.7.4(d), choose $a_i \in B$ to satisfy $a_i g = u_i f$, and define $f' : P \to B$ by the rule $xf' = \sum (\alpha_i, x) a_i$. Then f' is R-linear: $(rx)f' = \sum (\alpha_i, rx) a_i = \sum r(\alpha_i, x) a_i = r(xf')$, and the diagram commutes: $xf'g = (\sum (\alpha_i, x) a_i)g = \sum (\alpha_i, x) u_i f = (\sum (\alpha_i, x) u_i) f = xf$. ∎

The case when the family (u_i) is finite is worth stating separately:

Corollary 4.7.6. *The module P is a direct sum of R^n if and only if there exist $u_1, \ldots, u_n \in P$, $\alpha_1, \ldots, \alpha_n \in \mathrm{Hom}_R(P, R)$ such that (4.7.12) holds.* ∎

This means in particular: if P is finitely generated, by n elements say, then it is projective iff it is a direct summand of R^n. For example, a cyclic module is a direct summand of R iff it is projective; but a left ideal of R may well be projective without being a direct summand of R (because it need not be principal). We also note that the equation $(\alpha_i, u_j) = \delta_{ij}$ no longer usually holds.

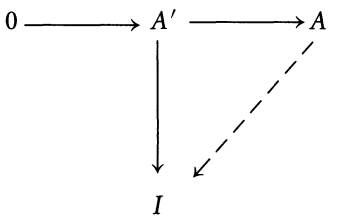
Example 1. *Let* k be a field (possibly skew). Then every k-module is free and hence projective; moreover every projective module (in fact every module) is free.

Example 2. $R = \mathbf{Z}$. Every projective module is free (because every subgroup of a free abelian group is free abelian). More generally, over a principal ideal domain R every submodule of a free R-module is free, hence 'projective' and 'free' mean the same in this case.

Example 3. For any coprime integers m, n we have $\mathbf{Z}/(mn) = \mathbf{Z}/(m) \oplus \mathbf{Z}/(n)$, by the Chinese remainder theorem (Theorem 4.5.2). Regarding the summands on the right as $\mathbf{Z}/(mn)$-modules, we see that they are projective, but they are not free, because the number of elements in a free $\mathbf{Z}/(mn)$-module is divisible by mn.

Example 4. A commutative integral domain in which all ideals are projective is called a *Dedekind domain* (see Section 10.5). Every commutative principal ideal domain is Dedekind, but not conversely; thus the ring of integers in $\mathbf{Q}(\sqrt{-5})$ is a non-principal Dedekind domain.

For injective modules there is a precise analogue of Theorem 4.7.4 (a), (b), (d) but not (c); this will not be needed here (see FA). For the present we note that a module I is injective iff every homomorphism $A' \to I$, where $A' \subseteq A$, can be extended to a homomorphism $A \to I$. Thus the diagram shown can be completed by a map $A \to I$ to make the triangle commutative.



This follows because the property stated is equivalent to the exactness of the sequence

$$\mathrm{Hom}(A, I) \to \mathrm{Hom}(A', I) \to 0.$$

The condition for I to be injective can still be simplified: we have to extend a homomorphism from a submodule of A into I to a homomorphism $A \to I$. Since we can ascend to A by adjoining one element at a time, we need only postulate extendability to modules with cyclic quotients. This idea is made precise in

Theorem 4.7.7 (Baer's criterion). *Let R be a ring and I be a left R-module. Then I is injective if and only if every homomorphism $\mathfrak{a} \to I$, where $\mathfrak{a}$ is a left ideal of R, can be extended to a homomorphism $R \to I$.*

Proof. The remark just made shows the condition to be necessary. Suppose that it holds and let an R-module A with a submodule A' be given, with a homomorphism $f : A' \to I$. We partially order the extensions of f with domain in A by writing $f_1 \leq f_2$ whenever the domain of f_2 contains that of f_1 and f_1, f_2 agree on the domain of f_1. These extensions form an inductive family, as is easily checked; hence by Zorn's lemma there is a maximal one $f'' : A'' \to I$. If $A'' \neq A$, take $a \in A \backslash A''$ and put $B = A'' + Ra$. The set $\mathfrak{a} = \{r \in R | ra \in A''\}$ is a left ideal of R and the map $f_0 : \mathfrak{a} \to I$ defined by $rf_0 = (ra)f''$ is a homomorphism from $\mathfrak{a}$ to I. By hypothesis it can be extended to R, i.e. there exists $u \in I$ such that $rf_0 = ru$. Now write

$$(x + ra)f' = xf'' + ru \quad (x \in A'', r \in R).$$

This is well-defined, for if $x + ra = 0$, then $r \in \mathfrak{a}$ and so $xf'' + ru = xf'' + (ra)f'' = (x + ra)f'' = 0$. Clearly it is a homomorphism of B into I and this contradicts the maximality of A''. Therefore $A'' = A$ and f has been extended to A. ∎

Let R be any integral domain; a left R-module M is said to be *divisible* if the equation in x:

$$u = ax \ (u \in M, a \in R^\times) \tag{4.7.13}$$

always has a solution in M. Every injective module M is divisible: the map $ra \mapsto ru (r \in R)$ is a homomorphism $Ra \to M$; if M is injective, this extends to a homomorphism $R \to M$, and if $1 \mapsto m$ in this homomorphism, then $x = m$ is a solution of (4.7.13). We shall examine the precise relationship between injective and divisible modules in FA; for the moment we shall show that the converse holds over a principal ideal domain:

Proposition 4.7.8. *Every injective module over an integral domain is divisible. Over a principal ideal domain the converse holds: every divisible module is injective.*

Proof. Only the converse remains to be proved. By Theorem 4.7.7 we must show that for a divisible module M, any homomorphism of a left ideal $\mathfrak{a}$ into M extends to a homomorphism $R \to M$. Since R is principal, $\mathfrak{a} = Ra$ for some $a \in R$; suppose that $a \mapsto u$. Then the homomorphism $\mathfrak{a} \to M$ is given by $f : ra \mapsto ru$. By divisibility the equation (4.7.13) has a solution $x = m$ say. Now the map $f' : r \mapsto rm$ is a homomorphism extending f, for we have $(ra)f' = ram = ru$. ∎

In particular this tells us what the injective $\mathbf{Z}$-modules are. Examples of divisible abelian groups are $\mathbf{Q}$ and $\mathbf{Z}(p^\infty)$, the group of all p^n-th roots of unity, $n = 1, 2, \ldots$, where p is a prime. It can be shown that every divisible abelian group is a direct sum of groups of the above types. We shall not stop to show this as it will not be needed in the sequel (see Fuchs (1970) and Exercise 5 below).

Exercises

1. Show that any finitely generated projective module is finitely presented. Show that any finitely related module is the direct sum of a finitely presented module and a free module.
2. Show that if M is a finitely generated module over a Noetherian ring R, then $M^* = \mathrm{Hom}_R(M, R)$ is again finitely generated.
3. Show that any quotient of a divisible module is again divisible.
4. Show that a torsion-free divisible module over a commutative integral domain is injective.
5. Show that a divisible abelian p-group is a direct sum of copies of $\mathbf{Z}(p^\infty)$; show that a torsion-free divisible abelian group is a direct sum of copies of $\mathbf{Q}$. Hence determine the structure of a general divisible abelian group.
6. Let I be an injective R-module and $\mathfrak{a}$ be an ideal in R. Show that the submodule of I annihilated by $\mathfrak{a}$ is injective as $(R/\mathfrak{a})$-module.

4.8 The Tensor Product of Modules

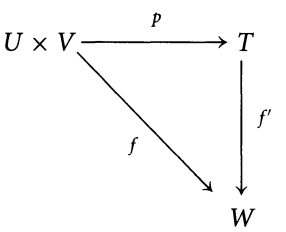
The tensor product may be defined for any pair of bimodules, but before doing this we shall examine the simpler case of modules over a commutative ring. This will help us to understand the general case; it is also enough for many applications.

Let K be a commutative ring and U, V, W be any K-modules. We shall write them as left modules, although, as we have seen, it is only a matter of notation whether a module over a commutative ring is regarded as a left or right module.

We want to consider bilinear mappings from U, V to W, i.e. mappings

$$f : U \times V \to W, \tag{4.8.1}$$

such that f is K-linear in each argument. Our object will be to construct a K-module T and a bilinear mapping $p : U \times V \to T$ which is universal for all bilinear mappings (4.8.1), in the sense that to any bilinear mapping f as in (4.8.1) there corresponds a unique linear mapping $f' : T \to W$ such that the accompanying triangle commutes.



A module T with these properties is called a *tensor product* of U and V and is denoted by $U \otimes_K V$ or simply $U \otimes V$. If it exists it is unique up to isomorphism, as universal object and we shall speak of *the* tensor product.

To prove the existence of T we form the free K-module A on the set $U \times V$ (without the module structure); in A we consider the submodule B generated by all the elements

$$(u+u', v) - (u, v) - (u', v), (u, v+v') - (u, v) - (u, v')(u, u' \in U, v, v' \in V)$$

$$(\alpha u, v) - \alpha(u, v), (u, \alpha v) - \alpha(u, v), \quad \alpha \in K. \tag{4.8.2}$$

There is a mapping $p : U \times V \to A/B$, obtained by taking the inclusion mapping $U \times V \to A$, followed by the natural homomorphism $A \to A/B$. This mapping p is bilinear, for the elements (4.8.2) generating B were just chosen to ensure this. We set $T = A/B$ and claim that T, with the mapping p, is the required tensor product. Let $f : U \times V \to W$ be any bilinear mapping; regarded as a set mapping, i.e. ignoring bilinearity, it may be extended to a unique homomorphism $f_1 : A \to W$, because A is free on the elements (u, v). We claim that $\ker f_1 \supseteq B$; for we have

$$[(u+u', v) - (u, v) - (u', v)]f_1 = (u+u', v)f - (u, v)f - (u', v)f = 0,$$

$$[(\alpha u, v) - \alpha(u, v)]f_1 = (\alpha u, v)f - \alpha[(u, v)f] = 0,$$

by the bilinearity of f, and similarly for the other relations. Hence f_1 may be taken via T, by the factor theorem, and this provides the required mapping $f' : T \to W$. This mapping f' is unique since its values are determined on the images of (u, v) in T and these form a generating set. Our conclusions may be summed up as follows:

Theorem 4.8.1. *Let U, V be modules over a commutative ring K. Then there exists a K-module $U \otimes V$ together with a bilinear mapping $p : U \times V \to U \otimes V$ which is universal for bilinear mappings from $U \times V$ to K-modules.* ■

The image of (u, v) in $U \otimes V$ is denoted by $u \otimes v$. Thus $U \otimes V$ is a K-module with generating set $\{u \otimes v | u \in U, v \in V\}$ and defining relations

$$(u+u') \otimes v = u \otimes v + u' \otimes v, \quad u, u' \in U,$$

$$u \otimes (v+v') = u \otimes v + u \otimes v', \quad v, v' \in V,$$

$$(\alpha u) \otimes v = u \otimes (\alpha v) = \alpha(u \otimes v), \quad \alpha \in K.$$

There is another way of expressing Theorem 4.8.1 which is often useful. Theorem 4.8.1 states in effect that for any K-modules U, V, W there is a natural bijection between the set of bilinear mappings $U \times V \to W$ and the set of homomorphisms $U \otimes V \to W$. Now a mapping $f : U \times V \to W$ is linear in the second variable iff for each $u_0 \in U$, the mapping $V \to U \times V \to W$ given by $v \mapsto (u_0, v) \mapsto (u_0, v)f$ is linear. Further, f is bilinear iff in addition the mapping $U \to \mathrm{Hom}_K(V, W)$ given by $u \mapsto (u, -)f$ is linear, i.e. $f \in \mathrm{Hom}(U, \mathrm{Hom}(V, W))$. Hence there is a natural bijection

$$\mathrm{Hom}_K(U, \mathrm{Hom}_K(V, W)) \cong \mathrm{Hom}_K(U \otimes_K V, W). \tag{4.8.3}$$

This is easily verified to be an isomorphism of K-modules. The property expressed in

(4.8.3) is known as *adjoint associativity*; later we shall see its general form for bimodules.

From the definition it is easy to check that tensor products satisfy the associative and commutative laws:

Proposition 4.8.2. *Let U, V, W be any K-modules, where K is a commutative ring. Then*

$$U \otimes V \cong V \otimes U, \tag{4.8.4}$$

$$U \otimes (V \otimes W) \cong (U \otimes V) \otimes W. \tag{4.8.5}$$

Proof. The rule $(u, v) \mapsto v \otimes u$ is a bilinear mapping $U \times V \to V \otimes U$, and hence gives rise to a homomorphism $\alpha : U \otimes V \to V \otimes U$, in which $u \otimes v \mapsto v \otimes u$. The general element of $U \otimes V$ has the form $\sum u_i \otimes v_i$ and it follows that $\alpha : \sum u_i \otimes v_i \mapsto \sum v_i \otimes u_i$. The same argument shows that $\beta : \sum v_i \otimes u_i \mapsto \sum u_i \otimes v_i$ is a homomorphism; clearly it is inverse to α, hence α is an isomorphism and (4.8.4) follows.

The proof of (4.8.5) is quite similar. We consider the mapping $\alpha : U \times V \times W \to U \otimes (V \otimes W)$ given by $(u, v, w) \mapsto u \otimes (v \otimes w)$. For fixed w this is bilinear in u, v and hence gives rise to a mapping $\alpha'' : (U \otimes V) \otimes W \to U \otimes (V \otimes W)$, in which $(u \otimes v) \otimes w \mapsto u \otimes (v \otimes w)$. The inverse mapping is constructed in the same way and this shows α'' to be an isomorphism, which proves (4.8.5). ∎

We observe that it is possible to define $U \otimes V \otimes W$ directly by the universal property for trilinear mappings, and a similar proof will show that it is isomorphic to either of the modules in (4.8.5). The same holds for more than three factors; this is just the generalized associative law (see Section 2.1). We shall therefore omit brackets in repeated tensor products.

Next we prove a 'distributive law':

Proposition 4.8.3. *For any K-modules U, V', V'' we have*

$$U \otimes (V' \oplus V'') \cong (U \otimes V') \oplus (U \otimes V''). \tag{4.8.6}$$

Proof. We show that the module on the right of (4.8.6) satisfies the universal property of the tensor product. A bilinear mapping from $U \times (V' \oplus V'')$ is given by $(u, v', v'') \mapsto (u \otimes v', u \otimes v'')$. If $f : U \times (V' \oplus V'') \to W$ is any bilinear mapping, then

$$(u, v', v'')f = (u, v')f + (u, v'')f,$$

and the expression on the right can be regarded as a mapping from $(U \otimes V') \oplus (U \otimes V'')$. Thus f is uniquely factored by the standard bilinear mapping, and the result follows. ∎

The definition of the tensor product by a universal property is useful for proving the existence of mappings from $U \otimes V$ to a K-module, for we need only find the

appropriate bilinear mapping from $U \times V$. It also has the merit of generality; but the definition is not such that it allows the structure of $U \otimes V$ to be read off. For example, if r, s are coprime integers, then $\mathbf{Z}/(r) \otimes \mathbf{Z}/(s) = 0$. This is seen as follows. Since r, s are coprime, there exist $m, n \in \mathbf{Z}$ such that $mr + ns = 1$. Now for any $a \in \mathbf{Z}/(r)$, $b \in \mathbf{Z}/(s)$ we have

$$a \otimes b = mr(a \otimes b) + ns(a \otimes b) = m(ra \otimes b) + n(a \otimes sb) = 0.$$

It follows that $\mathbf{Z}/(r) \otimes \mathbf{Z}/(s) = 0$, because the tensor product is generated by elements of the form $a \otimes b$.

It is important to bear in mind that the general element of $U \otimes V$ is not of the form $u \otimes v$, but is a *sum* of such terms: $\sum u_i \otimes v_i$. For example, if V is a free K-module, with basis $e_1, \ldots, e_n$, then every element of $U \otimes V$ can be written uniquely in the form $\sum u_i \otimes e_i (u_i \in U)$, i.e. $U \otimes K^n \cong U^n$. To prove this fact, let us first take the case $n = 1$:

$$U \otimes K \cong U. \tag{4.8.7}$$

We have a bilinear mapping $\theta : (u, \lambda) \mapsto u\lambda$ from $U \times K$ to U, and if $f : U \times K \to W$ is any bilinear mapping, then $(u, \lambda)f = (u\lambda, 1)f$, hence $f = \theta f'$, where $f' : u \mapsto (u, 1)f$, and clearly f' is the only mapping with this property. Thus U satisfies the universal property of Theorem 4.8.1 and (4.8.7) follows. Now $U \otimes K^n \cong U^n$ follows by induction on n, using the distributive law (Proposition 4.8.3). Thus we obtain

Proposition 4.8.4. *For any K-module U over a commutative ring K, the tensor product with a free K-module of rank n is a direct sum of n copies of U:*

$$U \otimes K^n \cong U^n. \quad \blacksquare \tag{4.8.8}$$

By symmetry a corresponding result holds for the first factor, and combining the two, we obtain

Corollary 4.8.5. *If U and V are free K-modules of finite rank over a commutative ring K, say $U \cong K^m$, $V \cong K^n$, then $U \otimes V \cong K^{mn}$. In particular, this applies to finite-dimensional vector spaces over a field, and we then have* $\dim(U \otimes V) = \dim U . \dim V$. $\blacksquare$

Explicitly, if $e_1, \ldots, e_m$ is a basis for U and $f_1, \ldots, f_n$ is a basis for V, then the elements $e_i \otimes f_j$ $(i = 1, \ldots, m, j = 1, \ldots, n)$ form a basis for $U \otimes V$.

We record a property noted before (4.8.7), namely the independence property of the tensor product:

Proposition 4.8.6. *Let U be any K-module and V be a free K-module with basis $e_1, \ldots, e_n$ over a commutative ring K. Then every element of $U \otimes V$ is unique of the form*

$$\sum u_i \otimes e_i, \quad \text{where } u_i \in U. \quad \blacksquare \tag{4.8.9}$$

Caution is needed in applying this result. Thus if $\sum u_i \otimes v_i = 0$ in $U \otimes V$ and the v_i are linearly independent over K, then it does not follow that the u_i must vanish. If the submodule generated by the v_i is denoted by V' (so that the v_i form a basis for V'), then all we can conclude is that the u_i all vanish if $\sum u_i \otimes v_i = 0$ in $U \otimes V'$. Now the inclusion $V' \to V$ induces the homomorphism

$$U \otimes V' \to U \otimes V, \tag{4.8.10}$$

which however may not be injective. For example, the inclusion $2\mathbf{Z} \to \mathbf{Z}$ is injective, but it does not remain so on tensoring with $\mathbf{Z}/(2)$. If $\mathbf{Z}/(2)$, $\mathbf{Z}$, $2\mathbf{Z}$ are generated by e, f, f' respectively, then $(\mathbf{Z}/(2)) \otimes \mathbf{Z}$, $(\mathbf{Z}/(2)) \otimes 2\mathbf{Z}$ are both isomorphic to $\mathbf{Z}/(2)$, by (4.8.7), with generators $e \otimes f$, $e \otimes f'$ respectively. But f' maps to $2f$ and $e \otimes f' \mapsto e \otimes 2f = 2e \otimes f = 0$. Thus (4.8.10) is the zero mapping in this case. A more precise analysis of this phenomenon will be undertaken in FA. For the moment we note that (4.8.10) is certainly injective if V' is a direct summand in V, by Proposition 4.8.3; so in that case we can identify $U \otimes V'$ with its image in $U \otimes V$. We note that this always holds when K is a field.

Let us next consider the effect of the tensor product on homomorphisms. Given any K-linear maps $\alpha : U \to U'$, $\beta : V \to V'$, there is a unique K-linear map $\alpha \otimes \beta : U \otimes V \to U' \otimes V'$ such that the left-hand square of the diagram below commutes:

$$\begin{array}{ccccc} U \times V & \xrightarrow{\alpha \times \beta} & U' \times V' & \xrightarrow{\alpha' \times \beta'} & U'' \times V'' \\ \downarrow{\scriptstyle \lambda} & & \downarrow{\scriptstyle \lambda'} & & \downarrow{\scriptstyle \lambda''} \\ U \otimes V & \xrightarrow{\alpha \otimes \beta} & U' \otimes V' & \xrightarrow{\alpha' \otimes \beta'} & U'' \otimes V'' \end{array} \tag{4.8.11}$$

For the mapping $(u, v) \mapsto u\alpha \otimes v\beta$ from $U \times V$ to $U' \otimes V'$ is bilinear, and hence can be taken via $U \otimes V$, by the universal property of $U \otimes V$.

If $\alpha' : U' \to U''$, $\beta' : V' \to V''$ is another pair of homomorphisms, we obtain a commutative diagram (4.8.11). Since $(u, v)(\alpha \times \beta)(\alpha' \times \beta') = (u\alpha\alpha', v\beta\beta')$ for any $u \in U$, $v \in V$, we have $(\alpha \times \beta)(\alpha' \times \beta') = \alpha\alpha' \times \beta\beta'$, and it follows from the diagram (4.8.11) that

$$\alpha\alpha' \otimes \beta\beta' = (\alpha \otimes \beta)(\alpha' \otimes \beta'). \tag{4.8.12}$$

In the special case $V'' = V' = V$, $\beta' = \beta = 1$, (4.8.12) reduces to

$$\alpha\alpha' \otimes 1 = (\alpha \otimes 1)(\alpha' \otimes 1), \tag{4.8.13}$$

and together with the obvious equation $1 \otimes 1 = 1$ this shows that the assignment $U \mapsto U \otimes V$ is a functor from K-modules to K-modules, for any given V. By symmetry the assignment $V \mapsto U \otimes V$ is also a functor for fixed U. Thus the tensor product is a bifunctor.

The above diagram shows that there is a correspondence between pairs of maps $(\alpha, \beta) \in \mathrm{Hom}_K(U, U') \times \mathrm{Hom}_K(V, V')$ and maps $\alpha \otimes \beta \in \mathrm{Hom}_K(U \otimes V, U' \otimes V')$. So we have a mapping $(\alpha, \beta) \mapsto \alpha \otimes \beta$ which is clearly bilinear; by the universal property of the tensor product it induces a linear mapping

$$\mathrm{Hom}_K(U, U') \otimes \mathrm{Hom}_K(V, V') \to \mathrm{Hom}_K(U \otimes V, U' \otimes V'). \tag{4.8.14}$$

We remark that for a pair of mappings $\alpha : U \to U'$, $\beta : V \to V'$ the expression $\alpha \otimes \beta$ is ambiguous: it may mean the element of the left of (4.8.14) or the induced homomorphism from $U \otimes V$ to $U' \otimes V'$, and one of these is mapped to the other in (4.8.14). It will usually be clear from the context which interpretation is intended; in some important cases the mapping (4.8.14) is an isomorphism and the ambiguity disappears. For example, when U and V are free of finite rank, say $U \cong K^m$, $V \cong K^n$, then (4.8.14) reduces to $U'^m \otimes V'^n \cong (U' \otimes V')^{mn}$, by a double application of Proposition 4.8.3, together with the relation

$$\mathrm{Hom}_K(K^n, U) \cong U^n,$$

which follows by associating with $(u_1, \ldots, u_n) \in U^n$ the map $e_i \mapsto u_i$, where $e_1, \ldots, e_n$ is the standard basis of K^n. In particular, when $U' = U$, $V' = V$, we obtain

Proposition 4.8.7. *If U, V are free modules of finite rank (over a commutative ring K), then the mapping* (4.8.14) *induces the isomorphism*

$$\mathrm{End}_K(U) \otimes \mathrm{End}_K(V) \cong \mathrm{End}_K(U \otimes V). \qquad \blacksquare$$

When we come to consider tensor products over a non-commutative ring, the corresponding construction leads in the first instance to abelian groups rather than modules. Thus let R be any ring, U be a right R-module and V be a left R-module, and for any abelian group W consider mappings $f : U \times V \to W$ which are *biadditive*, i.e. additive in each argument, and *R-balanced*, i.e.

$$(ur, v)f = (u, rv)f \quad \text{for all } u \in U, v \in V, r \in R.$$

A mapping which is biadditive and R-balanced will again be called *R-bilinear*, or simply *bilinear*, if the ring R is clear from the context. We can again construct $U \otimes_R V$, now merely an abelian group, universal for R-balanced biadditive maps from $U \times V$ to abelian groups. The existence is proved as before, $U \otimes V = A/B$, where A is the free abelian group on $U \times V$ and B is the subgroup generated by

$$(u + u', v) - (u, v) - (u', v), u, \quad u' \in U,$$

$$(u, v + v') - (u, v) - (u, v'), \quad v, v' \in V,$$

$$(ur, v) - (u, rv), \quad r \in R.$$

Suppose now that U is an (S, R)-bimodule and V is an (R, T)-bimodule, for some rings S, T. Then the tensor product $U \otimes V$ just defined may be regarded as an

(S,T)-bimodule in the following way. Take $s \in S$ and consider the mapping $\lambda_s : U \times V \to U \otimes V$ defined by

$$\lambda_s : (u, v) \mapsto su \otimes v.$$

Clearly this is biadditive and balanced; e.g. to prove the latter, we have $s(ur) \otimes v = (su)r \otimes v = su \otimes rv$, by the bimodule property of U. It follows that λ induces a homomorphism $U \otimes V \to U \otimes V$ which is simply denoted by s; thus we have

$$s\Big(\sum u_i \otimes v_i\Big) = \sum su_i \otimes v_i. \tag{4.8.15}$$

If we do this for each $s \in S$, we obtain a left S-module structure on $U \otimes V$, for we have, for any $s, s' \in S$,

$$(ss')(u \otimes v) = (ss')u \otimes v = s(s'u) \otimes v = s[s'u \otimes v] = s[s'(u \otimes v)],$$

and of course $1(u \otimes v) = u \otimes v$. Similarly we can define a right T-module structure on $U \otimes V$ such that $(u \otimes v)t = u \otimes vt$ for $t \in T$, and $U \otimes V$ is an (S,T)-bimodule, because

$$s[(u \otimes v)t] = s[u \otimes vt] = su \otimes vt = (su \otimes v)t = [s(u \otimes v)]t.$$

Given any (S,T)-bimodule W, we can as before regard any homomorphism $f : U \times V \to W$ which is S-linear in the first, T-linear in the second argument and R-balanced, as defining for each $u \in U$ a T-linear map $f_u : v \mapsto (u, v)f$. The set of all these T-linear maps has a natural (S,R)-bimodule structure induced from $\mathrm{Hom}_T(V, W)$ and the map $u \mapsto f_u$ is a homomorphism of (S,R)-bimodules; $ur \mapsto f_{ur}$ and $(ur, v)f = (u, rv)f$ because f is R-balanced. Thus the natural homomorphism (4.8.3) leads to an isomorphism of S-bimodules, again called *adjoint associativity*:

$$\mathrm{Hom}_T(U \otimes_R V, W) \cong \mathrm{Hom}_R(U, \mathrm{Hom}_T(V, W)) \quad ({}_SU_R, {}_RV_T, {}_SW_T) \tag{4.8.16}$$

By symmetry we likewise have an isomorphism of T-bimodules:

$$\mathrm{Hom}_S(U \otimes_R V, W) \cong \mathrm{Hom}_R(V, \mathrm{Hom}_S(U, W)). \tag{4.8.17}$$

Like the hom-functor, the tensor product is not an exact functor; however it is right exact:

Proposition 4.8.8. *For any ring R, the tensor product $U \otimes_R V$ is right exact in each variable.*

Proof. By symmetry it will be enough to show that $- \otimes V$ is right exact. Given an exact sequence of right R-modules:

$$U' \xrightarrow{\alpha} U \xrightarrow{\beta} U'' \to 0,$$

we have to show that for any right R-module V, the sequence

$$U' \otimes V \xrightarrow{\alpha'} U \otimes V \xrightarrow{\beta'} U'' \otimes V \to 0$$

is exact, where $\alpha' = \alpha \otimes 1$, $\beta' = \beta \otimes 1$. Clearly β' is surjective and $\alpha'\beta' = 0$, i.e. im $\alpha' \subseteq \ker \beta'$ and it remains to show that equality holds. Since im $\alpha = \ker \beta = X$, say, it is clear that im α' is the subgroup of $U \otimes V$ generated by all products $x \otimes v$ $(x \in X, v \in V)$. Further, each $u'' \in U''$ can be written as $u'' = u\beta$ for some $u \in U$, which is unique (mod X), so we have a bilinear map $U'' \times V \to (U \otimes V)/(\text{im } \alpha')$ given by $(u'', v) \mapsto u \otimes v$, where $u \in U$ is such that $u\beta = u''$. We thus obtain a homomorphism $f : U'' \otimes V \to (U \otimes V)/(\text{im } \alpha')$ which maps $u\beta \otimes v$ to the residue class of $u \otimes v$ (mod im α'), and so has the form $(\beta \otimes 1)f = \beta' f$ on $U \otimes V$. Hence it vanishes on $\ker \beta'$ and so im $\alpha' = \ker \beta'$, as claimed. ■

The following description of the relations in a general tensor product is often useful:

Proposition 4.8.9. *Let R be a ring and U be a right R-module generated by a family (u_λ), $\lambda \in I$ with defining relations $\sum u_\lambda a_{\lambda\mu} = 0$, $\mu \in J$. If V is a left R-module with a family (x_λ) of elements indexed by I, almost all zero, such that*

$$\sum u_\lambda \otimes x_\lambda = 0 \quad \text{in } U \otimes V, \tag{4.8.18}$$

then there exist elements $y_\mu \in V$, almost all zero, such that

$$x_\lambda = \sum a_{\lambda\mu} y_\mu. \tag{4.8.19}$$

Proof. By hypothesis U has a presentation

$$0 \to L \xrightarrow{\alpha} F \xrightarrow{\beta} U \to 0,$$

where F is free on a family (f_λ), $\lambda \in I$, and L is the submodule generated by the elements $\sum f_\lambda a_{\lambda\mu}$. Tensoring with V and observing that this operation is right exact, we obtain an exact sequence

$$L \otimes V \xrightarrow{\alpha'} F \otimes V \xrightarrow{\beta'} U \otimes V \to 0.$$

By hypothesis, $(\sum f_\lambda \otimes x_\lambda)\beta' = \sum u_\lambda \otimes x_\lambda = 0$, hence by exactness, since L is generated by the elements $\sum f_\lambda a_{\lambda\mu}$,

$$\sum f_\lambda \otimes x_\lambda = \Big(\sum f_\lambda a_{\lambda\mu} \otimes y_\mu\Big)\alpha'$$

for some elements $y_\mu \in V$, almost all zero. Now α' is the homomorphism induced by the inclusion $L \to F$ and F is free on the f_λ. Equating coefficients in $F \otimes V$, we obtain (4.8.19). ■

Here it is important to bear in mind the hypothesis that (u_λ) is a generating set of U and $\sum u_\lambda a_{\lambda\mu} = 0$ is a family of defining relations. If (4.8.18) holds for some elements in $U \otimes V$ we cannot conclude that (4.8.19) follows; in fact this comes close to a criterion for U to be flat (i.e. $U \otimes -$ to be exact). We note that the result may be stated in matrix form as follows: Let U be a right R-module with

presentation matrix A, relative to a generating family u (written as a row), so that $uA = 0$. If x is a column vector over V with almost all components 0, such that $u \otimes x = 0$, then there exists a column vector y over V with almost all components 0 such that $x = Ay$.

Exercises

1. Define $U \otimes V \otimes W$ directly by the universal property for trilinear mappings and show that it is isomorphic to $U \otimes (V \otimes W)$. Do the same for more than three factors.
2. Give a direct proof that $\mathrm{Hom}(U, \mathrm{Hom}(V, W)) \cong \mathrm{Hom}(V, \mathrm{Hom}(U, W))$ for K-modules U, V, W.
3. Let A be a commutative integral domain with field of fractions K. Show that for any A-module U the kernel of the natural homomorphism $U \to K \otimes U$ is the torsion submodule of U.
4. Show that tensor products preserve infinite direct sums, but not necessarily direct products. Thus prove that $(\coprod U_i) \otimes V \cong \coprod (U_i \otimes V)$, but the natural mapping $(\prod U_i) \otimes V \to \prod (U_i \otimes V)$ need not be an isomorphism. (Hint. Take $U_i = \mathbf{Z}/(2^i)$, $i \in \mathbf{N}$, $V = \mathbf{Q}$.)
5. Let $U = \mathbf{Z}/(2)$ with basis u and $V = \mathbf{Z}^2$ with basis e_1, e_2. Verify that $v_1 = e_1 + e_2, v_2 = e_1 - e_2$ are linearly independent and $u \otimes v_1 - u \otimes v_2 = 0$, but $u \neq 0$.
6. Let U, V be K-modules (K a commutative ring) and $U^* = \mathrm{Hom}(U, K)$, $V^* = \mathrm{Hom}(V, K)$ be their duals. Find a natural homomorphism $U^* \otimes V^* \to (U \otimes V)^*$ and give conditions on U and V under which this is an isomorphism.
7. Show that there is a natural homomorphism $U^* \otimes V \to \mathrm{Hom}(U, V)$ under which $\alpha \times v$ corresponds to the map $u \mapsto \langle \alpha, u \rangle v$. Verify that this is an isomorphism if U is free of finite rank.
8. Show that for any K-module U, the mapping $\alpha \times u \mapsto \langle \alpha, u \rangle$ defines a homomorphism $U^* \otimes U \to K$. Verify that if U is free of finite rank, this defines a linear map $\mathrm{End}(U) \to K$, which is just the trace.
9. Write down the matrix multiplication implicit in (4.8.12) and if A, B are $m \times m$ resp. $n \times n$, deduce that $\det(A \otimes B) = (\det A)^n (\det B)^m$, $\mathrm{tr}(A \otimes B) = \mathrm{tr}\, A . \mathrm{tr}\, B$.
10. Show that for infinite-dimensional spaces (4.8.14) need not be surjective.

4.9 Duality of Finite Abelian Groups

Let G be an abelian group, written multiplicatively. By a *character* on G one understands a homomorphism of G into the multiplicative group of non-zero complex numbers, $\chi : G \to \mathbf{C}^\times$. If $\chi(x) = 1$ for all $x \in G$, χ is said to be *trivial*. The characters of G can again be multiplied and form a group $G^* = \mathrm{Hom}(G, \mathbf{C}^\times)$, called the *character group* or *dual* of G. For example, if $G = \mathbf{C}_n$ is cyclic of order n, with generator s, and $\chi : \mathbf{C}_n \to \mathbf{C}^\times$ is a character, then $\omega = \chi(s)$ satisfies the equation

$\omega^n = 1$, and conversely, every function $\chi(s^r) = \omega^r$, where ω is an n-th root of 1, is a character of $\mathbf{C}_n$. Further, the mapping $\chi \mapsto \chi(s)$ is a homomorphism from G^* to $\mathbf{C}^\times$. In particular, choosing for $\chi(s)$ a primitive n-th root of 1, we obtain a generator for $\mathbf{C}_n^*$, clearly of order n; thus we see that $\mathbf{C}_n^* \cong \mathbf{U}_n$, where $\mathbf{U}_n$ is the group of n-th roots of 1 in $\mathbf{C}$. Thus $\mathbf{C}_n^*$ is again cyclic of order n.

Let us now write our abelian group additively and put $\mathbf{T} = \mathbf{R}/\mathbf{Z}$. We observe that $\mathbf{T}$, the additive group of real numbers (mod 1), is isomorphic to the multiplicative group of complex numbers of modulus 1, via the mapping $x \mapsto \exp(2\pi i x)$. Since $\mathbf{T}$ is divisible, it is injective as $\mathbf{Z}$-module and it follows that the functor $A \mapsto A^* = \mathrm{Hom}_{\mathbf{Z}}(A, \mathbf{T})$ is exact, so for any subgroup B of A we have the exact sequence

$$0 \to (A/B)^* \to A^* \to B^* \to 0. \tag{4.9.1}$$

Using these facts, we easily obtain

Theorem 4.9.1. *Every finite abelian group is isomorphic to its dual:*

$$A^* \cong A. \tag{4.9.2}$$

Proof. For the cyclic case this has been checked above. In general we can write $A = A_1 \times \ldots \times A_r$, where each A_i is cyclic, by the basis theorem for abelian groups. Now it is easily seen that

$$\mathrm{Hom}(B \times C, \mathbf{T}) \cong \mathrm{Hom}(B, \mathbf{T}) \times \mathrm{Hom}(C, \mathbf{T}),$$

i.e. $(B \times C)^* \cong B^* \times C^*$. By induction it follows that $A^* \cong A_1^* \times \ldots \times A_r^*$ and we have seen that $A_i^* \cong A_i$, therefore $A^* \cong A$, as claimed. ∎

We note that the isomorphism depends essentially on the choice of basis in A; there is no natural transformation from A^* to A. However, we observe that there is a natural transformation

$$\alpha : A \to A^{**}, \tag{4.9.3}$$

of A into its *bidual* (i.e. second dual), defined as follows. Given $x \in A$, we define $\alpha_x \in A^{**}$ by the rule

$$\alpha_x : \chi \mapsto \chi(x). \tag{4.9.4}$$

This is a homomorphism from A^* to $\mathbf{T}$, as is easily checked. We thus have an element α_x of A^{**}, for each $x \in A$, and it is clear that the mapping $x \mapsto \alpha_x$ is a homomorphism from A to A^{**}, indicated by (4.9.3). To show that α is a natural transformation, take a homomorphism of abelian groups $f : A \to B$. It induces a homomorphism $f^* : B^* \to A^*$ (since the functor * is contravariant) and hence $f^{**} : A^{**} \to B^{**}$, and we have a diagram as shown below:

$$\begin{array}{ccc} A & \xrightarrow{f} & B \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \alpha} \\ A^{**} & \xrightarrow{f^{**}} & B^{**} \end{array}$$

Here f^{**} is defined by the equation

$$\alpha_x f^{**} = \alpha_{xf}; \tag{4.9.5}$$

to say that α is natural means that this diagram commutes, which is just the condition (4.9.5). To find the kernel of α, we note that $\alpha_x = 0$ means $\chi(x) = 0$ for all $\chi \in A^*$. Now if $x \neq 0$, then there is a character χ_1 on the subgroup generated by x which does not vanish on x, and by (4.9.1), χ_1 can be extended to a character on A. Therefore we can find $\chi \in A^*$ such that $\chi(x) \neq 0$, and this shows α_x in (4.9.4) to be non-zero when $x \neq 0$. Hence the mapping (4.9.3) is injective, for any abelian group A.

Suppose now that A is finite; then A and A^{**} have the same order, by Theorem 4.9.1, and hence α is then an isomorphism. If we only know that A^* is finite, then so is A^{**}, and since (4.9.3) is injective in any case, we again find that A is finite. So we have

Theorem 4.9.2. *Let A be any abelian group. If either A or its dual A^* is finite, then so is the other and the bidual A^{**} is naturally isomorphic to A.* ∎

We observe that for a finite abelian group A, of exponent m say, every character takes values in $\mathbf{U}_m$, the group of m-th roots of 1; hence A^* may then also be defined as $\mathrm{Hom}(A, \mathbf{U}_m)$. For this reason all that has been said about finite abelian groups still applies if we take $\mathbf{Q}/\mathbf{Z}$ instead of $\mathbf{R}/\mathbf{Z}$; we note that $\mathbf{Q}/\mathbf{Z}$ is just the torsion subgroup of $\mathbf{R}/\mathbf{Z}$.

We conclude by noting the orthogonality relations of characters. Here G is taken to be multiplicative and the values are taken in $C^\times$.

Theorem 4.9.3. *Let G be a finite abelian group of order n. Then*

(i) *for any $\chi, \psi \in G^*$,*

$$\frac{1}{n}\sum_{x\in G} \chi(x)\psi(x^{-1}) = \begin{cases} 1 & \text{if } \chi = \psi, \\ 0 & \text{if } \chi \neq \psi; \end{cases}$$

(ii) *for any $x, y \in G$,*

$$\frac{1}{n}\sum_{\chi\in G^*} \chi(x)\chi(y^{-1}) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{if } x \neq y. \end{cases}$$

Proof. (i) For $\chi = \psi$ the result is clear because $\chi(x)\chi(x^{-1}) = \chi(xx^{-1}) = 1$. Otherwise take $a \in G$ such that $\chi(a) \neq \psi(a)$ and replace x by ax. Since ax runs over G as x does, we have

$$\sum \chi(x)\psi(x^{-1}) = \sum \chi(ax)\psi(x^{-1}a^{-1}) = \chi(a)\psi(a^{-1})\sum \chi(x)\psi(x^{-1}).$$

Hence $[1 - \chi(a)\psi(a^{-1})]\sum \chi(x)\psi(x^{-1}) = 0$, and here the first factor is not zero, by the choice of a, hence $\sum \chi(x)\psi(x^{-1}) = 0$. This proves (i). Now (ii) follows by applying the result to G^* and using Theorem 4.9.2. ∎

Exercises

1. Let G be a finite abelian group and H be a subgroup. Show that the annihilator of H in G^* is a subgroup of G^* of order $(G : H)$.
2. Show that for any field K (even skew) the correspondence $V \mapsto V^* = \mathrm{Hom}_K(V, K)$ is an exact contravariant functor from left to right vector spaces.
3. Let R be a commutative ring and T be an R-module which is injective and such that $M^* = \mathrm{Hom}_R(M, T) \neq 0$ whenever $M \neq 0$ (T is an *injective cogenerator*). For $N \subseteq M$ define $N^\perp$ as the submodule of M^* annihilating N and for $P \subseteq M^*$ define $P^\perp$ as the submodule of M annihilating P. Show that any submodule N of M satisfies $N^{\perp\perp} = N$.
4. Let φ be a non-trivial character on the additive group of a finite field F. Show that every character α of F has the form $\alpha = \alpha_b$, where $\alpha_b(x) = \varphi(bx)$.
5. Let α be a character of the additive group and β a character of the multiplicative group of $\mathbf{F}_q$, the field of q elements. Show that

$$\left| \sum_{x \in \mathbf{F}_q} \alpha(x)\beta(x) \right| = q^{1/2}$$

unless α or β is trivial; the sum is 0 if α is trivial, -1 if β is trivial and $q - 1$ if both are trivial. (Hint. Evaluate the square of the expression on the left.)

Further Exercises for Chapter 4

1. Let R be a ring such that in any left R-module any maximal linearly independent set is a basis. Give a direct proof that R is a skew field.
2. Let R be a ring such that in any torsion-free left R-module, given any linearly dependent set, one (suitably chosen) element of the set is a linear combination of the rest. Show that R is a local ring (i.e. the set of all its non-units is an ideal).
3. Let M be a left R-module with annihilator $\mathrm{Ann}_R(M) = \{x \in R | xM = 0\}$. Show that if M is Noetherian, then $R/\mathrm{Ann}_R(M)$ is left Noetherian.
4. Show that any module of finite composition length is finitely generated.
5. Let $0 \to M_1 \to \ldots \to M_r \to 0$ be an exact sequence of modules. Show that if each module M_i has finite length t_i, then $\sum(-1)^i t_i = 0$. Show that this still holds if the length is replaced by the multiplicity of a given type of simple module in a composition series.
6. Show that if f is an idempotent endomorphism of a module M, then $M = \mathrm{im}\, f \oplus \ker f$.
7. Show that any injective endomorphism of an Artinian module is an automorphism, and dually any surjective endomorphism of a Noetherian module is an automorphism. Give examples to show that 'Artinian' and 'Noetherian' cannot be interchanged.
8. Show that a module M is finitely generated iff the set of all finitely generated proper submodules of M has a maximal member.
9. Show that in a finitely generated module M any generating set contains a finite subset which generates M.

10. Show that every non-trivial commutative ring has a homomorphism onto a field.
11. Let R be a ring which is not right Noetherian. Show that among the right ideals of R that are not finitely generated there is a maximal one.
12. For which integers n is $\mathbf{Z}/(n)$ semisimple as $\mathbf{Z}$-module?
13. A module M is said to be an *essential extension* of a submodule N if N meets every non-zero submodule non-trivially. Show that a semisimple module cannot be an essential extension of a proper submodule.
14. Show that a module is an essential extension of its socle iff every non-zero submodule contains a simple submodule. Deduce that over an Artinian ring every module is an essential extension of its socle. Give an example of a Noetherian module which is not an essential extension of its socle.
15. Let $M = \oplus_I P_i$ be a direct sum of modules (not necessarily simple) and for $J \subset I$ write $\sum_J$ for the sum of all P_i with $i \in J$. Show that $\sum_J \cap \sum_K = \sum_{J\cap K}$ for any subsets J, K of I.
16. Show that the submodules of a module M have unique complements iff M is semisimple and each α-socle is simple.
17. A module is called *semi-Artinian* if every non-zero quotient contains a simple submodule. Show that a module which is Noetherian and semi-Artinian has a finite composition length.
18. Show that a left R-module is simple iff it is isomorphic to $R/\mathfrak{a}$, for some maximal left ideal $\mathfrak{a}$ of R.
19. (K. R. Goodearl) Given rings $A \subseteq B$ and a bimodule ${}_A U_B$, consider the triangular matrix ring $R = \begin{pmatrix} A & U \\ 0 & B \end{pmatrix}$. For any modules M_A, N_B and a B-module homomorphism $f : M \otimes_A B \to N$, we can define $M \oplus N$ as a right R-module by the rule

$$(m, n)\begin{pmatrix} a & u \\ 0 & b \end{pmatrix} = (ma, f(m \otimes u) + nb).$$

Verify that this is indeed an R-module and show that every R-module is of this form.

20. Show that the ideals of $\begin{pmatrix} A & M \\ 0 & B \end{pmatrix}$ are of the form $\mathfrak{a} \oplus N \oplus \mathfrak{b}$, where $\mathfrak{a}$ is an ideal in A, $\mathfrak{b}$ is an ideal in B and N is an (A, B)-bimodule in M such that $N \supseteq \mathfrak{a}M + M\mathfrak{b}$.
21. Show that if R_i is a ring with centre C_i, then the direct product $\prod R_i$ has centre $\prod C_i$.
22. Let R be a commutative ring in which 1 has infinite additive order (e.g. a field of characteristic 0). By comparing traces of matrices show that R has IBN.
23. Show that $\mathrm{Hom}_R(R, M) \cong M$, for any ring R and any R-module M.
24. Show that $\mathrm{Hom}(\mathbf{Z}/(r), \mathbf{Z}/(s)) \cong \mathbf{Z}/(d)$, where $d = (r, s)$.
25. Show that $\mathbf{Q} \otimes_{\mathbf{Z}} \mathbf{Q} \cong \mathbf{Q}$, $A \otimes_{\mathbf{Z}} \mathbf{Q} = 0$ for any abelian torsion group A.

26. Let R be a principal ideal domain. Show that a submodule of a free R-module is free. (Hint. Use the projections on R and apply transfinite induction.)
27. (R. Baer) Fix a prime p and in $\mathbf{Z}^{\mathbf{N}}$ denote by B the subgroup of elements (a_i) such that for any $m \geq 0$ almost all the a_i are divisible by p^m. Assuming that $\mathbf{Z}^{\mathbf{N}}$ is free, show that B is free of uncountable rank; show that B/pB is elementary abelian of countable order and obtain a contradiction. Deduce that $\mathbf{Z}^{\mathbf{N}}$ is not free.
28. Show that over a Noetherian ring any direct sum of injective modules is injective (the Noetherian assumption is necessary as well as sufficient, see FA 4.6).
29. Show that for any ring R and any finitely generated projective R-module P, $\mathrm{Hom}_R(P, M) \cong \mathrm{Hom}_R(P, R) \otimes_R M$. (Hint. Establish a natural transformation and take $P = R^n$.)

5 Algebras

Historically the first rings to be studied (in the second half of the 19th century) were the rings of integers in algebraic number fields. At about the same time the theory of algebras began to develop; its most important landmarks were the Wedderburn structure theorems for semisimple algebras, and the study of the radical. The theories merged when it was realized that the Wedderburn theorems could be stated more generally for Artinian rings. This is the form in which the results will be presented here, in Section 5.2 and 5.3; the formulation for general rings (Jacobson radical and density theorem) will be reserved for FA. As a preparation for this study we examine the form the tensor product takes for algebras in Section 5.4, and in Section 5.5 we introduce scalar invariants. In Section 5.6 algebras are used to define an important number-theoretic function, the Möbius function.

5.1 Algebras; Definition and Examples

Most of our rings have a coefficient ring; this is often given as a field, so that the ring under consideration is a vector space. But it is convenient to frame the definitions more generally; all we need to impose on the coefficient ring is commutativity. At a later stage we shall see how even this restriction can be lifted.

Thus let K be a commutative ring; by an *algebra* over K or *K-algebra* we understand a ring A which is also a K-module, such that the multiplication in A is bilinear. Explicitly we have in addition to the ring laws,

$$\alpha x.y = x.\alpha y = \alpha(xy), \quad \text{for any } x, y \in A, \alpha \in K. \tag{5.1.1}$$

For example, K itself is always a K-algebra in a natural way. We also note that any ring R may be regarded as a $\mathbf{Z}$-algebra by putting

$$na = a + a + \ldots + a(n \text{ terms}), \quad (-n)a = -na, \quad a \in R, n \in \mathbf{N}. \tag{5.1.2}$$

It is easily verified that R becomes a $\mathbf{Z}$-algebra in this way.

A *homomorphism* of K-algebras is a ring homomorphism which is K-linear, and a *subalgebra* of a K-algebra A is a subring admitting multiplication by all the elements of K. For example, the *centre* of A, defined as $C = \{z \in A | zx = xz \text{ for all } x \in A\}$ is a subalgebra of A, as is easily checked.

From (5.1.1) we find by taking $y = 1$ that $x.\alpha 1 = \alpha x$, while $x = 1$ gives $\alpha 1.y = \alpha y$. Therefore

$$\alpha 1.x = x.\alpha 1 = \alpha x, \quad x \in A, \alpha \in K.$$

Taking $x = \beta 1$, we obtain $\alpha 1.\beta 1 = \alpha(\beta 1) = \alpha\beta.1$. These rules, together with the equations $(\alpha + \beta).1 = \alpha 1 + \beta 1$ and $1.1 = 1$, which hold in any K-module, show that the map $\alpha \mapsto \alpha 1$ is a homomorphism of K into the centre of A. Conversely, given any ring R and a homomorphism f from K to the centre of R, we can define R as a K-algebra by putting

$$\alpha x = (\alpha f)x, \quad x \in R, \alpha \in K.$$

This shows a K-algebra A to be nothing more than a ring with a homomorphism from K to the centre of A. In particular, any ring whose centre contains K may be regarded as a K-algebra.

Sometimes a K-algebra is defined more generally as a K-module A with a bilinear multiplication. If A contains an element 1 such that $x.1 = 1.x = x$ for all $x \in A$, the algebra A is said to be *unital*. What we described above is the special case when A is unital and associative, and so is a ring. Occasionally we shall consider the term 'algebra' in this more general sense, but when nothing is said to the contrary, K-algebras are understood to be associative and unital.

We remark that for a non-unital K-algebra, i.e. lacking a 1, we can always adjoin a 1 by forming the direct sum $K \oplus A$ with the multiplication

$$(\alpha, a)(\beta, b) = (\alpha\beta, \alpha b + \beta a + ab), \quad a, b \in A, \alpha, \beta \in K. \tag{5.1.3}$$

The resulting K-algebra is denoted by A^1; it has the one $(1, 0)$ and is associative if A is.

A K-algebra A is said to be *augmented* if there is an algebra homomorphism $\varepsilon : A \to K$. The kernel of ε is an ideal of A, called the *augmentation ideal*. Augmented algebras may be described as follows:

Theorem 5.1.1. *A K-algebra A is augmented if and only if the homomorphism $\alpha \mapsto \alpha 1$ $(\alpha \in K)$ is an embedding and A contains an ideal $\mathfrak{a}$ such that*

$$A = \mathfrak{a} \oplus K.1. \tag{5.1.4}$$

Proof. Assume that A is an augmented K-algebra and write $\mathfrak{a} = \ker \varepsilon$ for the augmentation ideal. Any $x \in A$ may be written in the form

$$x = (x - x\varepsilon) + x\varepsilon,$$

where $x\varepsilon \in K$ and $(x - x\varepsilon)\varepsilon = x\varepsilon - x\varepsilon = 0$, hence $\mathfrak{a} + K.1 = A$, and $\mathfrak{a} \cap K.1 = 0$, because ε restricted to K is injective, so we have the direct sum (5.1.4).

Conversely, if A contains an ideal $\mathfrak{a}$ and a copy $K.1$ of K such that (5.1.4) holds, then the projection on $K.1$ is an augmentation, as is easily verified. ■

We observe that for any algebra A without a 1 the algebra A^1, formed as in (5.1.3), is augmented. Conversely, given an augmented algebra as in (5.1.4), we have $A = \mathfrak{a}^1$; the verification is straightforward and may be left to the reader.

If our coefficient ring is a field k, any k-algebra A is a vector space over k and so has a basis $\{u_i\}$ say. Every element of A can then be uniquely expressed in the form $\sum \alpha_i u_i$ ($\alpha_i \in k$, almost all 0). In particular, the multiplication in A is described completely by the products $u_i u_j$. These take the form

$$u_i u_j = \sum \gamma_{ijr} u_r, \quad \gamma_{ijr} \in k.$$

The elements γ_{ijr} are called the *structure constants* of A. They determine the algebra structure completely, by bilinearity, but it must be remembered that (i) the γ_{ijr} depend on the choice of the basis in A, and (ii) the γ_{ijr} cannot be assigned arbitrarily, for they will need to satisfy the equations ensuring that A is associative and unital. If every non-zero element has an inverse (and $1 \neq 0$), A is a skew field, which is also called a *division algebra* when A is finite-dimensional over k.

Let us give some examples of K-algebras, where K is any commutative ring.

Example 1. Let V be any K-module and put $A = \mathrm{End}_K(V)$. We can regard A as a K-module by defining for $f \in A$, $\alpha \in K$, $x \in V$, $\alpha f : x \mapsto \alpha(xf) = (\alpha x)f$. With this definition it is easily checked that A is indeed a K-algebra. In particular, when V is the free K-module of rank n, $V = K^n$, then $A \cong \mathfrak{M}_n(K)$, the full $n \times n$ matrix ring over K. This is easily verified directly, but follows also from Theorem 5.1.3 below and Corollary 4.4.2.

Example 2. $\mathfrak{T}_n(K)$, the set of all upper triangular $n \times n$ matrices over K, i.e. matrices $C = (c_{ij})$, where $c_{ij} = 0$ for $i > j$.

Example 3. Let A be a finite-dimensional vector space over a field k, with basis $u_1, \ldots, u_n$ say. Then for any elements $c_{ijk} \in k$, a k-algebra on A (not necessarily associative or commutative, or with a one) is defined by setting $u_i u_j = \sum c_{ijk} u_k$. Thus A may be defined by its 'multiplication table'.

Example 4. Let $M = \{u_i | i \in I\}$ be a monoid and take A to be the free K-module on M as basis, with multiplication $u_i u_j = u_k$ as in M. By linearity this defines a multiplication on A and it is not hard to see that A becomes a K-algebra in this way. The associativity of A follows from the associativity of M, and the unit element of M is also the 1 for A. This algebra is usually denoted by KM and is called the *monoid algebra* on M; in particular, for a group G we obtain the *group algebra* KG. The mapping $\varepsilon : \sum \alpha_i u_i \mapsto \sum \alpha_i$ is an augmentation for KM. The following are particular cases of this construction.

(i) Let $M = \{1, x, x^2, \ldots\}$ consist of all powers of a single element x. Then M is the infinite cyclic monoid and KM consists of all polynomials in x over K, with the usual multiplication, thus $KM \cong K[x]$.

(ii) Let G be the infinite cyclic group, with generator x. The group algebra is $K[x, x^{-1}]$, the ring of all *Laurent polynomials* in the indeterminate x.

(iii) Let $G = \mathbf{C}_n$ be the cyclic group of order n, with generator u. Then $K\mathbf{C}_n$ consists of all 'polynomials' $\alpha_0 + \alpha_1 u + \ldots + \alpha_{n-1}u^{n-1}$ with the usual multiplication, but subject to $u^n = 1$.

(iv) Let M be the free monoid on $x_1, \ldots, x_r$, i.e. the set of all 'words' $x_{i_1}x_{i_2}\ldots x_{i_n}$ $(1 \leq i_\nu \leq r)$, with juxtaposition as multiplication, including the empty word as the neutral element. The monoid algebra on K is called the *free associative algebra* on $x_1, \ldots, x_r$ as free generating set and is denoted by $K\langle x_1, \ldots, x_r\rangle$.

Example 5. Any subring of $\mathbf{C}$, the complex numbers, contains $\mathbf{Z}$ and so is a $\mathbf{Z}$-algebra. Given $\alpha \in \mathbf{C}$, consider $\mathbf{Z}[\alpha]$, the subring generated by α. As abelian group this is not usually finitely generated; it is so precisely when α satisfies an equation

$$x^n + a_1x^{n-1} + \ldots + a_n = 0, \quad \text{where } a_i \in \mathbf{Z}. \tag{5.1.5}$$

For if α satisfies (5.1.5), then we can express all powers of α as linear combinations of $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$, so $\mathbf{Z}[\alpha]$ is the finitely generated as abelian group. The converse also holds, as is not hard to verify (see Section 9.4). An equation or polynomial with highest coefficient 1 is called *monic*, and a complex number satisfying a monic equation (5.1.5) with integer coefficients is called an *algebraic integer*. Since the conjugates satisfy the same equation, they are again algebraic integers, though they need not lie in $\mathbf{Z}[\alpha]$. For example, $\mathbf{Z}[i]$, where i is a root of $x^2 + 1 = 0$, is a ring, the ring of *Gaussian integers*.

Example 6. Finite-dimensional algebras over a field. Let k be a field and A be an n-dimensional k-algebra with basis $u_1, \ldots, u_n$. As we have seen, A is completely determined by the n^3 structure constants γ_{ijr}. If we choose the basis so that $u_1 = 1$, then

$$\gamma_{1ir} = \gamma_{i1r} = \delta_{ir},$$

while the associativity is expressed by the equations

$$\sum_\nu \gamma_{ij\nu}\gamma_{\nu rs} = \sum_\nu \gamma_{ivs}\gamma_{jr\nu}.$$

Example 7. Any Boolean algebra may be regarded as an $\mathbf{F}_2$-algebra, as we saw in Section 4.1.

Let A be an n-dimensional algebra over a field k and consider the right multiplication in A,

$$\rho_a : x \mapsto xa, \quad x \in A. \tag{5.1.6}$$

This is a k-linear mapping and hence can be represented by an $n \times n$ matrix over k. By a *matrix representation*, or simply a *representation*, one understands a k-algebra homomorphism into a full matrix ring over k. Now the right multiplication ρ_a is a representation of A, as is easily checked. It is called the *regular representation* of A; more precisely it is the *right* regular representation, the *left* regular representation

being given by $\lambda_a : x \mapsto ax$. Strictly speaking, λ is an *anti*-representation, because we have

$$\lambda_{ab} = \lambda_b \lambda_a. \tag{5.1.7}$$

To describe ρ_a explicitly, if $a = \sum \alpha_i u_i$, then

$$\rho_a : u_i \mapsto \sum u_i \alpha_j u_j = \sum_{j\nu} \alpha_j \gamma_{ij\nu} u_\nu.$$

Hence ρ_a is represented by the matrix

$$(\rho_a)_{ij} = \sum_\nu \alpha_\nu \gamma_{i\nu j}$$

relative to the basis $\{u_j\}$. We further note that ρ is *faithful*, i.e. its kernel as a homomorphism is 0. For if $\rho_a = 0$, then $a = 1.a = 1\rho_a = 0$. We thus obtain an analogue of Cayley's theorem for groups:

Theorem 5.1.2. *Let A be an n-dimensional algebra over a field k. Then A is isomorphic to a subalgebra of the matrix ring $\mathfrak{M}_n(k)$.* ∎

The regular representation can of course be defined for any K-algebra as a subring of $\mathrm{End}_K(A)$, or indeed any ring. It turns out that the rings of left and right multiplication are each others' centralizers in $\mathrm{End}(R)$. This is an important result, best stated for general rings:

Theorem 5.1.3. *Let R be any ring and regard R as left R-module, by left multiplication. Then*

$$\mathrm{End}_R({}_RR) \cong R. \tag{5.1.8}$$

Similarly, $\mathrm{End}_R(R_R) \cong R^o$. Moreover, if R is a K-algebra (over a commutative ring K), then (5.1.8) *is a K-algebra isomorphism.*

Proof. Consider the mapping $\rho : a \mapsto \rho_a$ from R to $\mathrm{End}_R({}_RR)$, where ρ_a is given by (5.1.6). The image is in $\mathrm{End}_R({}_RR)$ because $(bx)a = b(xa)$ for any $b \in R$, and it is easily verified that $\rho_{a+b} = \rho_a + \rho_b$, $\rho_{ab} = \rho_a \rho_b$, $\rho_1 = 1$, and if R is a K-algebra, then $\rho_{\alpha a} = \alpha \rho_a$ for $\alpha \in K$. We have already seen that ρ is injective, and it only remains to prove surjectivity. Let $\theta \in \mathrm{End}_R({}_RR)$, say $1\theta = c$. Then for all $x \in R$, $x\theta = (x.1)\theta = x.1\theta = xc = x\rho_c$ and this shows ρ to be surjective; therefore it is an isomorphism. The assertion for λ follows similarly, but we have an anti-isomorphism this time, because of (5.1.7). ∎

The result may also be expressed by saying that for any ring R, the multiplication rings ρ_R, λ_R are subrings of $\mathrm{End}(R)$ and are centralizers of each other.

We can now give the description promised in Section 4.4 of a class of rings whose only Morita-equivalents are the full matrix rings.

Theorem 5.1.4. *Let R be a ring such that every finitely generated projective R-module is free. Then the rings Morita-equivalent to R are precisely the full matrix rings $\mathfrak{M}_n(R)$, $n = 1, 2, \ldots$.*

Proof. We saw in Section 4.4 that R_n is always Morita-equivalent to R. Conversely, suppose that S is Morita-equivalent to R; then the categories ${}_S\mathrm{Mod}$ and ${}_R\mathrm{Mod}$ are equivalent, and finitely generated projective modules in these categories correspond to each other, for being projective is a categorical property and so is being finitely generated: a module is finitely generated iff it cannot be written as the union of a chain of proper submodules. By hypothesis every finitely generated projective left R-module has the form R^n, for some $n \geq 1$. If S corresponds to R^n under the category equivalence, then these modules have isomorphic endomorphism rings. Now $\mathrm{End}_R({}_RR) \cong R$, by Theorem 5.1.3, hence by Corollary 4.4.2, $S \cong \mathrm{End}_S({}_SS) \cong \mathrm{End}_R(R^n) \cong \mathfrak{M}_n(R)$. ∎

A ring with IBN having the property of Theorem 5.1.4 will be called *projective-free*.

Exercises

1. Show that every K-algebra (for any commutative ring K) on a single generator is commutative and is a homomorphic image of the polynomial ring $K[x]$.
2. Verify that Equations (5.1.2) define a $\mathbf{Z}$-algebra structure on any ring.
3. Show that every 2-dimensional $\mathbf{R}$-algebra has a basis 1, u, where u^2 is either 0 or 1 or -1, and verify that no two of these are isomorphic. Classify all 2-dimensional k-algebras which are (i) unital, (ii) non-unital.
4. Let A be a non-unital k-algebra, where k is a field of characteristic not 2, and suppose that $x^2 = 0$ for all $x \in A$. Show that $xyz = 0$ for all $x, y, z \in A$.
5. Show that the group algebra of every finite non-trivial group has zerodivisors.
6. Let S be a semigroup and kS be its semigroup algebra over a field k (defined as for monoids). Show that if kS has a 1, then so does S.
7. Let R be any ring and K be a commutative ring. Show that $R \otimes_{\mathbf{Z}} K$ is a K-algebra and that the map $\lambda : R \to R \otimes_{\mathbf{Z}} K$ given by $r \mapsto r \otimes 1$ is a ring homomorphism such that for any ring homomorphism $f : R \to A$, where A is a K-algebra, there is a unique K-algebra homomorphism $f' : R \otimes K \to A$ such that $f = \lambda f'$.
8. Let R be a ring such that $1 \neq 0$ and $axa = a$ has a unique solution for each $a \in R$. Show that R is a skew field.
9. Find a non-trivial ring R such that $\mathfrak{M}_2(R) \cong R$. (Hint. Use Exercise 2 of Section 4.6 and Theorem 5.1.4.)
10. Let k be a field. Show that (i) any non-trivial unital k-algebra contains an isomorphic copy of k in its centre and (ii) any ring R whose centre contains a subfield isomorphic to k can be defined as a k-algebra. Examine what goes wrong if R contains a subfield isomorphic to k, but not contained in the centre of R.
11. Show that a Boolean ring R is simple iff $R \cong \mathbf{2}$. Deduce that an ideal I in a Boolean ring R is maximal iff $R/I \cong \mathbf{2}$.

12. Verify that a subset of a Boolean algebra is an ideal iff it is an ideal in the corresponding Boolean ring. Show that an ideal in $\mathscr{P}(S)$ is maximal iff for each subset X of S the ideal contains either X or its complement. (Hint. Use Exercise 11.)
13. Show that over an algebraically closed field of characteristic 0 there are two algebras of dimension 2 and five of dimension 3 (always with 1). How many algebras are there of dimension 4?
14. (J. Dieudonné) Show that the k-algebra generated by x, y with defining relations $y^2 = yx = 0$ is left but not right Noetherian.

5.2 The Wedderburn Structure Theorems

We now come to one of the central results of ring theory, giving an explicit description of simple and semisimple Artinian rings. A ring R is called *simple* if $R \neq 0$ and R has no ideals other than 0 or R. R is called *left semisimple* if it is semisimple as left module over itself. It should be noted that a simple ring is *not* a special case of a left semisimple ring; in fact we shall see in a moment that a left semisimple ring is also right semisimple and (left and right) Artinian, whereas there are simple rings that are not Artinian.

First we note a lemma due to Issai Schur [1905], basic in much that follows.

Lemma 5.2.1 (Schur's lemma). *Let R be any ring. If M is a simple R-module, then $\mathrm{End}_R(M)$ is a skew field.*

Proof. We have to show that every non-zero endomorphism of M is an automorphism. Let $\alpha : M \to M$ be a non-zero endomorphism. Then $\ker \alpha$ is a proper submodule of M, because $\alpha \neq 0$, hence by simplicity, $\ker \alpha = 0$. Similarly, $\mathrm{im}\ \alpha$ is a submodule, non-zero and hence equal to M. Thus α is bijective and its inverse is easily seen to be an endomorphism. Hence every non-zero endomorphism of M has an inverse and so $\mathrm{End}_R(M)$ is a skew field. ∎

We begin by describing simple Artinian rings.

Theorem 5.2.2 (Wedderburn's first structure theorem). *For any ring R the following conditions are equivalent:*

(a) *R is simple and left Artinian;*
(b) *R is left semisimple non-zero and all simple left R-modules are isomorphic;*
(c) *$R \cong \mathfrak{M}_n(D)$, where D is a skew field and $n \geq 1$;*
(a°)–(c°) *the right-hand analogues of* (a)–(c).

Moreover, the integer n in (c) *is unique and D is unique up to isomorphism.*

Proof. (a) ⇒ (b). Let Rc be a minimal left ideal of R. By the simplicity of R we have $R = RcR = \sum Rca_i$, where a_i ranges over R. The left ideal Rca_i is a homomorphic image of Rc, by the map $rc \mapsto rca_i$; so by the minimality of Rc, either $Rca_i = 0$ or

$Rca_i \cong Rc$. Hence R is a sum of left ideals isomorphic to Rc and, by Theorem 4.3.4, a direct sum, so (b) holds. Further, any simple left R-module is a quotient of R by a left ideal, hence isomorphic to a minimal left ideal.

(b) $\Rightarrow$ (c). Since R is finitely generated (by 1) as a left R-module, and semisimple by hypothesis, it is a direct sum of finitely many minimal left ideals, all isomorphic among themselves. Take a minimal left ideal U and suppose that $R \cong U^n$. By Schur's lemma, $D = \mathrm{End}_R(U)$ is a skew field; by Corollary 4.4.2, $\mathrm{End}_R(U^n) \cong \mathfrak{M}_n(D)$; and by Theorem 5.1.3, $\mathrm{End}_R({}_RR) \cong R$; hence $R \cong \mathfrak{M}_n(D)$, as claimed. Here n is uniquely determined as the composition length of ${}_RR$, while D is unique up to isomorphism as the endomorphism ring of the unique simple left R-module type.

(c) $\Rightarrow$ (a). $D_n = \mathfrak{M}_n(D)$ has finite dimension as left D-space; every left ideal is a subspace, so the descending chain condition holds and D_n is left Artinian. To show that D_n is simple, take any $a = (a_{ij}) \neq 0$, say $a_{rs} \neq 0$. Then $e_{ir}ae_{sj}a_{rs}^{-1} = e_{ij}$, hence the ideal generated by a contains all the e_{ij} and so coincides with D_n. This shows D_n to be simple.

Finally, since condition (c) is left–right symmetric, (a°) and (b°) also hold, and again imply (c). ■

Since the centre of a skew field D is evidently a field, D_n is an algebra over a field, so we obtain

Corollary 5.2.3. *Any simple Artinian ring is an algebra over a field.* ■

This result actually holds without the Artinian hypothesis (Exercise 3).

Next we turn to semisimple rings.

Theorem 5.2.4 (Wedderburn's second theorem). *Every left semisimple ring is a finite direct product of full matrix rings over skew fields:*

$$R \cong \mathfrak{M}_{n_1}(D_1) \times \ldots \times \mathfrak{M}_{n_r}(D_r), \tag{5.2.1}$$

where the n_i and the isomorphism types of the D_i are determined by R. Conversely, every ring of the form (5.2.1) *is semisimple; in particular, every left semisimple ring is right semisimple and (left and right) Artinian. Moreover, two minimal left ideals of R are isomorphic if and only if they lie in the same factor on the right of* (5.2.1).

Proof. Since R is left semisimple and finitely generated as left ideal, we have $R = H_1 \oplus \ldots \oplus H_r$, where $H_i \cong I_i^{n_i}$, I_i being a minimal left ideal and different I's being non-isomorphic. By Schur's lemma, $\mathrm{End}_R(I_i) = D_i$ is a skew field, $\mathrm{End}_R(H_i) \cong \mathfrak{M}_{n_i}(D_i)$ and using Theorem 5.1.3 and Corollary 4.3.8, we have

$$R \cong \mathrm{End}_R({}_RR) \cong \prod \mathfrak{M}_{n_i}(D_i).$$

Here n_i and the isomorphism type of D_i is determined by the type component H_i of R, which is itself unique, as we have seen in Section 4.3.

Conversely, for any skew field D and any $n \geq 1$, we have $\mathfrak{M}_n(D) \cong I^n$, where I is a minimal left D-module, represented for example by a single column of the matrix ring $\mathfrak{M}_n(D)$. Hence $\prod \mathfrak{M}_{n_i}(D) \cong \oplus I_i^{n_i}$ is left semisimple. It has finite composition

length and so is left Artinian. By the evident symmetry of the matrix ring it is also right semisimple and right Artinian. ■

From this result we see that there is no need to distinguish between 'left' and 'right' semisimple. For finite-dimensional algebras we have a sharper conclusion, using a sharper form of Schur's lemma for algebraically closed fields (cf. Section 7.3):

Lemma 5.2.5. *Let k be an algebraically closed field, R be a k-algebra and U be a simple R-module which is a finite-dimensional k-space. Then* $\mathrm{End}_R(U) \cong k$.

Proof. As R-module, U is also a k-module and by hypothesis it is finite-dimensional, so each R-endomorphism α of U is represented by a matrix $\rho(\alpha)$. Since k is algebraically closed, $\rho(\alpha)$ has an eigenvalue λ in k. Thus $\rho(\alpha) - \lambda.1$ is singular, and it defines an endomorphism of U, which by Schur's lemma (Lemma 5.2.1) can only be the zero map. Hence $\rho(\alpha) = \lambda.1$ and so $\mathrm{End}_R(U) \cong k$. ■

Proposition 5.2.6. *Any semisimple finite-dimensional algebra R over an algebraically closed field k is a direct product of full matrix rings over k.*

Proof. If we go through the proof of Theorem 5.2.4, we now find that each I_i is a finite-dimensional k-space, hence by Lemma 5.2.5, $\mathrm{End}_R(I_i) \cong k$ and the conclusion follows from Theorem 5.2.4. ■

Theorems 5.2.2 and 5.2.4 were proved by Joseph H. Maclagan Wedderburn in 1908 for finite-dimensional k-algebras. In 1928 Emil Artin made the observation that these theorems were valid more generally for any rings satisfying both chain conditions on right ideals. Emmy Noether remarked in 1929 that Theorem 5.2.2 needed only the descending chain condition. Then in 1939 Charles Hopkins proved that in any ring the ascending chain condition is a consequence of the descending chain condition (see Section 5.3 below; this result was also obtained independently at about the same time by Jacob Levitzki, but owing to wartime conditions only published in 1945). Thus any Artinian ring is necessarily Noetherian (here the presence of a unit element is essential). For this reason the Wedderburn theorems are now usually stated for Artinian rings. Issai Schur first proved the lemma that bears his name in his dissertation in 1901, where he gave a simplified treatment of group representations, a topic that will be taken up in FA.

From the description of semisimple rings in Theorem 5.2.4 we easily derive other conditions which are sometimes useful:

Theorem 5.2.7. *For any ring R the following conditions are equivalent:*

(a) *R is semisimple;*
(b) *every left R-module is semisimple;*
(c) *every left R-module is projective;*
(d) *every left R-module is injective;*
(a^0)–(d^0) *the right-analogues of* (a)–(d).

Proof. (a) $\Leftrightarrow$ (b). (a) is a special case of (b). When (a) holds, then every direct sum of copies of R is semisimple, hence so is every homomorphic image, but this includes every left R-module, by Theorem 4.6.3.

(b) $\Leftrightarrow$ (c) $\Leftrightarrow$ (d). By Theorem 4.3.4, (b) holds iff every short exact sequence of left R-modules splits, and this is equivalent to each of (c), (d). Now the right-hand analogues follow by the symmetry of (a). ■

Let R be a simple Artinian ring, say $R \cong \mathfrak{M}_r(D)$. Every left R-module is a direct sum of copies of the unique simple left R-module. If the number of copies in the direct sum is α, we may define the *dimension* of M over R as

$$[M : R] = \frac{1}{r} \cdot \alpha. \tag{5.2.2}$$

This is either a rational number with denominator dividing r or ∞. Here the normalization has been chosen so that $[R : R] = 1$. In a similar way the dimension of a module over a semisimple ring could be defined as an m-tuple of rational numbers, but this will not be needed.

Exercises

1. Show that in a full matrix ring over a skew field each row is a minimal right ideal and each column is a minimal left ideal.
2. For any ring R and any $n \geq 1$, show that the centre of $\mathfrak{M}_n(R)$ is isomorphic to the centre of R.
3. Show that the centre of a simple ring is a field.
4. Show that the centre of a semisimple ring is a direct product of fields. Show that if the centre of a semisimple ring is a field, then R is simple.
5. Let R be a semisimple ring such that the endomorphism ring of any simple R-module is commutative. Show that $R^{\circ} \cong R$, even though R° need not equal R (the latter equality holds iff R is commutative).
6. Show that the dimension of a module M over a simple Artinian ring R, as defined in (5.2.2), is a non-negative integer n iff M is free of rank n.
7. Let $R = \prod_1^r \mathfrak{M}_{n_i}(D_i)$. Show that every finitely generated left R-module M is defined up to isomorphism by an r-tuple of rational numbers $\alpha = (\alpha_1, \ldots, \alpha_r)$ such that $n_i\alpha_i \in \mathbf{N}$. What is the condition on α for M to be free?
8. Show that every homomorphic image of a semisimple ring is again semisimple, but that this need not hold for every subring.
9. Show that every left or right ideal in a semisimple ring is generated by an idempotent, which is central iff the ideal is two-sided.
10. Let R be a simple ring. Show that any two non-zero elements of R have the same additive order λ, which is either 0 or a prime number. Verify that R may be defined as a P-algebra, where P is the prime field of characteristic λ.
11. Show that if R is left Artinian (or left Noetherian) and $n \geq 1$, then so is $\mathfrak{M}_n(R)$.

5.3 The Radical

In general an Artinian ring need not be semisimple, but we shall find that there is a uniquely determined 'largest' homomorphic image which is semisimple. The kernel of this homomorphism is called the *radical*. To see the form this takes let us consider for a moment the commutative case. From Theorem 5.2.4 it is clear that a commutative Artinian ring is semisimple iff it is the direct product of a finite number of fields. The simplest case is that where R is a homomorphic image of a polynomial ring $k[x]$, k being a field. This is a principal ideal domain, so every ideal has the form (f), where f is a polynomial in x over k. Let $f = p_1^{\alpha_1} \dots p_r^{\alpha_r}$ be a complete factorization of f into irreducible factors p_i (where the p_i are distinct). By the Chinese remainder theorem (Theorem 4.5.2), we have

$$R = k[x]/(f) = \prod_{i=1}^{r} k[x]/(p_i^{\alpha_i}),$$

and this is a direct product of fields iff $\alpha_1 = \dots = \alpha_r = 1$. It follows that the largest semisimple homomorphic image of R is $k[x]/(p_1 \dots p_r)$. We note that $(p_1 \dots p_r)^m \equiv 0 (\mathrm{mod}\, f)$, where $m = \max(\alpha_1, \dots, \alpha_r)$. Thus $p_1 \dots p_r$ is nilpotent (mod f) and we see that R is semisimple whenever it has no nilpotent elements apart from 0. We shall soon see that this condition always holds for commutative Artinian rings, but it certainly does not hold generally in this form, since e.g. the 2×2 matrix ring over a field is semisimple and yet contains the nilpotent element e_{12}. What is required is a nilpotent *ideal*; the ideal generated by e_{12} is not nilpotent because it contains the non-zero idempotent $e_{11} = e_{12}e_{21}$. An ideal $\mathfrak{a}$ is said to be *nilpotent* if $\mathfrak{a}^n = 0$ for some n, where $\mathfrak{a}^n$ is the set of all finite sums of terms $a_1 a_2 \dots a_n$, $a_i \in \mathfrak{a}$. We shall find that the radical in an Artinian ring can be characterized as the sum of all nilpotent ideals. This will be proved in Theorem 5.3.5 below, but to do so it will be convenient to define the notion of radical in quite general rings. First we consider what happens in the Artinian case.

Theorem 5.3.1. *Every Artinian ring R contains a unique ideal $\mathfrak{N}$ such that* (i) *$R/\mathfrak{N}$ is semisimple and* (ii) *any ideal $\mathfrak{c}$ of R such that $R/\mathfrak{c}$ is semisimple satisfies $\mathfrak{c} \supseteq \mathfrak{N}$.*

The ideal $\mathfrak{N}$ determined in this way is called the *radical* of R.

Proof. If $\mathfrak{a}$ is any maximal ideal of R, then the quotient $R/\mathfrak{a}$ is simple Artinian, because the ideals of $R/\mathfrak{a}$ correspond to the ideals of R that contain $\mathfrak{a}$. Given any maximal ideals $\mathfrak{a}_1, \mathfrak{a}_2, \dots$ of R, we can form the descending chain

$$\mathfrak{a}_1 \supset \mathfrak{a}_1 \cap \mathfrak{a}_2 \supset \mathfrak{a}_1 \cap \mathfrak{a}_2 \cap \mathfrak{a}_3 \supset \dots, \tag{5.3.1}$$

where at each stage $\mathfrak{a}_{r+1}$ is chosen so that $\mathfrak{a}_1 \cap \dots \cap \mathfrak{a}_r \not\subset \mathfrak{a}_{r+1}$, to ensure that the inclusion in (5.3.1) is proper. Since R is Artinian, the chain (5.3.1) must break off, so for some r, $\mathfrak{N} = \mathfrak{a}_1 \cap \mathfrak{a}_2 \cap \dots \cap \mathfrak{a}_r$ contains all maximal ideals of R. Since the

$\mathfrak{a}_i$ are distinct and maximal, they are pairwise comaximal, so by Theorem 4.5.2 we have

$$R/\mathfrak{N} \cong \prod_{i=1}^{r} R/\mathfrak{a}_i.$$

Each factor on the right is simple Artinian, hence $R/\mathfrak{N}$ is semisimple. If $\mathfrak{c}$ is any ideal of R such that $R/\mathfrak{c}$ is semisimple, then $\mathfrak{c}$ can be expressed as an intersection of maximal ideals of R; we need only take the ideals of R corresponding to the maximal ideals of $R/\mathfrak{c}$. Thus $\mathfrak{c} = \cap \mathfrak{b}_i$, where each $\mathfrak{b}_i$ is maximal. Since $\mathfrak{b}_i \supseteq \mathfrak{N}$, it follows that $\mathfrak{c} \supseteq \mathfrak{N}$. Thus $\mathfrak{N}$ has the required properties, and it is clear that $\mathfrak{N}$ is uniquely determined by (i) and (ii). ■

The proof of Theorem 5.3.1 shows that the radical of an Artinian ring R may be defined as the intersection of all maximal ideals of R. From the structure of semisimple rings it is clear that the radical may also be defined as the intersection of all maximal left ideals, or equivalently as the intersection of all maximal right ideals. This property was used by Jacobson in 1945 (taking up an earlier idea of Perlis), to study the radical in general rings. In fact there are several equivalent properties of the radical which will also be needed, and we begin by introducing them.

Lemma 5.3.2. *Let R be any ring. For an element $a \in R$ the following conditions are equivalent:*

(a) *for each simple left R-module M, $aM = 0$;*
(b) *a belongs to each maximal left ideal of R;*
(c) *$1 - xa$ has a left inverse for all $x \in R$;*
(d) *$1 - xay$ has an inverse for all $x, y \in R$;*
(a^0)–(d^0) *the right-hand analogues of* (a)–(d).

Proof. (a) ⇒ (b). Let $\mathfrak{m}$ be a maximal left ideal of R. Then $R/\mathfrak{m}$ is a simple left R-module, hence $a(R/\mathfrak{m}) = 0$ by (a), i.e. $aR \subseteq \mathfrak{m}$, and so $a \in \mathfrak{m}$.

(b) ⇒ (c). Assume that (b) holds, but not (c). Then for some $x \in R$, $1 - xa$ has no left inverse, i.e. $1 \notin R(1 - xa)$. By Krull's theorem (Theorem 4.2.6) there is a maximal left ideal $\mathfrak{m} \supseteq R(1 - xa)$, thus $1 - xa \in \mathfrak{m}$, and by (b) $a \in \mathfrak{m}$, hence $xa \in \mathfrak{m}$ and $1 = 1 - xa + xa \in \mathfrak{m}$, a contradiction.

(c) ⇒ (a). Let M be a simple left R-module. Given $u \in M$, if $au \neq 0$, then $Rau = M$ by simplicity, hence $u = xau$ for some $x \in R$, i.e. $(1 - xa)u = 0$, and by (c) it follows that $u = 0$. Hence $au = 0$ for all $u \in M$ and (a) holds.

Now (d) ⇒ (c) trivially; we complete the proof by showing that (a) + (c) ⇒ (d). By (a), $aM = 0$ for any simple module M, hence for any $y \in R$, $ayM \subseteq aM = 0$, i.e. ay satisfies (a) and hence (c). Thus $1 - xay$ has a left inverse $1 - b$ say:

$$(1 - b)(1 - xay) = 1. \tag{5.3.2}$$

This may be written as $b = (b - 1)xay$, hence $bM = 0$ for any simple M, therefore b satisfies (a) and hence also (c), so $1 - b$ has a left inverse $1 - c$:

$$(1 - c)(1 - b) = 1. \tag{5.3.3}$$

By (5.3.2), (5.3.3),

$$1 - c = (1 - c)(1 - b)(1 - xay) = 1 - xay,$$

hence $c = xay$ and now (5.3.2), (5.3.3) show $1 - b$ to be the inverse of $1 - xay$.

This proves that (a)–(d) are equivalent; now the symmetry follows because (d) is left–right symmetric. ∎

We note that if $1 - a$ has the inverse $1 - a'$, then

$$a + a' = aa' = a'a. \tag{5.3.4}$$

Such an element a' is uniquely determined by a; it is called the *quasi-inverse* of a. In any ring R the set $\mathbf{J}(R)$ of all $a \in R$ satisfying one (and hence all) of Lemma 5.3.2 (a)–(d) is called the *Jacobson radical* of R. From the remarks following Theorem 5.3.1 it is clear that $\mathbf{J}(R)$ coincides with the radical in any Artinian ring. We now give another proof of this fact which helps to clarify the relation between the two ways of defining the radical.

Theorem 5.3.3. *Let R be a left (or right) Artinian ring and $J = \mathbf{J}(R)$ be its Jacobson radical. Then R/J is semisimple and J is the least ideal with semisimple quotient.*

Proof. Pick maximal left ideals $\mathfrak{l}_1, \mathfrak{l}_2, \ldots$ in R such that $\mathfrak{l}_1 \cap \mathfrak{l}_2 \cap \ldots \cap \mathfrak{l}_{r-1} \not\subset \mathfrak{l}_r$ as far as possible. Then

$$R \supset \mathfrak{l}_1 \supset \mathfrak{l}_1 \cap \mathfrak{l}_2 \supset \ldots \supset \mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_r, \tag{5.3.5}$$

and since R is Artinian, the chain eventually breaks off, say at the r-th stage. This means that every maximal left ideal contains $\mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_r$, which therefore coincides with J. If any $\mathfrak{l}_i$, $1 \leq i \leq r$, contains the intersection of the others, we omit it; so we may take the intersection $\mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_r$ to be *irredundant*, i.e. no $\mathfrak{l}_i$ can be omitted. Writing $\mathfrak{a}_i = \bigcap_{j \neq i} \mathfrak{l}_j$, we now have $\mathfrak{a}_i \not\subset \mathfrak{l}_i$ and it follows that $\mathfrak{a}_i + \mathfrak{l}_i = R$, because $\mathfrak{l}_i$ is maximal. Furthermore $\mathfrak{a}_i \cap \mathfrak{l}_i = \cap_i \mathfrak{l}_i = J$ and

$$\mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_{i-1}/\mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_i \cong (\mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_{i-1} + \mathfrak{l}_i)/\mathfrak{l}_i = R/\mathfrak{l}_i,$$

where $\mathfrak{l}_1 \cap \ldots \cap \mathfrak{l}_{i-1}$ and $\mathfrak{l}_i$ are comaximal by the maximality of $\mathfrak{l}_i$. Since $R/\mathfrak{l}_i$ is a simple module, this shows that the chain (5.3.5) has simple quotients and hence has finite composition length. Moreover, $\mathfrak{a}_i/J \cong (\mathfrak{a}_i + \mathfrak{l}_i)/\mathfrak{l}_i = R/\mathfrak{l}_i$, hence $\mathfrak{a}_i/J$ is simple, and the sum $\sum(\mathfrak{a}_i/J)$ is direct, for if $x_i \in \mathfrak{a}_i$ satisfies $\sum x_i \equiv 0 (\mathrm{mod}\ J)$, then $x_1 \in \mathfrak{a}_1 \cap \mathfrak{l}_1 = J$, hence $x_1 \equiv 0 (\mathrm{mod}\ J)$, and similarly for each x_i. This shows $R/J = \oplus(\mathfrak{a}_i/J)$ to be semisimple.

J is the least ideal with this property, for if $\mathfrak{c}$ is such that $R/\mathfrak{c}$ is semisimple, then there are left ideals $\mathfrak{b}_i$ minimal above $\mathfrak{c}$ such that $R/\mathfrak{c} = \oplus(\mathfrak{b}_i/\mathfrak{c})$. Clearly if $\mathfrak{d}_1 = \sum_{i \neq 1} \mathfrak{b}_i$, then $R/\mathfrak{d}_1 = (\mathfrak{d}_1 + \mathfrak{b}_1)/\mathfrak{d}_1 \cong \mathfrak{b}_1/(\mathfrak{d}_1 \cap \mathfrak{b}_1) = \mathfrak{b}_1/J$ and this is simple, hence $\mathfrak{d}_1$ is maximal, and similarly for $\mathfrak{d}_i = \sum_{j \neq i} \mathfrak{b}_j$. By definition of J, $\mathfrak{d}_i \supseteq J$, hence $\mathfrak{c} = \cap \mathfrak{d}_i \supseteq J$, as required. ∎

This describes the structure of R/J in the Artinian case. It remains to examine J itself; in particular we wish to identify J in the Artinian case as the maximal nilpotent

ideal. We first note that every nilpotent element has a quasi-inverse. For, given $n \geq 1$, any x satisfies

$$(1-x)(1+x+x^2+\ldots+x^{n-1}) = (1+x+x^2+\ldots+x^{n-1})(1-x) = 1-x^n.$$

This shows that if $x^n = 0$ for some n, then $1-x$ has an inverse, i.e. x has a quasi-inverse. By a *nilideal* one understands an ideal in which every element is nilpotent; every nilpotent ideal is clearly a nilideal, but the converse need not hold (see Exercise 7, also Section 6.3).

Proposition 5.3.4. *Every left or right nilideal of a ring R is contained in $\mathbf{J}(R)$. In particular, this holds for every nilpotent (left or right) ideal.*

Proof. This follows directly from the above remarks and condition (d) of Lemma 5.3.2. ∎

We can now give the promised description of the radical in an Artinian ring.

Theorem 5.3.5. *In any left (or right) Artinian ring R the sum of all nilpotent ideals is itself a nilpotent ideal, the radical J, and R/J is semisimple. Moreover, R itself is semisimple if and only if it has no nilpotent ideals other than zero.*

Proof. By Proposition 5.3.4, $J = \mathbf{J}(R)$ contains all nilpotent left or right ideals and R/J is semisimple, by Theorem 5.3.3, while no smaller ideal has this property. To complete the proof, it therefore only remains to show that J itself is nilpotent. We have $J \supseteq J^2 \supseteq \ldots$ and since R is left Artinian, equality must hold at some stage: $J^k = J^{k+1} = \ldots$. We put $I = J^k$, so that $I^2 = I$; if $I \neq 0$, let $\mathfrak{n}$ be a minimal left ideal subject to $I\mathfrak{n} \neq 0$. Then for some $a \in \mathfrak{n}$, $Ia \neq 0$ and $I(Ia) = I^2 a = Ia$, hence by the minimality of $\mathfrak{n}$, $Ia = \mathfrak{n}$. In particular, $a = xa$ for some $x \in I$, so $(1-x)a = 0$; but $x \in J$, hence by Lemma 5.3.2(c), $a = 0$, and so $Ia = 0$, which is a contradiction. Therefore $I = J^k = 0$, as claimed. ∎

Lemma 5.3.2 has an important consequence which is often useful. We state and prove it in its usual form, for finitely generated modules, though it holds in fact for all non-zero modules (see Bass [1960]).

Lemma 5.3.6 (Nakayama's lemma). *Let R be any ring and M be a finitely generated non-zero left R-module. Then $\mathbf{J}(R)M \neq M$.*

Proof. Since M is finitely generated non-zero, it has a maximal proper submodule M' (Proposition 4.2.5), and M/M' is simple. By Lemma 5.3.2(a), $a(M/M') = 0$ for all $a \in \mathbf{J}(R)$, hence $\mathbf{J}(R)M \subseteq M'$ and so $\mathbf{J}(R)M \neq M$. ∎

We record another form of Nakayama's lemma, which is also used:

Corollary 5.3.7. *Let R be any ring, M be a finitely generated left R-module, N be a submodule of M and $\mathfrak{a}$ be an ideal of R contained in $\mathbf{J}(R)$. If $\mathfrak{a}M + N = M$, then $N = M$.*

Proof. We need only apply Lemma 5.3.6 to M/N. By hypothesis $M/N = \mathfrak{a}(M/N) \subseteq \mathbf{J}(R)(M/N)$ and M/N is finitely generated, hence $M/N = 0$, i.e. $N = M$. ∎

The next result is a useful application of Nakayama's lemma.

Theorem 5.3.8. *Let R be any ring and let I be an ideal contained in $\mathbf{J}(R)$. Given two finitely generated projective left R-modules P, Q, if $P/IP \cong Q/IQ$, then $P \cong Q$.*

Proof. Any map $\bar{f} : P/IP \to Q/IQ$ can be lifted to a map $f : P \to Q$ to make the diagram

$$\begin{array}{ccc} P & \longrightarrow & P/IP \\ \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle \bar{f}} \\ Q & \longrightarrow & Q/IQ \end{array}$$

commutative, because P is projective and $Q \to Q/IQ$ is surjective. Since $\bar{f}$ is surjective, Nakayama's lemma shows that f is surjective. Since Q is projective, $N = \ker f$ is a direct summand of P; it is finitely generated, as homomorphic image of Q, and $N/IN = 0$ because $\ker \bar{f} = 0$. Hence $N = 0$, again by Nakayama's lemma; thus f is an isomorphism, as claimed. ∎

We conclude this section by examining the relation between Noetherian and Artinian rings. A Noetherian ring need not be Artinian, as the example of the rational integers $\mathbf{Z}$ shows. However we shall find that every Artinian ring is Noetherian (Hopkins' theorem). This is a consequence of the following more general result:

Theorem 5.3.9. *Let R be a left Artinian ring and M be a left R-module. Then the following conditions are equivalent:*

(a) *M is Artinian,*
(b) *M is Noetherian,*
(c) *M has a composition series,*
(d) *M is finitely generated.*

Proof. The quotient $A = R/J$ of R by the radical $J = \mathbf{J}(R)$ is semisimple; moreover, J *is nilpotent, say $J^k = 0$. We form the chain of submodules*

$$M \supseteq JM \supseteq J^2M \supseteq \ldots \supseteq J^kM = 0. \tag{5.3.6}$$

Each quotient $F_i = J^{i-1}M/J^iM$ is annihilated by J and so may be regarded as an A-module, and R-submodules and A-submodules are the same. As A-module, F_i is semisimple, by Theorem 5.2.7. Now if M is Artinian, then each F_i is Artinian and as semisimple Artinian A-module it has a composition series. By composing all

these series we obtain a composition series for M. Hence (a) $\Rightarrow$ (c); similarly (b) $\Rightarrow$ (c), clearly (c) $\Rightarrow$ (d), (b) and now (d) $\Rightarrow$ (a) by Theorem 4.2.3. ■

Applying the result to R itself, we obtain

Corollary 5.3.10 (Hopkins' theorem). *Every left Artinian ring is left Noetherian.* ■

In general a ring may well be left Artinian without being right Artinian, as we saw in Exercise 8 of Section 4.4. There are also Artinian modules that fail to be Noetherian (see Hartley [1977]; Cohn [1997]).

Exercises

1. Find the radical of $\mathbf{Z}/(n)$, for different n.
2. Find the radical of the ring of all upper triangular $n \times n$ matrices over a field k.
3. Show that a reduced Artinian ring is a direct product of skew fields.
4. Show that for any ring R and any $n \geq 1$, $\mathbf{J}(R_n) = \mathbf{J}(R)_n$.
5. Let R be a ring and $\mathfrak{a}$ be an ideal in R. Show that if $\mathbf{J}(R/\mathfrak{a}) = 0$, then $\mathbf{J}(R) \subseteq \mathfrak{a}$.
6. For any ring R, show that if $a \in R$ is such that $\bar{a} \in R/\mathbf{J}(R)$ is a unit, then so is a.
7. Let A be the commutative k-algebra generated by $x_1, x_2, \ldots$ with defining relations $x_n^{n+1} = 0$ $(n = 1, 2, \ldots)$. Show that the ideal generated by $x_1, x_2, \ldots$ is a nilideal but not nilpotent.
8. By Theorem 5.3.9 every finitely generated module M over an Artinian ring R has a finite composition series. Find a bound for the composition length in terms of the number of generators of M and invariants of R.
9. (S. A. Amitsur) Let A be an algebra (not necessarily Artinian) over a field k. Show that any element of $\mathbf{J}(A)$ algebraic over k is nilpotent. Deduce that if A has countable dimension over an uncountable field, then $\mathbf{J}(A)$ is a nilideal. (Hint. Show that if a is transcendental over k and λ_i are distinct elements of k, then the elements $(a - \lambda_i)^{-1}$ are linearly independent over k.)
10. Verify that the ring $k[[x]]$ of formal power series in x over a field k is an integral domain with non-zero Jacobson radical.
11. (A. A. Klein) In a ring R, let $\mathfrak{a}$ be a right ideal satisfying the identity $x^n = 0$. Choose $a \in \mathfrak{a}$ such that $a \neq 0$ and by evaluating $(a^{n-1}y + a)^n$ for $y \in R$ show that $a^{n-1}Ra^{n-1} = 0$. Deduce Levitzki's theorem: if R has a non-zero right ideal which is a nilideal of bounded index, then R has a non-zero nilpotent ideal.

5.4 The Tensor Product of Algebras

Let K be a commutative ring. We recall that a K-algebra is a K-module A with a bilinear mapping $\mu_0 : A \times A \to A$, the multiplication in A. By Theorem 4.8.1 it comes to the same thing to have a linear mapping

$$\mu : A \otimes A \to A, \tag{5.4.1}$$

and we shall also refer to μ in (5.4.1) as the *multiplication* in A. Explicitly we have $(\sum x_i \otimes y_i)\mu = \sum x_i y_i$ for $x_i, y_i \in A$. The associativity of A is expressed by the commutativity of the diagram

$$\begin{array}{ccc} A \otimes A \otimes A & \xrightarrow{\mu \otimes 1} & A \otimes A \\ \downarrow{\scriptstyle 1 \otimes \mu} & & \downarrow{\scriptstyle \mu} \\ A \otimes A & \xrightarrow{\mu} & A \end{array}$$

and the existence of a unit element e in A is expressed by the equations $(x \otimes e)\mu = (e \otimes x)\mu = x$, while commutativity is expressed by $(x \otimes y)\tau\mu = (x \otimes y)\mu$, where τ is the transposition

$$\tau : x \otimes y \mapsto y \otimes x. \tag{5.4.2}$$

Given two K-algebras A, B, we can define their tensor product as a K-algebra in a natural fashion. Let $\tau : B \otimes A \to A \otimes B$ be defined as in (5.4.2), for $x \in B$, $y \in A$; this gives rise to the permutation map $\tau_1 : 1 \otimes \tau \otimes 1 : A \otimes B \otimes A \otimes B \to A \otimes A \otimes B \otimes B$, where

$$(a_1 \otimes b_1 \otimes a_2 \otimes b_2)\tau_1 = a_1 \otimes a_2 \otimes b_1 \otimes b_2.$$

If we combine this with the multiplications μ of A and ν of B, we obtain a linear mapping

$$\pi = \tau_1(\mu \otimes \nu) : A \otimes B \otimes A \otimes B \to A \otimes B. \tag{5.4.3}$$

We claim that this multiplication is associative whenever μ and ν are. Put $C = A \otimes B$; then (5.4.3) can be written $\pi : C \otimes C \to C$ and for any $a_i \in A$, $b_i \in B$ we have

$$\begin{aligned}(a_1 \otimes b_1 \otimes a_2 \otimes b_2 \otimes a_3 \otimes b_3)(\pi \otimes 1)\pi &= (a_1 a_2 \otimes b_1 b_2 \otimes a_3 \otimes b_3)\pi \\ &= (a_1 a_2)a_3 \otimes (b_1 b_2)b_3.\end{aligned}$$

Applying $(1 \otimes \pi)\pi$, we obtain $a_1(a_2 a_3) \otimes b_1(b_2 b_3)$, which is the same, by the associativity in A and B. Since the elements on the left span $C \otimes C \otimes C$, the associativity of π follows. In a similar way we can show that C is commutative whenever A and B are. When A has a unit element e, then the mapping

$$b \mapsto e \otimes b \quad (b \in B) \tag{5.4.4}$$

is a homomorphism from B to $A \otimes B$; likewise for a one in B, and if A, B have ones e, f say, then $e \otimes f$ is a one for C. We sum up these results in

Theorem 5.4.1. *Let K be a commutative ring and A, B be any K-algebras. Then $A \otimes B$ is again a K-algebra, which is associative, commutative or has a unit element whenever this is so for both A and B.* ∎

In what follows we shall assume that our algebras are associative and have a one. Even when A has a one e, (5.4.4) need not be an embedding (see Exercise 5), but this is so if $K = k$ is a field and $A \neq 0$, for in that case A has a basis including e and the subspace spanned by e is a direct summand in A, so we can apply Proposition 4.8.3.

Example 1. Let k be a field and E be an extension field of k. If A is any k-algebra, then $A \otimes E$ is an algebra over E, of dimension $[A : k]$. Explicitly, if $u_1, \ldots, u_n$ is a k-basis of A, then the elements $u_1 \otimes 1, \ldots, u_n \otimes 1$ form a basis of $A \otimes E$ over E. Regarded as an E-algebra, $A \otimes E$ is denoted by A_E and is called the algebra obtained from A by extension of the ground field to E.

Example 2. Let $A = \mathfrak{M}_r(K)$ be a full matrix ring over K. Then A has a basis e_{ij} $(i, j = 1, \ldots, r)$ over K (the 'matrix units'), with the multiplication rule

$$e_{ij}e_{kl} = \delta_{jk}e_{il}, \quad \sum e_{ii} = 1.$$

For any K-algebra B, the tensor product $A \otimes B$ is a free B-module with the same basis as A, hence

$$\mathfrak{M}_r(K) \otimes B \cong \mathfrak{M}_r(B). \tag{5.4.5}$$

By combining Theorem 5.2.2 with (5.4.5) we see that every central simple algebra is a tensor product of a central division algebra and a full matrix algebra.

If in (5.4.5), $B = \mathfrak{M}_s(K)$, then the right-hand side becomes $\mathfrak{M}_r(\mathfrak{M}_s(K)) \cong \mathfrak{M}_{rs}(K)$, for the elements of $\mathfrak{M}_r(\mathfrak{M}_s(K))$ are $r \times r$ matrices whose entries are $s \times s$ matrices over K. Thus we find

$$\mathfrak{M}_r(\mathfrak{M}_s(K)) \cong \mathfrak{M}_{rs}(K). \tag{5.4.6}$$

As a module isomorphism this follows already from Corollary 4.8.5; the above argument shows that it is an algebra isomorphism.

Example 3. In any field k of characteristic not 2, take $a, b \in k$ and define a 4-dimensional algebra with basis 1, u, v, uv and multiplication rules

$$u^2 = a, v^2 = b, vu = -uv.$$

This algebra is called a *quaternion algebra* and will be denoted by $(a, b; k)$. In this notation Hamilton's quaternions take the form $(-1, -1; \mathbf{R})$.

There is a simple criterion for an algebra to be a tensor product which is often useful. Let C be an algebra over a field k. Two subspaces U, V of C are said to be *linearly disjoint* over k if for any linearly independent elements u_i in U and v_j in V the elements u_iv_j in C are linearly independent over k. Clearly this just means that the natural mapping $U \otimes V \to C$ induced by the mapping $(u, v) \mapsto uv$ is injective. Now the criterion can be stated as follows:

Proposition 5.4.2. *Let C be an algebra over a field k. Given subalgebras A, B of C, if* (i) *A and B are linearly disjoint,* (ii) *$AB = C$ and* (iii) *A and B commute elementwise, then $C \cong A \otimes B$.*

Proof. The mapping $(x, y) \mapsto xy$ from $A \times B$ to C is bilinear and so induces a k-linear mapping $A \otimes B \to C$. It is injective by (i), surjective by (ii) and a homomorphism by (iii). ∎

Next we shall describe centralizers in a tensor product. In any ring R the *centralizer* of a subset X is the set

$$\mathbf{C}(X) = \{a \in R | ax = xa \quad \text{for all } x \in X\}.$$

It is easily verified that $\mathbf{C}(X)$ is a subring of R; when R is a K-algebra, $\mathbf{C}(X)$ is again a K-algebra.

We shall also need a formula for intersections in a tensor product. Let U, V be any K-modules (where K is a commutative ring) and let U', V' be direct summands in U, V respectively, so the $U' \otimes V$, $U \otimes V'$, $U' \otimes V'$ may be regarded as submodules of $U \otimes V$, by Proposition 4.8.3. Then

$$U' \otimes V \cap U \otimes V' = U' \otimes V'. \tag{5.4.7}$$

For we can write $U = U' \oplus U''$, $V = V' \oplus V''$; hence

$$\begin{aligned} U \otimes V &= (U' \otimes V') \oplus (U' \otimes V'') \oplus (U'' \otimes V') \oplus (U'' \otimes V'') \\ &= W_1 \oplus W_2 \oplus W_3 \oplus W_4, \end{aligned} \tag{5.4.8}$$

say. Now the left-hand side of (5.4.7) is $(W_1 \oplus W_2) \cap (W_1 \oplus W_3)$ and by (5.4.8) this is clearly W_1, so (5.4.7) holds.

Proposition 5.4.3. *Let A_1, A_2 be algebras over a field k, let B_i be a subalgebra of A_i and B_i' be the centralizer of B_i in A_i $(i = 1, 2)$. Then the centralizer of $B_1 \otimes B_2$ in $A_1 \otimes A_2$ is $B_1' \otimes B_2'$.*

Proof. Let us denote by C the centralizer of $B_1 \otimes B_2$ in $A_1 \otimes A_2$. It is clear that

$$B_1' \otimes B_2' \subseteq C, \tag{5.4.9}$$

where we have identified $B_1' \otimes B_2'$ with its image in $A_1 \otimes A_2$. This is justified by the above remark because k is a field.

It remains to prove equality in (5.4.9). Using a k-basis $\{v_i\}$ of A_2, we can write every element of $A_1 \otimes A_2$ uniquely in the form $\sum a_i \otimes v_i$, where $a_i \in A_1$. Given any $b \in B_1$, we have

$$\Big(\sum a_i \otimes v_i\Big)(b \otimes 1) - (b \otimes 1)\Big(\sum a_i \otimes v_i\Big) = \sum \big(a_i b - b a_i\big) \otimes v_i. \tag{5.4.10}$$

If $\sum a_i \otimes v_i \in C$, then the left-hand side of (5.4.10) is 0, hence $a_i b - b a_i = 0$ for all i, and this holds for all $b \in B_1$; therefore $a_i \in B_1'$ and so $\sum a_i \otimes v_i \in B_1' \otimes A_2$. Thus $C \subseteq B_1' \otimes A_2$ and similarly, $C \subseteq A_1 \otimes B_2'$; it follows that

$$C \subseteq (A_1 \otimes B_2') \cap (B_1' \otimes A_2) = B_1' \otimes B_2',$$

by (5.4.7). Together with (5.4.9) this shows that $C = B_1' \otimes B_2'$ as claimed. ∎

If we take $B_i = A_i$, then B_i' is the centre of A_i and we obtain

Corollary 5.4.4. *If A_1, A_2 are algebras over a field with centres Z_1, Z_2 respectively, then the centre of $A_1 \otimes A_2$ is $Z_1 \otimes Z_2$.* ∎

As we saw in Section 5.2, the centre of a simple algebra A over a field k is a field, E say. Clearly $E \supseteq k$; if equality holds, A is said to be *central* simple. While a general treatment of central simple algebras will be reserved for FA, here are a few results that we shall use later.

A k-algebra D which is a skew field of finite dimension over k will be called a *division algebra* over k, and the dimension will be written $[D : k]$. For example, over the complex numbers $\mathbf{C}$ the only division algebra is $\mathbf{C}$ itself; this is a consequence of the fact that $\mathbf{C}$ is algebraically closed, i.e. every algebraic equation over $\mathbf{C}$ has a root in $\mathbf{C}$. More generally, we have

Proposition 5.4.5. *The only division algebra over an algebraically closed field F is F itself.*

Proof. Let D be a division algebra over F. Given $a \in D$, the powers of a are linearly dependent over F, because D is finite-dimensional, say

$$f(a) = a^n + c_1 a^{n-1} + \ldots + c_n = 0 \quad (c_i \in F)$$

is an equation of least degree for a. Since F is algebraically closed, $f(x)$ has a zero λ in F, and hence $f(x)$ has the factor $x - \lambda : f(x) = (x - \lambda)g(x)$, where g is of degree $n - 1$. We have $f(a) = (a - \lambda)g(a) = 0$ and by the minimality of $n, g(a) \neq 0$, hence $a - \lambda = 0$, so $a = \lambda \in F$ and this shows that $D = F$. ∎

Let A be a k-algebra, E be an extension field of k and consider the algebra A_E obtained by extending the ground field. If $u_1, \ldots, u_n$ is a k-basis of A, then this is also an E-basis of A_E, as we saw earlier. Thus we have

$$[A_E : E] = [A : k].$$

For a central simple algebra we have more precise information about the dimension. We shall need the fact, proved in Section 7.3, that any field can be extended to an algebraically closed field (Theorem 7.3.4 below).

Theorem 5.4.6. *The dimension of any central simple k-algebra, if finite, is a perfect square.*

Proof. Let E be an algebraically closed field containing k and form $A_E = A \otimes E$. By Theorem 5.2.2, $A_E = D_n$, where D is a skew field; we have $[A_E : E] = n^2[D : E]$, hence D is finite-dimensional over E, and by Proposition 5.4.5, $D = E$. It follows that $[A : k] = [A_E : E] = n^2$, as claimed. ∎

To conclude this section we shall show that the collection of central simple k-algebras is closed under tensor products. This will lead to the notion of a Brauer

group, which will be required briefly in Chapter 8. We shall need a lemma on vector spaces.

Lemma 5.4.7. *Let D be a skew field with an automorphism α and let V be a right D-space with basis $u_1, \ldots, u_n$. Define an action of α on V by setting $(\sum u_i\lambda_i)\alpha = \sum u_i(\lambda_i\alpha)$, so that the u_i are fixed under α. Then any subspace of V admitting α has a basis left fixed by α.*

Proof. Suppose the subspace W of V admits α and is r-dimensional. The quotient space V/W is spanned by the residue classes of $u_1, \ldots, u_n$, so we can choose a basis from them, which by renumbering may be taken as $u_{r+1}, \ldots, u_n$. Modulo W we can express the remaining u's in terms of them:

$$u_i \equiv \sum_{r+1}^{n} u_k\gamma_{ki} \pmod{W}, \quad i = 1, \ldots, r.$$

Put $z_i = u_i - \sum u_k\gamma_{ki}$ $(i = 1, \ldots, r)$; then $z_1, \ldots, z_r$ form a basis of W. They are linearly independent, for if $\sum z_i\lambda_i = 0$, then $\sum u_i\lambda_i - \sum u_k\gamma_{ki}\lambda_i = 0$, hence $\lambda_1 = \ldots = \lambda_r = 0$, and so the z's form a basis, because W is r-dimensional. We now apply α and obtain

$$z_i\alpha = u_i - \sum u_k(\gamma_{ki}\alpha). \tag{5.4.11}$$

Since W admits α, $z_i\alpha$ is a linear combination of the z_j, so we have

$$z_i\alpha = \sum z_j\lambda_{ji} = \sum u_j\lambda_{ji} - \sum u_k\gamma_{kj}\lambda_{ji}. \tag{5.4.12}$$

Equating (5.4.11) and (5.4.12), we obtain

$$u_i - \sum u_j\lambda_{ji} - \sum u_k\Big(\gamma_{ki}\alpha - \sum \gamma_{kj}\lambda_{ji}\Big) = 0.$$

By the linear independence of the u's, all coefficients must vanish; in particular, $\lambda_{ji} = \delta_{ji} (i = 1, \ldots, r)$, hence by (5.4.12), $z_i\alpha = z_i$, so the z's form the required basis of W. ∎

We can now achieve our aim announced earlier:

Theorem 5.4.8. *Let A, B be central simple algebras over a field k. Then the tensor product $A \otimes B$ is again a central simple k-algebra.*

Proof. By Theorem 5.2.2 and (5.4.5), B has the form $D_m \cong k_m \otimes D$, where D is a central division algebra; hence $A \otimes B \cong A \otimes k_m \otimes D \cong A_m \otimes D = P$, say. Now P is a D-module under the action $x \mapsto a^{-1}xa$, for $a \in D^\times$, and a k-basis of A will be a basis of P as D-module and fixed under the action of D. By Corollary 5.4.4, P has centre k and we have to show that P is simple. Let $\mathfrak{a}$ be a non-zero ideal of P; then $\mathfrak{a}$ admits the action of D, therefore by Lemma 5.4.7 it has a basis fixed under the action of D, i.e. centralizing D. But the centralizer of D in P is A_m, so $\mathfrak{a}$ has a basis in A_m; since A_m is simple, $\mathfrak{a} \supseteq A_m$, so $\mathfrak{a}$ contains 1 and hence equals P, which is therefore simple, as claimed. ∎

For any central simple k-algebra of the form $A = k_n \otimes D$ (D a division algebra) let us call D the *division algebra component* or *skew field component* of A; two central simple algebras are said to be *similar* if they have isomorphic division algebra components. Given a tensor product of central simple algebras, $P = A \otimes B$, the similarity class of P clearly depends only on the similarity classes of A and B, not on A, B themselves. These similarity classes form a monoid, by Theorem 5.4.8 and the associativity of the tensor product. In fact this monoid is a group, with the class of A^O, the opposite algebra of A, as inverse of the class of A. To see this we shall show that $A^O \otimes A \cong k_n$, where $n = \dim A$. We can interpret k_n as the ring of linear transformations of A, and we have a linear mapping $A^O \otimes A \to k_n$ in which $\sum a_i \otimes b_i$ corresponds to the mapping $x \mapsto \sum a_i x b_i$. This is easily verified to be a homomorphism; the kernel is 0, since $A^O \otimes A$ is simple, by Theorem 5.4.8. Now a comparison of dimensions shows that it must be an isomorphism, hence $A^O \otimes A \cong k_n$, as claimed, and it follows that the similarity class of A^O is the inverse to the class of A. We thus obtain

Theorem 5.4.9. *For any field k, the similarity classes of central simple k-algebras form a group.* ∎

This group, first introduced by Richard Brauer in 1928, is called the *Brauer group* of k and is denoted by $\mathbf{B}_k$. A closer study of this group will be made in FA.

Exercises

1. Show that if A, B are finitely generated K-algebras, then so is $A \otimes B$.
2. Let A be a K-algebra and x be an indeterminate over A. Show that $A[x] \cong A \otimes K[x]$.
3. Let $\mathbf{Z}[i]$ be the ring of Gaussian integers (obtained by adjoining a root i of $x^2 + 1 = 0$ to $\mathbf{Z}$). Show that $\mathbf{R} \otimes_{\mathbf{Z}} \mathbf{Z}[i] \cong \mathbf{C}$.
4. Let B be an algebra with unit element 1. Verify that the map from A to $A \otimes B$ defined by $x \mapsto x \otimes 1$ is a homomorphism.
5. Let $A = \mathbf{Z}/(2)$ as $\mathbf{Z}$-algebra and $B = \mathbf{Z}$. Verify that the natural homomorphism $B \to A \otimes B$ is not an embedding.
6. Fill in the details in the proof of Theorem 5.4.1.
7. Let A, B be algebras with unit elements $\neq 0$ over a field, but not assumed to be associative. Show that if $A \otimes B$ is associative, then so are A and B.
8. Let A be a k-algebra, where k is a field, and let E be an extension field of k. Show that $A \otimes E$ may be defined by taking the multiplication table of A in terms of a basis over k and regarding it as a multiplication table over E.
9. Let A, B be K-algebras. Given a right A-module U and a left $(A \otimes B)$-module V, show that $(U \otimes_K B) \otimes_{A \otimes B} V \cong U \otimes_A V$. (Hint. Use (4.8.7) and the associative law.)

5.5 The Regular Representation; Norm and Trace

Let A be a finite-dimensional algebra over a field k. We have seen in Section 5.1 that A has a faithful matrix representation over k, the regular representation ρ. If $u_1, \ldots, u_n$ is a basis of A, then the regular representation $a \mapsto \rho(a) = (\rho_{ij}(a))$ is given by the equations

$$u_i a = \sum \rho_{ij}(a) u_j.$$

The matrix of the regular representation is still dependent on the choice of basis in A, but we get an invariant in A by taking the characteristic polynomial

$$\det(xI - \rho(a)) = x^n + \lambda_1 x^{n-1} + \ldots + \lambda_n. \tag{5.5.1}$$

Here the second and last coefficients are just the trace and the determinant of $\rho(a)$, apart from sign. We define the *trace* of a, $T_{A/k}(a)$, and the *norm* of a, $N_{A/k}(a)$, as

$$T_{A/k}(a) = -\lambda_1 = \mathrm{tr}(\rho(a)), \quad N_{A/k}(a) = (-1)^n \lambda_n = \det(\rho(a)).$$

More briefly we often write $T(a), N(a)$ when no confusion is possible. The properties of the trace and determinant of a matrix lead immediately to the formulae

$$T(a+b) = T(a) + T(b), T(\lambda a) = \lambda T(a),$$

$$T(ab) = T(ba), T(1) = n, \quad a, b \in A, \lambda \in k. \tag{5.5.2}$$

$$N(ab) = N(a)N(b), N(\lambda a) = \lambda^n N(a), N(1) = 1.$$

As a first application we obtain a criterion for zerodivisors in A in terms of the norm.

Proposition 5.5.1. *Let A be a finite-dimensional algebra over a field k and let $c \in A$. Then the following conditions are equivalent:*

(a) *c is a non-zerodivisor in A,*
(b) *c is a unit in A,*
(c) $N(c) \neq 0$.

In particular, every finite-dimensional algebra without zerodivisors is a division algebra.

Proof. (a) $\Rightarrow$ (b). If $[A : k] = n$, then the elements $1, c, c^2, \ldots, c^n$ are linearly dependent over k, so we have an equation

$$\gamma_0 c^n + \gamma_1 c^{n-1} + \ldots + \gamma_n = 0, \quad \gamma_i \in k, \text{ not all } 0.$$

Choose an equation of least degree r; then since c is a non-zerodivisor, $\gamma_r \neq 0$, and on dividing by it, we may assume that $\gamma_r = 1$. Then $-(\gamma_0 c^{r-1} + \ldots + \gamma_{r-1})$ is an inverse for c in A.

(b) $\Rightarrow$ (c). If c is a unit, then $1 = N(cc^{-1}) = N(c)N(c^{-1})$, hence $N(c) \neq 0$.

(c) $\Rightarrow$ (a). If $N(c) \neq 0$, then the matrix $\rho(c)$ of the regular representation is non-singular, hence the mapping $x \mapsto xc$ is injective, i.e. c is a non-zerodivisor. ∎

To establish the usual transitivity formulae we shall need a lemma. We assume that the reader has met the matrix reduction over an algebraically closed field.

Lemma 5.5.2. *Let F/k be a field extension of degree r and denote by ρ the regular representation of F in k; thus $\rho : F \to \mathfrak{M}_r(k)$ is a homomorphism. Denote by ρ_n the induced homomorphism $\mathfrak{M}_n(F) \to \mathfrak{M}_{nr}(k)$. Then for any $C \in \mathfrak{M}_n(F)$,*

$$\operatorname{tr}(\rho_n(C)) = \operatorname{tr}(\rho(\operatorname{tr}(C))), \tag{5.5.3}$$

$$\det(\rho_n(C)) = \det(\rho(\det(C))). \tag{5.5.4}$$

Proof. By definition $\rho(c)$ is an $r \times r$ matrix over k; its entries will be written $\rho_{\lambda\mu}(c)$. Writing $C = (c_{ij})$, we have on the left of (5.5.3), $\sum \rho_{\lambda\lambda}(c_{ii})$ and on the right $\sum \rho_{\lambda\lambda}(\sum c_{ii})$, and these two expressions are equal, by the linearity of the functions $\rho_{\lambda\lambda}$.

Next consider (5.5.4). Since both sides are unchanged on replacing C by $P^{-1}CP$, we may, on passing to an algebraic closure of F, replace C by a triangular matrix. Now $\rho_n(C)$ is a block triangular matrix, with $r \times r$ blocks $\rho(c_{ii})$ along the main diagonal and zeros below it; hence on taking a Laplace expansion by the last r rows and using induction on n, we obtain $\det \rho(c_{11}) . \det \rho(c_{22}) \ldots \det \rho(c_{nn})$ on the left of (5.5.4). On the right we have $\det \rho(c_{11}c_{22}\ldots c_{nn})$, and these results agree, because ρ is a homomorphism. Thus (5.5.4) holds over an algebraic closure of F, and hence over F itself. ∎

Let F/k be a field extension of degree r and A be an F-algebra of dimension n. Then we may regard A as a k-algebra of dimension rn (see Proposition 7.1.2 below). We can therefore define norm and trace for A as an F-algebra and as a k-algebra; they are related by the Transitivity Formulae:

$$T_{F/k}(T_{A/F}(c)) = T_{A/k}(c), \tag{5.5.5}$$

$$N_{F/k}(N_{A/F}(c)) = N_{A/k}(c). \tag{5.5.6}$$

Proof. If $\rho_{A/F}$ denotes the regular representation of A over F, regarded as a homomorphism $A \to \mathfrak{M}_n(F)$, then we clearly have

$$\rho_{A/k}(c) = \rho_{F/k}(\rho_{A/F}(c)). \tag{5.5.7}$$

In (5.5.3) let us take $C = \rho_{A/F}(c)$; the left-hand side can by (5.5.7) be written $\operatorname{tr}(\rho_{A/k}(c))$ and this is just $T_{A/k}(c)$. The right-hand side is $\operatorname{tr} \rho_{F/k}(T_{A/F}(c)) = T_{F/k}(T_{A/F}(c))$, and so (5.5.5) is established. Now (5.5.6) follows in the same way from (5.5.4), putting again $C = \rho_{A/F}(c)$. ∎

As an example let us calculate the trace and norm in a field extension. Let E/k be a field extension of degree r and take $c \in E$ of degree s over k. The minimal equation for c has the form

$$x^s + \lambda_1 x^{s-1} + \ldots + \lambda_s = 0. \tag{5.5.8}$$

The extension $k(c)/k$ is of degree s, and the matrix $\rho(c)$ of c in the regular representation satisfies (5.5.8), hence a comparison of degrees shows that the minimal polynomial is also the characteristic polynomial of $\rho(c)$. More explicitly we can see this by taking the basis $1, c, \ldots, c^{s-1}$ of $k(c)/k$. In terms of this basis we have

$$c^i \,.\, c = \begin{cases} c^{i+1} & \text{if } i < s-1, \\ -\lambda_1 c^{s-1} - \ldots - \lambda_s & \text{if } i = s-1. \end{cases}$$

Hence the matrix for c is

$$\begin{pmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -\lambda_s & -\lambda_{s-1} & -\lambda_{s-2} & \ldots & -\lambda_2 & -\lambda_1 \end{pmatrix}.$$

This is just the companion matrix of the minimal polynomial. Its trace is $-\lambda_1$ and its determinant is $(-1)^s\lambda_s$. Since $[E : k(c)] = r/s$, we have, by transitivity,

$$T_{E/k}(c) = \frac{r}{s}(-\lambda_1), \quad N_{E/k}(c) = [(-1)^s\lambda_s]^{r/s}.$$

These formulae show that for a separable field extension the definitions of norm and trace given here coincide with those given in Section 7.9 below.

More generally, given any representation of a k-algebra, $\sigma : A \to \mathfrak{M}_r(k)$, we can define its trace and norm as

$$\mathrm{Tr}_\sigma(x) = \mathrm{Tr}(\sigma(x)), \quad \mathrm{Nm}_\sigma(x) = \det(\sigma(x)),$$

and they will again satisfy laws corresponding to (5.5.2). The trace defines a quadratic form $\mathrm{Tr}_\sigma(x^2)$ on A, which is non-singular iff

$$\det(\mathrm{Tr}_\sigma(u_i u_j)) \neq 0, \tag{5.5.9}$$

for a basis $u_1, \ldots, u_n$ of A. Such a representation will sometimes allow us to recognize when the algebra is semisimple.

Proposition 5.5.3. *Let σ be a representation of an algebra A over a field k. If the quadratic form Tr_σ is non-singular, then A is semisimple.*

Proof. We first note that if an element c is nilpotent, then so is $\sigma(c)$, hence its trace is then zero, thus $\mathrm{Tr}_\sigma(c) = 0$. Suppose that A is not semisimple; then its radical is non-zero and if a basis $v_1, \ldots, v_n$ of A is chosen so that $v_1, \ldots, v_r$ is a basis of $\mathbf{J}(A)$ $(r \geq 1)$, then $v_1 v_i$ is nilpotent for all i; hence the first row of the matrix $\mathrm{Tr}_\sigma(v_1 v_i)$ is zero, and so the form Tr_σ is singular. ∎

This sufficient condition for semisimplicity is also necessary for a faithful representation in characteristic 0 (see Exercise 4).

In any commutative algebra the sum and product of two nilpotent elements are again nilpotent, but neither of these assertions holds in the general case. In these circumstances it is somewhat surprising that the nilpotence of an algebra follows from the existence of a nilpotent basis. This is the content of the next result, proved by Wedderburn in 1937. Note that, although the proof depends on a trace argument, it holds in any characteristic.

Theorem 5.5.4. *Let A be a finite-dimensional algebra over a field k and B be a subalgebra with a basis consisting of nilpotent elements. Then B is nilpotent.*

Proof. We remark that the conclusion shows that B cannot contain 1, but this is not obvious at the outset. Let C be the subalgebra of A generated by B and 1; thus if $u_1, \ldots, u_n$ is a nilpotent basis of B, then C is spanned by $1, u_1, \ldots, u_n$. To prove that B is nilpotent it will be enough to show that B is contained in every maximal ideal of C, for then $B \subseteq \mathbf{J}(C)$ and the conclusion follows by the nilpotence of $\mathbf{J}(C)$ (Theorem 5.3.5). Suppose then that there is a maximal ideal $\mathfrak{m}$ of C not containing B. Then $\mathfrak{m} + B = C$ and $C/\mathfrak{m} \cong B/(B \cap \mathfrak{m})$. If $\overline{C} = C/\mathfrak{m}$ and $\bar{x}$ denotes the residue class of x (mod $\mathfrak{m}$), then the simple algebra $\overline{C}$ is spanned by $\bar{u}_1, \ldots, \bar{u}_n$ and for suitable renumbering, $\bar{u}_1, \ldots, \bar{u}_r$ is a basis. Each $\bar{u}_i$ is again nilpotent, so we have a simple algebra with a nilpotent basis; we shall show that this leads to a contradiction. Let E be an algebraically closed field containing k; then $\overline{C}_E$ is a simple E-algebra, hence isomorphic to $\mathfrak{M}_t(E)$ for some t. Let Tr be the trace function on the matrix ring $\mathfrak{M}_t(E)$; for any nilpotent element u, $\mathrm{Tr}(u) = 0$, and since there is a basis of nilpotent elements, $\mathrm{Tr}(x) = 0$ for all $x \in \mathfrak{M}_t(E)$, by linearity. But $\mathrm{Tr}(e_{11}) = 1 \neq 0$, so we have reached a contradiction. Hence B is nilpotent. ∎

We remark that the regular representation on $M_t(E)$ would give $\mathrm{Tr}(e_{11}) = t$, and so cannot be used in finite characteristic. In fact the trace function used here is the 'reduced trace' which we shall meet again in FA.

Exercises

1. Show that the norm and trace of the full matrix algebra $\mathfrak{M}_n(k)$ are given by $N(A) = (\det A)^n$, $T(A) = n.\mathrm{tr}(A)$. Find the norm and trace of the algebra $\mathfrak{T}_n(k)$ of upper triangular matrices.
2. Find the norm and trace for the group algebra of $\mathbf{C}_n$, the cyclic group of order n. Likewise for the Klein 4-group $V = \mathrm{gp}\{a, b | a^2 = b^2 = (ab)^2 = 1\}$.
3. Show that in characteristic 0 a matrix A is nilpotent iff $\mathrm{tr}(A^r) = 0$ for $r = 1, 2, \ldots$.
4. Let σ be a faithful representation of a k-algebra A, where char $k = 0$ (e.g. the regular representation). Show that the set $\mathfrak{n} = \{x \in A | \mathrm{Tr}_\sigma(xa) = 0 \text{ for all } a \in A\}$ is a nilpotent ideal of A. Deduce that if A is semisimple, then Tr_σ is non-singular.
5. Show that a field extension F/k of finite degree is separable (see Section 7.4 below) iff there is an element $c \in F$ such that $T(c) \neq 0$.
6. A field extension E/k is *purely inseparable* if char $k = p$ is prime and for each $x \in E$ there exists $q = p^r$ such that $x^q \in k$. Show that for a purely inseparable field extension of degree $r > 1$, $N(c) = c^r$, $T(c) = 0$.

5.6 Möbius Functions

With each partially ordered set we can associate an algebra, its incidence algebra, which turns out to be useful in answering enumeration questions. The following brief account is based on the elegant treatment by Gian-Carlo Rota [1964].

Let P be a partially ordered set and consider the collection $\mathscr{I}_P$ of square matrices with entries in $\mathbf{R}$ (or any given commutative integral domain), whose rows and columns are indexed by P:

$$A = (a_{ij}), \quad \text{where } a_{ij} = 0 \text{ unless } i \leq j \quad (i, j \in P). \tag{5.6.1}$$

For example if $P = \{a, b, c\}$ with $a < c$, $b < c$, and the matrices of $\mathscr{I}_P$ are indexed by P in the order a, b, c, then $\mathscr{I}_P$ consists of all upper triangular 3×3 matrices (5.6.1) with $a_{12} = 0$. For any finite P an easy induction shows that the elements of P can be arranged as a sequence so that i precedes j whenever $i \leq j$, but this is not essential for our purpose (it means in effect that A in (5.6.1) can be written as an upper triangular matrix).

If P is finite, so that the matrices in $\mathscr{I}_P$ have finitely many rows and columns, $\mathscr{I}_P$ is closed under the usual addition and multiplication of matrices and so forms a linear algebra over $\mathbf{R}$, called the *incidence algebra* of P. Thus let $A = (a_{ij})$, $B = (b_{ij})$, $C = (c_{ij}) \in \mathscr{I}_P$; if $C = AB$, then

$$c_{ik} = \sum_j a_{ij} b_{jk}, \tag{5.6.2}$$

and here it is enough to confine the summation to indices j such that $i \leq j \leq k$, by (5.6.1). For example, for a totally ordered set P, $\mathscr{I}_P$ is just the set of all upper triangular matrices. We can also allow P to be infinite, provided that each interval $[i, k]$ is finite, for then the summation (5.6.2) contains only finitely many non-zero terms, for any pair i, k. A partially ordered set in which all intervals are finite will be called *locally finite*. We begin by noting the conditions for a matrix in $\mathscr{I}_P$ to be invertible.

Proposition 5.6.1. *Let P be a locally finite partially ordered set and $\mathscr{I}_P$ be its incidence algebra. Then $A \in \mathscr{I}_P$ is invertible if and only if all its diagonal elements are invertible.*

Proof. Suppose that $A = (a_{ij})$ has the inverse $A^{-1} = (a'_{ij})$. Then by (5.6.2),

$$a'_{ii} a_{ii} = 1; \tag{5.6.3}$$

hence the condition is necessary. When it holds, we can define a'_{ii} by (5.6.3). Given $i, k \in P$, $i < k$, assume that a'_{ij} has already been defined for all j such that $i \leq j < k$; then we can determine a'_{ik} uniquely from

$$\sum a'_{ij} a_{jk} = 0, \tag{5.6.4}$$

for the only unknown term in (5.6.4) is a'_{ik} and it occurs with the coefficient a_{kk}, which is invertible by hypothesis. In this way we can determine a'_{ik} for all $i \leq k$,

and together with the equation $a'_{ik} = 0$ when $i \leq k$ does not hold this defines $A' = (a'_{ij})$ uniquely to satisfy $A'A = I$. Hence A' is the required inverse. ■

We note that when $A = (a_{ij})$ is given and $A^{-1} = (a'_{ij})$, then a'_{hk} depends only on the values of a_{ij} for i, j in the interval $[h, k]$. As an example of Proposition 5.6.1 consider the zeta-matrix $Z = (z_{ij})$ defined by

$$z_{ij} = \begin{cases} 1 & \text{if } i \leq j, \\ 0 & \text{otherwise.} \end{cases}$$

By Proposition 5.6.1, Z has an inverse Z^{-1}, usually denoted by M and called the *Möbius matrix*. Its importance stems from its use in the inversion formula:

Theorem 5.6.2 (Möbius inversion formula). *Let P be a locally finite partially ordered set with a greatest element ω. Given any function $f(i)$ on P, define g by the equation*

$$g(i) = \sum_{j \geq i} f(j), \tag{5.6.5}$$

Then on writing $Z^{-1} = (m_{ij})$, we have

$$f(i) = \sum m_{ij} g(j) = \sum g(h) m_{hi}. \tag{5.6.6}$$

Proof. By putting $f = (f(i)), g = (g(i))$, regarded as column vectors, we can write (5.6.5) as $g = Zf$, where the summation is over the interval $[i, \omega]$. It follows that $f = Z^{-1}g = Mg$, and this is the first equation (5.6.6). The second equation follows similarly, by regarding f, g as row vectors. ■

As an illustration consider the set $\mathbf{N}$ of natural numbers, partially ordered by divisibility; this is infinite, but locally finite, with greatest element 1, if we set $i \leq j$ whenever $j|i$. The Z-matrix is $z_{ij} = 1$ if $j|i$ and 0 otherwise. Since Z only depends on the fraction i/j, we shall write $z_{ij} = \zeta(i/j)$; then $\zeta(n) = 1$ for all $n \in \mathbf{N}$ and $\zeta(\alpha) = 0$ for $\alpha \in \mathbf{Q} \backslash \mathbf{Z}$. The inverse $M = (m_{ij})$ can again be written $\mu(i/j)$, where $\mu(n)$, called the *Möbius function*, is given by

$$\mu(n) = \begin{cases} (-1)^r & \text{if } n \text{ is the product of } r \text{ distinct primes,} \\ 0 & \text{if } n \text{ is divisible by the square of a prime.} \end{cases} \tag{5.6.7}$$

To establish (5.6.7), we note the formula $\sum z_{ij} m_{jk} = \delta_{ik}$; translated to ζ and μ, this becomes

$$\sum_{rs=n} \zeta(r)\mu(s) = \delta_{n1}.$$

Hence we obtain

$$\mu(1) = 1, \quad \sum_{d|n} \mu(d) = 0 \quad \text{for } n > 1.$$

Given $n \geq 1$, let us single out a prime p dividing n and write $n = mp^t$, where m is prime to p. Then

$$0 = \sum_{d|n} \mu(d) = \sum_{c|m} [\mu(c) + \mu(cp) + \ldots + \mu(cp^t)],$$

and this will be satisfied if for each factor c of m the sum shown vanishes. Taking $t = 1$, we find $\mu(cp) = -\mu(c)$, and for $t > 1$, $\mu(cp^t) = 0$, and this leads to (5.6.7), by an induction on n.

In this case Theorem 5.6.2 reduces to the classical

Möbius Inversion Formula. *Let f be any function on* **N**. *If g is defined by the equation*

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

then f is given in terms of g by the formula

$$f(n) = \sum_{d|n} g(d)\mu(n/d) \quad \textit{or also} \quad f(n) = \sum_{d|n} \mu(d)g(n/d).$$

The simplest partially ordered set with more than one element is the 2-element set **2**. Its incidence algebra consists of all upper triangular 2×2 matrices; in particular,

$$Z = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad Z^{-1} = M = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

We can regard **2** as $\mathscr{P}(U)$, where U is a 1-element set. More generally, let S be any finite set and consider $\mathscr{P}(S)$. We assert that for $X \subseteq Y \subseteq S$,

$$m_{XY} = (-1)^{|Y|-|X|}. \tag{5.6.8}$$

For if we determine m_{XY} by (5.6.4), we have $m_{XY} = -\sum m_{XZ}$, where the summation is over all Z such that $X \subseteq Z \subset Y$. Put $|Y| - |X| = r$ and assume the result for values less than r. There are $\binom{r}{i}$ subsets Z such that $X \subseteq Z \subset Y$ and $|Z| = |X| = i$; hence by the induction hypothesis,

$$m_{XY} = -1 + \binom{r}{1} - \binom{r}{2} + \ldots + (-1)^r \binom{r}{r-1} = (-1)^r - (1-1)^r = (-1)^r,$$

and so (5.6.8) follows.

The Möbius inversion formula of Theorem 5.6.2 can be applied to $\mathscr{P}(S)$ as follows. Let f_X be any **Z**-valued function on $\mathscr{P}(S)$ and put

$$g_X = \sum_{Y \supseteq X} f_Y.$$

Then by (5.6.6) and (5.6.8) we have

$$f_X = \sum_{Y \supseteq X} (-1)^{|Y|-|X|} g_Y. \tag{5.6.9}$$

Let us put

$$G_r = \sum_{|Z|=r} g_Z. \tag{5.6.10}$$

Then on taking $X = \varnothing$ in (5.6.9) we find, if $|S| = n$,

$$f_\varnothing = G_0 - G_1 + G_2 - \ldots + (-1)^n G_n. \tag{5.6.11}$$

This is known as the *sieve principle* or the *principle of inclusion-exclusion.* If we think of S as a number of properties which the objects of some collection may or may not possess, and f_X is the number of objects having precisely the properties in X, then g_X is the number of objects having at least all the properties in X (and possibly others). Now (5.6.11) states that to find the number of objects having none of the given properties we take the number of objects, subtract, for each property, the number of objects having that property; then for each pair of properties, add the objects having both these properties (because they will have been subtracted twice); for each triple of properties, subtract. . . . To give an example, of two dozen suitors arriving at the Royal Court, 12 were short, 11 were fair, 14 were plain, 3 were short and fair, 7 were short and plain, 4 were fair and plain, and none were all three. By (5.6.10) we see that of the 24 suitors, $24 - (12 + 11 + 14) + (3 + 7 + 4) = 1$ was tall, dark and handsome.

Exercises

1. An interviewer questioned 47 people and reported: 22 were male, 18 were married, 19 were retired, 5 were male and married, 4 retired and married and 2 male and retired. How would you test these data for their consistency?
2. Show that the Möbius matrix for $\{1, 2, \ldots, n\}$ with the natural order is $I - N$, where $N = (n_{ij})$, $n_{ij} = \delta_{ji+1}$.
3. Find the Möbius matrices of the partially ordered sets in Figure 5.1.

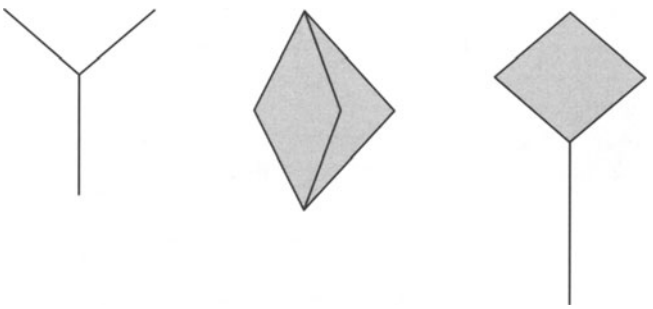


Figure 5.1

4. Let S_1, S_2 be two partially ordered sets and $S = S_1 \times S_2$ with the ordering $(x_1, x_2) \leq (y_1, y_2)$ iff $x_i \leq y_i$ $(i = 1, 2)$. Show that if μ_i is the Möbius function for S_i, then the Möbius function μ for S is given by $\mu(i_1, i_2; j_1, j_2) = \mu_1(i_1, j_1)\mu_2(i_2, j_2)$. Deduce another proof of (5.6.8).
5. A real-valued function f on $\mathbf{N}$ is called *multiplicative* if $f(ab = f(a)f(b)$ whenever a, b are coprime. Prove from the definition that the Möbius function on M is multiplicative, and hence derive (5.6.7).
6. For any real-valued multiplicative function f on $\mathbf{N}$ show that

$$\sum_{d|n} \mu(d)f(d) = \prod_{p|n} [1 - f(p)],$$

where p runs over all primes dividing n.

Show that the Euler function $\varphi(n)$ indicating the number of positive integers less than and prime to n satisfies $\sum_{d|n} \varphi(d) = n$, and is multiplicative. Deduce that $\varphi(n) = n\prod(1 - p^{-1})$, where the product is taken over all primes p dividing n. Show further that

$$\varphi(n) = n\prod_{d|n} \mu(d)\frac{n}{d}.$$

7. Prove that the elements of a finite partially ordered set P can be arranged as a sequence $a_1, a_2, \ldots, a_n$ such that $a_i < a_j$ implies $i < j$, and show that the number of ways of doing it is equal to the number of maximal chains in P^*, the set of lower segments in P.
8. Show that the inverse of the ζ-function $\zeta(s) = \sum n^{-s}$ is $\zeta(s)^{-1} = \sum \mu(n)n^{-s}$, where $\mu(n)$ is the Möbius function.
9. Show that for any finite partially ordered set, $Z = I + N$, where N is a nilpotent matrix. Hence obtain the formula $M = I - N + N^2 - \ldots$ for the Möbius matrix. What happens in the case of an infinite (but locally finite) set?
10. Show that for any locally finite partially ordered set the number of chains from i to j is the (i, j)-entry of $(2 - Z)^{-1}$.

Further Exercises for Chapter 5

1. Let A be an n-dimensional algebra (without 1) over a field k. Show that A has a faithful representation in $\mathfrak{M}_{n+1}(k)$ and give an example where $n + 1$ cannot be replaced by n. (Hint. Take A to be nilpotent.)
2. Find all ideals of $\mathfrak{M}_n(\mathbf{Z})$, for varying n.
3. Let K be any commutative ring and $u \in K^n$, $v \in {}^nK$. Show that the set of all $n \times n$ matrices A such that $uA = 0 = Av$ is a subalgebra of $\mathfrak{M}_n(K)$ (possibly without 1). Deduce that the set of 'semimagic squares', i.e. matrices whose row sums and column sums are 0, is an algebra. Is this true of the set of all 'magic squares' (in which the sums along the two diagonals are also 0)?
4. Show that a finite commutative local ring with n units has at most $(n + 1)^2$ elements.

5. Show that a ring with p^3 elements (p prime) is either commutative or isomorphic to $\mathfrak{T}_2(\mathbf{F}_p)$.
6. Show that if an element a in a ring R has two right inverses a', a'', then $a'' + a'a - 1$ is another right inverse. Deduce that any element with more than one right inverse has infinitely many.
7. Show that a ring R such that $nR = 0$ for a square-free integer n is a direct product of algebras over fields.
8. Let A be a finite-dimensional k-algebra, where k is a field of prime characteristic p, and define a trace function on A as a linear function $f : A \to k$ such that $f(a^p) = f(a)^p$. Show that any trace function vanishes on the radical.
9. Let A be an algebra satisfying $x^3 = 0$ ($x \in A$). Show that $x^2yx^2 = xyx^2yx = 0$ ($x, y \in A$).
10. Let A be a commutative algebra (not necessarily with 1) and define its *duplicate* as $A' = (A \otimes A)/D$, where $D = \{x \otimes y - y \otimes x | x, y \in A\}$. Show that $A' = A^2 \oplus C$, where C is isomorphic to the annihilator of A; in particular, for algebras with $1, A' \cong A$. What happens if this definition is applied to a non-commutative algebra?
11. Show that the tensor product of augmented algebras is again augmented.
12. Show that an algebra (possibly without 1) is augmented iff it has an ideal of codimension 1 not containing A^2.
13. A vector $a = (a_1, \ldots, a_n) \in \mathbf{R}^n$ is called *stochastic* if $0 \leq a_i \leq 1$ and $\varepsilon(a) = \sum a_i = 1$. A matrix is *stochastic* if its rows are stochastic and an algebra A is *stochastic* if it has a basis (called a *natural* basis) u_i such that in the regular representation $\rho(u_i)$ is stochastic. Show that a basis is natural iff it is obtained from a natural basis by transformation by a stochastic matrix. Show that an algebra is stochastic iff it is augmented and the convex hull of the elements satisfying $\varepsilon(x) = 1$ is a simplex closed under multiplication.
14. Show that an algebra A is semisimple provided that the trace with respect to some representation σ satisfies $\det(\mathrm{Tr}_\sigma(u_i v_j)) \neq 0$, where $\{u_i\}$, $\{v_j\}$ are any two bases of A. By taking as bases for the group algebra kG of a finite group G the elements $u_1, u_2, \ldots$ of G and $u_1^{-1}, u_2^{-1}, \ldots$, show that kG is semisimple whenever char k does not divide the order of G. (Maschke's theorem, see FA, **6.2.**)
15. Let P be a finite partially ordered set with Möbius matrix M. Show that the Möbius matrix of the set with the opposite ordering (and the same indexing of rows and columns) is M^T, the transpose of M.
16. (L. Weisner) Let L be a finite lattice and $b > 0$ in L. Show that the Möbius matrix $M = (m_{xy})$ of L satisfies, for any $a \in L$,

$$\sum_{x \vee b = a} m_{0x} = 0.$$

17. Let S be a set of n elements and for $X \subseteq S$ let $f(X)$ be the number of permutations of S whose set of fixed points is precisely X. Show that the number of permutations fixing every point of X is $\sum f(Y)$, where the sum is over all

$Y \supseteq X$. Deduce that the number of derangements of S, i.e. permutations without fixed point, is

$$n! \sum_{\nu=0}^{n} (-1)^{\nu}/\nu!.$$

(Thus the probability that a permutation has no fixed point is, for large n, close to e^{-1}.)

18. Let G be a finite abelian group and consider the Möbius matrix $M = (m_{AB})$ for $\mathrm{Lat}(G)$. Show that m_{AB} vanishes unless $A \subseteq B$, and in that case it depends only on the quotient group B/A. Writing $m_{AB} = \mu(B/A)$, show that if $G = G_1 \times \ldots \times G_r$ is the representation of G as the direct product of its primary components, then $\mu(G) = \mu(G_1) \ldots \mu(G_r)$, and for a p-group C, $\mu(C) = (-1)^k p^{k(k-1)/2}$ or 0 according as C is elementary abelian, of order p^k say, or not.

6 Multilinear Algebra

Polynomial rings form a simple example of a graded algebra; such algebras occur frequently and in Section 6.1 we define this concept. Another important example is given by free algebras, which are discussed in Section 6.2, as well as the related notions of tensor algebra and symmetric algebra on a K-module. A graded algebra has an important invariant, its Hilbert series, essentially a power series whose coefficients indicate the dimensions of the components. In Section 6.3 we show that the Hilbert series of a commutative Noetherian ring is a rational function and also prove the Golod–Shafarevich theorem, giving a sufficient condition for a graded algebra to be infinite-dimensional. The applications, to construct a finitely generated algebra which is nil but not nilpotent, and a finitely generated infinite p-group, are sketched in the exercises. Finally Section 6.4 deals with exterior algebras, providing a simple derivation of determinants, and giving a brief geometrical application.

6.1 Graded Algebras

In a polynomial ring the *degree* of a polynomial is a useful concept which allows us to analyse the ring in different ways. Thus we may write the polynomial ring as a direct sum of its homogeneous components. There are many other rings which share this property, and it is convenient to describe its general form before applying it. Throughout, the coefficient ring K is an arbitrary commutative ring. The case when there is no coefficient ring is included by taking $K = \mathbf{Z}$, since every ring is a $\mathbf{Z}$-algebra. To avoid trivialities, we shall usually assume that K is non-trivial.

Definition. *A graded K-algebra* is a family (A_n) of K-modules indexed by $\mathbf{Z}$, with a bilinear mapping

$$A_m \times A_n \to A_{m+n}, \tag{6.1.1}$$

denoted by $(a, b) \mapsto ab$, such that for any $a \in A_m$, $b \in A_n$, $c \in A_p$,

$$a(bc) = (ab)c \tag{6.1.2}$$

and there is an element $e \in A_0$ such that $ae = ea = a$ for all $a \in A_n$ ($n \in \mathbf{Z}$). A_n is called the *component* of degree n; if $a \in A_n$, we say that a is *homogeneous* of degree n and write $\deg a = n$.

We shall assume that distinct A_n have only 0 in common, so that every non-zero element has a unique degree, while 0 has all degrees. From the definition it is clear that A_0 is a K-algebra and each A_n is an A_0-bimodule.

Given a graded K-algebra (A_n), let us put $A = \oplus_n A_n$; we can define a multiplication on A, using (6.1.1) and the distributive laws. Thus if $a = a_m + a_{m+1} + \ldots + a_M$ and $b = b_n + b_{n+1} + \ldots + b_N$, where $a_i, b_i \in A_i$, then

$$ab = a_m b_n + a_m b_{n+1} + a_{m+1} b_n + \ldots + a_M b_N.$$

It is clear that multiplication is associative, by (6.1.2), and distributive, with e as unit element; thus A is a K-algebra. Sometimes A is called an internally graded algebra, corresponding to the externally graded algebra (A_n). Of course we may be given an algebra which is internally graded, i.e. A may have the form $A = \oplus_n A_n$, where (A_n) is externally graded.

Example 1. Let $A = K[x_1, \ldots, x_d]$ be a polynomial ring in d indeterminates, and denote by A_n the submodule of all homogeneous polynomials of total degree n; then A is internally graded by the A_n. We observe that in this case $A_n = 0$ for $n < 0$; this is expressed by saying that A is *positively* graded. In particular, for $d = 1$ we have $A = K[x]$; here $A_n = Kx^n$.

Example 2. The ring of 'Laurent polynomials' $k[t, t^{-1}]$ is an example of a graded ring with non-zero components for all $n \in \mathbf{Z}$. By contrast, the ring of all Laurent series has no obvious grading (it has a filtration, see Further Exercise 4).

Example 3. Any ring R may be graded by putting $A_0 = R$, $A_n = 0$ for $n \neq 0$. Such a ring is said to be *concentrated in degree* 0.

Let A be a graded ring. A *graded A-module* is a family of K-modules (M_n) such that $A_r M_s \subseteq M_{r+s}$ for all $r, s \in \mathbf{Z}$. We can again form an internally graded module by writing $M = \oplus M_n$; an element x of M is called *homogeneous* of *degree* n if $x \in M_n$. Every element of M can be uniquely expressed as a sum of finitely many homogeneous elements. We remark that for any graded ring A, a graded A-module M which is finitely generated can always be generated by a finite set of homogeneous elements; we need merely take all the components of the given set of generators.

The tensor product (over K) of two graded K-algebras is again a graded K-algebra in an obvious sense: given $A = (A_n)$, $B = (B_n)$, we have $A \otimes B = C$, where

$$C_n = \oplus_r A_r \otimes B_{n-r}. \tag{6.1.3}$$

For example, the polynomial ring $K[x, y]$ may be constructed as the tensor product of $K[x]$ and $K[y]$. We observe that when A and B are both positively graded, the sum in (6.1.3) is necessarily finite, but this is not essential for the definition to make sense.

A graded algebra A is said to be *anticommutative* if

$$a_m b_n = (-1)^{mn} b_n a_m \quad \text{for } a_m \in A_m, b_n \in A_n. \tag{6.1.4}$$

Thus to interchange a and b we change the sign if m and n are both odd and leave it the same otherwise. For example, if $A_n = 0$ for all odd n, then (6.1.4) reduces to the commutative law. Some writers omit the 'anti' and call (6.1.4) the commutative law for graded algebras; although there is some justification for this practice, we shall not follow it but refer to (6.1.4) as the *anticommutative law*. We remark that if A is anticommutative and K admits division by 2 (e.g. if K is a field of characteristic not 2), then $a^2 = 0$ for any a of odd degree; this is sometimes made part of the general definition.

Let $A = (A_n)$, $B = (B_n)$ be two graded algebras. By a *linear mapping of degree* r from A to B we understand a family of K-linear mappings $f_n : A_n \to B_{n+r}$. A linear mapping of degree 0 from A to B which is a ring homomorphism is called a *homomorphism of graded algebras*. Similarly we can define linear mappings of degree r between graded A-modules.

The definition of subalgebra and ideal should be clear: if we regard A as internally graded, a *graded* subalgebra B of A (also called a *homogeneous* subalgebra of A) is a subalgebra B in the ordinary sense such that $B = \sum (B \cap A_n)$; likewise a graded ideal $\mathfrak{a}$ of A is an ideal in the usual sense such that $\mathfrak{a} = \sum (\mathfrak{a} \cap A_n)$. This condition can be expressed by saying that for any finite family $a_n \in A_n (n \in \mathbf{Z})$ we have $\sum a_n \in B$ (or $\mathfrak{a}$) iff each a_n lies in B (or $\mathfrak{a}$). The reader should verify that the image and kernel of a homomorphism of graded algebras are a graded subalgebra and ideal in this sense. Conversely, given a graded algebra A and a graded ideal $\mathfrak{a}$ in A, the quotient $A/\mathfrak{a}$ is again graded, and the natural homomorphism $A \to A/\mathfrak{a}$ is a homomorphism of graded algebras. Finally we observe that if X is a subset of A consisting of homogeneous elements, then the ideal $\mathfrak{a}$ of A generated by X is graded, for each element of $\mathfrak{a}$ can be written as a sum of terms uxv, again in $\mathfrak{a}$, where $x \in X$ and u, v are homogeneous.

In a positively graded ring A, the set $A_+ = \sum_1^\infty A$ is an ideal called the *augmentation ideal*; it is such that $A = A_0 \oplus A_+$. If $A_0 \cong K$ under the mapping $a \mapsto ae$ ($a \in K$), A is an augmented K-algebra in the sense of Section 5.1.

In the commutative case Noetherian graded rings have the following convenient description.

Proposition 6.1.1. *A positively graded commutative ring A is Noetherian if and only if A_0 is Noetherian and A is finitely generated as a ring over A_0.*

Proof. If A_0 is Noetherian and A is generated by $x_1, \ldots, x_d$ over A_0, then it is a homomorphic image of the polynomial ring $A_0[x_1, \ldots, x_d]$ and hence is Noetherian by the Hilbert basis theorem (see Section 10.4 below). Conversely, assume that A is Noetherian; then so is $A_0 = A/A_+$ and A_+ is finitely generated as an ideal. By an earlier remark we may take a family of homogeneous generators $x_1, \ldots, x_d$ of A_+, of degrees $r_1, \ldots, r_d$ say, and write C for the subring generated by the x_i over A_0. To complete the proof we show that $A_n \subseteq C$, by induction on n. Clearly $A_0 \subseteq C$; if $n > 0$ and $A_r \subseteq C$ for $r < n$, take $y \in A_n$. We can write y as a linear combination of the x's, say $y = \sum a_i x_i$, where a_i is homogeneous of degree $n - r_i \geq 0$. Since $r_i > 0$, we can apply induction to conclude that $a_i \in C$ and so $A_n \subseteq C$. It follows that $A = C$ and so A is finitely generated as a ring over A_0. ∎

In all that has been said about graded algebras, the set $\mathbf{Z}$ used to index the components could be replaced by any additive group or even a monoid. The only other case we shall actually need is that of algebras graded by $\mathbf{Z}/(2)$, the group of two elements, briefly: mod 2 graded algebras. Such an algebra has the form $A = A_0 \oplus A_1$, where $A_0A_1 + A_1A_0 \subseteq A_1$, $A_0A_0 + A_1A_1 \subseteq A_0$; it is often called a *superalgebra*.

Exercises

1. Let $f : A \to B$ be a linear mapping of degree r between graded algebras. Show that if f is a homomorphism, then r must be 0.
2. Verify that for any homomorphism of graded algebras, the kernel and image are graded.
3. Let R be a K-algebra of the form $R = K \oplus \mathfrak{a}$, where $\mathfrak{a}$ is an ideal (i.e. R is an augmented K-algebra). Describe R as a graded algebra, with a grading monoid of two elements.
4. Show that $\mathbf{C}$ may be regarded as a mod 2 graded algebra over $\mathbf{R}$. What is the generalization to binomial field extensions, i.e. extensions generated by a root of $x^n = a$?
5. Define a graded tensor product $A \otimes B$ of graded algebras A, B as a tensor product with multiplication $(a_i \otimes b_r)(a'_j \otimes b'_s) = (-1)^{jr} a_i a'_j \otimes b_r b'_s$. Show that the graded tensor product of anticommutative algebras is anticommutative.

6.2 Free Algebras and Tensor Algebras

We have already encountered the free algebra $K\langle X\rangle$ on a set X over a field K as Example 4(iv) of Section 5.1. There it was described as the monoid algebra of a the free monoid on X, but it may also be regarded as the set of all polynomials in the elements of X over K, care being taken to preserve the order of the terms from X. In what follows we shall take K to be any non-trivial commutative ring and to simplify the notation, take X to be finite (as in Section 5.1), $X = \{x_1, \ldots, x_d\}$. Then each element of $K\langle X\rangle$ can be uniquely written as

$$f = \sum \alpha_{i_1 \ldots i_r} x_{i_1} \ldots x_{i_r} = \sum \alpha_I x_I, \tag{6.2.1}$$

where $x_I = x_{i_1} \ldots x_{i_r}$ and the sum is over all distinct sequences $I = (i_1, \ldots, i_r)$, $1 \leq i_r \leq d$, with coefficients α_I in K and almost all 0. If $g = \sum \beta_I x_I$ is another element of $K\langle X\rangle$, then

$$fg = \sum \alpha_I \beta_J x_I x_J. \tag{6.2.2}$$

We can embed X in $K\langle X\rangle$ by identifying x_i with the element (6.2.1) for which $\alpha_I = 1$ if $I = i$ and 0 otherwise. Relative to this mapping $K\langle X\rangle$ has the following 'universal property' characterizing free algebras: given any mapping into a K-algebra

A, $\varphi : X \to A$, say $x_i\varphi = a_i$, there is just one way of extending φ to a homomorphism from $K\langle X\rangle$ to A, namely by mapping

$$\sum \alpha_I x_I \mapsto \sum \alpha_I a_I,$$

where $a_I = a_{i_1} \ldots a_{i_r}$ if $I = (i_1, \ldots, i_r)$. In the special case where X consists of a single element x, the free algebra on x is just the polynomial ring $K[x]$, which of course is commutative, but when X has more than one element, $K\langle X\rangle$ is non-commutative, because e.g. $x_1x_2 \neq x_2x_1$.

In the above construction of $K\langle X\rangle$ we formed first products and then sums. It is also possible to carry out these procedures in the opposite order, and it will be instructive to do this in a more general context. Let U be a K-module and denote by U^n the n-fold tensor product of U with itself over K:

$$U^n = U \otimes U \otimes \ldots \otimes U \quad (n \text{ factors}). \tag{6.2.3}$$

In particular, $U^1 = U$ and for $n = 0$ we shall interpret U^0 as K. By the associative law for tensor products we have for any $r, s \geq 0$,

$$U^r \otimes U^s \cong U^{r+s}. \tag{6.2.4}$$

We shall use this isomorphism to define a multiplication on the direct sum

$$\mathbf{T}(U) = \bigoplus_{n=0}^{\infty} U^n. \tag{6.2.5}$$

In other words, (U^n) is a graded ring for the multiplication (6.2.3) and $\mathbf{T}(U)$ is the corresponding internally graded ring. This algebra $\mathbf{T}(U)$ is called the *tensor K-algebra* on the K-module U. Every element $a \in \mathbf{T}(U)$ is unique of the form $a = \sum a_n$, where $a_n \in U^n$ is the homogeneous component of degree n of a. If $b = \sum b_n$, where $b_n \in U^n$, is another element, then $ab = \sum a_m b_n$ and the homogeneous component of degree n of ab is $\sum a_p b_{n-p}$. We state the universal property of $\mathbf{T}(U)$ in

Theorem 6.2.1. *Let U be a K-module, where K is a commutative ring. Then there is a homomorphism $\lambda : U \to \mathbf{T}(U)$ such that every K-linear mapping $f : U \to A$ into a K-algebra can be uniquely factored by λ.*

This is also expressed by saying that the tensor algebra $\mathbf{T}(U)$ is the *universal K-algebra* for K-linear mappings of U into K-algebras.

Proof. We shall take λ to be the embedding identifying U with its image U^1 in $\mathbf{T}(U)$. Given a K-linear mapping $f : U \to A$, we define, for each $n \geq 1$, a mapping

$$f : (u_1, \ldots, u_n) \mapsto (u_1 f) \ldots (u_n f).$$

Since f is K-linear, this mapping is multilinear and so gives rise to a linear mapping $f^{(n)} : U^n \to A$. On putting these mappings $f^{(n)}$ together we obtain a mapping $\bar{f} : \mathbf{T}(U) \to A$, which is a homomorphism because it is linear and for any

$a = a_1 \otimes \ldots \otimes a_r$, $b = b_1 \otimes \ldots \otimes b_s$ we have $a\bar{f} = a_1 f \ldots a_r f, b\bar{f} = b_1 f \ldots b_s f$, and $ab = a_1 \otimes \ldots \otimes a_r \otimes b_1 \otimes \ldots \otimes b_s$, hence

$$\begin{aligned}(ab)\bar{f} &= (a_1 \otimes \ldots \otimes a_r \otimes b_1 \otimes \ldots \otimes b_s)\bar{f} \\ &= (a_1 f) \ldots (a_r f)(b_1 f) \ldots (b_s f) \\ &= (a\bar{f})(b\bar{f}).\end{aligned}$$

Moreover, it is determined by f because it must agree with f on the generating set U^1 of $\mathbf{T}(U)$. ■

If U is the free K-module on $X = \{x_1, \ldots, x_d\}$, $\mathbf{T}(U)$ clearly coincides with the free algebra $K\langle X\rangle$.

Theorem 6.2.1 provides us with a 'free' K-algebra on a given K-module; this means that the coefficients (from K) lie in the centre of the algebra, but sometimes we have a more general situation. To describe it we need to generalize the notion of a K-algebra. Let E be a general ring (not necessarily commutative); by an *E-ring* we understand a ring R which is an E-bimodule such that $x(yz) = (xy)z$ for any x, y, z in R or E. Even when E is commutative, an E-ring is more general than an E-algebra, because the action of E need not centralize the ring. The difference may be succinctly described by saying that whereas a K-algebra (K commutative) is a ring R with a homomorphism from K to the centre of R, an E-ring is a ring R with a homomorphism from E to R. Frequently it is convenient to assume that both E and R are K-algebras and that the homomorphism $E \to R$ is K-linear. In this case R will be called an E-ring over K or sometimes an E_K-ring. When the coefficient ring E is a K-algebra, K is usually embedded in E as a subring of the centre, and by identification we shall assume K to be a central subring of E. An E-bimodule over K or E_K-bimodule is understood to be an E-bimodule U such that $au = ua$ for all $u \in U, a \in K$.

From any E_K-bimodule U we can form a tensor E_K-ring $\mathbf{T}(U)$ on U as before. We again define U^n by (6.2.3), where the tensor products are over E and $U^0 = E$, $U^1 \cong U$. Now (6.2.4) is clear and $\mathbf{T}(U)$ can again be defined by (6.2.5), this time as an E_K-ring. As before we have

Theorem 6.2.2. *Let E be a K-algebra, where K is a commutative ring, and let U be an E_K-bimodule. Then the tensor E_K-ring $T(U)$ is the universal E_K-ring for bimodule homomorphisms of U into E_K-rings. Thus there is a homomorphism $\lambda : U \to T(U)$ such that every E-bimodule mapping $f : U \to A$, where A is an E_K-ring, can be uniquely factored by λ.*

Proof. The proof is similar to that of Theorem 6.2.1 and may be left to the reader. ■

When $E = K$, this reduces to the case of tensor K-algebras considered in Theorem 6.2.1. To give a more general example, for any set X the free E_K-bimodule F on X is a direct sum of copies of E indexed by X. Now $\mathbf{T}(F)$, the tensor E_K-ring on F, also

denoted by $E_K\langle X\rangle$, may be described as the ring generated by E and X with the defining relations

$$ax = xa \quad \text{for all } x \in X, a \in K. \tag{6.2.6}$$

For simplicity we shall often suppress explicit reference to K in what follows. As in Section 3.3 we can easily verify that the construction of $\mathbf{T}(U)$ from U is in fact a functor from E-bimodules to E-rings. If $\alpha : U \to V$ is a homomorphism of E-bimodules and $\lambda : V \to \mathbf{T}(V)$ is the canonical embedding, then $\alpha\lambda$ is a homomorphism from U to $\mathbf{T}(V)$ and by the universal property of $\mathbf{T}(U)$ we obtain a homomorphism $\mathbf{T}(\alpha)$: $\mathbf{T}(U) \to \mathbf{T}(V)$, which is easily seen to possess the functorial property.

As an application of Theorem 6.2.2 let us consider derivations. A *derivation* over E of an E-ring A is an E-linear mapping δ of A into an A-bimodule such that

$$(xy)^\delta = x.y^\delta + x^\delta.y \quad \text{for all } x, y \in A.$$

For example, the polynomial ring $K[t]$ has the derivation $f \mapsto f'$, where $f' = df/dt$ is the usual derivative. It is often useful to consider a slightly more general situation. Let $\alpha : C \to A$, $\beta : C \to B$ be two homomorphism between E-rings and let M be an (A, B)-bimodule. By an *(α, β)-derivation* over E we understand a mapping $\delta : C \to M$ which is linear and satisfies

$$(xy)^\delta = x^\alpha.y^\delta + x^\delta.y^\beta \quad \text{for all } x, y \in C. \tag{6.2.7}$$

When $A = B = C$ and $\alpha = \beta = 1$, this reduces to the previous case. If we put $A = B = M = E = k[t]$ and for a fixed $p \in K$ define $\alpha : C \to K$ as the evaluation mapping $c \mapsto c(p)$, then the mapping $c \mapsto c'(p)$ (the derivative of c, evaluated at p) is an (α, α)-derivation.

We note that an (α, β)-derivation maps 1 into 0. For if we put $x = y = 1$ in (6.2.7) and observe that $1^\alpha = 1^\beta = 1$, we find that $1^\delta = 0$. It is easily checked that the kernel of δ is a subalgebra of C, called the algebra of *δ-constants*.

Given an (A, B)-bimodule M and homomorphisms $\alpha : C \to A$, $\beta : C \to B$, each $m \in M$ defines a mapping

$$x \mapsto x^\alpha m - mx^\beta, \quad x \in X. \tag{6.2.8}$$

This is easily verified to be an (α, β)-derivation; it is called the *inner derivation* induced by m. An *outer* derivation is one that is not inner.

The study of derivations can be reduced to that of homomorphisms as follows. Let $\begin{pmatrix} A & M \\ 0 & B \end{pmatrix}$ be the ring of all matrices

$$\begin{pmatrix} a & x \\ 0 & b \end{pmatrix}, \quad a \in A, b \in B, x \in M,$$

with the usual matrix addition and multiplication. The (A, B)-bimodule property ensures that we get an E-ring in this way:

$$\begin{pmatrix} a & x \\ 0 & b \end{pmatrix}\begin{pmatrix} a' & x' \\ 0 & b' \end{pmatrix} = \begin{pmatrix} aa' & ax' + xb' \\ 0 & bb' \end{pmatrix},$$

and the associative law holds because in a product of three factors,

$$a(a'x'' + x'b'') + x(b'b'') = (aa')x'' + (ax' + xb')b''.$$

Now it is easily verified that the mapping of C into $\begin{pmatrix} A & M \\ 0 & B \end{pmatrix}$ defined by

$$x \mapsto \begin{pmatrix} a^\alpha & x^\delta \\ 0 & b^\beta \end{pmatrix} \quad (x \in C) \tag{6.2.9}$$

is an E-ring homomorphism iff α, β are E-ring homomorphisms from C to A, B respectively and δ is an (α, β)-derivation from C to M over E.

Let U be any E-bimodule and let $\alpha : U \to A$, $\beta : U \to B$ be E-linear mappings into E-rings A, B. By Theorem 6.2.2 they extend to unique homomorphisms α', β' of $\mathbf{T}(U)$ into A, B respectively. Further, let $\delta : U \to M$ be an E-linear mapping of U into an (A, B)-bimodule M. Then for $x \in U$, (6.2.9) defines an E-linear mapping of U into $\begin{pmatrix} A & M \\ 0 & B \end{pmatrix}$ which again extends to a unique homomorphism of $\mathbf{T}(U)$ into $\begin{pmatrix} A & M \\ 0 & B \end{pmatrix}$. Suppose that $x \in \mathbf{T}(U)$ maps to $\begin{pmatrix} a & x' \\ 0 & b \end{pmatrix}$ (it is easily seen that the (2,1)-entry must be 0). Then $x \mapsto a$ is a homomorphism of $\mathbf{T}(U)$ into A extending α, hence it must be α'. Similarly $x \mapsto b$ must be β', while $x \mapsto x'$ is an (α', β')-derivation δ' say, by the above remarks. This derivation extends δ and is uniquely determined by it. Hence we have

Proposition 6.2.3. *Let U be an E-bimodule, let A, B be any E-rings and M be an (A, B)-bimodule. Given any E-linear mappings $\alpha : U \to A$, $\beta : U \to B$, $\delta : U \to M$, there exist unique homomorphisms $\alpha' : \mathbf{T}(U) \to A$, $\beta' : \mathbf{T}(U) \to B$ extending α and β respectively, and there is a unique (α', β')-derivation over E, $\delta' : \mathbf{T}(U) \to M$, extending δ.* ∎

The result can be applied to free algebras by taking $E = K$ and U the free K-module on $x_1, \ldots, x_d$. Another important special case is that where $E = K$, $M = A = B = \mathbf{T}(U)$, $\beta = 1$. The conclusion then takes the following form.

Corollary 6.2.4. *Let U be a K-module and $\mathbf{T}(U)$ be its tensor algebra. Given any K-linear mappings $\alpha : U \to \mathbf{T}(U)$, $\delta : U \to \mathbf{T}(U)$, there is a unique endomorphism α' of $\mathbf{T}(U)$ extending α and a unique $(\alpha', 1)$-derivation δ' of $\mathbf{T}(U)$ extending δ.* ∎

Exercises

1. Verify that $U \mapsto \mathbf{T}(U)$ is a functor.
2. Given a surjective homomorphism $\mu : U \to V$ of K-modules, show that $\mathbf{T}(\mu) : \mathbf{T}(U) \to \mathbf{T}(V)$ is surjective.

3. Let K be an integral domain and X be any set. By considering degrees, show that $K\langle X\rangle$ is again an integral domain.
4. Prove Leibniz's formula for $(1, 1)$-derivations:

$$(ab)\delta^n = \sum \binom{n}{i} (a\delta^i)(b\delta^{n-i}).$$

More generally, show that for a $(1, \beta)$-derivation δ,

$$(ab)\delta^n = \sum a\delta^i . bf_i^n(\beta, \delta),$$

where $f_i^n(\beta, \delta)$ is the coefficient of t^i in the expansion of $(t\beta + \delta)^n$.
5. Let $\alpha : C \to A$, $\beta : C \to B$ be K-algebra homomorphisms and M be an (A, B)-bimodule. Show that the (α, β)-derivations into M form a K-module $\mathrm{Der}(C, M)$. Taking $C = K\langle X\rangle$, use Proposition 6.2.3 to deduce that $\mathrm{Der}(K\langle X\rangle, M) \cong M^X$. Verify (6.2.9) for any (α, β)-derivation from C to M.
6. Let K be non-trivial and suppose that $K\langle X\rangle$ and $K\langle Y\rangle$ are isomorphic as K-algebras. Use Exercise 5 to show that X and Y are equipotent. ($|X|$ is called the *rank* of the free algebra $K\langle X\rangle$; this exercise shows it to be uniquely determined by $K\langle X\rangle$.)
7. Show that for any commutative ring K, the free algebra $K\langle X\rangle$ admits a unique antiautomorphism leaving K and X elementwise fixed. Show more generally that any antiautomorphism of a K-algebra E leaving K elementwise fixed extends to an antiautomorphism of $E_K\langle X\rangle$ leaving X elementwise fixed.

6.3 The Hilbert Series of a Graded Ring or Module

In the study of finite-dimensional spaces or algebras over a field the dimension provides a useful comparison, based on the formula for a linear mapping from a space U:

$$\dim \ker f + \dim \mathrm{im}\, f = \dim U. \tag{6.3.1}$$

It is an easy consequence that for a short exact sequence of vector spaces

$$0 \to V' \to V \to V'' \to 0 \tag{6.3.2}$$

we have the formula

$$\dim V = \dim V' + \dim V''. \tag{6.3.3}$$

Let us generally define an *additive function* on a class of additive groups as a $\mathbf{Z}$-valued function λ such that for any short exact sequence (6.3.2), we have

$$\lambda(V) = \lambda(V') + \lambda(V''). \tag{6.3.4}$$

As (6.3.3) shows, the dimension is an additive function; more generally, for modules of finite composition length the length is an additive function. The defining property extends to longer sequences as follows:

Proposition 6.3.1. *Let λ be an additive function on modules. Then for any exact sequence of modules*

$$0 \to M_0 \to M_1 \to \ldots \to M_r \to 0 \tag{6.3.5}$$

we have

$$\sum (-1)^i \lambda(M_i) = 0.$$

Proof. Let $f_i : M_i \to M_{i+1}$ be the maps in (6.3.5); by exactness we have $\ker f_i = \operatorname{im} f_{i-1} = N_i$ say, where $N_0 = N_{r+1} = 0$. From the short exact sequence $0 \to N_i \to M_i \to N_{i+1} \to 0$ we obtain $\lambda(M_i) = \lambda(N_i) + \lambda(N_{i+1})$, and the result follows on taking alternating sums. ∎

When we are dealing with a positively graded module $M = \oplus M_n$ over a graded ring we have an infinite sequence $\lambda(M_n)$ of numbers and we define the *Hilbert series*, also called the *Poincaré series*, of M as the formal power series

$$H(M) = \sum \lambda(M_n) t^n.$$

It is a remarkable fact that for finitely generated modules over a commutative Noetherian ring this function is rational, for any additive function λ.

Theorem 6.3.2 (Hilbert–Serre). *Let $A = \oplus A_n$ be a positively graded commutative Noetherian ring, generated as A_0-algebra by $x_1, \ldots, x_d$, homogeneous of positive degrees $r_1, \ldots, r_d$ respectively. If $M = \oplus M_n$ is any finitely generated positively graded A-module, then each homogeneous component M_n is finitely generated as A_0-module and the Hilbert series of M, for any additive function λ, has the form*

$$H(M) = \frac{f(t)}{\prod_{i=1}^{d} (1 - t^{r_i})}, \quad \text{where } f \in \mathbf{Z}[t]. \tag{6.3.6}$$

Proof. By hypothesis, M is finitely generated over A, by $u_1, \ldots, u_k$ say, where each u_i may be taken homogeneous, of degree s_i, say. Any $v \in M_n$ has the form $v = \sum u_i a_i$, where a_i is homogeneous of degree $n - s_i$, hence M_n is spanned over A_0 by all $u_i c_i$, where c_i runs over all products of $x_1, \ldots, x_d$ of degree $n - s_i$.

To prove (6.3.6) we have to show that $\prod (1 - t^{r_i}) H(M)$ is a polynomial in t. We shall use induction on d, the number of generators of A over A_0. When $d = 0$, $A = A_0$ and M is a finitely generated A_0-module, hence $M_n = 0$ for all large n, so $H(M)$ is a polynomial in t and the result holds.

Now assume that $d > 0$ and that the result holds for $d - 1$. Multiplication by x_d is a linear mapping of M of degree $r = r_d$, so we have an exact sequence

$$0 \to N_n \to M_n \xrightarrow{x_d} M_{n+r} \to L_{n+r} \to 0. \tag{6.3.7}$$

Write $L = \oplus L_n, N = \oplus N_n$; these modules are a quotient and a submodule of M

respectively and so are finitely generated. Both are annihilated by x_d and so are $A_0[x_1, \ldots, x_{d-1}]$-modules. Applying λ to (6.3.7) we obtain

$$\lambda(N_n) - \lambda(M_n) + \lambda(M_{n+r}) - \lambda(L_{n+r}) = 0.$$

If we multiply by t^{n+r} and sum over n, the first term gives $t^r H(N)$; similarly for the second term, while the third term gives $H(M)$ except for the first r terms; thus we obtain $H(M)$ by adding a suitable polynomial. The same applies to the fourth term and so we have altogether $t^r H(N) - t^r H(M) + H(M) - H(L) = g$, i.e.

$$(1 - t^r)H(M) + [t^r H(N) - H(L)] = g, \tag{6.3.8}$$

where g is a polynomial in t. Now N, L are $A_0[x_1, \ldots, x_{d-1}]$-modules, as we saw; by induction on d, $\prod_{i=1}^{d-1}(1 - t^{r_i})[H(L) - t^r H(N)]$ is a polynomial, hence by (6.3.8), so is $\prod_{i=1}^{d}(1 - t^{r_i})H(M)$, and (6.3.6) follows. ∎

Consider the special case where all the x_i are of degree 1, thus $r_i = 1$ for all i:

Corollary 6.3.3. *If in Theorem 6.3.2 $x_1, \ldots, x_d$ are all of degree 1 and $\delta = \delta(M)$ is the order of the pole of $H(M)$ at $t = 1$, then $\delta \leq d$ and there exists a unique polynomial h of degree $\delta - 1$ over $\mathbf{Q}$ such that $\lambda(M_n) = h(n)$ for all large n.*

Proof. By Theorem 6.3.2 we have $H(M) = (1 - t)^{-d} f(t)$. By cancelling any power of $1 - t$ dividing f, we can write this as $H(M) = (1 - t)^{-\delta} g(t)$, where $g(1) \neq 0$. Thus if $g = \sum a_i t^i$, then $a = \sum a_i \neq 0$.

Since $(1 - t)^{-\delta} = \sum \binom{\delta + j - 1}{\delta - 1} t^j$, we have $H(M) = \sum_{ij} \binom{\delta + j - 1}{\delta - 1} a_i t^{i+j}$; hence if g is of degree N, then

$$\lambda(M_n) = \sum_{i=0}^{N} a_i \binom{n + \delta - i - 1}{\delta - 1} \quad \text{for } n \geq N.$$

The sum on the right is a polynomial h in n; the i-th term in the sum has leading term $a_i n^{\delta - 1}/(\delta - 1)!$, hence the leading term of h is $a n^{\delta - 1}/(\delta - 1)!$, and this is non-zero, by hypothesis. Clearly h is unique, since its value is prescribed for all $n \geq N$. ∎

When the base ring A_0 is Artinian as well as Noetherian, each component M_i is of finite composition length $l(M_i)$. This length function is clearly additive and we may apply Theorem 6.3.2 and Corollary 6.3.3 with $\lambda = l$. In that case the polynomial h of Corollary 6.3.3 is called the *Hilbert polynomial* or *characteristic polynomial* of M. We note that it is integer-valued for integer arguments, but not necessarily with integer coefficients. It can be shown that $h(\xi)$ is a linear combination of binomial coefficients $\binom{\xi}{r}$, $r = 0, 1, \ldots, \delta - 1$ (see Exercises 2 and 3).

To illustrate Corollary 6.3.3 let us take k to be a field and $\lambda(M)$ to be the dimension of M as vector space over k. For the polynomial ring $A = k[x_1, \ldots, x_d]$, $\lambda(A_n)$ is

the number of products of the x's of degree n. A typical such product is

$$x_1^{\nu_1} \ldots x_d^{\nu_d}. \tag{6.3.9}$$

We form (6.3.9) by taking a row of $n+d-1$ blank squares and putting down x_1 in the first ν_1 squares, then leaving a blank square, then x_2 in the next ν_2 squares etc. Such a product is uniquely determined by choosing the $d-1$ squares to remain blank, hence their number is

$$\lambda(A_n) = \binom{n+d-1}{d-1},$$

and the Hilbert series for A is

$$H(A) = \sum_n \binom{n+d-1}{d-1} t^n = (1-t)^{-d}. \tag{6.3.10}$$

As an example let us take $d = 3$ and use x, y, z; for our graded module we take $M = A/(x^2+y^2-z^2)$. The component M_n for $n \geq 2$ is spanned by $x^r y^{n-r}$ and $x^r y^{n-r-1} z$, in all $2n+1$ monomials, hence $\lambda(M_n) = 2n+1$ and this holds for $n = 0, 1$ as well. Thus $H(M) = \sum (2n+1)t^n = 2(1-t)^{-2} - (1-t)^{-1} = (1+t)(1-t)^{-2}$. Here $\delta(M)$ is 2; it can be shown that in general $\delta(M)$ is the dimension of the variety whose ideal is the annihilator of M, in this case the cylinder $x^2+y^2 = z^2$ (see Hartshorne (1977) I.7).

The Hilbert series for a polynomial ring can be obtained more simply by using

Proposition 6.3.4. *If A, B are any graded k-algebras (over a field k), and λ is the dimension function, then the Hilbert series satisfies*

$$H(A \otimes B) = H(A)H(B).$$

Proof. Put $C = A \otimes B$; since $C_n = \sum A_i \otimes B_{n-i}$, we have $\lambda(C_n) = \sum \lambda(A_i)\lambda(B_{n-i})$, and so $H(C) = \sum H(C_n)t^n = \sum \lambda(A_i)\lambda(B_{n-i})t^n = H(A)H(B)$. ∎

It is clear that for a polynomial ring in one variable we have $H(k[x]) = \sum t^n = (1-t)^{-1}$, hence $H(k[x_1, \ldots, x_d]) = (1-t)^{-d}$.

For the free algebra $F = k\langle x_1, \ldots, x_d \rangle$ it is clear that $\lambda(F_n) = d^n$, hence (see also Exercise 1):

$$H(F) = \sum d^n t^n = (1-dt)^{-1}. \tag{6.3.11}$$

Another important result, due to Golod and Shafarevich who use it to construct infinite 'class-field towers' in algebraic number theory, is an estimate for the Hilbert series of an algebra, which in some cases enables one to recognize from the presentation that the algebra is infinite-dimensional.

Let k be a field and A be any k-algebra, generated by $x_1, \ldots, x_d$ say. The defining relations will be polynomials in the x's, and if A is graded, these defining relations may be taken to be homogeneous. We may always assume that there are no relations of degree 1, for if $\sum a_i x_i = 0$ is a non-trivial relation, we can make a linear trans-

formation to new variables $y_1, \ldots, y_d$ such that $y_1 = \sum a_i x_i$. Then $y_1 = 0$ is one of the defining relations, so A is already generated by $y_2, \ldots, y_d$. Thus a graded algebra always has a set of defining relations which are homogeneous of degree ≥ 2. Now our intuition tells us that the fewer relations there are, the larger dim A will be, and here relations of low degree carry more weight than those of higher degree. How can these ideas be made precise to give a usable estimate? The theorem of Golod and Shafarevich provides a very satisfactory answer. To state it we shall use the following convention: we compare two power series over $\mathbf{Q}$ by writing $\sum a_n t^n \geq \sum b_n t^n$ to mean: $a_n \geq b_n$ for all n.

Theorem 6.3.5 (Golod and Shafarevich). *Let $F = k\langle x_1, \ldots, x_d\rangle$ be the free algebra on d generators over a field k, graded by the degree in the x's, and let $A = F/\mathfrak{a}$ be the quotient of F by an ideal $\mathfrak{a}$ with a homogeneous generating set of r_n elements of degree n $(n = 2, 3, \ldots)$. Then A is infinite-dimensional, provided that*

$$\left(1 - dt + \sum r_n t^n\right)^{-1} \geq 1. \tag{6.3.12}$$

We observe that the series in brackets has integer coefficients and constant term 1, hence its inverse is again a series with integer coefficients.

Proof. (Vinberg) Assume that (6.3.12) holds and write $r = \sum r_n t^n$; we first show that $g = (1 - dt + r)^{-1}$ is not a polynomial. By definition $(1 - dt + r)g = 1$, hence r and g cannot both be polynomials. Now

$$(1 + r)g = 1 + dtg, \tag{6.3.13}$$

and $r \geq 0$ by definition, while the same is true of g, by (6.3.12). If g were a polynomial, r cannot be one, and neither can $(1 + r)g$, whereas the right-hand side of (6.3.13) clearly is a polynomial. This contradiction shows that g cannot be a polynomial.

Let V be the vector space spanned by the defining relations of $\mathfrak{a}$, so by definition, $H(V) = r = \sum r_n t^n$, and since $H(F) = (1 - dt)^{-1}$, the inequality (6.3.12) takes the form

$$(H(F)^{-1} + H(V))^{-1} \geq 1. \tag{6.3.14}$$

Take a graded complement B of $\mathfrak{a}$ in F (as vector space):

$$F = \mathfrak{a} \oplus B. \tag{6.3.15}$$

We claim that

$$\mathfrak{a} = \mathfrak{a}F_1 + BV. \tag{6.3.16}$$

For any element of $\mathfrak{a}$ has the form $\sum f_\lambda p_\lambda g_\lambda$, where $f_\lambda, g_\lambda \in F$, $p_\lambda \in V$ and $f_\lambda, g_\lambda, p_\lambda$ are all homogeneous. If $\deg g_\lambda > 0$, we can write $g_\lambda = \sum h_i x_i$ and then find that $f_\lambda p_\lambda g_\lambda = \sum f_\lambda p_\lambda h_i x_i \in \mathfrak{a}F_1$. If $\deg g_\lambda = 0$, the term reduces to $f_\lambda p_\lambda$ and this lies in FV, but $FV = \mathfrak{a}V + BV$ and $\mathfrak{a}V \subseteq \mathfrak{a}FF_1 \subseteq \mathfrak{a}F_1$. Thus (6.3.16) follows. Computing dimensions, we find

$$H(\mathfrak{a}) \leq H(\mathfrak{a})dt + H(B)H(V),$$

i.e.

$$H(\mathfrak{a})(1-dt) \leq H(B)H(V).$$

Using the equation $H(F) = H(\mathfrak{a}) + H(B)$ from (6.3.15) to eliminate $H(\mathfrak{a})$ and remembering (6.3.11), we obtain

$$(H(F) - H(B))H(F)^{-1} \leq H(B)H(V),$$

which simplifies to

$$H(B)(H(F)^{-1} + H(V)) \geq 1. \tag{6.3.17}$$

Denoting the left-hand side of (6.3.17) by G, we can express $H(B)$ in the form

$$H(B) = G(H(F)^{-1} + H(V))^{-1}. \tag{6.3.18}$$

By (6.3.17), $G \geq 1$, while the second factor on the right of (6.3.18) is ≥ 1 by (6.3.14); moreover, $H(B)$ is not a polynomial, for as (6.3.18) shows, it is the product of two series ≥ 1, of which the second is not a polynomial. Therefore $H(B)$ is not a polynomial, and since $H(A) = H(B)$, by (6.3.15) and the definition of the algebra A, it follows that A is infinite-dimensional. ■

The main application of this result in algebra is the construction of a finitely generated nilalgebra which is not nilpotent, i.e. an algebra A (without 1) with a finite generating set, such that every element of A is nilpotent, but $A^n \neq 0$ for all n (Exercise 5). This answers a question raised by Kurosh in 1941. By an adaptation of this example a finitely generated infinite p-group can be constructed, thus answering an earlier question of Burnside in 1902, see Exercise 6.

Exercises

1. Let $F = k\langle x_1, \ldots, x_d\rangle$. Show that $F = k \oplus \sum x_i F$; deduce that $H(F) = 1 + dtH(F)$ and hence show again that $H(F) = (1-dt)^{-1}$. What is the corresponding formula when x_i is of degree i?
2. Define the differencing operator Δ by $\Delta f(x) = f(x+1) - f(x)$. Show that $\Delta\binom{x}{n} = \binom{x}{n-1}$ and generally, for any polynomial f of degree n, Δf has degree $n-1$.
3. Show that a polynomial over $\mathbf{Q}$ has integer values for all integer arguments iff it is an integral linear combination of binomial coefficients $\binom{x}{n}$. Show that this holds even for polynomials which take integer values for all integer arguments $\geq N$, for some N. (Hint. Use Exercise 2 and induction.)
4. (Golod) With the notation of Theorem 6.3.5, show that if for some ε,

$$0 < 2\varepsilon < d, \quad r_n \leq \varepsilon^2(d - 2\varepsilon)^{n-2}, \tag{6.3.19}$$

then

$$\left(1 - dt + \sum r_n t^n\right)^{-1} \geq \frac{1 - (d - 2\varepsilon)t}{(1 - (d - \varepsilon)t)^2}.$$

Show that this expression has positive coefficients, and deduce that $F/\mathfrak{a}$ is infinite-dimensional whenever r_n satisfies (6.3.19). (Note that for fixed n, the estimate (6.3.19) for r_n is greatest when $\varepsilon = d/n$.)

5. (Golod) Let $F = k\langle x_1, \ldots, x_d\rangle$, where $d \geq 2$ and denote the augmentation ideal by F_+. Show that the general polynomial g in F_+ of degree ν is a linear combination of $q = d + d^2 + \ldots + d^\nu$ monomials, and that g^N can be written as a linear combination of $\binom{N+q-1}{q-1}$ forms, where these forms are independent of the coefficients of g, and have degrees between N and νN.

   Deduce the existence of a sequence $U', U'', \ldots$ of spaces and integers $N_1, N_2, \ldots$ such that $\nu N_\nu < N_{\nu+1}$ and the homogeneous components of $U^{(\nu)}$ are 0 except for degrees in the range $[N_\nu, \nu N_\nu]$. Show that if $\mathfrak{a}^{(\nu)}$ is the ideal of F generated by $V^{(\nu)} = U' + \ldots + U^{(\nu)}$ and $r_n = \dim(V^{(\nu)})_n)$, then (i) r_n satisfies (6.3.19) for $\varepsilon = d/n$ and (ii) for any $g \in F_+$ of degree $\leq \nu$, $g^{N\nu} \in \mathfrak{a}^{(\nu)}$. Deduce that $F_+/\mathfrak{a}$, where $\mathfrak{a} = \cup\mathfrak{a}^{(\nu)}$, is a finitely generated nilalgebra, which is infinite-dimensional, and hence not nilpotent. (Hint. Observe that $\binom{N+q-1}{q-1} \leq (N+q)^q$ and choose $N = N_\nu$ to satisfy $(N+q)^q \leq \varepsilon^2(d - 2\varepsilon)^{N-2}$.)

6. (Golod) Let $A = F/\mathfrak{a}$ be the algebra with nil but not nilpotent augmentation ideal $A_+ = F_+/\mathfrak{a}$ constructed in Exercise 5, with $k = \mathbf{F}_p$ as ground field, where p is a given prime number. Show that every element of the form $1 + c$, $c \in A_+$ has multiplicative order a power of p. Deduce that the group G generated by the $1 + x_i \pmod{\mathfrak{a}}$ is an infinite p-group. (Hint. If G were finite, the degrees of its members would be bounded, by m say. Now pick a monomial $u_1 \ldots u_{m+1} \neq 0$ in A and consider the element $\Pi(1 + u_i)$ of G.)

6.4 The Exterior Algebra on a Module

In this section we shall describe an algebra which is useful in the study of subspaces of a vector space. But the definition can be formulated more generally for modules. Let U be a K-module, where K is any commutative ring, and let $\mathbf{T}(U)$ be the tensor algebra on U, as graded algebra. In $\mathbf{T}(U)$ consider the ideal $\mathfrak{a}$ generated by all elements u^2, $u \in U$. Since the elements u^2 are homogeneous in $\mathbf{T}(U)$, the ideal $\mathfrak{a}$ is graded and therefore the quotient $\mathbf{T}(U)/\mathfrak{a}$ is again graded. It is called the *exterior algebra* on U and is written $\Lambda(U)$. The component of degree r of $\Lambda(U)$ is denoted by $\Lambda^r(U)$; its elements are called *r-vectors* or *multivectors*. The multiplication in $\Lambda(U)$ is denoted by $x \wedge y$. Thus $\Lambda(U)$ is the K-algebra generated by the module U with the defining relations

$$u \wedge u = 0 \quad \text{for all } u \in U. \tag{6.4.1}$$

For any $u, v \in U$ we have $uv + vu = (u+v)^2 - u^2 - v^2$ in $\mathbf{T}(U)$, hence (6.4.1) entails the relation $u \wedge v + v \wedge u = 0$, i.e.

$$u \wedge v = -v \wedge u \quad \text{for all } u, v \in U. \tag{6.4.2}$$

By induction on the degree we find that $\Lambda(U)$ is anticommutative; in fact the definition shows that when K is a field of characteristic not 2 then $\Lambda(U)$ is the free anticommutative algebra on the K-module U.

On every graded algebra we can define a linear mapping σ which multiplies each homogeneous element of degree r by $(-1)^r$. Since $(-1)^r(-1)^s = (-1)^{r+s}$, this is an algebra homomorphism; in fact it is an automorphism, since its square is the identity mapping. Any mapping $\neq 1$ whose square is 1 is called an *involution*, and σ is called the *standard involution* on $\Lambda(U)$. By an *antiderivation* of $\Lambda(U)$ we understand a $(\sigma, 1)$-derivation of $\Lambda(U)$ into itself, i.e. a linear mapping δ such that

$$(a \wedge b)^\delta = a^\delta \wedge b + a^\sigma \wedge b^\delta \quad \text{for all } a, b \in \Lambda(U). \tag{6.4.3}$$

Such a mapping is entirely determined by its effect on the generating set U of $\Lambda(U)$, and a mapping $\delta : U \to \Lambda(U)$ defines an antiderivation of $\Lambda(U)$ provided that

$$u^\delta \wedge u - u \wedge u^\delta = 0 \quad \text{for all } u \in U. \tag{6.4.4}$$

For we can always extend δ to an antiderivation of $\mathbf{T}(U)$, again written δ, and this will induce an antiderivation of $\Lambda(U)$ provided that δ maps $\mathfrak{a}$, the kernel of the mapping $\mathbf{T}(U) \to \Lambda(U)$, into itself. Now by (6.4.4), $(u^2)^\delta = u^\delta u + u^\sigma u^\delta = u^\delta u - uu^\delta \in \mathfrak{a}$ and more generally, for any $a, b \in \mathbf{T}(U)$, $(au^2b)^\delta = a^\delta u^2 b + a^\sigma (u^2)^\delta b + a^\sigma (u^\sigma)^2 b^\delta$, and this again lies in $\mathfrak{a}$. Thus δ maps $\mathfrak{a}$ into itself and so induces an antiderivation on $\Lambda(U)$.

To give an example, if $u^\delta \in U$, then (6.4.4) reduces to $2u \wedge u^\delta = 0$; on the other hand, if $u^\delta \in K$, i.e. if $\delta \in U^* = \mathrm{Hom}_K(U, K)$, then condition (6.4.4) always holds. Thus any element of the dual module U^* defines an antiderivation of degree -1 on $\Lambda(U)$; this is sometimes called *interior* multiplication.

The structure of $\Lambda(U)$ may be described as follows.

Theorem 6.4.1. *If the K-module U is spanned by $e_1, \ldots, e_n$, then $\Lambda(U)$ is spanned by the elements*

$$e_{i_1} \wedge \ldots \wedge e_{i_r} \quad \text{where } i_1 < \ldots < i_r, \quad r = 0, 1, \ldots. \tag{6.4.5}$$

In particular, whenever U is finitely spanned by n elements, say, then $\Lambda(U)$ can be spanned by 2^n elements. Moreover, when U is a free K-module on $e_1, \ldots, e_n$ as basis, then the elements (6.4.5) *form a basis of $\Lambda(U)$.*

Proof. From the definition of $\Lambda(U)$ as a quotient of the tensor algebra on U it follows that $\Lambda(U)$ is spanned by the products (6.4.5), where the suffixes range from 1 to n without restriction. By (6.4.2) we can permute adjacent factors by changing the sign. Thus we can bring the suffixes into ascending order and, if two are equal, the corresponding product vanishes by (6.4.1). This shows that $\Lambda(U)$ is spanned by the products (6.4.5) with strictly ascending suffix sets, as claimed.

Now assume that U is a free K-module with $e_1, \ldots, e_n$ as basis. We must show that the products (6.4.5) are linearly independent. Since $\Lambda(U)$ is graded, every relation is a sum of homogeneous relations; let

$$\sum a_{i_1 \ldots i_r} e_{i_1} \wedge \ldots \wedge e_{i_r} = 0 \tag{6.4.6}$$

be a non-trivial relation of least degree. Take a dual basis $\lambda_1, \ldots, \lambda_n$ for the e's, i.e. a family of linear functionals on U such that $\langle \lambda_i, e_j \rangle = \delta_{ij}$, and denote the corresponding antiderivations on $\Lambda(U)$ again by $\lambda_1, \ldots, \lambda_n$. Suppose that (6.4.6) contains a non-zero term in which $i_1 = 1$. Applying λ_1 to (6.4.6), we obtain

$$\sum a_{1 i_2 \ldots i_r} e_{i_2} \wedge \ldots \wedge e_{i_r} = 0$$

and this is a non-trivial relation in $\Lambda(U)$ of lower degree. This is a contradiction, and it shows that no non-trivial relation (6.4.6) exists with $i_1 = 1$. Similarly for $i_1 = 2, \ldots, n$, hence there are no non-trivial relations (6.4.6) and so the elements (6.4.5) form a basis, as claimed. ∎

If U is a free K-module of infinite rank, with a totally ordered basis $\{e_i\}$, then it follows in exactly the same way that the set of elements (6.4.5) where the i_r now range over the whole index set, form a basis of $\Lambda(U)$. All this applies in particular when K is a field, so that every K-module is free. In that case we have the following criterion for linear dependence:

Corollary 6.4.2. *Let U be a vector space over a field k. Then the elements $u_1, \ldots, u_r \in U$ are linearly dependent if and only if*

$$u_1 \wedge \ldots \wedge u_r = 0 \quad \text{in } \Lambda(U).$$

Proof. If the u's are linearly dependent, one of them can be expressed in terms of the rest, say $u_1 = \sum_2^r a_i u_i$, hence

$$u_1 \wedge u_2 \wedge \ldots \wedge u_r = \sum_2^r \alpha_i u_i \wedge u_2 \wedge \ldots \wedge u_r,$$

and the right-hand side vanishes because each term has a repeated factor. Conversely, if $u_1, \ldots, u_r$ are linearly independent, they can be completed to a basis of U and then $u_1 \wedge \ldots \wedge u_r \neq 0$ by Theorem 6.4.1 (and the remark following it), because it forms part of a basis of $\Lambda(U)$. ∎

The following universal property of $\Lambda(U)$ is clear from the definition.

Proposition 6.4.3. *Let U be a K-module and A be a K-algebra. Then any linear mapping $\alpha : U \to A$ such that $(u^\alpha)^2 = 0$ for all $u \in U$ can be extended uniquely to a homomorphism of $\Lambda(U)$ into A.* ∎

Taking in particular $A = \Lambda(V)$, for any K-module V, we see that any linear mapping $f : U \to V$ induces a homomorphism $\Lambda(f) : \Lambda(U) \to \Lambda(V)$. It is easily

verified that $\Lambda(fg) = \Lambda(f)\Lambda(g)$ and $\Lambda(1) = 1$, so that Λ is a functor from K-modules to graded K-algebras.

We next show how determinants can be quite naturally defined in terms of $\Lambda(U)$. Let U be a free K-module of rank n and $f : U \to U$ be a linear mapping of U into itself. In terms of a basis $e_1, \ldots, e_n$ of U, f is given by an $n \times n$ matrix $\alpha = (\alpha_{ij})$, where

$$e_i f = \sum \alpha_{ij} e_j.$$

We saw that f induces an endomorphism $\Lambda(f)$ of $\Lambda(U)$, in particular $\Lambda^n(f)$ maps $\Lambda^n(U)$ into itself, and since $\Lambda^n(U)$ is free of rank 1, $\Lambda^n(f)$ is determined by a scalar factor. Let us find this scalar; we have

$$(e_1 \wedge \ldots \wedge e_n)f = \sum \alpha_{1i_1}\alpha_{2i_2} \ldots \alpha_{ni_n} e_{i_1} \wedge \ldots \wedge e_{i_n}.$$

The sum on the right is over all suffix sequences $(i_1, \ldots, i_n)$, but $e_{i_1} \wedge \ldots e_{i_n}$ is zero unless $(i_1, \ldots, i_n)$ is a permutation of $(1, 2, \ldots, n)$, and then it is $\pm e_1 \wedge \ldots \wedge e_n$ with $+$ or $-$ according as this permutation is even or odd. We shall denote this sign by $\varepsilon(i_1, \ldots, i_n)$ and recall the definition of the determinant of a matrix α:

$$\det \alpha = \sum \varepsilon(i_1, \ldots, i_n)\alpha_{1i_1}\alpha_{2i_2} \ldots \alpha_{ni_n}.$$

A comparison shows that

$$(e_1 \wedge \ldots \wedge e_n)f = (\det \alpha)e_1 \wedge \ldots \wedge e_n. \tag{6.4.7}$$

In fact we can take (6.4.7) as the definition of $\det \alpha$; this associates the determinant with an endomorphism of U rather than a matrix. Of course it is well known that if f is given by a matrix A in one coordinate system and by B in another, then $B = P^{-1}AP$, for some invertible matrix P, so that $\det A = \det B$.

From this definition it is particularly easy to prove the multiplication theorem of determinants. Writing $e_T = e_1 \wedge \ldots \wedge e_n$, we have for any linear mappings f, g with matrices α, β respectively, $e_T(fg) = (e_T f)g$; hence by applying (6.4.7) to both sides, we find that

$$\det(\alpha\beta)e_T = (\det \alpha)(\det \beta)e_T,$$

and so $\det(\alpha\beta) = (\det \alpha)(\det \beta)$.

More generally, any endomorphism f of U induces an endomorphism $\Lambda^r(f)$ of $\Lambda^r(U)$, for $1 \leq r \leq n$; the corresponding matrix $\alpha^{(r)}$ has $\binom{n}{r}$ rows and columns, the entries being r-th order minors of the matrix α. This matrix $\alpha^{(r)}$ is called the r-th *compound matrix* of α and as before we see that

$$(\alpha\beta)^{(r)} = \alpha^{(r)}\beta^{(r)}. \tag{6.4.8}$$

This is an identity of degree r in the entries of α and β, called the *Binet–Cauchy identity*; the case $r = n$ is just the multiplication theorem for determinants.

Let U again be the free K-module with basis $e_1, \ldots, e_n$ and let us find the multiplication table for $\Lambda(U)$ in terms of the standard basis (6.4.5). For any subset I of

$\{1, 2, \ldots, n\}$ we denote by e_I the product of the e_i $(i \in I)$ in ascending order; thus if $I = \{1, 2, \ldots, r\}$ then $e_I = e_1 \wedge \ldots \wedge e_r$. Given any subsets H, K of $\{1, \ldots, n\}$, if $H \cap K \neq \varnothing$, then it is clear that $e_H \wedge e_K = 0$. When $H \cap K = \varnothing$, we have

$$e_H \wedge e_K = \varepsilon_{HK} e_{H \cup K}, \tag{6.4.9}$$

where $\varepsilon_{HK} = \pm 1$. Here the number of inversions necessary to arrange the suffixes of $e_H \wedge e_K$ in ascending order is just the number of pairs $i \in H, j \in K$ such that $i > j$. If this number is n, then $\varepsilon_{HK} = (-1)^n$.

If H is any subset of $T = \{1, \ldots, n\}$ and H' is the complementary subset, then $e_H \wedge e_{H'} = \varepsilon_{HH'} e_T$. Suppose that $|H| = r$; on applying a linear mapping f with matrix α, we find

$$e_H . \Lambda^r(f) = \sum \alpha_{HI} e_I,$$

where I runs over all r-element subsets of T and α_{HI} is the minor corresponding to the rows indexed by H and the columns indexed by I. Thus by applying f to both sides of the equation

$$\varepsilon_{HH'} e_T = e_H \wedge e_{H'},$$

we obtain

$$\begin{aligned}(\det \alpha)\varepsilon_{HH'} e_T &= e_H \Lambda^r(f) \wedge e_{H'} \Lambda^{n-r}(f) \\ &= \sum \alpha_{HI} \alpha_{H'J} e_I \wedge e_J,\end{aligned}$$

where I runs over all r-element subsets of T and J over all $(n - r)$-element subsets of T. It is clear that

$$e_I \wedge e_J = \begin{cases} \varepsilon_{II'} & \text{if } J = I', \\ 0 & \text{otherwise,} \end{cases}$$

where I' is the complement of I. Hence we obtain the formula

$$\det \alpha = \sum_I \varepsilon_{II'} \varepsilon_{HH'} \alpha_{HI} \alpha_{H'I'}, \tag{6.4.10}$$

for a fixed r-element subset H and its complement H', where the summation is over all r-element sets I and its complement I'. This is called the *Laplace expansion* of $\det \alpha$ in terms of the r rows indexed by H. In terms of compound matrices this can be written briefly as

$$\alpha^{(r)} \alpha_{(n-r)} = (\det \alpha) . I, \tag{6.4.11}$$

for an appropriately defined matrix of cofactors $\alpha_{(n-r)}$.

In a vector space U over a field, any non-zero vector can be transformed into any other non-zero vector, i.e. the automorphisms of U act transitively on the non-zero vectors. By contrast the automorphisms of $\Lambda(U)$ do not always act transitively on the set of all non-zero r-vectors for $r > 1$, e.g. the 2-vectors of the form $u \wedge v$ form a proper subset of $\Lambda^2(U)$ (at least when $\dim U > 3$), which is mapped into itself by all automorphisms. An element of $\Lambda(U)$ is said to be *decomposable* if it

can be written as a product of vectors: $u_1 \wedge \ldots \wedge u_r$. These decomposable r-vectors were used to describe r-dimensional subspaces by Hermann Grassmann, who introduced exterior algebras in 1844; the exterior algebra on the dual space U^* is also called the *Grassmann algebra* of U. As an example let us derive the condition for a 2-vector to be decomposable.

Proposition 6.4.4. *Let U be a vector space over a field K of characteristic not 2, with basis $e_1, \ldots, e_n$. Then any 2-vector in $\Lambda(U)$ can be written in the form*

$$w = \sum c_{ij} e_i \wedge e_j, \quad \text{where } c_{ii} = 0, c_{ij} = -c_{ji}, \tag{6.4.12}$$

and w is decomposable if and only if, for any four subscripts i, j, k, l,

$$c_{ij}c_{kl} + c_{ik}c_{lj} + c_{il}c_{jk} = 0. \tag{(6.4.13)}$$

Proof. By definition any 2-vector has the form $\sum c_{ij} e_i \wedge e_j$, but there are no terms in $e_i \wedge e_i$, so we may take $c_{ii} = 0$, and for fixed i, j we have $\lambda e_i \wedge e_j + \mu e_j \wedge e_i = (\lambda - \mu) e_i \wedge e_j = c e_i \wedge e_j - c e_j \wedge e_i$, where $c = (\lambda - \mu)/2$. Thus every 2-vector can be expressed as in (6.4.12). Now let $w = p \wedge q$, where $p = \sum a_i e_i$, $q = \sum b_i e_i$; then $w = \sum (a_i b_j - a_j b_i) e_i \wedge e_j$. Consider the determinant

$$\begin{vmatrix} a_i & a_j & a_k & a_1 \\ b_i & b_j & b_k & b_1 \\ a_i & a_j & a_k & a_1 \\ b_i & b_j & b_k & b_1 \end{vmatrix}$$

If we make a Laplace expansion by the first two rows and observe that the determinant vanishes, we obtain (6.4.13), bearing in mind that $a_i b_j - a_j b_i = c_{ij}$. Conversely, given n^2 elements c_{ij} of K where $c_{ii} = 0$ and $c_{ij} = -c_{ji}$, assume that (6.4.13) holds for all choices of subscripts. If all the c_{ij} vanish, the 2-vector (6.4.12) is zero and hence decomposable. Otherwise let $c_{12} \neq 0$ say, and put $a_j = (c_{1j})$, $b_i = (c_{i2})$; then $p = \sum a_j e_j$ and $q = \sum b_i e_i$ are such that $p \wedge q$ has coordinates

$$\begin{aligned} c'_{ij} = a_i b_j - a_j b_i &= c_{1i}c_{j2} - c_{1j}c_{i2} \\ &= -c_{12}c_{ij} \quad \text{by (6.4.13).} \end{aligned}$$

Hence $w = -c_{12}^{-1} p \wedge q$ is decomposable. ∎

We observe that in three dimensions no relations are needed: every 2-vector is then decomposable. In a four-dimensional space there is just one relation

$$c_{01}c_{23} + c_{02}c_{31} + c_{03}c_{12} = 0. \tag{6.4.14}$$

As the proof of Proposition 6.4.4 shows, in n dimensions we need only impose the conditions (6.4.13) for fixed i, j such that $c_{ij} \neq 0$, and all pairs k, l; thus $\binom{n-2}{2}$ relations are needed and the rest are a consequence.

Proposition 6.4.4 has an interesting application to lines in projective 3-space. The straight lines in three-dimensional projective space are two-dimensional subspaces of a four-dimensional space, and so can be described by decomposable 2-vectors. From Proposition 6.4.4 we see that each line is described by six homogeneous coordinates c_{ij} $(i, j = 0, 1, 2, 3)$ satisfying (6.4.14), the *Plücker* or *line* coordinates. This equation represents a quadric W in five dimensions (since there are six coordinates), the *Klein quadric*; many properties of lines in space can be obtained very simply by regarding the lines as points of this quadric. To give an example, this interpretation shows that the lines in 3-space form a four-dimensional variety. A subvariety of codimension one is called a *line complex*; such a subvariety can be obtained by imposing certain algebraic relations on the line coordinates, and a theorem of Klein asserts that any line complex in projective 3-space can be defined by a single equation besides (6.4.14). For a proof we note that the coordinate ring of a line complex is obtained from the coordinate ring A of W by dividing by a minimal prime ideal (minimal because the subvariety is of codimension one, i.e. maximal). Now it can be shown that A is a unique factorization domain, hence every minimal ideal is principal (see Theorem 10.2.10 below) and this means that the quotient of A by a minimal prime ideal is obtained by imposing a single equation, as claimed (this simple proof is due to Nagata [1957]; see also Cohn [1973]). For a discussion of the properties of lines using W, see Semple and Roth (1949).

The decomposable r-vectors correspond to the r-dimensional subspaces of U; the coordinates can generally be taken to be antisymmetric in all subscripts (at least in characteristic 0). They are again called the *Plücker coordinates* of the subspace, and they can again be characterized by relations like (6.4.13). Their derivation is straightforward and we shall not stop to give it here (see Exercise 8), but we note the following relation between the multivectors of different subspaces.

Proposition 6.4.5. *Let U be an n-dimensional space over a field K. To each r-dimensional subspace V of U $(1 \leq r \leq n)$ there corresponds a decomposable r-vector v, determined up to an element of $K^\times$, such that*

$$x \in V \Leftrightarrow x \wedge v = 0.$$

Given subspaces V, W with vectors v, w, we have $v \wedge w \neq 0$ if and only if $V \cap W = 0$, and when this is so, then $v \wedge w$ is the vector corresponding to $V + W$.

Proof. Let $e_1, \ldots, e_r$ be a basis for V; the corresponding r-vector is $v = e_1 \wedge \ldots \wedge e_r$. Any other basis of V is related to the e's by a linear transformation α and the change of basis changes v by a factor $\det \alpha$, by (6.4.7). Now $x \in V$ iff x is linearly dependent on $e_1, \ldots, e_r$ and this is so iff $x \wedge v = 0$, by Corollary 6.4.2. Given subspaces V, W with multivectors v, w, if $V \cap W \neq 0$, then for a suitable choice of basis v and w have a common factor and so $v \wedge w = 0$. Otherwise we obtain a basis for $V + W$ by taking a basis for V and one for W, so $v \wedge w$ is a multivector for $V + W$. ∎

Exercises

1. Show that the elements of even degree of $\Lambda(U)$ form a commutative subalgebra.
2. Show that any exterior algebra satisfies the identity $[[x, y], z] = 0$, where $[x, y] = xy - yx$.
3. Find the first power of t in $\left(1 - dt + \binom{d+1}{2}t^2\right)^{-1}$ which has a negative coefficient (there must be one, by Theorems 6.3.5 and 6.4.1).
4. In a K-module U consider an equation $a = \sum \alpha_i u_i$, where $a, u_i \in U$ and $\alpha_i \in K$. Show that

$$u_1 \wedge \ldots \wedge u_{i-1} \wedge a \wedge u_{i+1} \wedge \ldots \wedge u_n = \alpha_i(u_1 \wedge \ldots \wedge u_n).$$

5. Let U be an n-dimensional vector space. Show that the decomposable multivector w_i corresponding to a subspace W_i $(i = 1, 2)$ satisfies $w_2 = z \wedge w_1$ for some multivector z iff $W_2 \supseteq W_1$.
6. Let U be a vector space with basis $e_1, \ldots, e_n$. Show that if $u_1, \ldots, u_n \in U$ satisfy $\sum u_i \wedge e_i = 0$, then $u_i = \sum a_{ij} e_j$, where $a_{ij} = a_{ji}$.
7. Let A be the coordinate ring of the quadric $x_0 x_3 = x_1 x_2$. Show that A is an integral domain, but not a unique factorization domain.
8. Let V be a vector space with basis $e_1, \ldots, e_n$ over a field of characteristic 0. Show that any r-vector can be written as $\sum a_{i_1 \ldots i_r} e_{i_1} \wedge \ldots \wedge e_{i_r}$, where $a_{i_1 \ldots i_r}$ is antisymmetric, and it is decomposable iff

$$a_{i_1 \ldots i_r} a_{j_1 \ldots j_r} - \sum a_{i_1 \ldots i_{\lambda-1} j_1 i_{\lambda+1} \ldots i_r} a_{i_\lambda j_2 \ldots j_r} = 0.$$

9. Prove (6.4.7) by verifying that $\Lambda^n(f)$ satisfies the axioms for a determinant: $\det A$ is an alternating multilinear function of the columns of the matrix A which is 1 for $A = I$.
10. Show that the determinant of the r-th compound of an $n \times n$ matrix α is a power of $\det \alpha$: $\det \alpha^{(r)} = (\det \alpha)^m$. By comparing degrees show that $m = \binom{n-1}{r-1}$. (Sylvester–Franke theorem. Hint. Use (6.4.11).)

Further Exercises for Chapter 6

1. Let A be a graded K-algebra. Find the condition for the mapping $a_n \mapsto \lambda_n a_n$ (where $\deg a_n = n$ and $\lambda_n \in K$) to be (i) an endomorphism and (ii) a derivation.
2. Let K be a commutative ring. Show that the centre of $K\langle X\rangle$ is K whenever

$$|X| > 1.$$

3. Show that a homomorphism $K\langle X\rangle \to K$, for any ring K, maps nilpotent elements to nilpotent elements. If K has no nilpotent elements apart from 0, the same is true for $K\langle X\rangle$. (Hint. Take the free monoid on X to be ordered lexicographically.)

4. By a *filtered* K-algebra one understands a K-algebra A with a sequence of submodules

$$\ldots \supseteq A_n \supseteq A_{n+1} \supseteq \ldots, \quad A_i A_j \subseteq A_{i+j}, \quad \cap A_n = 0, \quad \cup A_n = A.$$

Given such an algebra, put $G_n = A_n/A_{n+1}$ and define a multiplication $G_i \times G_j \to G_{i+j}$ by taking for $\alpha \in G_i$, $\beta \in G_j$ any representative a of α, b of β and defining $\alpha\beta$ as the residue class of ab (mod A_{i+j+1}). Verify that this definition of $\alpha\beta$ depends only on α, β and not on a, b and that with this definition (G_n) becomes a graded K-algebra (called the *graded ring associated with* A and denoted by $\mathrm{gr}(A)$). Verify that for any ideal $\mathfrak{a}$ of a ring R, if $\cap \mathfrak{a}^n = 0$, then the powers of $\mathfrak{a}$ define a filtration and so give rise to a graded ring.
5. Let A be a filtered K-algebra and $\mathrm{gr}(A)$ be its associated graded algebra. Show that if $\mathrm{gr}(A)$ is an integral domain, then so is A. Does the converse hold?
6. Show that for any vector space V over a field, each element of $\Lambda(V)$ is either invertible or nilpotent (a ring with this property is said to be *completely primary*).
7. Let V be a vector space over a field. Show that if a p-vector w in $\Lambda(V)$ is divisible by $u_1, \ldots, u_r$, where the u_i are linearly independent, then it is divisible by $u_1 \wedge \ldots \wedge u_r$. Deduce that any p-vector w can be written as $u_1 \wedge \ldots \wedge u_r \wedge v$, where v has no linear factor.
8. Let $F = k\langle x_1, \ldots, x_d\rangle$; given an equation $\sum x_i u_i = 0$, where $u_i \in F$, show that $u_i = 0$. Deduce that in any element $f \in F$ the right-hand cofactor of x_i is uniquely determined (this is called the *transduction* with respect to x_i).
9. Show that in $k\langle x, y\rangle$ the elements xy^n $(n = 1, 2, \ldots)$ generate a subalgebra which is free on these generators.
10. Let X be a finite set and denote by $K\langle\langle X\rangle\rangle$ the set of all sums (6.2.1), allowing infinitely many non-zero coefficients. Show that the multiplication law (6.2.2) defines a K-algebra structure on $K\langle\langle X\rangle\rangle$, containing $K\langle X\rangle$ as subalgebra. Show that a mapping $X \to K\langle\langle X\rangle\rangle$ extends to an endomorphism of $K\langle\langle X\rangle\rangle$ provided that the image of any $x \in X$ has zero constant term (this algebra $K\langle\langle X\rangle\rangle$ is the free power series ring in X over K, see FA Chapter 11).
11. Let K be an integral domain. Show that $K\langle\langle X\rangle\rangle$ is also an integral domain. Show further that $f \in K\langle\langle X\rangle\rangle$ is invertible iff its constant term is invertible in K, whereas an element p of $K\langle X\rangle$ is invertible iff p reduces to its constant term and the latter is invertible in K.
12. Let V be a vector space over a field k. Show that for each $r = 1, 2, \ldots, n$ there is a pairing $\langle , \rangle : \Lambda(V^*) \times \Lambda(V) \to k$ such that

$$\langle f_1 \wedge \ldots \wedge f_r, x_1 \wedge \ldots \wedge x_r\rangle = \det(\langle f_i, x_j\rangle).$$

Hence find an injective linear mapping $\Lambda(V^*) \to (\Lambda(V))^*$ which is an isomorphism if V is finite-dimensional.
13. Using Exercise 12 show that if V is n-dimensional, then every $(n-1)$-vector in $\Lambda(V)$ is decomposable. Show that (in characteristic not two) a 2-vector p is decomposable iff $p \wedge p = 0$.

14. Let V be a left k-space and V^* be its dual. If $f \in \mathrm{End}(V)$ has the matrix α relative to a certain basis of V, then f^* relative to the dual basis of V^* again has the matrix α if V^* is regarded as a right k-space (and f^* acts on the left), but it has the transposed matrix α^T if V^* is regarded as left k-space (with f^* acting on the right). Deduce that $\det \alpha = \det \alpha^T$.
15. Let V be a vector space and $a \in V$. Show that the mapping $x \mapsto a \wedge x$ of V into $\Lambda(V)$ extends to an antiderivation of $\Lambda(V)$; obtain its expression in terms of multiplication by a.
16. Show that the r-th compound matrix $\alpha^{(r)}$ satisfies

$$\alpha_{11}^{(n-1)}\alpha_{ii}^{(n-1)} - \alpha_{1i}^{(n-1)}\alpha_{i1}^{(n-1)} = \alpha_{1i,1i}^{(n-2)}.\det \alpha,$$

where subscripts indicate the rows and columns omitted.
17. Show that a square matrix A over a commutative ring K is a zerodivisor iff $\det A$ is 0 or a zerodivisor in K.
18. For any square matrix A (over a commutative ring K) define $e_i(A)$ to be the i-th elementary symmetric function of its eigenvalues. Show that $e_i(A) = \mathrm{tr}(A^{(i)})$, where $A^{(i)}$ is the i-th compound, and the characteristic polynomial of A is $\sum(-1)^r e_r(A)\lambda^{n-r}$.

    Deduce that if $K = \mathbf{C}$ and A^H is the Hermitian conjugate (i.e. transpose of complex conjugate), then

$$\det(A^H A + \lambda I) = \sum_0^n \mathrm{tr}(A^{(i)H}A^{(i)})\lambda^{n-i}.$$

7 Field Theory

Fields form one of the basic algebraic concepts, for which there is an extensive theory, dealing mainly with the form taken by field extensions. In Section 7.1 field extensions are described, and the special case of splitting fields is introduced in Section 7.2, leading to the notion of algebraic closure (Section 7.3). In Section 7.4 we examine the problems arising in finite characteristic. One of the main tools in this study is Galois theory and this forms the subject of Sections 7.5 and 7.6, while Sections 7.10 and 7.11 bring its application to the solution of equations. The special case of finite fields is studied in Section 7.8, using information on the roots of unity (Section 7.7); Section 7.9 is devoted to generators and some invariants of extensions.

7.1 Fields and their Extensions

We recall from Section 4.1 that a *field* is a commutative ring in which $1 \neq 0$ and every non-zero element has an inverse. It follows that a field F has exactly two ideals, 0 and F, and hence every homomorphism from a field to a non-trivial ring is injective (because the kernel is a proper ideal). Obvious examples of fields are the rational numbers $\mathbf{Q}$ and, for every prime number p, the field $\mathbf{F}_p = \mathbf{Z}/p$ of p elements.

Let us also recall the notion of characteristic, basic in all that follows. In any field F consider the multiples of $1 : 1, 1+1, 1+1+1, \ldots$. We abbreviate these expressions as $1, 2.1, 3.1, \ldots$ and remark that two cases can arise:

(i) $n.1 \neq 0$ *for all* $n > 0$. *Then* F *is said to have characteristic* 0. In this case the homomorphism $n \mapsto n.1$ provides an embedding of $\mathbf{Z}$ in F; since F is a field, we can invert the elements $n.1$ and so form the subfield P generated by 1. This is the *prime subfield* of F; clearly it is isomorphic to $\mathbf{Q}$. It is contained in every subfield of F and so is the least subfield of F.

(ii) For some $n > 0$ we have $n.1 = 0$. Since F has no zerodivisors, it follows easily that the least such n is a prime p, say, which is called the *characteristic* of F in this case. Now $0, 1, 2.1, \ldots, (p-1).1$ already form a subfield P, isomorphic to $\mathbf{F}_p$, which thus is the prime subfield.

We see that the characteristic of F, often written char F, is either 0 or a prime number p, and our results can be summed up as follows:

Theorem 7.1.1. *Every field F has a least subfield P, the prime subfield of F, which is contained in every subfield of F. Either* (i) char $F = 0$ *and* $P \cong \mathbf{Q}$, *or* (ii) char $F = p$, *a prime number, and then* $P \cong \mathbf{F}_p$. ■

Let F be a field and k be a subfield; we shall often describe k as the *ground field*, F as an *extension* of k and write F/k for the field F considered as an extension of k (the risk of confusion with quotient rings $R/\mathfrak{a}$ is small since fields have no non-trivial quotients). Thus F is a k-algebra, in particular it is a vector space over k. Its dimension is called the *degree* of F over k and is written $[F : k]$; this notation will sometimes be used even if F is only a vector space. The dimension is a positive integer or ∞; for example, $[\mathbf{C} : \mathbf{R}] = 2$, $[\mathbf{R} : \mathbf{Q}] = \infty$, and of course for every field k, $[k : k] = 1$. By a *finite* extension one understands an extension of finite degree. We note that for every extension F/k, F and k have the same prime subfield, and hence the same characteristic.

Let F/k be a finite extension, of degree n say. Then there is a basis $u_1, \ldots, u_n$ for F over k and relative to this basis each element $a \in F$ can be uniquely expressed in the form

$$a = \sum \alpha_i u_i, \quad \text{where } \alpha_i \in k. \tag{7.1.1}$$

In order to know F completely we need only know k and the expressions (7.1.1) for the n^2 products $u_i u_j$, but it is not particularly easy to work with these n^2 expressions and we shall soon find simpler ways of describing the extension F/k. A *k-homomorphism* (of extensions of k) is a homomorphism which is k-linear, or equivalently, one which reduces to the identity on k.

The following product formula is a basic tool in the study of field extensions.

Proposition 7.1.2. *Let F/k be a field extension. Given any vector space V over F, we can regard V also as a vector space over k; its dimension $[V : k]$ is finite if and only if both $[V : F]$ and $[F : k]$ are finite, and when this is so, then*

$$[V : k] = [V : F][F : k]. \tag{7.1.2}$$

Proof. Suppose first that F/k is finite, with basis $u_1, \ldots, u_m$ say, and that V is finite-dimensional over F, with basis $v_1, \ldots, v_n$. For clarity we shall use latin subscripts for the range $1, \ldots, m$ of the first basis, and greek subscripts for the range $1, \ldots, n$ of the F-basis of V. Thus the u_i form a basis for F/k and the v_λ form an F-basis of V; we shall prove that the set of products $u_i v_\lambda$ is a k-basis for V. This will prove that $[V : k] = mn$, and (7.1.2) follows in this case.

Let $x \in V$; since V is spanned by the v_λ over F, we have

$$x = \sum a_\lambda v_\lambda \quad \text{for some } a_\lambda \in F. \tag{7.1.3}$$

We express each a_λ in terms of the u_i : $a_\lambda = \sum \alpha_{\lambda i} u_i$, where $\alpha_{\lambda i} \in k$ and insert these values in (7.1.3):

$$x = \sum \alpha_{\lambda i} u_i v_\lambda .$$

This shows that the mn products $u_i v_\lambda$ span V over k. To prove their linear independence, let us assume a relation

$$\sum \alpha_{\lambda i} u_i v_\lambda = 0. \tag{7.1.4}$$

On writing $a_\lambda = \sum \alpha_{\lambda i} u_i$, we see that this is of the form $\sum a_\lambda v_\lambda = 0$, where $a_\lambda \in F$, and hence, by the linear independence of the v_λ we have $a_\lambda = 0$ ($\lambda = 1, \ldots, n$). This means that $\sum \alpha_{\lambda i} u_i = 0$ for all λ, and since the u_i are linearly independent over k, we conclude that $\alpha_{\lambda i} = 0$ for all λ, i. Hence the relation (7.1.4) is trivial and it follows that the $u_i v_\lambda$ are linearly independent over k; this then shows that they form a k-basis for V.

We observe that in order to prove that the $u_i v_\lambda$ are linearly independent we needed only the linear independence of the sets $\{u_i\}$, $\{v_\lambda\}$, not the fact that they were bases. This means that from any m linearly independent elements of F over k and any n linearly independent elements of V over F we can form mn linearly independent elements of V over k. Suppose now that V is of finite dimension N over k; then N is a bound for the number of elements in a linearly independent set of vectors in V over k. Hence $mn \leq N$ and it follows that both $[V : F]$ and $[F : k]$ are finite. Thus $[V : k]$ is finite iff both $[V : F]$ and $[F : k]$ are finite, and when this is so, (7.1.2) holds. ■

We note particularly the case where V is an extension field E of F. Thus whenever we have a tower of fields,

$$k \subseteq F \subseteq E, \tag{7.1.5}$$

then we have the *product formula*:

$$[E : k] = [E : F][F : k], \tag{7.1.6}$$

whenever either side is finite.

As in the case of Lagrange's theorem for groups (Theorem 2.1.3), this has the useful consequence that for any tower of fields as in (7.1.5), the degree $[F : k]$ is a divisor of $[E : k]$. For example, an extension E/k of prime degree has no proper subextensions. However, an extension without proper subextensions need not be of prime degree; examples are easily constructed with the help of Galois theory, as we shall see in Section 7.6.

Consider any field extension F/k. Given elements $\alpha_1, \ldots, \alpha_n$ of F, we write $k[\alpha_1, \ldots, \alpha_n]$ for the *subring* and $k(\alpha_1, \ldots, \alpha_n)$ for the *subfield* generated by $\alpha_1, \ldots, \alpha_n$ over k. For an extension of finite degree we actually have equality:

$$k(\alpha_1, \ldots, \alpha_n) = k[\alpha_1, \ldots, \alpha_n], \tag{7.1.7}$$

for the right-hand side is a finite-dimensional k-algebra without zerodivisors, and so is a field.

Let us take a closer look at a *simple* extension, $n = 1$. Such an extension has the form $k(\alpha)$, where $\alpha \in F$. Consider first the subring $k[\alpha]$; we have a unique k-homomorphism λ from the polynomial ring $k[x]$ to $k[\alpha]$ which maps x to α. This is clearly surjective and it gives rise to the exact sequence

$$0 \to \ker \lambda \to k[x] \to k[\alpha] \to 0. \tag{7.1.8}$$

There are two cases to consider:

(i) λ is injective; it is then an isomorphism: $k[\alpha] \cong k[x]$. In this case we say that α is *transcendental* over k. The field of fractions $k(\alpha)$ of $k[\alpha]$ is called a *purely transcendental* extension of k.

(ii) λ is not injective; in this case α is called *algebraic* over k. For complex numbers these terms are used with reference to the rational numbers $\mathbf{Q}$ as ground field; e.g. $\sqrt{2}$ and $\sqrt{-1}$ are algebraic numbers, while e and π are transcendental.

We observe that in (7.1.8) the kernel of λ represents the set of all polynomials in x which vanish for $x = \alpha$. Thus α is algebraic precisely when it satisfies a non-trivial polynomial equation over k. By (7.1.8) we have $k[\alpha] \cong k[x]/\ker \lambda$; we claim that $k[\alpha]$ is finite-dimensional over k. For $\ker \lambda$, being a non-zero ideal, contains a polynomial $f = c_0 x^n + c_1 x^{n-1} + \ldots + c_n$ which is non-zero, say $c_0 \neq 0$. By definition we have

$$c_0 \alpha^n + \ldots + c_n = 0, \tag{7.1.9}$$

and since we can divide by c_0, this equation shows α^n to be linearly dependent on $1, \alpha, \ldots, \alpha^{n-1}$ over k. On multiplying (7.1.9) by α, we see that α^{n+1} is linearly dependent on $\alpha, \alpha^2, \ldots, \alpha^n$ and hence on $1, \alpha, \ldots, \alpha^{n-1}$. An easy induction shows that all powers of α are linearly dependent on $1, \alpha, \ldots, \alpha^{n-1}$ and since $k[\alpha]$ is spanned by the positive powers of α over k, it follows that $k[\alpha]$ is finite-dimensional over k. By the previous remark we see that $k[\alpha]$ is already a field. When (7.1.9) holds, α is called a *root* of the equation $f = 0$; it is a *zero* (or also a *root*) of f.

To study the kernel of λ more closely, we recall that the polynomial ring $k[x]$ is a principal ideal domain. Explicitly, let $\mathfrak{a} \neq 0$ be an ideal in $k[x]$ and take a non-zero polynomial p of least degree in $\mathfrak{a}$; then any $f \in \mathfrak{a}$ can be written as $f = pq + r$ for polynomials q, r, where $\deg r < \deg p$. Hence $r = f - pq \in \mathfrak{a}$ and by the minimality of $\deg p$, r must vanish, so that $f = pq$ and $\mathfrak{a} = (p)$. Hence every ideal of $k[x]$ is principal and so $\ker \lambda$ is generated by the monic polynomial p (i.e. with highest coefficient 1) of least degree with α as zero. If p has degree n, then $k[\alpha]$ is spanned by $1, \alpha, \ldots, \alpha^{n-1}$ and these elements are linearly independent over k, for any dependence would give an equation of lower degree satisfied by α. Thus $1, \alpha, \ldots, \alpha^{n-1}$ is a basis of $k[\alpha]$ and we see that $[k[\alpha] : k] = n = \deg p$. We note that p may always be taken monic and it is then uniquely determined, for if α satisfied two different monic equations of degree n, it would also satisfy their difference, an equation of lower degree, contradicting the minimality of n. The monic polynomial p of least degree satisfied by α is called the *minimal polynomial* for α over k, and $\deg p$ is called the *degree* of α over k.

We also note that p is *irreducible*, i.e. an atom in $k[x]$, where an *atom* is an element not zero or a unit, which cannot be written as a product of two non-units. For p is not a unit, because $\deg p \geq 1$, and if $p = fg$, then $0 = p(\alpha) = f(\alpha)g(\alpha)$, hence $f(\alpha) = 0$ or $g(\alpha) = 0$, so either f or g has degree n and the other factor has degree 0, i.e. it is a unit.

We summarize these results as follows:

Proposition 7.1.3. *Let E/k be any field extension and $\alpha \in E$. Either*

(i) *α is transcendental over k; then $k(\alpha) \cong k(x)$, where x is an indeterminate, or*
(ii) *α is algebraic over k. In that case $k(\alpha) = k[\alpha]$ and if p is the minimal polynomial for α over k, of degree n, then p is irreducible over k and $k(\alpha) \cong k[x]/(p)$, $[k[\alpha] : k] = n$. Moreover, p is uniquely determined as the monic polynomial of least degree satisfied by α.* ∎

Of course over a larger field p may become reducible; e.g. over $k[\alpha]$, p has the factor $x - \alpha$, and this does not lie in $k[x]$ unless $\alpha \in k$ and p has degree 1.

An extension E/k is called *algebraic* if all its elements are algebraic over k; otherwise it is called *transcendental*; thus a transcendental extension may well contain algebraic elements. It is clear that every finite extension is algebraic, for if $[E : k] = n$, then for any $\alpha \in E$, the extension $k(\alpha)/k$ has a degree dividing n, by Proposition 7.1.2. However, not every algebraic extension is finite, as we shall see in Section 7.3. But there is a converse for finitely generated extensions:

Proposition 7.1.4. *Let E/k be an extension generated by a finite number of algebraic elements over k. Then $[E : k]$ is finite; in particular, the sum and product of algebraic elements are algebraic.*

Proof. Suppose that E is generated by $\alpha_1, \ldots, \alpha_r$ over k. For $r = 0$ the result is clear; when $r > 0$, $[k(\alpha_2, \ldots, \alpha_r) : k] < \infty$ by induction on r and since α_1 is algebraic over k, it is algebraic over $k(\alpha_2, \ldots, \alpha_r)$, so $[E : k(\alpha_2, \ldots, \alpha_r)] < \infty$. Hence by the product formula (Proposition 7.1.2) we find that $[E : k] < \infty$, as claimed. It follows that any element of E is algebraic over k; in particular, this holds for $\alpha_1 + \alpha_2, \alpha_1\alpha_2$. ∎

Corollary 7.1.5. *A field extension is algebraic if it is generated by algebraic elements.*

Proof. Assume that E/k is generated by a set A of algebraic elements and let $c \in E$. Then c lies in the subextension generated by a finite subset of A, say $c \in k(\alpha_1, \ldots, \alpha_r)$, where $\alpha_i \in A$. Since the α_i are algebraic, $k(\alpha_1, \ldots, \alpha_r)$ is finite over k and so c is algebraic over k, as claimed. ∎

As an application of the product formula (7.1.6) we briefly indicate how to prove the impossibility of certain geometrical constructions. In Euclid's *Elements* the only allowable constructions are by ruler and compasses. This enables us to construct from a given length all multiples and submultiples, i.e. all lengths commensurate with a given length. This means that all rational numbers can be constructed.

Next, using the fact that a triangle inscribed in a circle on a diameter as base is right-angled (Thales' theorem), we can extract square roots of differences of squares (by Pythagoras' theorem); hence we can find arbitrary square roots by the formula

$$c = [(c+1)/2]^2 - [(c-1)/2]^2.$$

Each time we extract a square root, we enlarge our field by an extension of degree 2, if at all. Therefore all extensions reached in this way have as degree a power of 2. It follows that a real number α is constructible by ruler and compasses – it 'admits quadrature' – only when its degree over $\mathbf{Q}$ is a power of 2. This condition, that $[\mathbf{Q}(\alpha) : \mathbf{Q}] = 2^m$, though necessary, is not sufficient for constructibility; we shall meet the precise condition later, in Exercise 1 of Section 7.11.

With the help of the above remark several ruler-and-compass constructions proposed in ancient Greece can be shown to be insoluble.

(i) Quadrature of the circle. This is the problem of constructing a square equal in area to the area of a circle of radius 1. What has been said shows that π would have to be algebraic over $\mathbf{Q}$ of degree 2^m, for some $m \geq 0$. In fact it was proved by Ferdinand von Lindemann in 1882 that π is transcendental over $\mathbf{Q}$, so the quadrature of the circle is impossible. The proof that π is transcendental will not be given here, see Jacobson (1985) or Lang (1984).

(ii) Duplication of the cube (Delian problem). The oracle at Delos revealed that to avert a plague, it would be necessary to double the size of a certain altar, built in the shape of a cube. Since a cube of side a has volume a^3, this entailed the solution of the equation $(ax)^3 = 2a^3$, or $x^3 = 2$. This is an equation of degree 3, irreducible over $\mathbf{Q}$, so $2^{1/3}$ does not admit quadrature (it is reported that at first a new altar in the shape of a cube with twice the side of the old altar was built, whereupon the plague got worse).

(iii) Trisection of an angle. Let α be a given angle and write $\alpha = 3\beta$. Then we must solve $\cos(3\beta) = \cos\alpha = \lambda$, say, or $4\cos^3\beta - 3\cos\beta = \lambda$, i.e.

$$4x^3 - 3x - \lambda = 0.$$

For example, for $\lambda = 1/2$ ($\alpha = 60°$) the equation becomes, on putting $y = 2x$,

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

This equation is irreducible over $\mathbf{Q}$, as we see by putting $y = z + 1$ and applying Eisenstein's criterion (Theorem 7.2.7 below) to the resulting equation. It follows that the angle $60°$ cannot be trisected by ruler and compasses.

Of course none of these impossibility proofs affect the practical constructibility, which is possible to any degree of accuracy; e.g. the 'engineer's method' of trisecting an angle involves moving the compasses and so is excluded here. Problems (ii) and (iii) were shown to be insoluble in 1837 by Pierre Wantzel.

Exercises

1. Find a basis for $\mathbf{Q}(\sqrt{2}, \sqrt{3})/\mathbf{Q}$. Find the minimal polynomial for $\sqrt{2}+\sqrt{3}$ and use it to show that $\mathbf{Q}(\sqrt{2}, \sqrt{3}) = \mathbf{Q}(\sqrt{2}+\sqrt{3})$.
2. Let $k(x)$ be the field of rational functions in an indeterminate x. Show that every element of $k[x]$ which is not in k is transcendental over k.
3. Show that every automorphism of a field leaves the prime subfield elementwise fixed.
4. Let F/k be a field extension and σ be a homomorphism of F into a field F'. Show that $[F^\sigma : k^\sigma] = [F : k]$.
5. Show that an endomorphism of F/k, as k-algebra, is a field endomorphism of F leaving k elementwise fixed.
6. Show that if F/k is a field extension of finite degree, then every endomorphism of F/k is an automorphism. Give examples to show that this may fail to hold for arbitrary extensions; does it hold for all algebraic extensions?
7. Show that a field extension is algebraic iff every subalgebra is a field.
8. Show that for $k \subseteq E \subseteq F$, if F/E and E/k are algebraic, then so is F/k.
9. Show that $k(x)/k((x^3+2)/(x^2+x-1))$, where x is an indeterminate, is algebraic and find a basis.
10. Verify in detail that the magnitudes constructible by ruler and compasses are just those obtainable using only square roots and rational operations.
11. Give an algebraic proof that every angle can be bisected by using only ruler and compasses.
12. Show that in any field extension F/k, the set F_0 of elements of F that are algebraic over k is a subfield.

7.2 Splitting Fields

We have seen in Section 7.1 that every algebraic element over a field k is a zero of a polynomial. Conversely, suppose that we are given a polynomial f over k; will f have a zero in k, or in a suitable extension of k? There may be no zero in k itself; for example the real polynomial x^2+1 has no real zeros, but it does have a zero in the field of complex numbers. We shall see that essentially the same method used to construct $\sqrt{-1}$ works in the general case.

Theorem 7.2.1 (Kronecker). *Let f be a polynomial of positive degree over a field k. Then there exists an extension E/k in which f has a zero.*

Proof. The quotient ring $R = k[x]/(f)$ is a non-zero finite-dimensional k-algebra, its dimension being $\deg f$. Let $\mathfrak{a}$ be a proper ideal of maximal dimension in R; then $\mathfrak{a}$ is a maximal ideal, hence $E = R/\mathfrak{a}$ is a field, and still a k-algebra. Thus we have a homomorphism $k \to E$, necessarily injective. By identifying k with its image in E we may regard the latter as an extension of k. If $x \mapsto \alpha$ in the homomorphism $k[x] \to R \to E$, then $f(\alpha) = 0$, so f has a zero in E. ∎

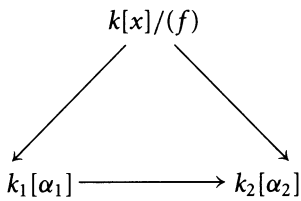
This proof, depending on the construction of a maximal ideal in $k[x]/(f)$, is not very explicit; in particular, we should like to know when the ideal (f) will itself be maximal, and to what extent the construction is unique. Both questions are answered by the next result. We recall that an ideal $\mathfrak{a}$ in a commutative ring R is said to be *prime* if $\mathfrak{a} \neq R$ and $xy \in \mathfrak{a} \Rightarrow x \in \mathfrak{a}$ or $y \in \mathfrak{a}$; a *prime element* is an element which generates a prime ideal. In a principal ideal domain (more generally, in a unique factorization domain) an element is prime iff it is irreducible, as is easily checked (see Section 10.2 below).

Proposition 7.2.2. *Let f be a non-zero polynomial over a field k. Then $k[x]/(f)$ is a field if and only if f is irreducible.*

Given two isomorphic fields k_1, k_2, let f_1 be an irreducible polynomial over k_1 and f_2 be the corresponding polynomial over k_2. Suppose that f_i has a zero α_i in an extension E_i of k_i $(i = 1, 2)$. Then the isomorphism from k_1 to k_2 can be extended in just one way to an isomorphism from $k_1(\alpha_1)$ to $k_2(\alpha_2)$ such that α_1 maps to α_2.

Proof. We know that $R = k[x]/(f)$ is finite-dimensional over k, so by Proposition 5.5.1, it will be a field iff it is an integral domain, which is the case iff (f) is a prime ideal, i.e. iff f is prime, which is the case precisely when f is irreducible.

For the second part, let k be a field isomorphic to k_1 and k_2 and let $\varphi_i : k \to k_i$ be two isomorphisms such that $\varphi_1^{-1}\varphi_2$ is the given isomorphism between k_1 and k_2. Further denote by f_i the polynomial over k_i which corresponds to f over k under the isomorphism φ_i. The isomorphism φ_i extends to a unique homomorphism $\varphi_i' : k[x] \to k_i[\alpha_i]$ that maps x to α_i. Since $f_i(\alpha_i) = 0$, this homomorphism maps $f(x)$ to 0 and so can be factored by the natural homomorphism $k[x] \to k[x]/(f)$ to give a homomorphism $\lambda_i : k[x]/(f) \to k_i[\alpha_i]$, by the factor theorem (cf. Theorem 3.3.3).



Clearly this map is surjective, and since f is irreducible, $k[x]/(f)$ is a field. Hence $\ker \lambda_i = 0$, so λ_i is injective and so is an isomorphism. The residue class of $x \pmod f$ maps to α_i under this isomorphism, hence $\lambda_1^{-1}\lambda_2$ is an isomorphism from $k_1[\alpha_1]$ to $k_2[\alpha_2]$ extending $\varphi_1^{-1}\varphi_2$ and mapping α_1 to α_2. It is plainly unique since it is prescribed on a generating set of $k_1[\alpha_1]$. ■

With the help of this result we see more clearly what happens in the construction of a root for a given equation $f = 0$ (Theorem 7.2.1). Given a polynomial f over k, we split off an irreducible factor p, say $f = pg$. Then $F = k[x]/(p)$ is a field, and denoting the residue class of x in F by α, we have $p(\alpha) = 0$, hence $f(\alpha) = p(\alpha)g(\alpha) = 0$.

For example, over the real field **R**, the polynomial $x^2 + 1$ is irreducible, and it leads to the construction of the complex numbers in the form $\mathbf{C} = \mathbf{R}[x]/(x^2 + 1)$.

We note that the construction of an extension E in which a given polynomial f has a zero is not usually unique, but depends on which irreducible factor of f is chosen. Thus if we are given $f = x^4 - 4$ over **Q**, we have the factorization

$$x^4 - 4 = (x^2 + 2)(x^2 - 2),$$

where the factors on the right are irreducible over **Q**, and correspondingly there are two non-isomorphic extensions of **Q** containing a root of $x^4 - 4 = 0$. On the other hand, the equation $x^4 + 1 = 0$ is irreducible over **Q** and so all extensions generated by a root of this equation over **Q** are isomorphic.

Over the complex numbers every non-constant polynomial has a zero; this is the content of the 'fundamental theorem of algebra', which will be proved in Section 7.3. It follows that every irreducible polynomial over **C** is linear, by the remainder theorem. Now whatever the field, by repeating Kronecker's construction (Theorem 7.2.1) we can split any polynomial into linear factors. We shall now carry out this process in detail, but first we need a definition.

Given a polynomial f over a field k, suppose that in some extension E/k, f can be expressed as a product of linear factors

$$f = a_0(x - \alpha_1)\dots(x - \alpha_n) \quad (\alpha_1, \dots, \alpha_n \in E, a_0 \in k^\times). \tag{7.2.1}$$

Then we shall say: f *splits* over E, and E is called a *splitting field* of f over k. If f splits over E but over no smaller field, then E is called a *minimal splitting field* of f over k. Given any splitting field E of f over k, we need only take $k(\alpha_1, \dots, \alpha_n)$, where the α_i are as in (7.2.1), to obtain a minimal splitting field. We now show that splitting fields always exist and the minimal ones are unique up to isomorphism.

Theorem 7.2.3. *Let k be a field. For any non-constant polynomial f over k there exists a splitting field. If $\deg f = n$, then any minimal splitting field E satisfies $[E : k] \leq n!$.*

Further, let k' be a field isomorphic to k, f' be the polynomial over k' corresponding to f over k and E' be a minimal splitting field for f' over k'. Then the given isomorphism between k and k' can be extended to an isomorphism between E and E'.

Proof. Existence: We begin by factorizing f over k:

$$f = p_1 p_2 \dots p_r, \tag{7.2.2}$$

where each p_i is irreducible over k. If each p_i is linear, then k itself is a splitting field of f; otherwise we adjoin a zero of some non-linear factor p_i, using Theorem 7.2.1. In the resulting extension we can write p_i as the product of a linear factor and another factor, so we have increased the number of linear factors in a complete factorization of f. If there is another non-linear factor left, we repeat the procedure; after a finite number of steps we obtain an extension in which f splits into linear factors and this is the desired splitting field. To estimate the degree of the extension, at the i-th stage we have at least i linear factors, so the irreducible factor to be split has degree at most $n - i$, and the degree of the extension increases by a factor of at most $n - i$. We start at stage 0, so the degree is at most $n(n - 1)\dots 2.1 = n!$ as claimed.

Uniqueness: We use induction on the degree $[E : k]$. If this is 1, then $E = k$, $E' = k'$ and there is nothing to prove. If $[E : k] > 1$, then some p_i in (7.2.2), say p_1, has degree greater than 1, and f' has a corresponding factorization $f' = p_1' \dots p_r'$ over k'. Take any zero α of p_1 in E and any zero α' of p_1' in E'; by Proposition 7.2.2, $k[\alpha] \cong k'[\alpha']$ and $[E : k] = [E : k[\alpha]][k[\alpha] : k] > [E : k[\alpha]]$. Clearly E is a minimal splitting field for f' over $k'[\alpha']$; hence by induction, $E \cong E'$ under the isomorphism mapping $k[\alpha]$ to $k'[\alpha']$. ∎

Example 1. $x^4 + 1 = 0$. Let θ be a root over $\mathbf{Q}$; then so are $-\theta, \theta^{-1}$ and $-\theta^{-1}$, and these roots are distinct, for $\theta \neq -\theta$, because $\theta \neq 0$ and if $\theta = \pm\theta^{-1}$, then $\theta^2 = \pm 1$, hence $\theta^4 + 1 = 2 \neq 0$. Thus we have a complete factorization

$$x^4 + 1 = (x - \theta)(x + \theta)(x - \theta^{-1})(x + \theta^{-1})$$

over $\mathbf{Q}(\theta)$; the latter is therefore a minimal splitting field. We remark that every automorphism of $\mathbf{Q}(\theta)$ can be described by a permutation of the roots of $x^4 + 1 = 0$. The group of all these permutations is the Galois group of the extension $\mathbf{Q}(\theta)/\mathbf{Q}$, to be studied in Section 7.6. The same holds over $\mathbf{F}_p$, the field of p elements, when $p \neq 2$. For $p = 2$, $x^4 + 1 = (x + 1)^4$ splits already over $\mathbf{F}_2$.

Example 2. $x^3 - 2 = 0$. Over the complex numbers we have

$$x^3 - 2 = (x - \alpha)(x - \omega\alpha)(x - \omega^2\alpha),$$

where $\alpha = \sqrt[3]{2}$, $\omega = (-1 + \sqrt{-3})/2$. Clearly $\mathbf{Q}(\alpha, \omega\alpha) = \mathbf{Q}(\omega, \alpha)$ is a splitting field, but neither $\mathbf{Q}(\alpha)$ nor $\mathbf{Q}(\omega)$ will do, for $\mathbf{Q}(\alpha)$ is real and $\mathbf{Q}(\omega)$ does not contain α. We note that $[\mathbf{Q}(\alpha, \omega) : \mathbf{Q}] = 6$, for ω, α have degrees 2, 3 respectively over $\mathbf{Q}$; hence $[\mathbf{Q}(\alpha, \omega)) : \mathbf{Q}]$ is a multiple of 6 and it equals 6 because it is spanned by $1, \alpha, \alpha^2, \omega, \omega\alpha, \omega\alpha^2$ over $\mathbf{Q}$.

Splitting fields naturally lead to an important class of extensions.

Definition. A field extension E/k is said to be *normal* if it is algebraic and every irreducible polynomial over k which has a zero in E splits completely in E.

It should be noted that normality is a property of extensions, not of fields. If $k \subseteq F \subseteq E$ and E/k is normal, then E/F is also normal, but F/k need not be. Thus in Example 2 above, $\mathbf{Q}(\alpha, \omega)/\mathbf{Q}$ is normal and so is $\mathbf{Q}(\alpha, \omega)/\mathbf{Q}(\alpha)$ but $\mathbf{Q}(\alpha)/k$ is not, because $x^3 - 2$ has a zero in $\mathbf{Q}(\alpha)$ without splitting completely. Our next result provides a description of the finite normal extensions.

Proposition 7.2.4. *A finite extension E/k is normal if and only if it is a minimal splitting field of a polynomial over k.*

Proof. Suppose that E is a minimal splitting field for a polynomial f over k and denote the zeros of f in E by $\alpha_1, \dots, \alpha_n$. Given any irreducible polynomial p over k, which has a zero β in E, we have to show that p splits in E. Let β' be another zero of p in some extension of E; we shall show that $\beta' \in E$. Since p is irreducible over k, we have $k(\beta) \cong k(\beta')$, by Proposition 7.2.2. Now E is clearly also a splitting

field of f over $k(\beta)$, while $E(\beta')$ is a splitting field of f over $k(\beta')$, and as minimal splitting fields they are isomorphic and $[k(\beta) : k] = [k(\beta') : k]$, $[E : k(\beta)] = [E(\beta') : k(\beta')]$. Hence

$$\begin{aligned}[E : k] &= [E : k(\beta)][k(\beta) : k] \\ &= [E(\beta') : k(\beta')][k(\beta') : k] \\ &= [E(\beta') : k].\end{aligned}$$

But E is a subspace of $E(\beta')$; since both have the same degree over k, we have $E = E(\beta')$ and so $\beta' \in$E, as claimed.

Conversely, let E/k be a finite normal extension. We can write $E = k(\alpha_1, \ldots, \alpha_n)$, taking e.g. a basis of E over k for the α_i. Let p_i be the minimal polynomial of α_i over k; by hypothesis each p_i splits into linear factors over E. Hence the same is true of $f = p_1 \ldots p_n$ and E is generated by the zeros of f; hence it is a minimal splitting field of f over k. ∎

In exactly the same way one can show that a general extension is normal iff it is a minimal splitting field for a set of polynomials.

Given any finite extension F/k, say $F = k(\alpha_1, \ldots, \alpha_n)$, let p_i be the minimal polynomial for α_i over k and put $f = p_1 \ldots p_n$. Any normal extension of k containing F must be a splitting field of f over k; if E is a minimal splitting field, it is normal over k and is contained in any normal extension containing F. Thus E/k may be described as the least normal extension containing F/k; it is called the *normal closure* of F/k. The construction shows that it is unique up to F-isomorphism. It may also be described as the field obtained by adjoining to F all the zeros of all irreducible polynomials over k which have a zero in F. Let us note another useful property of normal extensions:

Corollary 7.2.5. *Given a tower of finite extensions $k \subseteq F \subseteq E$, if E/k is normal, then any k-homomorphism $\varphi : F \to E$ extends to a k-automorphism of E.*

Proof. By Proposition 7.2.4, E is a minimal splitting field of a polynomial f over k, hence also over F. Write $F' = F\varphi$; then E is also a minimal splitting field of f over F'. By Theorem 7.2.3, the isomorphism $\varphi : F \to F'$ can be extended to an automorphism of E, which must leave k pointwise fixed, because φ does. ∎

Two elements or two subextensions of a normal extension E/k are said to be *conjugate* if there is a k-automorphism of E which transforms one into the other. We observe that a subextension is normal precisely when it is equal to all its conjugates (see Corollary 7.6.4 below). This definition shows the truth of

Corollary 7.2.6. *In a finite normal extension F/k the conjugates of a given element are permuted transitively by the k-automorphisms of F.* ∎

We conclude this section with a useful test for irreducibility, due to Gotthold Eisenstein.

Theorem 7.2.7 (Eisenstein's criterion). *Given* $f = a_0 + a_1x + \ldots + a_nx^n \in \mathbf{Z}[x]$, *and any prime* p, *suppose that* a_n *is prime to* p, $a_0, \ldots, a_{n-1}$ *are divisible by* p *but* a_0 *is not divisible by* p^2. *Then* f *is irreducible over* $\mathbf{Q}$.

Proof. Suppose that f is reducible over $\mathbf{Q}$ and hence (by Gauss's Lemma, Lemma 7.7.1 below) over $\mathbf{Z}$, say $f = gh$, where g, h are of positive degrees r, s respectively $(r + s = n)$ with integer coefficients: $g = \sum b_ix^i$, $h = \sum c_jx^j$. Then $a_0 = b_0c_0$, so just one of b_0, c_0 is divisible by p, say $p|b_0$. But $a_n = b_rc_s$ is not divisible by p, so neither are b_r, c_s. Thus the first but not the last coefficient of g is divisible by p. Let b_i be the first coefficient of g not divisible by p; then $i > 0$ and

$$a_i = b_ic_0 + b_{i-1}c_1 + \ldots + b_0c_i.$$

Modulo p this reads $b_ic_0 \equiv 0 \pmod{p}$, which is a contradiction since b_i, c_0 are both prime to p. Hence f must be irreducible. ■

Exercises

1. Find the degree of a minimal splitting field of f over $\mathbf{Q}$ in the following cases: (i) $f = x^4 - 1$, (ii) $f = x^4 + 1$, (iii) $f = x^4 + 2$, (iv) $f = x^4 + 4$.
2. Find the degree of a minimal splitting field of $x^6 + 1$ over $\mathbf{Q}$ and over $\mathbf{F}_2$.
3. Show that a minimal splitting field over k for a polynomial of degree n is generated over k by any $n - 1$ of its zeros.
4. Show that $x^4 - 2x^2 - 2$ is irreducible over $\mathbf{Q}$, and find two pairs of zeros which generate non-isomorphic extensions.
5. Find the normal closure of $\mathbf{Q}(8^{1/n})$ over $\mathbf{Q}$.
6. Show that if E and F are normal extensions of k within a field U, then EF, the subfield generated by E and F, and $E \cap F$ are normal over k.
7. Show that the field generated by a root of $x^4 - 2 = 0$ over $\mathbf{Q}$ is not normal. Deduce that a normal extension of a normal extension need not be normal.
8. Show that E is a normal closure of a separable extension F/k iff $E \otimes_k F$ is a direct product of $[F : k]$ fields isomorphic to E.

7.3 The Algebraic Closure of a Field

In Section 7.2 we defined a splitting field for any polynomial. More generally let $\mathscr{F}$ be any set of polynomials over a field k. An extension E/k will be called a *splitting field* for the set $\mathscr{F}$ if every $f \in \mathscr{F}$ splits completely over E. As before we can define minimal splitting fields and it is easily seen that E is a minimal splitting field for $\mathscr{F}$ iff each $f \in \mathscr{F}$ splits over E and E is generated over k by the set of all zeros of all the members of $\mathscr{F}$. If $\mathscr{F}$ is finite, say $\mathscr{F} = \{f_1, \ldots, f_r\}$, then we can replace it by the product $f = f_1 \ldots f_r$. Any splitting field for $f_1, \ldots, f_r$ over k is just a splitting field for f, and we are in the case considered in Section 7.2. As we noted after Proposition 7.2.4, an extension E/k is normal iff it is a minimal splitting field for some set of polynomials over k.

A field k is said to be *algebraically closed* if every polynomial over k splits already in k. This can also be expressed by saying that k is its own splitting field for the set of all polynomials over it.

Every algebraically closed field is infinite. This follows by arguments used for Euclid's theorem to show the existence of an infinity of prime numbers:

Theorem 7.3.1. *Every algebraically closed field is infinite.*

Proof. If k is a finite field, consider the polynomial

$$f = 1 + \prod_{a \in k} (x - a).$$

f has positive degree and $f(a) = 1$ for all $a \in k$, so f has no zeros in k, hence k cannot be algebraically closed. ∎

We shall see later (in Section 7.8) that the product $\prod (x - a)$ is $x^q - x$, where q is the number of elements of the finite field.

Let k be any field. By an *algebraic closure* of k we understand an extension E/k which is algebraic and algebraically closed. Our aim in the section will be to show that every field k has an algebraic closure and this is unique up to isomorphism over k. We begin by establishing the isomorphism property for splitting fields.

Proposition 7.3.2. *Let k be any field and $\mathscr{F}$ be a set of polynomials over k. Then any two minimal splitting fields of $\mathscr{F}$ over k are isomorphic.*

For the finite case this was proved in Theorem 7.2.3. In the general case we shall give two proofs.

First proof. Denote by E, E' two minimal splitting fields. Let $\mathscr{F}$ be indexed by Λ and for any subset I of Λ write E_I for the subfield of E generated by the zeros of all $f_\lambda (\lambda \in I)$ over k; in particular $E_\Lambda = E$. Now consider the set of all pairs (E_I, φ_I), where $\varphi_I : E_I \to E'$ is a k-homomorphism; this set is partially ordered by the rule: $(E_I, \varphi_I) \leq (E_J, \varphi_J)$ iff $E_I \subseteq E_J$, $\varphi_I = \varphi_J | E_I$. Our set is clearly inductive: given any chain $\{(E_I, \varphi_I)\}$, we take as its upper bound the union $\cup E_I$ with the union of the φ_I as homomorphism. As upper bound of the empty chain we have $(k, 1)$. By Zorn's lemma there is a maximal element (E_J, φ_J). If $J \neq \Lambda$, take $\lambda \in \Lambda \backslash J$ and let $E_{J'}$ be the minimal splitting field of f_λ in E over E_J. By Proposition 7.2.2 this exists and the homomorphism $\varphi_J : E_J \to E'$ can be extended to a homomorphism of $E_{J'}$ into E'. But this contradicts the maximality of (E_J, φ_J); hence $J = \Lambda$ and we have a homomorphism $E \to E'$. Now every $f \in \mathscr{F}$ splits over E and the zeros of all these polynomials are mapped by φ again to zeros of all the $f \in \mathscr{F}$ in E'; but these zeros generate E' over k, hence φ is an isomorphism. ∎

The second proof is shorter and uses tensor products of fields, which were defined in Section 5.4.

Second proof. With E, E' as before form the tensor product $E \otimes E'$ over k. This is a non-zero commutative ring. Its quotient by a maximal ideal $\mathfrak{p}$ is a field

$F = (E \otimes E')/\mathfrak{p}$ and we have homomorphisms of E, E' into F. Denote their images by E_1, E_1' respectively. Each of E_1, E_1' as a minimal splitting field is generated by the zeros of all the polynomials in $\mathscr{F}$; hence $E_1 = E_1'$ and it follows that $E \cong E'$. ∎

Now it is not hard to show the existence of a splitting field. Suppose first that the set of polynomials is countable, $f_1, f_2, \ldots$. We put $E_0 = k$ and define E_n recursively as a minimal splitting field of f_n over E_{n-1}. Then $E_{n-1} \subseteq E_n$ and the union of all the E_n, formed as in Proposition 3.2.9, is the required field. This method can be adapted to deal with the general case. Thus given a family of polynomials indexed by any set Λ, take any finite subset I of Λ and denote by f_I the product of the f_λ for $\lambda \in I$. By Theorem 7.2.3 we have a minimal splitting field E_I of f_I over k and for $I \subseteq J$ there is an embedding $E_I \to E_J$. By making an obvious identification we can regard E_I as a subfield of E_J and we thus have a set $\{E_I\}$ of fields, partially ordered by inclusion. Moreover, this set is *directed*, i.e. given E_I, E_J there is a finite set $K(= I \cup J)$ such that $E_I \subseteq E_K$, $E_J \subseteq E_K$. As a consequence we can define a field structure on the union E of all the E_I such that each E_I is a subfield. Given $x, y \in E$, say $x \in E_I$, $y \in E_J$, we can find E_K to contain E_I, E_J and in E_K we can form $x + y, xy, x^{-1}$ (if $x \neq 0$); the verification that E forms a field is straightforward and may be left to the reader. We sum up the result as follows:

Proposition 7.3.3. *For every set of polynomials over a field k there is a minimal splitting field, unique up to isomorphism.* ∎

Remark. We have shown that any two minimal splitting fields of a given set of polynomials are isomorphic, by an isomorphism which reduces to the identity on the ground field. However, this isomorphism is by no means unique; in fact, the study of the different possible isomorphisms constitutes the subject of Galois theory, which will be treated in Section 7.6 (and in the infinite case, in Chapter 11).

We can now prove the existence of an algebraic closure.

Theorem 7.3.4. *Let k be a field. Then a minimal splitting field Ω for the set of all polynomials over k is an algebraic closure of k.*

Proof. Since Ω is generated by algebraic elements over k, it is algebraic, by Corollary 7.1.5. Let f be a polynomial over Ω and denote by E the subfield of Ω generated over k by the coefficients of f. Since E is finitely generated over k, it has finite degree. Let E' be a minimal splitting field of f over E; then $[E' : k] = [E' : E][E : k]$, and this is again finite. Hence the zeros of f in E' are zeros of some polynomial g over k; but g splits over Ω, so these zeros lie in Ω and f splits over Ω, as required. ∎

Although Theorem 7.3.4 provides algebraically closed fields in profusion, it does not tell us whether a given field such as **C**, the complex numbers, is algebraically closed. As already mentioned, this is the case, and there are many proofs of this fact. A very simple one, using complex function theory, is based on Liouville's theorem: if a polynomial f has no complex zero, then $f(z)^{-1}$ is finite throughout the complex plane and bounded as $z \to \infty$, hence (by Liouville's theorem) a

constant. Other more topological proofs are based on the notion of the degree of a mapping. If f does not vanish, then the map

$$z \mapsto f(z) \tag{7.3.1}$$

of the 2-sphere (Riemann sphere) into itself omits at least one point from the image, and hence can be deformed into the constant map. But if f has degree n (as a polynomial), this map can also be deformed into the map $z \mapsto z^n$. This would mean that the latter map can be deformed into the constant map, which contradicts the fact that the degree is preserved by deformation.

Whatever method is chosen, the completeness of the complex numbers has to be used at some stage. In Section 8.8 we shall present another proof, which is more algebraic and uses a minimum of topological properties of the real or complex numbers.

Exercises

1. Show that the algebraic closure of a countable field is again countable.
2. Let k be a field and F/k be an algebraic extension. Show that if every finite algebraic extension of k admits a k-homomorphism into F, then F is an algebraic closure of k. (Hint. Take any α algebraic over F and form the normal closure of $k(\alpha)$.)
3. Show that over a finite field there are irreducible polynomials of arbitrarily high degree.

7.4 Separability

Let k be a field of prime characteristic p. The mapping

$$x \mapsto x^p \tag{7.4.1}$$

is an endomorphism of k, for we have

$$(xy)^p = x^p y^p,\ 1^p = 1,\ (x+y)^p = x^p + y^p. \tag{7.4.2}$$

Here the last equation holds because the binomial coefficient $\binom{p}{r}$ is divisible by p for $1 \leq r \leq p-1$. The mapping (7.4.1) is sometimes called the *Frobenius mapping*; as endomorphism of a field it is necessarily injective. If it is also surjective, and hence an automorphism, the field k is said to be *perfect*. Thus k is perfect iff every element is a p-th power, where $p = \operatorname{char} k$; in addition, every field of characteristic 0 is perfect, by definition. As an example of perfect fields of non-zero characteristic we have

Theorem 7.4.1. *Every finite field is perfect.*

Proof. If k is finite, then any injective mapping of k into itself must be bijective, by Dirichlet's box principle (see Section 1.1). ∎

As an example of an imperfect field consider $k(x)$, the field of rational functions in an indeterminate x over a field of prime characteristic p. Clearly the indeterminate x is not a p-th power in $k(x)$. If we write k^p for the image of k under the Frobenius mapping, then it is easily verified that

$$k(x)^p = k^p(x^p). \tag{7.4.3}$$

For example, if k is a finite field, then $k^p = k$ and $k(x)^p$ consist of all rational functions in x^p.

In order to study extensions of imperfect fields we need to consider the multiplicities of zeros of polynomials more closely. Let f be a monic polynomial of degree n over a field k. Over a minimal splitting field E of f we have

$$f = (x - \alpha_1)^{m_1} \dots (x - \alpha_t)^{m_t},$$

where $\alpha_1, \dots, \alpha_t$ are the distinct zeros of f, with α_i of multiplicity m_i. Let E' be another minimal splitting field of f; this is isomorphic to E, by Theorem 7.2.3, and if $\alpha_i \mapsto \alpha_i'$ under the isomorphism, then we have in E'

$$f = (x - \alpha_1')^{m_1} \dots (x - \alpha_t')^{m_t}.$$

Thus the multiplicities are independent of the choice of the splitting field.

We can test f for multiple zeros by finding its formal derivative f'. We recall that for $f = a_0 + a_1 x + \dots + a_n x^n$ this is defined as $f' = a_1 + 2a_2 x + \dots + n a_n x^{n-1}$. We also recall the familiar rules (which suffice to define it uniquely): $(f+g)' = f' + g'$, $(af)' = af'(a \in k)$, $(fg)' = f'g + fg'$, $x' = 1$. In terms of the derivative we have the following test for multiple zeros, which can be carried out within the ground field (because finding the highest common factor (f, g) of two polynomials f, g is a rational operation):

Proposition 7.4.2. *Let k be any field and $f \in k[x]$, where $\deg f > 0$. Then the zeros of f (in some splitting field of f) are simple if and only if f is prime to its derivative f'.*

Proof. The result follows from the observation that for any α in any extension of k,

$$f = f(\alpha) + (x - \alpha) f'(\alpha) + (x - \alpha)^2 g,$$

hence $(x - \alpha)^2 | f$ iff $f(\alpha) = f'(\alpha) = 0$, i.e. f and f' have $(x - \alpha)$ as a common factor. ∎

A polynomial is said to be *separable* if all its zeros (in some splitting field) are simple. If we apply Proposition 7.4.2 to an irreducible polynomial we obtain

Corollary 7.4.3. *An irreducible polynomial f over a field k is separable unless $f' = 0$. In particular, over a field of characteristic 0 every irreducible polynomial is separable.*

Proof. Assume that $f' \neq 0$ and put $d = (f, f')$. Then $\deg d \leq \deg f' < \deg f$; hence d, as a proper factor of an irreducible polynomial f, has degree 0, i.e. $d = 1$ and so f is prime to f'; this ensures that f is separable, by Proposition 7.4.2. If $\operatorname{char} k = 0$, then

$f' = 0$ means that f is of degree 0, contradicting the irreducibility. This proves the second part. ∎

This tells us all we need to know in characteristic 0, but the case of prime characteristic is more complicated.

Lemma 7.4.4. *Let k be a field of prime characteristic p. For any polynomial f over k, $f' = 0$ if and only if $f(x) = g(x^p)$ for some polynomial g.*

Proof. Let $f = \sum a_i x^i$; if $f' = 0$, then $ia_i = 0$ and it follows that $a_i = 0$ for any i prime to p. So f must have the form

$$f = a_0 + a_p x^p + a_{2p} x^{2p} + \ldots + a_{sp} x^{sp} = g(x^p). \tag{7.4.4}$$

Conversely it is clear that any f of the form (7.4.4) satisfies $f' = 0$. ∎

To illustrate the lemma we remark that $x^p - a$ has zero derivative, and in fact if $a \notin k^p$, then this polynomial has no zero in k, for over its splitting field it takes the form $(x - \alpha)^p$, where $a = \alpha^p$.

An element α of an algebraic extension F/k is called *separable* over k if its minimal polynomial over k is separable. If every element of F is separable over k we call F/k a *separable extension*. With the help of perfect fields we can describe separable extensions.

Proposition 7.4.5. *Over a perfect field every algebraic extension is separable. Conversely, if every algebraic (or even every finite) extension of k is separable, then k is perfect.*

Proof. For characteristic 0 there is nothing to prove: in that case every field is perfect and every algebraic extension is separable. So we may take the characteristic to be a prime p.

Let k be a perfect field and f be any irreducible polynomial over k. If f is not separable, then $f' = 0$, hence $f(x) = g(x^p)$, say

$$f = a_0 + a_1 x^p + \ldots a_r x^{rp}.$$

By hypothesis k is perfect, so $a_i = b_i^p$ for some $b_i \in k$, and now

$$\begin{aligned} f &= b_0^p + b_1^p x^p + \ldots + b_r^p x^{rp} \\ &= (b_0 + b_1 x + \ldots + b_r x^r)^p; \end{aligned}$$

but this contradicts the irreducibility of f. Hence every irreducible polynomial is separable and it follows that every algebraic extension is separable.

Conversely, if every finite extension is separable, take $a \in k$ and consider a splitting field of $x^p - a$. If b is a zero, then $x^p - a = (x - b)^p$; thus all zeros coincide. It follows that an irreducible factor can only have one zero and so must be linear. Thus $b \in k$, and $a = b^p \in k^p$; this shows k to be perfect. ∎

Exercises

1. Show that every finite extension of a perfect field is perfect.
2. Show that if f is irreducible over k, then all its zeros have the same multiplicity.
3. Show that over a field of characteristic p, $x^p - a$ is either irreducible or a p-th power of a linear polynomial.
4. Show that a field of characteristic p cannot have n distinct n-th roots of 1 unless n is prime to p.
5. Let k be a field of prime characteristic p. Show that if an element α in an extension of k is separable over $k(\alpha^p)$, then $\alpha \in k(\alpha^p)$. Deduce another proof of the converse part of Proposition 7.4.5.
6. Let f be a polynomial over a field of characteristic 0. Show that f has a zero which is at least m-fold iff $f, f', \ldots, f^{(m-1)}$ have a common factor. Verify that this still holds for characteristic p if $p \geq m$; what happens if $p < m$?
7. Let F/k be any extension of prime characteristic p and write

$$k_0 = \{\alpha \in F | \alpha^{p^r} \in k \quad \text{for some } r \geq 0\}.$$

Show that k_0 is a subfield of F and if F is perfect, then so is k_0, but no smaller field containing k is perfect. Show that any automorphism of F/k leaves k_0 elementwise fixed.

7.5 Automorphisms of Field Extensions

Galois theory may be described as the study of field extensions by means of their automorphisms. A basic aim is to show that under suitable conditions an extension of degree n has just n automorphisms, and in the next section we shall find conditions for equality. Although Galois developed his theory to deal with roots of equations, the theory was expressed by Dedekind in terms of field extensions, by means of two key lemmas on automorphisms. The first gives an upper bound on the number of k-automorphisms of a finite extension of k; we shall state the result more generally for homomorphisms, since it is no harder to prove in this form.

Consider any finite family of homomorphisms $\sigma_i : E \to E'$ between fields, regarded as k-algebras. It will be convenient to write such homomorphisms as exponents. We obtain a k-linear mapping from E to E' by forming a linear combination of these homomorphisms with coefficients in E':

$$\sum_i \sigma_i \alpha_i : x \mapsto \sum_i x^{\sigma_i} \alpha_i \quad (x \in E, \alpha_i \in E').$$

Of course this need not be a homomorphism. We claim that it is a non-zero mapping except in trivial cases, namely (i) if all the α_i vanish, or (ii) if $\sigma_i = \sigma_j$ and $\alpha_i = -\alpha_j$.

Lemma 7.5.1 (Dedekind's lemma). *Any set of distinct homomorphisms of a field E into another field E' is linearly independent over E'.*

Proof. Let $\{\sigma_i\}$ be a family of distinct homomorphisms. If they are linearly dependent, let us take a minimal linearly dependent subset, $\sigma_1, \ldots, \sigma_r$ say, where σ_1 is linearly dependent on $\sigma_2, \ldots, \sigma_r$:

$$x^{\sigma_1} = \sum_2^r x^{\sigma_i}\alpha_i \quad \text{for all } x \in E, \text{ where } \alpha_i \in E'. \tag{7.5.1}$$

Replace x by xy in (7.5.1):

$$x^{\sigma_1}y^{\sigma_1} = \sum_2^r x^{\sigma_i}y^{\sigma_i}\alpha_i. \tag{7.5.2}$$

Now multiply (7.5.1) by y^{σ_1} and subtract the result from (7.5.2):

$$\sum_2^r x^{\sigma_i}(y^{\sigma_i} - y^{\sigma_1})\alpha_i = 0.$$

This holds for all $x \in E$, but by hypothesis $\sigma_2, \ldots, \sigma_i$ are linearly independent, so $(y^{\sigma_i} - y^{\sigma_1})\alpha_i = 0$. Since $\sigma_i \neq \sigma_1$, it follows that $\alpha_i = 0$ for all $i \neq 1$, and (7.5.1) reduces to the form $x^{\sigma_1} = 0$; putting $x = 1$, we find $1 = 0$, a contradiction. ∎

Corollary 7.5.2. *Given two extensions E, E' of k, if $[E : k] = n$, then there are at most n k-homomorphisms from E to E'.*

Proof. Let $u_1, \ldots, u_n$ be a k-basis of E and suppose that there are $n+1$ distinct k-homomorphisms $\sigma_0, \ldots, \sigma_n$ to E'. Then the n equations in the $n+1$ unknowns x_i:

$$\sum_i u_j^{\sigma_i}x_i = 0 \quad (j = 1, \ldots, n)$$

have a non-zero solution $x_i = c_i$ in E'. Any $a \in E$ has the form $a = \sum \alpha_j u_j$ $(\alpha_j \in k)$, hence

$$\sum_i a^{\sigma_i}c_i = \sum_i \left(\sum_j \alpha_j u_j\right)^{\sigma_i} c_i = \sum_i \sum_j \alpha_j (u_j^{\sigma_i})c_i = 0,$$

and this contradicts the lemma. ∎

We note that when $E' = E$, any homomorphism is necessarily an automorphism (as injective endomorphism of a finite-dimensional vector space). We are specially interested in the case where the number of k-automorphisms of E is exactly n; by Corollary 7.5.2 it cannot be larger. Let us consider some examples to illustrate the situation.

Example 1. $[\mathbf{C} : \mathbf{R}] = 2$. Here there are two automorphisms, the identity and complex conjugation.

Example 2. $\mathbf{Q}(\alpha)/\mathbf{Q}$, where α is the real root of $x^3 = 2$. Here $[\mathbf{Q}(\alpha) : \mathbf{Q}] = 3$, but there are no automorphisms other than the identity, for $\mathbf{Q}(\alpha)$ contains only one root of $x^3 = 2$, which is necessarily mapped to a root of the same equation by

any automorphism and must therefore stay fixed. The number of automorphisms falls short of the possible total of three because (as we shall soon see) the extension is not normal. If we go over to the normal closure $\mathbf{Q}(\alpha, \omega)$, where ω is a root of $x^3 = 1$ but $\omega \neq 1$, we find six automorphisms of $\mathbf{Q}(\alpha, \omega)$, and in fact $\mathbf{Q}(\alpha)$ has three $\mathbf{Q}$-homomorphisms into $\mathbf{Q}(\alpha, \omega)$.

Example 3. Let k be a field of characteristic $p \neq 0$, $F = k(t)$ be the field of rational functions in an indeterminate t and α be a root of the equation $x^p = t$. This equation has at most one root, and the automorphism is determined once the image of α is fixed. Here the number of automorphisms falls short because the extension is inseparable.

These examples suggest that in order to get the maximum number of automorphisms of F/k we need to take F/k normal and separable, and this will be carried out in the next section. As a matter of fact, from any extension F/k we can form the normal closure, and this will be separable whenever F/k is; but when F/k is inseparable, there is nothing we can do to get 'enough' automorphisms. These ideas are made precise in Theorem 7.5.4 below, which is preceded by a lemma on the extension of homomorphisms. Two extensions F/k and F'/k' will be called *isomorphic* if there is an isomorphism $F \cong F'$ and k maps to k' in this isomorphism.

Lemma 7.5.3. *Let F/k, E/k be finite extensions, where E/k is normal. If there is a k-homomorphism $\varphi : F \to E$, then any k-homomorphism of a subextension of F into E can be extended to a k-homomorphism of F into E.*

Proof. Let $k \subseteq D \subseteq F$ and let $\theta : D \to E$ be a k-homomorphism. If $D \neq F$, take $\alpha \in F$ and let f be the minimal polynomial of α over D; then f is irreducible over D and the corresponding polynomial $f\theta$ is irreducible over $D' = D\theta$. Let g be the minimal polynomial of α over k; since $g(\alpha) = 0$, we have $f|g$, hence (applying θ) we find that $f\theta|g$. Now g has a zero in E, namely $\alpha\varphi$; therefore it splits completely and so $f\theta$ also splits over E. If α' is any zero of $f\theta$ in E, then by Proposition 7.3.2, there is an isomorphism from D(α)/D to $D'(\alpha')/D'$, in which α maps to α'. So θ has been extended to a homomorphism of D(α) into E, and after a finite number of steps we reach the required homomorphism $F \to E$. ∎

We now come to the promised result, showing that any separable extension has enough homomorphisms into a normal closure:

Theorem 7.5.4. *Let F/k be an extension of finite degree n and F'/k' be an isomorphic extension, say $\varphi : F \cong F'$ is an isomorphism and*

$$\varphi_0 : k \cong k', \tag{7.5.3}$$

where $\varphi_0 = \varphi|k$. If E'/k' is a normal extension containing F'/k', then there are at most n homomorphisms $F \to E'$ which extend the isomorphism φ_0 in (7.5.3), *and the normal closure of F'/k' in E'/k' is the field generated by the images of these homomorphisms.*

Moreover, the following conditions on the extension F/k are equivalent:

(a) *there are exactly* $n = [F : k]$ *homomorphisms from F to E extending* φ_o,
(b) *F/k is separable,*
(c) *F/k is generated by separable elements.*

Proof. By Corollary 7.5.2 there are at most n homomorphisms $F \to E'$ extending φ_o. Let E'' be the subfield of E' generated by all the images of F. Then $E'' \supseteq F'$ and we have to show that E''/k' is normal. By construction, E'' is generated by all the homomorphic images in E' of elements of F. Take any $\alpha \in F$, let f be its minimal polynomial over k and denote by f' its image under φ_o. Then f' is irreducible over k' and has a zero in E', hence it splits completely over E'. If α' is any zero of f' in E', then $k(\alpha)/k$ is isomorphic to $k'(\alpha')/k'$ and by Lemma 7.5.3 this isomorphism extends to a homomorphism of F into E'; hence α' as image of α lies in E''. Moreover, E'' is generated by all such elements α'; thus E'' is generated over k' by all the zeros of the polynomials corresponding to minimal polynomials of elements of F, and so E''/k' is normal, as claimed.

(a) $\Rightarrow$ (b). To prove that F/k is separable we must show that each element of F is separable over k. Let $\alpha \in F$ have minimal polynomial f over k, of degree r. If α is not separable, then the number s of distinct zeros of f in a splitting field is less than r. In particular, there are only s homomorphisms of $k(\alpha)$ into E', for each such homomorphism is determined by the image of α. Each of these homomorphisms extends in at most $[F : k(\alpha)] = n/r$ ways to a homomorphism of F into E', by Corollary 7.5.2. Hence there are $s.n/r < n$ homomorphisms in all, which contradicts (a).

(b) $\Rightarrow$ (c) is clear. To prove (c) $\Rightarrow$ (a), let $F = k(\alpha_1, \ldots, \alpha_r)$, where each α_i is separable over k. If α_1 has the minimal polynomial f_1 of degree n_1 over k, then the corresponding polynomial over k' has n_1 zeros and we obtain n_1 homomorphisms $k(\alpha_1) \to E'$ by mapping α_1 to one of these zeros. Now $F/k(\alpha_1)$ is separably generated and by induction on the degree, any homomorphism of $k(\alpha_1)$ into E' extends in $n/n_1 = [F : k(\alpha_1)]$ ways to a homomorphism of F into E'. Hence the isomorphism φ_o extends in $n_1.n/n_1 = n$ ways to a homomorphism of F into E', as claimed. ∎

In the next section we shall study the case of separable normal extensions in more detail; for the moment we shall prove a result which ensures the existence of a sufficient supply of automorphisms under rather different conditions. This is known as Artin's theorem, although it also goes back to Dedekind (see Dedekind (1894) p. 50).

Theorem 7.5.5 (Artin's theorem). *Let E be a field, G be a group of automorphisms of E and k be the set of elements of E fixed by G. Then k is a subfield of E. Moreover E has finite degree over k if and only if G is finite; in that case*

$$[E : k] = |G|. \tag{7.5.4}$$

Proof. That k is a field is easily checked. By Corollary 7.5.2, $|G| \leq [E : k]$, and it remains to establish equality. For infinite G this is clear, so let $|G| = r$ and assume

that we have $r+1$ elements of E that are linearly independent over k : $u_0, \ldots, u_r$. Consider the r equations in the $r+1$ unknowns x_j:

$$\sum u_j^\sigma x_j = 0 \quad \text{for all } \sigma \in G. \tag{7.5.5}$$

They have a non-trivial solution $x_j = a_j$ in E. We pick a solution with the fewest non-zero terms; if $a_0 \neq 0$ say, we can solve for the term in u_0:

$$u_0^\sigma = \sum_{j=1}^{r} u_j^\sigma b_j \quad \text{for some } b_j \in E \text{ and all } \sigma \in G. \tag{7.5.6}$$

For $\sigma = 1$, (7.5.6) reads $u_0 = \sum u_j b_j$, so not all the b_j lie in k, by the linear independence of the u_j over k, say $b_1 \notin k$. By the definition of k there exists $\tau \in G$ such that $b_1^\tau \neq b_1$. Now replace σ in (7.5.6) by $\sigma\tau^{-1}$, apply τ and note that σ runs over G as $\sigma\tau^{-1}$ does:

$$u_0^\sigma = \sum_{j=1}^{r} u_j^\sigma b_j^\tau \quad \text{for all } \sigma \in G. \tag{7.5.7}$$

Subtracting (7.5.7) from (7.5.6), we obtain

$$\sum_{j=1}^{r} u_j^\sigma (b_j - b_j^\tau) = 0.$$

This is a shorter relation than (7.5.6), and is non-trivial, because $b_1^\tau \neq b_1$, so we have a contradiction. It follows that $[E : k] \leq r$, and this proves equality in (7.5.4). ■

Exercises

1. Show that Dedekind's lemma holds for any set of homomorphisms of a group G into the multiplicative group of a field, i.e. verify that the additive structure of E is not involved.
2. Use Artin's theorem to show that for any field E with n distinct automorphisms, if k is the fixed field of this set of automorphisms, then $[E : k] \geq n$.
3. Find the fixed field F of the rational function field $k(x)$ under the automorphisms $x \mapsto 1 - x$, $x \mapsto 1/x$; show that the degree is 6. Verify that $(x^2 - x + 1)^3/(x^2 - x)^2$ lies in F and use this fact to find an equation for x over F.
4. Let F be a perfect field; show that the set of elements fixed under all automorphisms of F is a perfect subfield.
5. Let E be a field, G be a group of automorphisms of E and k be the fixed field. Show that any $\alpha \in E$ is algebraic over k iff it lies in a finite G-orbit.
6. Let E/k be a normal extension. Show that an element of E, of degree r over k, has at most r conjugates over k, with equality iff it is separable.

7.6 The Fundamental Theorem of Galois Theory

Let E be any field and G be a group of automorphisms of E. With each subgroup H of G we associate the subset of elements of E fixed by H:

$$H^* = \{x \in E | x^\sigma = x \quad \text{for all } \sigma \in H\}.$$

This is easily seen to be a subfield of E, called the *fixed field* of the subgroup H. For example, $1^* = E$, and for any subgroup H of G, H^* clearly contains G^*.

Similarly, with each subfield F of E we associate the subset of elements of G leaving F elementwise fixed:

$$F^* = \{\sigma \in G | x^\sigma = x \quad \text{for all } x \in F\}.$$

Again it is easily verified that F^* is a subgroup of G, called the group of *F-automorphisms* of G. Thus, writing $\mathscr{F}(E)$ for the set of all subfields of E and $\mathscr{B}(G)$ for the set of all subgroups of G, we have mappings $\mathscr{B}(G) \to \mathscr{F}(E)$ and $\mathscr{F}(E) \to \mathscr{B}(G)$. These mappings satisfy the rules

Γ.1 $F_1 \subseteq F_2 \Rightarrow F_1^* \supseteq F_2^* \quad (F_i \in \mathscr{F}(E))$,
$H_1 \subseteq H_2 \Rightarrow H_1^* \supseteq H_2^* \quad (H_i \in \mathscr{B}(G))$;
Γ.2 $F \subseteq F^{**}, H \subseteq H^{**}$;
Γ.3 $F^{***} = F^*, H^{***} = H^*$.

Γ.1 states that the mappings are order-inverting, and follows almost immediately from the definitions, as does **Γ.2**. To prove **Γ.3**, we need only use **Γ.1** and **Γ.2**: if we replace H by F^* in **Γ.2** we get $F^* \subseteq F^{***}$, and if we operate with * on **Γ.2** and use **Γ.1** we find $F^{***} \subseteq F^*$, hence $F^{***} = F^*$; similarly $H^{***} = H^*$.

Any correspondence between two partially ordered sets satisfying **Γ.1** and **Γ.2** and hence **Γ.3** is called a *Galois connexion*. We are particularly interested to know how the subfield F and the subgroup H have to be restricted so that the correspondence becomes a bijection. By **Γ.3** we see that we must consider the subfields F satisfying $F^{**} = F$ and the subgroups H satisfying $H^{**} = H$. For the present we shall limit ourselves to the case of finite groups of automorphisms and leave the general case to Chapter 11. We shall see that every subgroup H of G then satisfies $H^{**} = H$, while every field F satisfies $F^{**} = F$ provided that we start as above with a field E and a group G of automorphisms which singles out certain subfields of E. However, it is not true that every finite extension is of this form; a finite extension E/F will be called a *Galois extension* if F is the fixed field of a group of automorphisms of E. The group of all F-automorphisms of E is then called the *Galois group* of the extension and is written $\mathrm{Gal}(E/F)$. With the help of the results in Section 7.5 it is not hard to describe all Galois extensions.

Proposition 7.6.1. *Let E/F be a finite field extension. Then* (i) *E/F is a Galois extension if and only if it is normal and separable;* (ii) *E/F is contained in a Galois extension if and only if it is separable.*

Proof. Let E/F be a Galois extension of degree n with group G and let L be its normal closure. Then E/F has at most n F-homomorphisms into L, but by Artin's theorem it

has that many automorphisms, hence by Theorem 7.5.4, E/F is separable and $L = E$, i.e. E/F is also normal.

Conversely, suppose that E/F is normal and separable. Then by Theorem 7.5.4, E/F has $[E : F]$ automorphisms. Let $G = F^*$ be the group of all F-automorphisms and put $F_1 = F^{**}$. Then $F_1 \supseteq F$ and by Artin's theorem, $|G| = [E : F_1]$; but by Theorem 7.5.4, $|G| \geq [E : F]$; hence $[E : F_1] \geq [E : F] = [E : F_1][F_1 : F]$. It follows that $F_1 = F$ and so E/F is Galois.

To prove (ii) we note that the condition is necessary because every subextension of a separable extension is separable. Conversely, assume that E/F is separable and let L be a normal closure of E over F. Then L/F is normal and separable, as minimal splitting field of a separable polynomial, and it contains E/F, so by (i), E/F is embedded in a Galois extension. ∎

We now come to the main theorem of Galois theory, establishing the Galois connexion between fields and automorphism groups.

Theorem 7.6.2. *Let E be a field, G be a finite group of automorphisms of E and k be the fixed field of G. For any field F between k and E define the group of F-automorphisms in G as*

$$F^* = \{\sigma \in G | x^\sigma = x \quad \text{for all } x \in F\},$$

and for any subgroup H of G define the fixed field of H as

$$H^* = \{x \in E | x^\sigma = x \quad \text{for all } \sigma \in H\}.$$

Then the mappings $F \mapsto F^*$, $H \mapsto H^*$ define a Galois connexion between the set of all subgroups of G and the set of all fields between k and E, which is an order-inverting bijection:

$$\begin{array}{cc} E & 1 \\ \uparrow & \downarrow \\ F & H \\ \uparrow & \downarrow \\ k & G \end{array}$$

Proof. We have to show that $H^{**} = H$ and $F^{**} = F$. Let $H^* = F$; then E/F is Galois by definition, and $H^{**} \supseteq H$. Moreover, $H^{***} = H^* = F$ and hence, by Artin's theorem, $|H| = [E : F] = |H^{**}|$; therefore $H^{**} = H$.

Secondly, given F, since E/k is normal separable, by Proposition 7.6.1, so is E/F. Therefore it is Galois, i.e. $F = H^*$ for some $H \subseteq G$. It follows that $F^{**} = H^{***} = H^* = F$. ∎

We note the correspondence between group orders and field degrees:

Corollary 7.6.3. *Let E/k be a Galois extension with group G, and let the subfield F correspond to the subgroup H. Then*

$$|H| = [E : F], \quad (G : H) = [F : k].$$

Proof. By Artin's theorem, $[E : k] = |G|$, $[E : F] = |H|$; now the result follows by division. ∎

Next we note that conjugate extensions correspond to conjugate subgroups, and normal extensions correspond to normal subgroups.

Corollary 7.6.4. *Let E/k be a Galois extension with group G and let H_1, H_2 be any subgroups, corresponding to subfields F_1, F_2. Then F_1 and F_2 are conjugate under an automorphism $\sigma \in G$ if and only if H_1 and H_2 are conjugate subgroups of G: $H_2 = \sigma^{-1}H_1\sigma$.*

In particular, F/k is normal if and only if $H = F^*$ is a normal subgroup of G and when this is so, then $\mathrm{Gal}(F/k) \cong G/H$.

Proof. By definition, $\tau \in H_1 \Leftrightarrow x^\tau = x$ for all $x \in F_1$. Let $F_2 = F_1^\sigma$; then for any $y \in F_2$, $y^{\sigma^{-1}} \in F_1$ hence $y^{\sigma^{-1}\tau} = y^{\sigma^{-1}}$ which means that $\sigma^{-1}\tau\sigma \in H_2$. Thus $\sigma^{-1}H_1\sigma \subseteq H_2$ and since $F_2^{\sigma^{-1}} = F_1$, it follows that $\sigma H_2\sigma^{-1} \subseteq H_1$, whence we find $\sigma^{-1}H_1\sigma = H_2$. Retracing our steps, we obtain the converse.

If F corresponds to H, then F/k is normal iff it coincides with all its conjugates, i.e. $F^\sigma = F$ for all $\sigma \in G$, and this holds precisely when $\sigma^{-1}H\sigma = H$ for all $\sigma \in G$, i.e. when H is normal in G. In this case we can define a homomorphism from G to $N = \mathrm{Gal}(F/k)$ by restricting each automorphism of E to F. Each $\sigma \in G$ maps F into itself and so defines an element, $\overline{\sigma}$ say, of N; clearly the mapping $\sigma \mapsto \overline{\sigma}$ is a homomorphism. By Corollary 7.2.5 each automorphism of F/k extends to an automorphism of E/k, so every element of N lies in the image. To determine the kernel we note that $\overline{\sigma} = 1$ iff σ fixes F, i.e. $\sigma \in H$. Thus $N \cong G/H$, as asserted. ∎

This result in particular makes it clear why not every field extension is Galois: to a given separable extension F/k there correspond in general several conjugate extensions, and the relation between them is described by a family of conjugate subgroups of $\mathrm{Gal}(E/k)$, where E is the normal closure of F/k. In fact it was this situation which first led to an understanding of the concept of a normal subgroup (which is due to Galois). We also record a condition for the conjugacy of elements; we recall that two elements of an extension are conjugate if they lie in the same G-orbit.

Corollary 7.6.5. *Two elements of a Galois extension E/k are conjugate if and only if they have the same minimal polynomial over k.*

Proof. This is an immediate consequence of Proposition 7.2.2 and Corollary 7.2.5. ∎

The following consequence of Dedekind's lemma is often useful.

Proposition 7.6.6. *Let E/k be a Galois extension and suppose that $u_1, \ldots, u_n$ is a k-basis of E. Then the matrix $(u_i^\sigma)(\sigma \in G)$ is non-singular.*

Proof. By Theorem 7.5.4, (u_i^σ) is a square matrix. If this matrix were singular, the system of linear equations

$$\sum u_i^\sigma x_i = 0$$

would have a non-trivial solution, but by Dedekind's lemma the only solution is the zero solution. Hence the result follows. ∎

Strictly speaking we should indicate which index in the matrix (u_i^σ) refers to rows and which refers to columns, but this does not matter because a matrix over a field is singular iff its transpose is singular.

We give some examples to illustrate these results.

Example 1. Quadratic extension. An extension F/k is called *quadratic* if $[F : k] = 2$. In this case F is generated by an element of degree 2 over k (in fact any element of $F \setminus k$ will do), and any polynomial of degree 2 with a zero in F splits in F, so the extension must be normal. If $\alpha \in F \setminus k$ has the minimal polynomial $x^2 + ax + b$, then over F we have $x^2 + ax + b = (x - \alpha)(x - \beta)$, where $\beta = -a - \alpha$. We see that the extension is separable unless char $k = 2$ and $a = 0$.

Example 2. The symmetric group as Galois group. Let k be any field and write $E = k(x_1, \ldots, x_n)$, where the x_i are independent indeterminates. Then E admits the symmetric group G of all permutations of $x_1, \ldots, x_n$ as automorphism group. Its fixed field F consists of all symmetric functions in the x_i over k. Let $e_1, \ldots, e_n$ be the elementary symmetric functions in the x_i; then clearly $F \supseteq k(e_1, \ldots, e_n)$ and $x_1, \ldots, x_n$ are roots of the equation

$$x^n - e_1 x^{n-1} + \ldots + (-1)^n e_n = 0. \tag{7.6.1}$$

Hence E is a minimal splitting field of this equation over $k(e_1, \ldots, e_n)$; as such its degree is at most $n!$ by Theorem 7.2.3. But we saw that there are $n!$ automorphisms of $E/k(e_1, \ldots, e_n)$, namely G. Hence $F = k(e_1, \ldots, e_n)$ and $[E : F] = n!$. In particular, it follows that every symmetric function in $x_1, \ldots, x_n$ is a rational function in $e_1, \ldots, e_n$. This is almost the fundamental theorem on symmetric functions, which asserts that every symmetric *polynomial* in the x_i is just a *polynomial* in the e_i.

This example also shows that every finite group occurs as the Galois group of some extension. For any finite group is isomorphic to a subgroup of some symmetric group G, by Cayley's theorem (Theorem 2.2.1). Take E/F as above with $\mathrm{Gal}(E/F) = G$; then each subgroup H of G is the Galois group of an extension E/K, where K is the fixed field of H. Here we were able to prescribe E, but not K, so the following question remains: given a field k and a group G, does there exist a Galois extension E/k with group G? In general the answer is 'no', e.g. for $k = \mathbf{C}$ (and G non-trivial), but for $\mathbf{Q}$ we can always find a Galois extension with the symmetric group, as we shall see in Section 7.11 below. However, many unsolved questions remain, even when $k = \mathbf{Q}$.

We can now answer a question raised in Section 7.1, by giving an example of an extension of prescribed degree, without proper subextensions. All we need is to find, for any given n, a finite group with a maximal subgroup of index n. This is provided by $S = \mathrm{Sym}_n$, the symmetric group of degree n, and T, the stabilizer of a symbol. Clearly $(S : T) = n$; if H is a subgroup such that $T \subset H \subseteq S$, then H contains T as well as a permutation moving the symbol fixed by T; hence it acts transitively and so, by the orbit formula (see Section 2.1), $n = (H : T)$. It follows that $H = S$ and this shows T to be a maximal subgroup. If as in the above example we take $F = k(e_1, \ldots, e_n)$ with indeterminates e_i and adjoin a single root of Equation (7.6.1), we obtain an extension of degree n without a proper subextension.

Example 3. In general, if we adjoin a root of an equation $f = 0$, we do not get a normal extension, i.e. adjoining one root will usually not be enough to split f into linear factors. If it does, i.e. if we get a splitting field of f by adjoining a single (suitably chosen) zero of f, the equation $f = 0$ is said to be *normal* over k. In Example 1 we saw that every quadratic equation is normal, and earlier we noted that $x^3 - 2 = 0$ is an example of an equation which is not normal over $\mathbf{Q}$.

As an example of a normal equation, consider

$$x^3 - 3x - 1 = 0.$$

It may be verified that if θ is any root, then so is $-(\theta + 1)^{-1}$. Taken over $\mathbf{Q}$, the equation is irreducible, hence $\sigma : \theta \mapsto -(\theta + 1)^{-1}$ defines an automorphism of $\mathbf{Q}(\theta)/\mathbf{Q}$. It is easily checked that $\sigma^2 : \theta \mapsto -1 - \theta^{-1}$ and $\sigma^3 = 1$. Thus σ has order 3 and we have found three automorphisms of $\mathbf{Q}(\theta)/\mathbf{Q}$. Since the degree is 3, this is the most we can have and this shows that $\mathbf{Q}(\theta)/\mathbf{Q}$ is a Galois extension.

Example 4. So far all our extensions have been finite; to end with, here is an example of what can happen in the infinite case. Let k be a field of characteristic 0 and $F = k(t)$ be the field of rational functions in an indeterminate t. On F we have two automorphisms, each of order 2:

$$\sigma : t \mapsto -t, \quad \text{fixed field} : k(t^2),$$

$$\tau : t \mapsto 1 - t, \quad \text{fixed field} : k(t^2 - t).$$

The least field containing both $k(t^2)$ and $k(t^2 - t)$ is clearly $k(t)$, while $k(t^2) \cap k(t^2 - t) = k$. To see this, let us take f in the intersection, say $f = g/h$, where g, h are coprime polynomials in t. By hypothesis, $f^{\sigma\tau} = f$, hence $g^{\sigma\tau}h = h^{\sigma\tau}g$. On comparing degrees, we find that $g^{\sigma\tau} = \lambda g$, where $\lambda \in k$ (because g and h are coprime), and a comparison of leading terms shows that $\lambda = 1$. Thus $g^{\sigma\tau} = g$ and $h^{\sigma\tau} = h$. Now $\sigma\tau : t \mapsto t - 1$; hence if α is a root of $g(t) = 0$, then so is $\alpha - 1$, and by induction, so is $\alpha - n$, for all $n \in \mathbf{N}$. Therefore if g has positive degree (and hence a zero in some extension field), it has infinitely many zeros, a contradiction. Thus $\deg g = 0$; similarly $\deg h = 0$, and it follows that k is the fixed field of the group generated by σ and τ. But in fact k is already the fixed field of the cyclic group generated by $\sigma\tau$, which is clearly smaller, so we no longer

have a bijection between subgroups and subfields. Nevertheless there is a Galois connexion for infinite extensions; this will be dealt with in Section 11.8.

Galois' main results in the theory of equations were published in a memoir in 1846 (some 14 years after his death, at the age of 20, in a duel). Even after this long delay, his results appeared very novel and it took many years before they were properly assimilated and further developed. The treatment of Galois theory by Dedekind (in the famous XIth supplement to the 1894 edition of Dirichlet's *Vorlesungen über Zahlentheorie*) emphasized the field rather than the equation and gave the theory its modern 'linear' form. It is also the basis of the account by Artin (1948) which has served as a model for most subsequent expositions of the theory, including this one.

Exercises

1. Let k be a field of characteristic not 2. Show that any extension of degree 2 over k is separable and can be generated by a root of an equation $x^2 = a$, where a is an element of k not a square. Conversely, any such equation is separable and has roots $\pm\alpha$; show that the k-automorphisms are 1 and $b + \alpha c \mapsto b - \alpha c$ $(b, c \in k)$.
2. Let k be of characteristic 2. Show that any separable extension of degree 2 over k can be generated by a root of an equation $x^2 + x + a = 0$, where $a \in k$; conversely, any such equation is separable over k and has roots, $\alpha, \alpha + 1$. Show that the automorphisms are 1 and $b + \alpha c \mapsto b + c + \alpha c$ $(b, c \in k)$.
3. Let k be of characteristic 2. Show that any inseparable element α of degree 2 over k satisfies an equation $x^2 + a = 0$, where a is an element in k not a square, and conversely, any such equation is inseparable over k with a single root.
4. Let E/k be a Galois extension, F be a field between k and E, and G be the subgroup of $\mathrm{Gal}(E/k)$ mapping F into itself. Show that G is the normalizer of $\mathrm{Gal}(E/F)$ in $\mathrm{Gal}(E/k)$ and describe $G/\mathrm{Gal}(E/F)$.
5. Let E/k be a Galois extension with group G and N_i be a subextension with group $G_i = \mathrm{Gal}(E/N_i)$ $(i = 1, 2)$. Show that $G = G_1 \times G_2$ iff N_i/k is Galois and $N_1 \cap N_2 = k$, $N_1 N_2 = E$.
6. Let $x_1, \ldots, x_n$ be independent indeterminates and let F be the fixed field in $k(x_1, \ldots, x_n)$ under the group of all permutations of all the x_i. Show that x_i has degree i over $F(x_{i+1}, \ldots, x_n)$ and deduce that the monomials $x_1^{r_1} \ldots x_n^{r_n}$ $(0 \leq r_i \leq i - 1)$ form a basis of $k(x_1, \ldots, x_n)$ over F.
7. Let k be a field of prime characteristic p and $F = k(t)$ be the rational function field. Show that the group of automorphisms generated by $\sigma : t \mapsto -t$, $\tau : t \mapsto 1 - t$ is finite. Find the fixed field F_0 and the minimal equation of t over F_0.
8. Let f be a polynomial over a field k and let E be a minimal splitting field for f. Show that the automorphisms of E/k permute the zeros of f transitively iff f is a power of an irreducible polynomial.
9. Show that a finite field extension F/k is Galois iff the only elements of F fixed under all automorphisms of F over k are the elements of k.

7.7 Roots of Unity

Let A be an integral domain. An element p of A is called a *prime* if it is not zero or a unit and if $p|ab$ implies $p|a$ or $p|b$, for any $a, b \in A$. Clearly it follows that the residue-class ring $A/(p)$ is again an integral domain. A domain in which every non-zero element is either a unit or a product of primes is called a *unique factorization domain* (UFD); in such a ring any prime factorization of a given element is unique up to the order of the factors and associates (see Section 10.2 below). Let A be a UFD and consider an element f of the polynomial ring $A[x]$. By taking out common factors of the coefficients we can write $f = af_1$, where $a \in A$ and f_1 is a polynomial whose coefficients have no common non-unit factor; such f_1 is called *primitive*. Such polynomials have the following property:

Lemma 7.7.1 (Gauss's lemma). *Over a unique factorization domain the product of two primitive polynomials is primitive.*

Proof. We begin by proving that for any integral domain A, a prime in A stays a prime in $A[x]$. Let p be a prime in A and $f, g \in A[x]$; we have to show that $p|fg \Rightarrow p|f$ or $p|g$. Write $\overline{A} = A/(p)$ and denote the residue-class mapping on $A[x]$ by $f \mapsto \overline{f}$. Since p is prime, $\overline{A}$ is again an integral domain, hence so is $\overline{A}[x]$. If $p|fg$, then $\overline{f}\overline{g} = 0$, hence $\overline{f} = 0$ or $\overline{g} = 0$, because $\overline{A}[x]$ is an integral domain. But this means that $p|f$ or $p|g$, as we wished to show. Now the assertion of the lemma is an immediate consequence. ∎

Given a ring A and a ring B containing A, we shall say that A is *inert* in B if for $c \in A$, such that $c = ab$ for some $a, b \in B$, there exists a unit u in B such that au, $u^{-1}b \in A$. This just means that every factorization of c in B can be 'pulled down' to A; in particular, if c is unfactorable in A, it remains so in B.

Theorem 7.7.2. *Let A be a unique factorization domain with field of fractions K. Then $A[x]$ is inert in $K[x]$.*

Proof. Let $f \in A[x]$ and suppose that $f = gh$ for some $g, h \in K[x]$. By taking a common denominator of the coefficients of g we can find $\alpha_o \in A$ such that $\alpha_o g$ has coefficients in A. On dividing by any common factor we find $\alpha \in K$ such that $g_1 = \alpha g$ is a primitive polynomial over A. Similarly there exists $\beta \in K$ such that $h_1 = \beta h$ is over A and primitive. Write $\alpha\beta = a/b$, where a, b are coprime elements of A. Then

$$\frac{a}{b} f = \alpha\beta gh = g_1 h_1,$$

hence $af = bg_1h_1$. We claim that a must be a unit. For if not, then there is a prime p dividing a. Hence $p|bg_1h_1$, but p cannot divide b because a, b are coprime, nor can it divide g_1 or h_1 because they are primitive. This contradicts Gauss's lemma and it shows that a is a unit. Now $\alpha\beta b = a$, and $\alpha g = g_1$, $\alpha^{-1}h = a^{-1}b\beta h = a^{-1}bh_1$ both have coefficients in A, and this is what we had to show. ∎

Applied to **Z** (which we know to be a UFD) this result shows that any polynomial over **Z** which can be factorized over **Q** can be factorized over **Z**. In particular, if a monic polynomial over **Z** can be factorized over **Q**, we can take all the factors to be monic with coefficients in **Z**. Another consequence is the familiar fact that any rational root of an equation with integer coefficients

$$a_0x^n + a_1x^{n-1} + \ldots + a_n = 0 \quad (a_i \in \mathbf{Z}) \tag{7.7.1}$$

has a denominator dividing a_0 and a numerator dividing a_n.

By an n-th root of 1 (or of *unity*) we understand any root of the equation

$$x^n = 1. \tag{7.7.2}$$

The basic fact is that this equation, like any equation of degree n, cannot have more than n roots in any field. For if $f = 0$ is an equation over a field, of degree n with $n + 1$ distinct roots $\alpha_0, \alpha_1, \ldots, \alpha_n$, then we have the factorization

$$f = a_0(x - \alpha_1)\ldots(x - \alpha_n). \tag{7.7.3}$$

On putting $x = \alpha_0$, we obtain 0 since α_0 is a root, but none of the factors on the right of (7.7.3) vanish, which is impossible in an integral domain.

We ask: when does Equation (7.7.2) have precisely n distinct roots? In a splitting field (7.7.2) has distinct roots provided that $x^n - 1$ is prime to its derivative nx^{n-1}, and clearly this is the case iff $nx^{n-1} \neq 0$. Thus (7.7.2) has distinct roots whenever the characteristic is prime to n. In the case when char $k = p$ and $n = n'p^r$, where $(n', p) = 1$, we have $x^n - 1 = (x^{n'} - 1)^{p^r}$, so (7.7.2) is then equivalent to the equation

$$x^{n'} = 1.$$

We shall therefore assume in what follows that the characteristic of the field is prime to n. We also recall the basis theorem for finite abelian groups (Corollary 2.4.2), stated in multiplicative form for convenience:

Every finite abelian group can be written as a direct product of cyclic groups:

$$A = B_1 \times \ldots \times B_r, \tag{7.7.4}$$

and the order of B_i divides that of B_{i+1} $(i = 1, \ldots, r - 1)$.

It is clear that the exponent ν of A is equal to the order of B_r, and the representation (7.7.4) shows that A has an element of order ν, namely any generator of B_r. We note that the order of A is at least equal to its exponent, with equality iff A is cyclic.

Theorem 7.7.3 *In any field k, the roots of the equation $x^n = 1$ (for any integer $n \geq 1$) form a cyclic group under multiplication, whose order is a divisor of n.*

Proof. Clearly the roots form a group G, say, namely a subgroup of the multiplicative group $k^\times$ of k. If G has exponent ν, then $\nu | n$ and G has at most ν elements, because the equation $x^\nu = 1$ has at most ν roots in k. Thus $|G| \leq \nu$ and it follows that we have equality and G is cyclic. ∎

By a *primitive n-th root* of 1 one understands a field element whose multiplicative order is precisely n. In the above proof the generators of G are precisely the primitive roots of 1 in k.

Alternative proof. We now give a second proof of Theorem 7.7.3, more elementary and more explicit. It will be enough to show that any finite subgroup of $k^\times$ is cyclic. If $|G| = n$, then all the elements of G satisfy (7.7.2). Thus k contains n n-th roots of 1 and we have to find a primitive n-th root. For any $d|n$ denote by $\psi(d)$ the number of primitive d-th roots of 1 in k. If there is a primitive d-th root, say ζ, then the roots of $x^d = 1$ are the different powers of ζ, and just $\varphi(d)$ of them are primitive d-th roots. Here $\varphi(d)$, Euler's function, is the number of integers in the range $1, 2, \ldots, d$ that are prime to d. Thus we see that $\psi(d)$ is either 0 or $\varphi(d)$. Now we observe the following:

(i) For each $d|n$ there are just $\varphi(n/d)$ integers x in the range $1 \leq x \leq n$ such that $(x, n) = d$, and as d ranges over all divisors of n, every number from 1 to n is counted exactly once. Hence

$$n = \sum \varphi(n/d) = \sum \varphi(d),$$

where the summation is over all the divisors d of n.

(ii) Every n-th root of 1 is a primitive d-th root for some $d|n$. Since there are altogether n n-th roots of 1 in k, by hypothesis, we find

$$n = \sum \psi(d),$$

where the summation is again over all the divisors of n. Thus

$$\sum_{d|n} [\varphi(d) - \psi(d)] = 0,$$

and all the summands are non-negative; hence they must all be zero, in particular, $\psi(n) = \varphi(n) > 0$. This shows that any group of order n in $k^\times$ contains a primitive n-th root of 1; in fact the number is $\varphi(n)$ and Theorem 7.7.3 follows. ∎

If $\xi_1, \ldots, \xi_r$ $(r = \varphi(n))$ are all the primitive n-th roots of 1, then the polynomial

$$\Phi_n(x) = \prod_{i=1}^{r} (x - \xi_i)$$

is called the *n-th cyclotomic polynomial*; its degree is $\varphi(n)$. The word 'cyclotomic' (circle-cutting) is used because in the complex field $\mathbf{C}$ the n-th roots of 1 are $e^{2\pi ik/n}$ $(k = 1, 2, \ldots, n)$ and are obtained geometrically by dividing the unit circle about the origin into n equal parts. To obtain an explicit expression for Φ_n we observe that every n-th root of 1 is a primitive d-th root, for a unique divisor d of n. Hence

$$x^n - 1 = \prod_{d|n} \Phi_d(x). \tag{7.7.5}$$

By Theorem 7.7.2, $\Phi_d(x)$ is again monic, with integer coefficients. Using the Möbius function introduced in Section 5.6, we can solve for Φ_d and obtain the formula

$$\Phi_n(x) = \prod_{d|n} (x^d - 1)^{\mu(n/d)}.$$

For example, $\Phi_1 = x - 1$, $\Phi_2 = x + 1$, $\Phi_3 = x^2 + x + 1$, $\Phi_4 = x^2 + 1$, $\Phi_5 = x^4 + x^3 + x^2 + x + 1$, $\Phi_6 = x^2 - x + 1$, $\Phi_{12} = x^4 - x^2 + 1$. It is a curious fact that the coefficients of ϕ_n are 0 or ± 1 for low values of n, in fact for all $n < 105$, but when $n = 105$, there is a coefficient -2 and as *I.* Schur has shown, the absolute values of the coefficients of Φ_n are unbounded as n increases.

Let k be a field of characteristic prime to n and ζ be a primitive n-th root of 1 over k. Then $k(\zeta)$ is a minimal splitting field of $x^n = 1$ over k; since $nx^{n-1} \neq 0$, it is prime to $x^n - 1$, so the roots are distinct and hence $k(\zeta)/k$ is a Galois extension; it is also called a *cyclotomic* extension. Let f be the minimal polynomial of ζ over k, of degree s say; then $[k(\zeta) : k] = s$. Clearly $f | \Phi_n$, hence $s \leq \varphi(n)$, with equality iff Φ_n is irreducible over k. Whether this is so naturally depends on k; we now show that it is the case for $k = \mathbf{Q}$:

Theorem 7.7.4. *The cyclotomic polynomial Φ_n is irreducible over $\mathbf{Q}$.*

Proof. Let ζ be a primitive n-th root of 1 over $\mathbf{Q}$, and denote its minimal polynomial over $\mathbf{Q}$ by f. Then $f | \Phi_n$, say $\Phi_n = ff_1$ and by Theorem 7.7.2, f may be taken to be monic and over $\mathbf{Z}$.

If p is any prime not dividing n, then ζ^p is also primitive, with minimal polynomial g, say. Thus $f(x)$ and $g(x^p)$ have a common zero ζ, and so

$$g(x^p) = f(x)h(x). \tag{7.7.6}$$

We claim that f and g have a common zero. For if not, then the zeros of f and g are distinct zeros of $x^n - 1$, and so

$$x^n - 1 = f(x)g(x)k(x). \tag{7.7.7}$$

Now both (7.7.6) and (7.7.7) have integer coefficients, so we can reduce them mod p. As a polynomial over $\mathbf{F}_p$, the right-hand side of (7.7.7) has distinct zeros, because its derivative nx^{n-1} is not zero in $\mathbf{F}_p$. But (7.7.6), taken mod p, becomes $f(x)h(x) \equiv g(x^p) \equiv g(x)^p$, so that f and g have a common zero mod p, a contradiction.

Thus we have shown that f and g have a common zero, and by irreducibility, taking them both to be monic, we find that $f = g$. This means that $f(\zeta) = 0$ implies $f(\zeta^p) = 0$. This holds for any p not dividing n, and by induction, for any m prime to n. Since $f(\zeta) = 0$, it follows that $f(\zeta^m) = 0$ for all m prime to n. Hence f has $\varphi(n)$ distinct zeros, and so $f = \Phi_n$, i.e. Φ_n is irreducible, as claimed. ∎

We note that over $\mathbf{F}_p$, Φ_n may become reducible. For example, if $n = 12$ and $p = 11$, then $\Phi_{12} = x^4 - x^2 + 1 = (x^2 - 5x + 1)(x^2 + 5x + 1)$, while for $p = 13$, $x^4 - x^2 + 1 = (x - 2)(x + 2)(x - 6)(x + 6)$. This is not surprising, for as we shall see below, Φ_{12} becomes reducible over $\mathbf{Q}(\sqrt{3})$ and $3 \equiv 5^2 \pmod{11}$, while $3 \equiv 4^2 \pmod{13}$.

Next we examine the Galois group of a cyclotomic extension. For any positive integer n, let us write $\mathbf{U}(n)$ for the multiplicative group of the prime residue classes mod n, i.e. the group of units of $\mathbf{Z}/(n)$. By the definition of $\varphi(n)$ we have $|\mathbf{U}(n)| = \varphi(n)$. Now if ζ is a primitive n-th root of 1 over $\mathbf{Q}$, then the other primitive n-th roots are of the form ζ^m, where m runs over the prime residue classes mod m. Thus the automorphisms of $\mathbf{Q}(\zeta)/\mathbf{Q}$ are given by

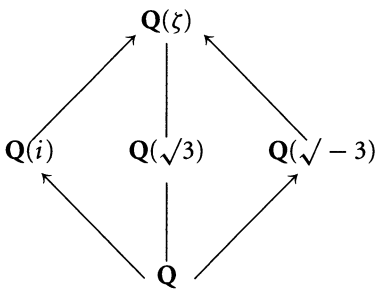
$$\alpha_m : \zeta \mapsto \zeta^m \quad (m \in \mathbf{U}(n)).$$

Clearly $\alpha_r\alpha_s = \alpha_{rs}$, so the mapping $m \mapsto \alpha_m$ is a homomorphism of $U(n)$ onto $\mathrm{Gal}(\mathbf{Q}(\zeta)/\mathbf{Q})$, and since both groups have the same order, they are isomorphic. Thus we have

Theorem 7.7.5. *The Galois group of the cyclotomic extension defined by Φ_n over $\mathbf{Q}$* is isomorphic to $\mathbf{U}(n)$, *the group of units in $\mathbf{Z}/(n)$.* ∎

This shows in particular that the Galois group of a cyclotomic extension is abelian. Kronecker first stated that conversely, every abelian extension of $\mathbf{Q}$ is contained in a cyclotomic extension; it was proved by Weber in 1886 and later more simply by Hilbert (see Weber (1896) Vol. II, p. 762).

As an example of a cyclotomic extension let us take $n = 12$, $\varphi(12) = 4$. The prime residue classes are 1, 5, 7, 11, and the group $\mathrm{Gal}(\mathbf{Q}(\zeta)/\mathbf{Q})$ is generated by the set $\{\alpha, \beta\}$, where $\alpha : \zeta \mapsto \zeta^5$, $\beta : \zeta \mapsto \zeta^7$. We note that $\alpha\beta : \zeta \mapsto \zeta^{35} = \zeta^{11}$, thus G is the Klein 4-group. There are three subgroups of order 2 corresponding to three quadratic extensions. We find them as follows: $\zeta^3 = i$ is left fixed by α and $\zeta^4 = (-1 + \sqrt{-3})/2$ is left fixed by β, hence we have the intermediate fields $\mathbf{Q}(i)$, $\mathbf{Q}(\sqrt{-3})$ and $\mathbf{Q}(\sqrt{3})$. Here $\mathbf{Q}(i)$ is generated by i, a primitive fourth root of 1, hence a root of $x^{12} = 1$; similarly for $\mathbf{Q}(\sqrt{-3})$. However, the field $\mathbf{Q}(\sqrt{3})$ is not generated by a 12-th root of 1, for the only such roots contained in it are ± 1.



Let us consider Φ_n over $\mathbf{F}_p$, where p is a prime not dividing n. Then $x^n - 1$ is separable over $\mathbf{F}_p$ and the primitive n-th roots of 1 in $\mathbf{F}_p$ are roots of $\Phi_n(x) = 0$, where this equation may now be reducible. We first determine when Φ_n has a linear factor over $\mathbf{F}_p$ or, what is the same, when $\mathbf{F}_p$ contains a primitive n-th root of 1.

Proposition 7.7.6. *The finite field $\mathbf{F}_p$ contains a primitive n-th root of 1 if and only if* $p \equiv 1 \pmod{n}$.

Proof. Any element c of $\mathbf{F}_p$ satisfies $c^{p-1} = 1$, so if c has order n, then $n|p-1$. Conversely, when $n|p-1$, then since $\mathbf{F}_p$ is cyclic of order $p-1$, it follows by Theorem 7.7.3 that $\mathbf{F}_p$ contains an element of order n. ■

From this remark it is easy to deduce a special case of Dirichlet's theorem on primes in arithmetic progressions. This states that for any coprime integers m, n there are infinitely many primes of the form $nr + m$ ($r \in \mathbf{N}$). Most proofs require the theory of the ζ-function or other analytical methods, but for $m = 1$ there is a simple direct proof.

Theorem 7.7.7. *For every positive integer n there are infinitely many primes congruent to* $1 (\mathrm{mod}\ n)$.

Proof. Assume that the number of such primes is finite, say $p_1, \ldots, p_r$. Put $a = np_1 \ldots p_r$ and let $t \in \mathbf{Z}$; then $\Phi_n(at) \equiv \Phi_n(0) \equiv \pm 1 \pmod{a}$, hence $a = np_1 \ldots p_r | \Phi_n(at) \mp 1$. As $t \to \infty$, $\Phi_n(at) \to \infty$, so we have $\Phi_n(at) \neq \pm 1$ for large enough t. Therefore $\Phi_n(at)$ is then divisible by a prime p, and since $n|a$ and $a|\Phi_n(at) \mp 1$, n is prime to p. This means that at is a primitive n-th root of 1 in F, i.e. it has multiplicative order $n \pmod{p}$, hence $p \equiv 1 \pmod{n}$, by Proposition 7.7.6, and $p \neq p_i$ because $\Phi_n(at) \equiv 0 \pmod{p}$, $\Phi_n(at) \equiv \pm 1 \pmod{p_i}$, $i = 1, \ldots, r$. Thus p is another prime $\equiv 1 \pmod{n}$. ■

Exercises

1. Let G be a finite group in which there are at most n elements of order dividing n, for any integer n. Show that G is cyclic.
2. Find Φ_n for $n = 18, 24, 50$.
3. Let p be a prime, m be a positive integer and $q = p^m$, $r = p^{m-1}$. Show that $\Phi_q(x) = 1 + x^r + x^{2r} + \ldots + x^{(p-1)r}$.
4. Show that Φ_m is irreducible over a minimal splitting field of Φ_n over $\mathbf{Q}$ iff the highest common factor of m and n is 1 or 2.
5. Prove that $\Phi_n(1) = \prod d^{\mu(n/d)}$, where d ranges over all divisors of n. Deduce that $\Phi_n(1)$ is 0, p or 1 according as n is 1, p^m (p a prime) or divisible by at least two primes.
6. Solve $x^2 = 2$ over $\mathbf{F}_5$.
7. Let p, q be distinct primes. Show that $x^q - 1$ splits into linear factors over $\mathbf{F}_p$ iff $p \equiv 1 \pmod{q}$.
8. Show that $\sum \mu(d) = \delta_{n1}$, $\sum |\mu(d)| = 2^k$, where each time the sum is over all divisors d of n, and k is the number of distinct prime factors of n.
9. Show that for a given positive number M the number of integers n such that $\varphi(n) < M$ is finite. Deduce that any finitely generated field extension of $\mathbf{Q}$ contains only finitely many roots of 1.

7.8 Finite Fields

A field with a finite number of elements is called a *finite field*, or a *Galois field*, after its discoverer. The alternative name is used in some books to avoid confusion with finite extensions.

Let V be an n-dimensional vector space over $\mathbf{F}_p$, the field of p elements. Taking a basis $u_1, \ldots, u_n$ of V, we can write every element of V uniquely as $\sum \alpha_i u_i$, where $\alpha_i \in \mathbf{F}_p$. Since each coefficient can assume p values independently, we obtain p^n elements in all:

Lemma 7.8.1. *An n-dimensional vector space over $\mathbf{F}_p$ has p^n elements.* ■

Clearly this result does not depend on p being prime, so it holds more generally for any finite field.

Any finite field F clearly has prime characteristic p and its prime subfield is $\mathbf{F}_p$. Hence F is a finite-dimensional space over $\mathbf{F}_p$ and it follows that the number of elements in F is a prime power p^n, where $p = \operatorname{char} F$ and $n = [F : \mathbf{F}_p]$.

If F is a field of q elements, then the multiplicative group $F^\times$ of F has $q - 1$ elements, hence every non-zero element of F satisfies the equation

$$x^{q-1} = 1. \tag{7.8.1}$$

By Theorem 7.7.3 it follows that the group $F^\times$ is cyclic. Any generator of $F^\times$ is called a *primitive element* of F. If ζ is a primitive element, then every element of $F^\times$ can be written as a power ζ^a, and these powers can be used as logarithms; we shall meet an example later.

Since (7.8.1) holds identically in $F^\times$, it follows that every element of F, zero or not, satisfies

$$x^q = x. \tag{7.8.2}$$

This equation has at most q roots in any field; therefore its roots are precisely the elements of F. Since $|F| = q$, it follows that (7.8.2) has distinct roots in F. This can also be checked directly by taking derivatives and using Proposition 7.4.2: $(x^q - x)' = qx^{q-1} - 1 = -1 \neq 0$. We can therefore write

$$x^q - x = \prod_{a \in F} (x - a), \tag{7.8.3}$$

and this shows that F is a minimal splitting field of $x^q - x$ over its prime subfield $\mathbf{F}_p$. This description of F shows that it is determined up to isomorphism by q. Thus for any integer q there can be at most one field of q elements, and only when q is a prime power. Conversely, when $q = p^n$, where p is a prime and $n \geq 1$, there is a field of q elements, namely the splitting field of (7.8.2) over $\mathbf{F}_p$. To show that this splitting field has exactly q elements we observe that the roots of (7.8.2) already form a field: if $a^q = a$ and $b^q = b$, then $(a - b)^q = a^q - b^q = a - b$, $(ab)^q = a^q b^q = ab$ and if $b \neq 0$, then $(b^{-1})^q = (b^q)^{-1} = b^{-1}$. We sum up the result as follows.

Theorem 7.8.2 (E. H. Moore, 1893). *For each prime p and each $n \geq 1$ there is exactly one field of $q = p^n$ elements (up to isomorphism), namely the minimal splitting field of $x^q - x$ over $\mathbf{F}_p$, and these are the only finite fields.* ■

The field of q elements is denoted by $\mathbf{F}_q$ or sometimes by GF(q) (for 'Galois field'). For $q = p$ this agrees with the notation $\mathbf{F}_p$ introduced earlier.

As an example consider $\mathbf{F}_9$; this is the minimal splitting field of $x^9 - x$ over $\mathbf{F}_3$. Its degree over $\mathbf{F}_3$ is 2, so we need to find an irreducible factor of $x^9 - x$ of degree 2. Since $\mathbf{F}_9$ consists of all the zeros of this polynomial, its irreducible factors must all be of degree 1 or 2 over $\mathbf{F}_3$. We have

$$x^9 - x = x(x-1)(x+1)((x^2+1)(x^2+x-1)(x^2-x-1).$$

Let α be a root of $x^2 + 1 = 0$. Then the elements of $\mathbf{F}_9$ can be written as $a + \alpha b$, where $a, b = 0, 1, 2$; in fact, since $\alpha^2 = -1$, we can regard the elements of $\mathbf{F}_9$ as 'complex numbers' over $\mathbf{F}_3$. In this example α is not primitive, because $\alpha^4 = 1$, but $\alpha + 1$ is primitive. Writing $\zeta = \alpha + 1$, we have the index table shown in Table 7.1, where the second row lists the power of ζ representing the entry in the first row, i.e. the logarithm to base ζ:

Table 7.1

x	1	2	α	2α	$\alpha+1$	$\alpha+2$	$2\alpha+1$	$2\alpha+2$
$\log x$	0	4	6	2	1	7	3	5

For example, $(\alpha + 1)(2\alpha + 1) = \zeta.\zeta^3 = \zeta^4 = 2$. Operating with powers of ζ simplifies multiplication, but addition is now more work. For example, the equation $(\alpha + 2) + (\alpha + 1) = 2\alpha$ becomes $\zeta^7 + \zeta = \zeta^2$. To facilitate addition one defines a function $Z(n)$ (the 'Zech logarithm') on the range $\{0, 1, \ldots, q-1\}$ by $\zeta^{Z(n)} = \zeta^n + 1$ for $\zeta^n \neq -1$, and uses the formula

$$\zeta^a + \zeta^b = \zeta^{Z(a-b)+b}.$$

For example, since $Z(6) = 1$, we have $\zeta^7 + \zeta = \zeta^{1+Z(6)} = \zeta^2$.

Let us determine the automorphisms of $\mathbf{F}_q$, where $q = p^n$. As a minimal splitting field of $x^q = x$, $\mathbf{F}_q$ is normal and separable over $\mathbf{F}_p$, hence $\mathbf{F}_q/\mathbf{F}_p$ is a Galois extension, and from Section 7.6 we know that there are $[\mathbf{F}_q : \mathbf{F}_p] = n$ automorphisms. In this special case we can describe them explicitly. Firstly we have the Frobenius mapping

$$\alpha : a \mapsto a^p, \tag{7.8.4}$$

which is an automorphism because $\mathbf{F}_q$ is finite (Theorem 7.4.1). The fixed field of α is the set of solutions of $x^p = x$, i.e. $\mathbf{F}_p$. Iterating α, we obtain

$$\alpha^r : a \mapsto a^{p^r},$$

and this is the identity on $\mathbf{F}_q$ iff $n|r$. Thus α has order n, and its powers are the n automorphisms of $\mathbf{F}_q/\mathbf{F}_p$; there can be no more since $|\mathrm{Gal}(\mathbf{F}_q/\mathbf{F}_p)| = [\mathbf{F}_q : \mathbf{F}_p] = n$.

Theorem 7.8.3. *Let $q = p^n$, where p is a prime and $n \geq 1$. Then $\mathbf{F}_q$ is a Galois extension of $\mathbf{F}_p$ with cyclic Galois group of order n, generated by the Frobenius mapping* (7.8.4). ■

For any $d \geq 1$, every finite field $\mathbf{F}_q$ has an extension of degree d, namely the minimal splitting field of $x^{q^d} - x$, and as in the proof of Theorem 7.8.2 we see that this extension is unique up to isomorphism. Since each finite extension of $\mathbf{F}_q$ is generated by a primitive element, its minimal polynomial has degree d over $\mathbf{F}_q$; thus $\mathbf{F}_q$ has an irreducible polynomial in each degree d. All these extensions are separable, by Proposition 7.4.5. Moreover, every finite extension is normal, for its normal closure has a cyclic Galois group, whose subgroups are therefore all normal. Thus every finite extension of a finite field is Galois.

To find the subextensions of $\mathbf{F}_q$, where $q = p^n$, we need only look for the subgroups of $\mathbf{C}_n$, the cyclic group of order n. It is well known that they correspond to the factors of n: if $n = dm$, then there is a unique subgroup of order d, generated by α^m, with corresponding subfield $\mathbf{F}_{p^m}$. Thus we obtain

Theorem 7.8.4. *For any prime power $q = p^m$ and any integer d the field $\mathbf{F}_q$ has an irreducible polynomial of degree d, and so has an extension of degree d, namely $\mathbf{F}_{q^d}$, and any two extensions of degree d are $\mathbf{F}_q$-isomorphic. Moreover, $\mathbf{F}_{p^m}$ is embeddable in $\mathbf{F}_{p^n}$ if and only if $m|n$; thus the subfields of $\mathbf{F}_{p^n}$ correspond to the factors of n. When $m|n$, $\mathbf{F}_{p^m}$ is the unique subfield of order p^m in $\mathbf{F}_{p^n}$ and $\mathbf{F}_{p^n}/\mathbf{F}_{p^m}$ is a Galois extension with cyclic Galois group of order n/m, generated by α^m, where α is the Frobenius mapping* (7.8.4). ■

One of the major applications of the theory of finite fields is coding theory, which forms the subject of Chapter 10 of FA.

Finite fields share with Boolean algebras the property of being functionally complete (Theorem 3.4.2); thus every function can be represented by a polynomial.

Proposition 7.8.5. *Every finite field is functionally complete. More precisely, any function on $\mathbf{F}_q$ can be represented by a unique polynomial of degree at most $q - 1$ in each variable.*

Proof. We have to show that every mapping $f : \mathbf{F}_q^n \to \mathbf{F}_q$ is a polynomial. For $n = 1$ this follows from the Lagrange interpolation formula. In fact we have the 'point function' $1 - (x - a)^{q-1}$, which is 1 at $a \in \mathbf{F}_q$ and 0 elsewhere, and which leads to the explicit formula

$$f(x) = \sum_{a \in \mathbf{F}_q} f(a)[1 - (x - a)^{q-1}]. \tag{7.8.5}$$

For general n we have similarly,

$$f(x_1, \ldots, x_n) = \sum f(a_1, \ldots, a_n) \prod_{i=1}^{n} [1 - (x_i - a_i)^{q-1}], \tag{7.8.6}$$

where the sum is taken over all $(a_1, \ldots, a_n) \in \mathbf{F}_q^n$. For the product on the right of (7.8.6) is 1 when $x_i = a_i$ $(i = 1, \ldots, n)$ and 0 otherwise. Further, (7.8.6) is of degree $< q$ in each x_i; if we had two such polynomials for f, their difference would be a polynomial of degree $< q$ vanishing on all of $\mathbf{F}_q$, i.e. on q values, which is only possible when this difference is zero. ■

We end this section with the remarkable result obtained by Wedderburn in 1905 that all finite division rings are commutative. The proof given here is due to Witt [1931]. We recall the following elementary fact: if q, m, n are positive integers such that $q > 1$ and $q^m - 1 | q^n - 1$, then $m | n$. For we can write $n = ma + b$, where $0 \leq b < m$, by the division algorithm. Then (mod $q^m - 1$) we have $q^n \equiv q^{ma} q^b \equiv q^b$, and this can only be 1 if $b = 0$.

Theorem 7.8.6 (Wedderburn). *Every finite division ring is commutative.*

Proof. Let E be a finite division ring. Its centre k must be a finite field, with q elements say, and taking a basis for E over k, we see that $|E| = q^n$, where $n = [E : k]$. Let us consider the multiplicative group $E^\times$ of E; it has order $q^n - 1$. If $\alpha \in E \backslash k$, then the centralizer of α is a proper subalgebra C of E, of order q^d say. The group $C^\times$ has order $q^d - 1$ and since it is a subgroup of E, we must have $q^d - 1 | q^n - 1$. By the remark made earlier this is possible only if $d | n$. Now the conjugates (in the group sense) of α in $E^\times$ correspond to the cosets of $C^\times$ in $E^\times$; hence the number of conjugates is $(q^n - 1)/(q^d - 1)$. This accounts for all elements of $E^\times$ outside the centre, while the centre has order $q - 1$. Since the conjugacy classes form a partition of $E^\times$, we obtain the equation

$$q^n - 1 = q - 1 + \sum \frac{q^n - 1}{q^d - 1}, \tag{7.8.7}$$

where the sum on the right is taken over various proper divisors d of n (possibly repeated). Now consider the cyclotomic polynomial $\Phi_n(x)$. It is a factor of $(x^n - 1)/(x^d - 1)$ for every proper factor d of n, hence the integer $r = \Phi_n(q)$ divides each term in the sum on the right of (7.8.7), as well as the left-hand side. Therefore

$$r | q - 1. \tag{7.8.8}$$

But $r = \Phi_n(q) = \prod (q - \zeta)$, where ζ runs over the primitive n-th roots of 1. Taking the roots to be complex numbers, we see that for $n > 1$, $|q - \zeta| > q - 1$, and hence $r > q - 1$. This contradicts (7.8.8), hence $n = 1$, and so $E = k$ is commutative. ■

Exercises

Unless otherwise specified, p is a prime and q is a prime power.

1. How many elements of $\mathbf{F}_q$ are generators? How many are primitive?
2. Show that for $q = p^n$, $x^q - x$ considered over $\mathbf{F}_p$ has irreducible factors of degree n but of no higher degree.

3. Show that an irreducible polynomial of degree r over $\mathbf{F}_q$ is a factor of $x^{q^n} - x$ iff $r|n$. Deduce that $x^{q^n} - x = \prod f_i(x)$, where f_i runs over all monic irreducible polynomials whose degrees divide n. Show that if t_r is the number of such polynomials, then $\sum rt_r = q^n$, and deduce a formula for t_r in terms of q, r and the Möbius function.
4. Show that $x^p - x - a$ $(a \neq 0)$ is irreducible over $\mathbf{F}_p$. Over $\mathbf{F}_q$, where $q = p^n$, show that $x^p - x - a$ is irreducible iff it has no linear factor. (Hint. If α is a zero, then so is $\alpha + 1$.)
5. Let $\mathfrak{p}$ be a non-zero prime ideal in the ring $\mathbf{Z}[i]$ of Gaussian integers (see Section 5.1). Show that $\mathbf{Z}[i]/\mathfrak{p}$ is a finite field; further show that the only fields that can occur are of p or p^2 elements, where p is a prime. Describe the quotient rings by the ideals (7), $(2 + i)$, (5); which are fields?
6. For any $r > 0$, denote by S_r the sum of the r-th powers of the elements of $\mathbf{F}_q$. If $q - 1$ does not divide r, find $y \in \mathbf{F}_q^\times$ such that $y^r \neq 1$, and deduce that $S_r = 0$. Prove the formula

$$S_r = \begin{cases} -1 & \text{if } q - 1 | r, \\ 0 & \text{otherwise.} \end{cases}$$

7. Put $k = \mathbf{F}_q$, where $q = p^s$, take $f_1, \ldots, f_r \in k[x_1, \ldots, x_n]$ such that $\sum \deg f_i < n$ and let V be the subset of k consisting of the common zeros of all the f_i. Verify that $P = \prod (1 - f_i^{q-1})$ is the characteristic function for V, i.e. for any $\xi = (\xi_1, \ldots, \xi_n) \in k^n$,

$$P(\xi) = \begin{cases} 1 & \text{if } \xi \in V, \\ 0 & \text{if } \xi \notin V. \end{cases}$$

   For any $f \in k[x_1, \ldots, x_n]$ write $S(f)$ for the sum $\sum f(\xi)$ taken over all $\xi \in k^n$. Deduce that $|V| \equiv S(P) \pmod{p}$, and by expressing P as a linear combination of monomials $x_1^{r_1} \ldots x_n^{r_n}$ with $\sum r_i < n(q - 1)$ and using Exercise 6, show that $S(P) = 0$. Deduce that the number of points on V is divisible by p. If the f_i are homogeneous, they have at least one common solution 0, hence V is then non-empty (Chevalley–Warning).
8. Show that every quadratic form in at least three variables over $\mathbf{F}_q$ vanishes at a point other than 0 (in particular, 'every conic over a finite field has a rational point').
9. Let ω be a primitive $(p - 1)$-th root of 1 in $\mathbf{F}_p$. Show that the affine group $G : x \mapsto ax + b$, $a, b \in \mathbf{F}_p$, $a \neq 0$, is generated by the mappings $x\tau = x + 1$ and $x\rho = \omega x$, and the subgroup generated by τ is normal in G. Prove that every automorphism of G is inner.

7.9 Primitive Elements; Norm and Trace

We have seen that a finite field extension F/k, regarded as a k-algebra, may be completely described in terms of k and the basis products $u_i u_j$. It is therefore in our interest to choose the basis in as simple a form as possible. In particular we

shall examine the case when F is *simple*, i.e. generated by a single element over k. Such a generating element for F is called a *primitive element* over k. We have seen in Section 7.1 that in this case F has a k-basis of the form $1, \alpha, \ldots, \alpha^{n-1}$. We shall find that every finite separable extension is simple, but some preparation is necessary.

First of all we recall the density principle for polynomials (sometimes called the principle of irrelevance of algebraic inequalities): for any non-zero polynomial f in $x_1, \ldots, x_r$ over an infinite field k there exist $c_1, \ldots, c_r \in k$ such that $f(c_1, \ldots, c_r) \neq 0$. For $r = 1$ this follows because a polynomial of degree n cannot vanish at more than n points; for $r > 1$ the result follows by looking at the coefficients of the powers x_1^i and using induction on r. Secondly we need an almost obvious remark on the Galois correspondence:

Lemma 7.9.1. *Let E/k be a Galois extension with group G and consider a subgroup H of G with corresponding subfield F. Given $\alpha_1, \ldots, \alpha_r \in F$, we have $k(\alpha_1, \ldots, \alpha_r) = F$ if and only if the group leaving each of $\alpha_1, \ldots, \alpha_r$ fixed is equal to H.*

Proof. The subgroup fixing $\alpha_1, \ldots, \alpha_r$ corresponds to $k(\alpha_1, \ldots, \alpha_r)$ and so is equal to H iff $k(\alpha_1, \ldots, \alpha_r) = F$. ∎

We can now show that finite separable extensions are simple.

Theorem 7.9.2 (Theorem of the primitive element). *If F is a finite separable extension of k then F has a primitive element,* i.e. *there exists $\beta \in F$ such that $F = k(\beta)$.*

Proof. When k is finite, then so is F and the conclusion certainly holds by Theorems 7.8.2 and 7.7.3. We may therefore take k to be infinite.

Let E be the normal closure of F over k, with Galois group G, and let H be the subgroup corresponding to F. If $F = k(\alpha_1, \ldots, \alpha_r)$, then H is the fixed group of $\alpha_1, \ldots, \alpha_r$ and we have to find $\beta \in F$ with fixed group exactly H. Since any $\beta \in F$ has fixed group containing H, this means that every element of $G \backslash H$ must move β.

Adjoin indeterminates $t_1, \ldots, t_r$ and consider $\lambda = \sum \alpha_i t_i$. Any $\sigma \in G$ acts by $\lambda^\sigma = \sum \alpha_i^\sigma t_i$, so σ moves λ iff $\sigma \notin H$. Hence

$$\varphi(t_1, \ldots, t_r) = \prod_{\sigma \notin H} (\lambda^\sigma - \lambda) \neq 0.$$

By density we can specialize t_i to $c_i \in k$ such that $\varphi(c_1, \ldots, c_r) \neq 0$. So if $\beta = \sum \alpha_i c_i$, then $\beta^\sigma \neq \beta$ for $\sigma \notin H$, and it follows that $F = k(\beta)$. ∎

A closer analysis shows that any finite extension $k(\alpha_1, \ldots, \alpha_r)/k$, where all but at most one of the α_i are separable, is still simple (Exercise 3), but there are examples of non-simple extensions generated by two elements (Exercise 5). The situation is clarified by the following criterion for an extension to be simple.

Theorem 7.9.3 (Steinitz). *A finite extension F/k is simple if and only if the number of fields between F and k is finite.*

Proof. Assume that $F = k(\alpha)$ and let f be the minimal polynomial for α over k. For any field E between F and k denote by p_E the minimal polynomial for α over E. It follows that $p_E | f$ and we have a correspondence $E \mapsto p_E$ between subfields and factors of f. Since a polynomial of degree n has at most 2^n monic factors (as we see by looking at a complete factorization in a splitting field), we need only show that the correspondence $E \mapsto p_E$ is injective, i.e. that E is determined by p_E.

Given E and p_E as above, let E' be the field obtained by adjoining the coefficients of p_E to k. Then $E' \subseteq E$ and p_E is irreducible over E', hence $[F : E] = [F : E'] = \deg p_E$, and it follows that $E' = E$. Thus E is determined by p_E and we have shown that when F/k is simple, of degree n, then there are at most 2^n intermediate fields.

Conversely, assume that there are only finitely many fields between F and k. The case where k is finite has been dealt with in Theorem 7.9.2, so we may take k to be infinite.

Given $\alpha, \beta \in F$ and $a \in k$, write $\gamma_a = \alpha + a\beta$. As a runs through k, we get an infinite family $\{\gamma_a\}$, but by hypothesis there are only finitely many fields between k and F, so there exist $a, b \in k, a \neq b$, such that $k(\gamma_a) = k(\gamma_b) = E$, say. Then $\gamma_a, \gamma_b \in E$, and solving the equations

$$\alpha + a\beta = \gamma_a, \quad \alpha + b\beta = \gamma_b,$$

for α, β, we find $\alpha, \beta \in E$ and hence

$$E = k(\alpha, \beta) = k(\gamma), \quad \text{where } \gamma = \gamma_a.$$

Here α, β were arbitrary in F, so we see that any extension $k(\alpha, \beta)$ of k is simple. Now choose $\alpha \in F$ such that its degree $[k(\alpha) : k]$ is maximal. If $k(\alpha) \neq F$, take $\beta \in F \backslash k(\alpha)$. By what has been shown, we have $k(\alpha) \subset k(\alpha, \beta) = k(\gamma)$ for some $\gamma \in F$, but this contradicts the maximality of $k(\alpha)$. Hence $k(\alpha) = F$, as required. ∎

This result shows again that every separable extension is simple, for it is contained in a Galois extension, by Proposition 7.6.1, and the intermediate fields of the latter correspond to the subgroups of the Galois group.

With every element α of an algebraic extension we can associate two elements of the ground field, its trace and norm, analogous to the trace and determinant of an endomorphism, and in fact equal to the latter for the endomorphism α_R of right multiplication. This may be regarded as a special case of the development in Section 5.5. Thus let F/k be a separable extension of degree n, L be any Galois extension of k containing F and $\sigma_1 = 1, \sigma_2, \ldots, \sigma_n$ be the n automorphisms of L over k defining distinct k-homomorphisms of F into L. If $G = \mathrm{Gal}(L/k)$ and H is the subgroup corresponding to F, then $(G : H) = n$ and $\sigma_1, \ldots, \sigma_n$ is a right transversal of H in G, i.e. G has the coset decomposition $G = \cup H\sigma_i$. We recall that for any $\alpha \in F$, the elements $\alpha^{\sigma_1} = \alpha, \alpha^{\sigma_2}, \ldots, \alpha^{\sigma_n}$ are the conjugates of α; they need not lie in F but do so whenever F/k is normal. We define mappings from F to k as follows:

$$N_{F/k}(\alpha) = N(\alpha) = \prod \alpha^{\sigma_i}, \tag{7.9.1}$$

$$T_{F/k}(\alpha) = T(\alpha) = \sum \alpha^{\sigma_i}. \tag{7.9.2}$$

$N(\alpha)$ is called the *norm* and $T(\alpha)$ the *trace* of α. Both lie in k, for any $\tau \in G$ permutes the n cosets: $H\sigma_i \mapsto H\sigma_i\tau$. Hence the elements on the right of (7.9.1), (7.9.2) are unchanged except for order, and so $N(\alpha)^\tau = N(\alpha)$, $T(\alpha)^\tau = T(\alpha)$. We again have the formulae for sums and products as in Section 5.5, since the σ_i are automorphisms:

$$N(\alpha\beta) = N(\alpha)N(\beta), \quad T(\alpha + \beta) = T(\alpha) + T(\beta),$$

$$N(\lambda\alpha) = \lambda^n N(\alpha), \quad T(\lambda\alpha) = \lambda T(\alpha), (\lambda \in k),$$

$$N(1) = 1, \quad T(1) = n.$$

We note that although we needed a Galois extension L to define the norm and trace, the actual values are independent of the choice of L: the norm and trace can be defined in terms of any Galois extension containing F and the outcome will be the same. This is easily verified directly but also follows from Proposition 7.9.4 below. We note the important

Transitivity Formulae. *If $E \supseteq F \supseteq k$ are separable extensions, then for any $\alpha \in E$,*

$$N_{E/k}(\alpha) = N_{F/k}(N_{E/F}(\alpha)), \quad T_{E/k}(\alpha) = T_{F/k}(T_{E/F}(\alpha)). \tag{7.9.3}$$

Proof of (7.9.3). Denote by L/k any Galois extension containing E/k. Let $\sigma_1, \ldots, \sigma_n$ be the k-homomorphisms of F into L and extend each σ_i to an automorphism of L, again denoted by σ_i (Corollary 7.2.5). Further let $\tau_1, \ldots, \tau_m$ be the F-homomorphisms of E into L. Every k-homomorphism φ of E into L, when restricted to F, agrees with some σ_i, hence $\varphi\sigma_i^{-1}$ leaves F pointwise fixed and so $\varphi\sigma_i^{-1} = \tau_j$ for some j. Thus $\varphi = \tau_j\sigma_i$ for a unique pair of indices (j,i). Now

$$T_{E/k}(\alpha) = \sum \alpha^{\tau_j\sigma_i} = \sum_j T_{F/k}(\alpha^{\tau_j}) = T_{F/k}(T_{E/F}(\alpha)),$$

and similarly for $N(\alpha)$. ∎

Proposition 7.9.4. *If F/k is a separable extension of degree n and $\alpha \in F$ has the minimal polynomial $x^r + a_1x^{r-1} + \ldots + a_r$, then*

$$N_{F/k}(\alpha) = [(-1)^r a_r]^{n/r}, \quad T_{F/k}(\alpha) = -(n/r)a_1. \tag{7.9.4}$$

Proof. This follows by applying the transitivity formula to the tower $k \subseteq k(\alpha) \subseteq F$; the latter also shows that $r|n$. ∎

The trace can be used to define an inner product on E which is often useful. Let us put

$$T(x, y) = T_{F/k}(xy). \tag{7.9.5}$$

By the properties of the trace this is k-bilinear and symmetric; thus we have an inner product defined on F with values in k. Its usefulness derives from the fact that it is non-singular, i.e. $T(x, a) = 0$ for all $x \in F$ implies $a = 0$ (see Section 8.1):

Proposition 7.9.5. *For any separable extension F/k, the inner product defined by* (7.9.5) *is non-singular.*

Proof. We need only show that the matrix $T(u_i, u_j)$ of the form (7.9.5) relative to a basis $u_1, \ldots, u_n$ of F over k is non-singular (see Section 8.1 below). Let $\sigma_1, \ldots, \sigma_n$ be the k-homomorphisms of F into a normal closure E; then

$$T(u_i, u_j) = \sum_\nu u_i^{\sigma_\nu} u_j^{\sigma_\nu}.$$

Hence the matrix $(T(u_i, u_j))$ has the form

$$(T(u_i, u_j)) = DD^T, \quad \text{where } D = (d_{i\nu}),\ d_{i\nu} = u_i^{\sigma_\nu}. \tag{7.9.6}$$

Now by Dedekind's lemma (Lemma 7.5.1), the system of equations

$$\sum_\nu u_i^{\sigma_\nu} x_\nu = 0$$

has only the trivial solution $x_\nu = 0$ over E, hence D and with it $(T(u_i, u_j))$ is non-singular, as claimed. ∎

The value of $(T(u_i, u_j))$ is called the *discriminant* of the extension F/k. It is not quite independent of the choice of basis; when we change the basis, the discriminant is multiplied by the square of an element of the ground field. So strictly speaking, the discriminant is a coset of the subgroup of squares in $k^\times$. If every element in k is a square, this tells us nothing; but for example, over a real field, if the discriminant for a given basis is positive, then it is positive for all bases.

In conclusion we note a useful property of norms and traces of finite fields.

Proposition 7.9.6. *Let k be a finite field. Then for any finite field extension F/k, the trace and norm are surjective.*

Proof. The trace is a k-linear function from F to k, and it is non-zero, by Proposition 7.9.5, hence it is surjective (because the image is one-dimensional). Here we have not used the fact that the field is finite, but merely that we have a Galois extension.

To prove that every element of k is a norm we observe that F/k is a Galois extension, with cyclic group generated by $\tau : x \mapsto x^q$, where $q = |k|$. Let $[F : k] = n$ and consider the norm homomorphism $N : F^\times \to k^\times$. For $a \in F$ we have $\mathrm{N}(a) = a^t$, where $t = 1 + q + \ldots + q^{n-1} = (q^n - 1)/(q - 1)$; hence $\ker N$ consists of the solutions in F of $a^t = 1$. Thus $|\ker N| \leq t$, and so $|\mathrm{im}\, N| \geq (q^n - 1)/t = q - 1$. Since $|k^\times| = q - 1$, we must have $\mathrm{im}\, N = k^\times$, and this shows N to be surjective. ∎

Exercises

1. Find a primitive element for $Q(\sqrt[2]{2}, \sqrt[3]{3}, \sqrt[5]{5})$.
2. Let F/k be a separable extension of degree n. Show that an element of F is primitive iff it has n conjugates in a normal closure of F/k.

3. By examining the proof of Theorem 7.9.3 show that if α is algebraic and β is separable over k, then $k(\alpha, \beta)/k$ is simple. Deduce that $k(\alpha_1, \ldots, \alpha_r)/k$ is simple provided that the α's are algebraic over k and all but at most one of them are separable.
4. Use Exercise 3 to show that if a non-simple finite extension of k exists, then there is one generated by two elements.
5. Let k be a field of prime characteristic p, $E = k(x, y)$ be the field of rational functions in two indeterminates x, y and write $F = k(x^p, y^p)$. Show that every element of E has degree 1 or p over F; deduce that E/F is not simple. Find infinitely many fields between E and F.
6. Let F/k be a separable extension of degree n. Show that n elements $a_1, \ldots, a_n$ of F form a basis iff $(T(a_i a_j))$ is non-singular.
7. Show that if F/k is a Galois extension of degree n, then $a_1, \ldots, a_n$ is a basis iff the matrix (a_i^σ) is non-singular, where the columns correspond to the different elements σ of the group $\mathrm{Gal}(F/k)$.
8. Let k be a finite field and F/k be a finite extension. Show that every k-linear map from F to k has the form $\lambda_\alpha : x \mapsto T(\alpha x)$, for a unique $\alpha \in F$.
9. For any $a_1, \ldots, a_n$ define the *Vandermonde matrix* as the matrix (a_i^j) $(j = 0, 1, \ldots, n-1)$ and call its determinant the *Vandermonde determinant*. Show that its value is $\prod_{i>j}(a_i - a_j)$.
10. Let f be a separable polynomial of degree n, irreducible over k, and let α be a zero of f in a splitting field. By taking a basis of the form $1, \alpha, \ldots, \alpha^{n-1}$ for $k(\alpha)$ over k, show that the discriminant of $k(\alpha)$ over k can be written as a product of two Vandermonde determinants (see Exercise 9). Deduce that the discriminant is the product of the squares of the differences of the roots of $f = 0$.

7.10 Galois Theory of Equations

We have seen that Galois theory may be described as the analysis of field extensions by means of automorphism groups. However, originally it was associated with the solution of equations, and in this section we shall describe the connexion. It is based on the fact that every polynomial equation $f(x) = 0$ over a field k defines a minimal splitting field E and if f is separable, E/k is a Galois extension. In particular, in characteristic 0 (the only case considered classically) all finite extensions are separable. Even in finite characteristic the separable case is the most important. For example, in algebraic number theory the residue class fields for prime ideals are finite fields, and hence perfect. The inseparable case arises mostly in algebraic geometry.

Consider a separable polynomial f over k, with minimal splitting field E. Then

$$E = k(\alpha_1, \ldots, \alpha_n),$$

where $\alpha_1, \ldots, \alpha_n$ are the roots of $f = 0$ in E. The Galois group $G = \mathrm{Gal}(E/k)$ is also called the *group* of the equation $f = 0$ over k. Any automorphism of E/k is completely determined by its effect on $\Sigma = \{\alpha_1, \ldots, \alpha_n\}$; moreover it must send any root to another root, so it maps Σ into itself, and being invertible, it therefore

defines a permutation of Σ . Thus any automorphism of E/k can be specified completely by the permutation of Σ it induces, and so $\mathrm{Gal}(E/k)$ is isomorphic to a subgroup of Sym_n, the symmetric group of degree n. Of course it will not in general be the whole of Sym_n, e.g. any α_i which lies in k must stay fixed.

As an example consider the equation $x^3 - 2 = 0$ over $\mathbf{Q}$. If α is a root and ω is a primitive cube root of 1, then $E = \mathbf{Q}(\alpha, \omega)$ is a minimal splitting field. Any permutation of the roots defines an automorphism of E, hence the group is the full symmetric group, in agreement with the fact that $[E : \mathbf{Q}] = 6$. Over $\mathbf{Q}(\omega)$ the same equation has the cyclic group of order 3: the roots $\alpha, \alpha\omega, \alpha\omega^2$ can be permuted cyclically, but not in any other way, while over $\mathbf{Q}(\alpha)$ we have the cyclic group of order 2: we can only interchange $\alpha\omega$ and $\alpha\omega^2$, while α remains fixed.

We first note a criterion for irreducibility in terms of the Galois group.

Theorem 7.10.1. *Let f be a separable polynomial over k. Then f is irreducible over k if and only if its group acts transitively on the roots.*

Proof. Let $\alpha_1, \ldots, \alpha_n$ be the roots of $f = 0$ in some splitting field E. Its group G acts by permutations on the set of roots. Assume first that f is irreducible over k; then for each $i = 2, \ldots, n$, $k(\alpha_1) \cong k(\alpha_i)$ under a k-isomorphism which maps α_1 to α_i, by Proposition 7.2.2. This k-isomorphism can be extended to a k-automorphism of E, by Corollary 7.2.5. Thus we obtain an element of G mapping α_1 to α_i and so G is transitive.

Conversely, if G is transitive, and p is the minimal polynomial for α_1 over k, then $p(\alpha_1) = 0$. By transitivity there exists for each i, $2 \leq i \leq n$, $\sigma \in G$ mapping α_1 to α_i. Hence $p(\alpha_i) = p(\alpha_1^\sigma) = [p(\alpha_1)]^\sigma = 0$, so α_i is also a root of $p = 0$. This holds for all i, so all the α_i are roots of $p = 0$, and since the roots are distinct, p must agree with f except for a constant factor. Thus f is irreducible, as claimed. ∎

To explore the connexion between the group and the equation we need to know how the group changes when the field is enlarged. We begin by translating the second isomorphism theorem of group theory, Theorem 2.3.3, to fields. For subfields K, L of a field E we shall write KL for the subfield of E generated by K and L.

Proposition 7.10.2. *Given a Galois extension E/k, let K and L be fields between E and k such that L/k is normal (and hence Galois). Write $\mathrm{Gal}(E/k) = G$ and let H, N be the subgroups corresponding to K and L, respectively. Then KL/K is normal with group $\mathrm{Gal}(KL/K) \cong HN/N$.*

Proof. Since E/KL has group $H \cap N$, which is normal in H, it follows that KL/K is normal with group $H/(H \cap N) \cong HN/N$. ∎

In this result we were limited to extensions K/k which are finite and separable; in fact a more general result holds:

Theorem 7.10.3 (On natural irrationalities). *Let f be a separable polynomial over a field k, let K be any field containing k and let E be a minimal splitting field of f over K. If L denotes the minimal splitting field of f over k contained in E, then E/K and L/k are*

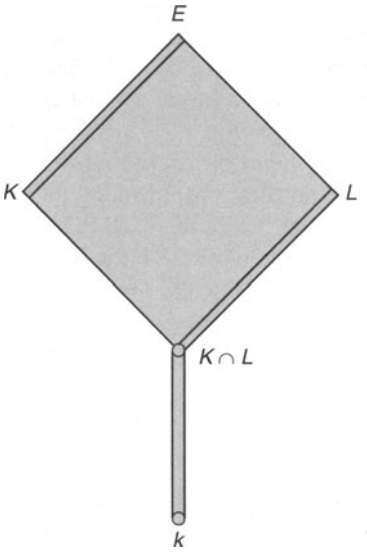


Figure 7.1

normal and Gal(E/K) *is isomorphic to the subgroup of* Gal(L/k) *corresponding to the subfield* $K \cap L$.

Briefly, the assertion is that 'extending the field reduces the Galois group to a subgroup'. An element of L is called a *natural irrationality* for f.

Proof. Since f is separable over k, it is separable over K and so E/K and L/k are separable, hence Galois. Let $\alpha_1, \ldots, \alpha_n$ be the zeros of f in E; then $E = K(\alpha_1, \ldots, \alpha_n)$, $L = k(\alpha_1, \ldots, \alpha_n)$ and Gal(E/K), Gal(L/k) may be regarded as groups of permutations of $\{\alpha_1, \ldots, \alpha_n\}$. Take $\sigma \in \mathrm{Gal}(E/K)$; its restriction $\sigma_o = \sigma|L$ is a homomorphism of L fixing $K \cap L \supseteq k$, hence a k-homomorphism of L and so an automorphism because L/k is normal. The mapping $\sigma \mapsto \sigma_o$ is a homomorphism from Gal(E/K) to Gal(L/k), as is easily checked. Its image is the subgroup corresponding to $K \cap L$, for every automorphism of L fixing $K \cap L$ extends to an automorphism of E, and it is injective, because if $\sigma_o = 1$, then σ leaves each α_i fixed, but that can only happen when $\sigma = 1$. ■

Corollary 7.10.4. *If f is separable over a field k, with group G of prime order, then the group of f over any extension field of k is either G or 1.*

Proof. The group must be a subgroup of G, by Theorem 7.10.3, and there are only the two possibilities stated, by Lagrange's theorem. ■

As another application of Theorem 7.10.3, consider the extension $k(\omega))/k$, where k is a field of characteristic 0 and ω is a primitive n-th root of 1. As we have seen, $\mathbf{Q}(\omega)/\mathbf{Q}$ is abelian, i.e. a Galois extension with abelian group, and k may be regarded

as an extension of $\mathbf{Q}$; hence $k(\omega)/k$ is an abelian extension. This is still true in prime characteristic p, as long as p does not divide n.

When we have a permutation group G, we can always form the subgroup G_+ of all even permutations; this is a subgroup of index 1 or 2 in G. Correspondingly we have, for every equation over k, an extension of degree 1 or 2 which we now identify.

Theorem 7.10.5. *Let k be a field of characteristic not 2 and f be a separable polynomial over k. Denote the roots of $f = 0$ in a splitting field by $\alpha_1, \ldots, \alpha_n$ and form*

$$\delta = \prod_{i<j} (\alpha_i - \alpha_j).$$

If G is the group of f, regarded as a permutation group of the roots, then the group G_+ of even permutations corresponds to the field $k(\delta)$.

Proof. Let D be the fixed field of G_+. Clearly $D \supseteq k$ and $\delta \in D$, hence $k(\delta) \subseteq D$. Further, $[D : k] = (G : G_+) = 1$ or 2; if this index is 1, then $k(\delta) = k$ and there is nothing to prove, so assume that $[D : k] = 2$ and take $\beta \in D \backslash k$. Then there is a permutation σ, necessarily odd, such that $\beta^\sigma \neq \beta$. But then $\delta^\sigma = -\delta$, hence $k \subset k(\delta) \subseteq D$, and since $[D : k] = 2$, it follows that $D = k(\delta)$, as required. ∎

We remark that the discriminant Δ of f (see Exercise 10 of Section 7.9) equals

$$(-1)^{\binom{n}{2}} \prod_{i \neq j} (\alpha_i - \alpha_j) = \delta^2.$$

Hence we obtain

Corollary 7.10.6. *The discriminant Δ of f is a square in k if and only if the group of f over k consists of even permutations only.* ∎

Example 1. The quadratic equation $x^2 + px + q = 0$ (in characteristic not 2) has discriminant $\Delta = p^2 - 4q$. The group over k has order 1 or 2 according as Δ is or is not a square in k.

Example 2. Consider the cubic $f = x^3 + px^2 + qx + r$. If f is reducible, it must have a linear factor: $f = (x - \alpha)g$, where $\alpha \in k$ and g is a quadratic, to which we can apply 1. If f is irreducible, the group is transitive, and so is either Sym_3 or Alt_3. Which it is depends on the discriminant $\Delta = -4p^3r + p^2q^2 + 18pqr - 4q^3 - 27r^2$.

We conclude this section by examining a special case going back to Lagrange: the cyclic extensions. An extension F/k is said to be *cyclic* if it is Galois, with cyclic Galois group. For example, every finite extension of a finite field is cyclic, as we have seen in Section 7.8. By a *radical* extension we understand a simple extension $k(\alpha)/k$, where α is a root of a binomial equation irreducible over k:

$$x^n = a, \quad \text{where } a \in k,$$

and where n is prime to char k. Clearly radical extensions will be of importance in the explicit solution of equations; our first observation is that they are closely related to cyclic extensions:

Proposition 7.10.7. *Let k be a field containing a primitive n-th root of 1 (and hence of characteristic prime to n). Then an extension F/k of degree n is radical if and only if F/k is cyclic.*

Proof. Let ω be a primitive n-th root of 1 in k. If $F = k(\alpha)$, where $\alpha^n = a \in k$, then the equation $x^n = a$ has distinct roots $\alpha, \alpha\omega, \ldots, \alpha\omega^{n-1}$ in F, and any automorphism of F/k has the form

$$\sigma : \alpha \mapsto \alpha\omega^{i(\sigma)},$$

where $i(\sigma)$ is an integer prime to n. It is easily checked that the mapping $\sigma \mapsto i(\sigma)$ is a homomorphism from $\mathrm{Gal}(F/k)$ to $\mathbf{C}_n$, in fact an isomorphism, since it is injective and both sides have the same order n. This shows F/k to be a cyclic extension.

Conversely, if F/k is cyclic of degree n, with group generated by σ, let us take $\alpha \in F$ and form the 'Lagrange resolvent'

$$(\omega, \alpha) = \alpha + \omega^{-1}\alpha^{\sigma} + \omega^{-2}\alpha^{\sigma^2} + \ldots + \omega^{1-n}\alpha^{\sigma^{n-1}}.$$

By Dedekind's lemma this is non-zero for some $\alpha \in F$, and

$$(\omega, \alpha)^{\sigma} = \omega(\omega, \alpha). \tag{7.10.1}$$

Hence $(\omega, \alpha)^n \in k$ and by (7.10.1), (ω, α) has distinct conjugates, so there are n distinct automorphisms of $k((\omega, \alpha))/k$, the minimal equation of (ω, α) has degree n over k and $k((\omega, \alpha)) = F$. ∎

In case n is a prime number p, we need not assume that ω lies in the ground field and we can say rather more:

Theorem 7.10.8. *Let k be a field, p be a prime number and consider the equation*

$$x^p = a, \quad \text{where } a \in k. \tag{7.10.2}$$

Either (7.10.2) *has a linear factor or it is irreducible over k, according as a is or is not a p-th power in k.*

Proof. If a is a p-th power, say $a = b^p$, then $x^p - a$ has the linear factor $x - b$. Thus if $x^p - a$ is irreducible over k, a cannot be a p-th power in k.

Conversely, suppose that $x^p - a$ is reducible over k. A splitting field of (7.10.2) contains a p-th root of a, say α. If char $k = p$, then $x^p - a = (x - \alpha)^p$ and by hypothesis some factor $(x - \alpha)^r$ lies in $k[x]$, where $0 < r < p$. Here the coefficient of x^{r-1} is $-r\alpha$, hence $\alpha \in k$ and so $a = \alpha^p$ is a p-th power in k. If char $k \neq p$, then any splitting field will also contain a primitive p-th root of 1, say ω. Now by hypothesis, $x^p - a$ has a non-trivial factorization over k: $x^p - a = fg$, where f, g are monic of positive degrees r, s respectively and $r + s = p$. The constant term of f is the product of r roots $\alpha\omega^{\nu}$ and hence is of the form $b = \alpha^r\omega_1$, where $\omega_1^p = 1$.

Thus $b^p = \alpha^{pr} = a^r$, and since $0 < r < p$, there exist $u, v \in \mathbf{Z}$ such that $ru + pv = 1$. It follows that $a = a^{ru}a^{pv} = b^{pu}a^{pv} = (b^u a^v)^p$, which shows a to be a p-th power in k, as claimed. ■

If k contains a primitive p-th root of 1, then $x^p - a$ is either irreducible or it splits completely into linear factors, by Proposition 7.10.7, but for general k this need not be so. For example, if $\alpha = 2^{1/3}$, then

$$x^3 - 2 = (x - \alpha)(x^2 + \alpha x + \alpha^2),$$

and here the second factor on the right is irreducible over $\mathbf{Q}(\alpha)$, for its zeros are not real.

We remark that in the case of Proposition 7.10.7 we have $(\omega^i, \alpha)^\sigma = \omega^i(\omega^i, \alpha)$, hence the n-th powers of the resolvents are in the ground field. For an equation of prime degree p, the p-th powers of the resolvents formed from the roots satisfy an equation of degree $p - 1$ whose coefficients are rational functions of a root of an equation of degree $(p - 2)!$ over the ground field. This is at the basis of Lagrange's method of solving cubics (see Exercise 8 of Section 7.11 below). However, when $p = 5$, the resolvent has degree $3! = 6$, so the method cannot be used for $p \geq 5$, and as we shall see in Section 7.11, the general equation of degree 5 or higher cannot be solved by radicals.

Exercises

1. Let f be an irreducible polynomial over k. Show that in a normal extension of k, f splits (if at all) into factors that are all of the same degree and are conjugate over k.
2. Let F/k be a Galois extension with group G. Show that if the adjunction of α to k reduces G to a subgroup H, then the degree of α over k is a multiple of $(G : H)$.
3. Show that an equation is normal iff all its roots can be expressed rationally in terms of a single one.
4. Show that if α satisfies the equation $x^3 - 3x + 1 = 0$, then so does $\alpha^2 - 2$. Hence find its group and solve the equation over $\mathbf{Q}$ in terms of radicals. (Hint. Put $x = u + v$.)
5. Show that an abelian transitive permutation group acts *regularly*, i.e. each permutation either moves all symbols or none. Deduce that an irreducible equation with an abelian group is normal.
6. Suppose that $x^4 - ax^2 + b = 0$ is irreducible over $\mathbf{Q}$. If k is an extension containing no root, show that over k the group is $\mathbf{C}_4$ if $a^2/b - 1$ is a square in k, the Klein 4-group if b is a square, $\mathbf{C}_2$ if both hold and the dihedral group of order 8 if neither holds. Find the group of $x^4 - 2$ over $\mathbf{Q}$.
7. Let k be a field and p be a prime different from char k. Show that for any $a \in k$, the equation $x^p = a$ splits into linear factors over k iff a is a p-th power and k contains a primitive p-th root of 1.
8. Suppose that char $k = p$ and let F/k be a cyclic extension of degree p. Show that if σ is a generating automorphism, then the linear transformation $S : \alpha \mapsto \alpha - \alpha^\sigma$

is nilpotent and hence find an element α in the kernel of S^2 but not of S. Show that $\beta = \alpha/(\alpha^\sigma - \alpha)$ satisfies $\beta^\sigma = \beta + 1$; deduce that β satisfies an equation $x^p - x - a = 0$.

9. Let $V = (\alpha_j^{i-1})$ be the Vandermonde matrix formed from the roots α_i of a separable equation of degree n. By evaluating $\det(V^T V)$ show that the discriminant is the determinant of (s_{i+j-2}), where s_r is the sum of the r-th powers of the roots.

7.11 The Solution of Equations by Radicals

The study of algebraic equations goes back to antiquity; the Babylonians knew two millennia BC how to solve quadratic equations as well as some particular cases of higher degrees, and examples of cubics were solved by Diophantos (AD 300), but the general solution of cubic and quartic equations was not accomplished until the 16th century.

The complete solution of the general cubic is due to Scipio Ferro (1515) and Niccolo Tartaglia, published by Girolamo Cardano in his *Ars Magna* (1545) (see Ore (1953)). This was soon followed by the general solution of the quartic equation (due to Lodovico Ferrari, later also James Gregory). Here 'solving' means giving a formula, involving repeated radicals, for the roots in terms of the coefficients (the custom of using letters for the coefficients had been introduced by François Viète in 1591) and attempts continued during the 17th century to obtain the solution of the general quintic by radicals. At the end of the 18th century Paolo Ruffini published what he claimed to be a proof of the impossibility of solving the general quintic by radicals; the proof was incomplete, although it contained some of the ideas utilized later. This result was finally proved by Niels Henrik Abel in 1826. Of course it must be borne in mind that the problem of solving an equation by radicals is theoretical (rather like a ruler-and-compass construction). In practice, given any equation with numerical coefficients, there are methods for calculating the roots to any degree of accuracy.

Within a few years of Abel's work Evariste Galois developed the correspondence between equations and the groups of permutations of their roots, which shed a very clear light on the subject. The theory of Galois forms the core of modern field theory, and it leads to an explicit description of field extensions.

Until Ernst Steinitz's fundamental paper of 1910 the complications arising in finite characteristic had not been contemplated, for although finite fields also went back to Galois, their algebraic extensions offered no difficulty. Today the problems of 'inseparability' arise mainly in algebraic geometry over a finite field; our account here will deal only with the basic facts, reserving a fuller treatment for Chapter 11.

An equation $f = 0$ is said to be *soluble by radicals* over a given field k, if there exists a finite tower of radical extensions

$$k = k_0 \subset k_1 \subset \ldots \subset k_r, \tag{7.11.1}$$

such that f splits completely in k_r; (7.11.1) is then called a *root tower* for f over k.

We can now establish the connexion with soluble groups; we recall from Section 2.4 that a finite group is soluble iff it has a composition series with factors of prime order, or equivalently, one with abelian factors.

Theorem 7.11.1 (Galois). *An equation $f = 0$ over a field of characteristic 0 is soluble by radicals if and only if the group of f over k is soluble.*

Proof. Assume that f has a soluble group, of order n say; we shall use induction on n to show that f is soluble by radicals. In the first place, $x^n - 1$ has an abelian, hence soluble, group of order $\varphi(n) < n$, by Theorem 7.7.5. By induction we can find a root tower (7.11.1) such that k_r contains a primitive n-th root of 1. Now f has a soluble group over k, and by Theorem 7.10.3, the group of f over k_r is a subgroup, which is still soluble, and so has a normal chain of subgroups with cyclic factors. This corresponds to a tower of fields

$$k_r \subset k_{r+1} \subset \ldots \subset k_m,$$

where k_i/k_{i-1} is cyclic of order dividing n and hence radical, by Proposition 7.10.7. Thus $f = 0$ is soluble by radicals.

Conversely, let $f = 0$ be soluble by radicals and consider a root tower (7.11.1) for f, where $[k_i : k_{i-1}]$ is prime for $i = 1, \ldots, r$. If $[k_r : k] = n$, let ω be a primitive n-th root of 1 and replace the tower (7.11.1) by

$$k \subseteq K_0 \subseteq K_1 \subseteq \ldots \subseteq K_r, \quad \text{where } K_i = k_i(\omega). \tag{7.11.2}$$

By Theorem 7.10.8, the steps which have not become trivial are still radical of prime degree. We now extend the tower to a Galois extension as follows. We know that K_1/k is Galois. If K_2/K_1 is a minimal splitting field of $x^p - a$, $a \in K_1$, we adjoin all the roots of all equations $x^p = a^\tau$, where τ ranges over $\mathrm{Gal}(K_1/k)$. This gives a Galois extension containing K_2 which can be reached by a root tower. Continuing in this way we get another root tower (7.11.2) in which K_r/k is Galois. Again by Proposition 7.10.7, K_i/K_{i-1} is cyclic while K_0/k is abelian; hence $\mathrm{Gal}(K_r/k)$ has a normal chain with abelian factors and so is soluble. Moreover, K_r contains a minimal splitting field E of f over k; now $\mathrm{Gal}(E/k)$ is a homomorphic image of $\mathrm{Gal}(K_r/k)$ and so is soluble. ■

It is easily verified that the symmetric groups of degree 3 and 4 are soluble, hence every equation of degree 3 or 4 is soluble. On the other hand, the symmetric group Sym_5 is insoluble, since Alt_5 is simple. More generally, Alt_n for $n \geq 5$ can be shown to be simple (see M. Hall (1959) p. 61). For the case $n = 5$ this may be seen directly as follows. We first list the 120 elements of Sym_5 by conjugacy classes:

C_1: 1;

C_2: $\binom{5}{2} = 10$ elements of type (1 2);

C_3: $2\binom{5}{2} = 20$ elements of type (1 2 3);

C_4: 30 elements of type (1 2 3 4);
C_5: 15 elements of type (1 2)(3 4);
C_6: 20 elements of type (1 2 3)(4 5);
C_7: 24 elements of type (1 2 3 4 5).

The alternating group contains C_1, C_3, C_5, C_7 and here C_1, C_3, C_5 remain conjugacy classes, while C_7 splits into two classes of 12 each. We thus obtain as class equation for Alt_5:

$$60 = 1 + 20 + 15 + 12 + 12. \tag{7.11.3}$$

For a normal subgroup we have to find a union of conjugacy classes including C_1, and thus some of the numbers on the right of (7.11.3), including 1, have to add up to a factor of 60; this is easily seen to be impossible except for the trivial cases 1, 60. This shows Alt_5 to be simple, and it shows incidentally that no Sym_n for $n \geq 5$ can be soluble, because Sym_n contains Alt_5 as a subgroup and every subgroup of a soluble group is again soluble.

As we saw in Section 7.6, the general quintic

$$x^5 - e_1x^4 + \ldots - e_5 = 0 \tag{7.11.4}$$

has the symmetric group over $k(e_1, \ldots, e_5)$, where the e_i are independent indeterminates; it is therefore insoluble. To be precise, we can reduce the group to Alt_5 by adjoining the square root of the discriminant, but no further reduction is possible because Alt_5 is simple.

It is also possible to construct equations with symmetric group over $\mathbf{Q}$. To do so we first show that when an equation over $\mathbf{Z}$ is reduced mod p, its group is replaced by a subgroup.

Theorem 7.11.2. *Let A be a unique factorization domain with field of fractions K, and let $\mathfrak{p}$ be a prime ideal of A, so that $\overline{A} = A/\mathfrak{p}$ is an integral domain, whose field of fractions is denoted by k. If f is a monic polynomial over A, $\overline{f}$ the corresponding polynomial over $\overline{A}$, where $\overline{f}$ is separable, then f is separable and if its group over K is G, then the group Γ of $\overline{f}$ over k is a subgroup of G, as a permutation group of the roots.*

Proof. It is clear that if f and its derivative f' have a common factor, then so do $\overline{f}$ and $\overline{f}'$, hence f must be separable. Let us denote its zeros in a splitting field E over K by $\alpha_1, \ldots, \alpha_n$ and with indeterminates $t_1, \ldots, t_n$ put $\lambda = \sum t_i\alpha_i$. For any permutation σ of $1, 2, \ldots, n$ we define the action of σ on λ by

$$\lambda^\sigma = \sum t_i\alpha_{i\sigma}. \tag{7.11.5}$$

Next form the polynomial

$$\varphi = \prod_\sigma (x - \lambda^\sigma),$$

where σ ranges over all $n!$ permutations of $1, 2, \ldots, n$. Clearly its coefficients are symmetric functions in the α_i and so belong to $\mathrm{K}(t_1, \ldots, t_n)$. In fact it is clear

from the construction that the coefficients are polynomials in the t_i with coefficients in A. We take a complete factorization of φ over $A[t_1, \ldots, t_n]$:

$$\varphi = \varphi_1 \varphi_2 \ldots \varphi_r. \tag{7.11.6}$$

By inertia (Theorem 7.7.2) the φ_i are irreducible over $K(t_1, \ldots, t_n)$. Since λ is a zero of φ, it must be a zero of some φ_i, say $\varphi_1(\lambda) = 0$. We claim that the group G of f over K is precisely the group of all permutations which map φ_1 into itself. For let $\sigma \in G$; then σ maps λ to λ^σ, satisfying the same irreducible equation. Hence φ_1 and φ_1^σ have a common factor and so coincide. Conversely, if $\varphi_1^\sigma = \varphi_1$, then λ^σ is again a zero of φ_1, hence $\sigma \in G$.

We now pass to the residue class ring $\overline{A}$ and obtain a factorization

$$\overline{\varphi} = \overline{\varphi}_1 \ldots \overline{\varphi}_r;$$

of course the factors $\overline{\varphi}_i$ on the right may well be reducible over k. The permutations of G are precisely the permutations transforming each $\overline{\varphi}_i$ into itself, while any other permutation maps φ_1 to φ_i $(i \neq 1)$. The permutations of Γ transform an irreducible factor of $\overline{\varphi}_1$, viz. that containing $x - \lambda$, into itself, and so must transform $\overline{\varphi}_1$ into itself. Hence Γ is a subgroup of G. ■

We shall apply this result with $A = \mathbf{Z}$. Given $f \in \mathbf{Z}[x]$ and a prime p, let us factorize $f \pmod p$. If $f \equiv f_1 \ldots f_r \pmod p$, where f_i is irreducible of degree d_i, then the group of f contains a permutation whose cycle structure is $d_1, d_2, \ldots, d_r$. For the group Γ of $f \pmod p$ is cyclic, as Galois group of an extension of finite fields. The orbits of Γ correspond to the irreducible factors of f, so the generating permutation of Γ consists of a d_1-cycle, a d_2-cycle, ..., a d_r-cycle, as claimed. It follows that if f has a factorization into irreducible factors of degrees $d_1, d_2, \ldots, d_r \pmod p$, then the group of f contains a permutation of type $d_1, d_2, \ldots, d_r$.

We shall also need a result on generating sets for Sym_n.

Lemma 7.11.3. *Any transitive subgroup of* Sym_n *containing a 2-cycle and an* $(n-1)$*-cycle is the whole group.*

Proof. Let H be the subgroup; by suitable numbering we can write the $(n-1)$-cycle as $\rho = (2\ 3 \ldots n)$. Since H is transitive, we can transform the 2-cycle to include 1, say $(1\ i)$. By conjugating with ρ we obtain $(1\ 2)$, $(1\ 3)$, ..., $(1\ n)$ and these transpositions generate the whole group. ■

We can now construct equations with symmetric group over $\mathbf{Q}$.

Theorem 7.11.4. *For every* $n \geq 1$ *there is an irreducible equation of degree* n *over* $\mathbf{Q}$ *whose group is the symmetric group of degree* n.

Proof. For any prime p and any $n \geq 1$ there are irreducible polynomials of degree n over $\mathbf{F}_p$, by Theorem 7.8.4. We choose three polynomials over $\mathbf{Z}$ as follows: f_1 is a product of an irreducible factor of degree $n-1$ by a linear factor (mod 2), and f_2 is a product of an irreducible quadratic factor and $n-2$ linear factors (mod 3).

Both f_1, f_2 may be taken to be monic; then $3f_1 - 2f_2$ is again monic and is congruent to $f_1 \pmod 2$ and to $f_2 \pmod 3$. The same is true of

$$f = 3f_1 - 2f_2 + 6f_3,$$

for any polynomial f_3 of degree $n-1$ over $\mathbf{Z}$. We now choose f_3 so that all coefficients of f after the first are divisible by 5, but the absolute term is not divisible by 25. Then f is irreducible over $\mathbf{Q}$, by Eisenstein's criterion (Theorem 7.2.7), hence its group is transitive. By Theorem 7.11.2 and the remark following it, applied with $p = 2$, the group contains an $(n-1)$cycle and a transposition, and therefore, by Lemma 7.11.3, it must be the whole of Sym_n. ■

By these methods it can also be shown that for any $n \geq 1$ there is an equation over $\mathbf{Q}$ whose group is the alternating group of degree n. But it is still an open question whether every finite group occurs as the Galois group of an extension of $\mathbf{Q}$.

Returning to the ruler-and-compass constructions (Section 7.1), we see that the elements of a Galois extension are constructible whenever the Galois group is a 2-group, for since any 2-group is soluble (Corollary 2.1.7), there is then a root tower in which all steps have degree 2. This has an application to the construction of regular n-gons.

We need to solve the cyclotomic equation $\Phi_n(x) = 0$, for if ζ is a root, then $\cos(2\pi/n) = (\zeta + \zeta^{-1})/2$. By Theorem 7.7.5, its group is abelian of order $\varphi(n)$. Writing $n = \prod p_i^{\nu_i}$, we have $\varphi(n) = \prod p_i^{\nu_i - 1}(p_i - 1)$, and this is a power of 2 precisely when each odd prime divisor occurs at most once and has the form $p = 2^r + 1$. If r has an odd factor, say $r = cd$, where d is odd, then $2^r + 1$ is divisible by $2^c + 1$, hence for a prime, r has to be a power of 2. The primes of the form $F_m = 2^{2^m} + 1$ are called *Fermat primes*; the first few are 3, 5, 17, 257, 65 537, but it is not known whether there are others (F_5 was proved composite by Euler in 1732). Our conclusions may be stated as follows:

Theorem 7.11.5. *A regular n-gon can be constructed by ruler and compasses precisely when each odd prime factor of n occurs only once and is a Fermat prime.* ■

This result was essentially known to Gauss, who gave an explicit construction of the regular heptadecagon (17-gon) in 1796, at the age of 19.

For irreducible equations of prime degree there is another criterion for solubility, also due to Galois. In these cases the group has a rather remarkable form, whose description depends on the following result on permutation groups:

Proposition 7.11.6. *Let G be a transitive permutation group of prime degree p. Then the following conditions are equivalent:*

(a) *G is soluble;*
(b) *G has a transitive normal subgroup T of order p which is its own centralizer;*
(c) *G can be written as a group of affine transformations over* $\mathbf{F}_p$*:*

$$x \equiv ax + b \pmod p, \quad \text{where } a \text{ is prime to } p, \tag{7.11.7}$$

which includes the subgroup T of all translations, $x \mapsto x + b$;
(d) *every element $\neq 1$ of G fixes at most one symbol.*

When this is so, G has a cyclic subgroup M such that $MT = G, M \cap T = 1$, and the order of M divides $p - 1$. Every non-trivial normal subgroup of G has the form NT, where N is a subgroup of M.

Proof. We remark that as a transitive group of prime degree p, G has order divisible by p. Thus it may be regarded as a subgroup of Sym_p which has order $p!$, so the highest power of p dividing $|G|$ is the first. Hence any Sylow p-subgroup of G has order p.

We begin by proving the equivalence of (a), (b) and (c) and then show that (c) $\Rightarrow$ (d) $\Rightarrow$ (b).

(a) $\Rightarrow$ (b). Let $N \lhd G$; since G is transitive, it permutes the orbits under the action of N transitively, so they must all have the same size. But the total number of symbols permuted is a prime, hence each orbit of N has either 1 symbol or p symbols, i.e. N is either the trivial group or it acts transitively. Now G is soluble, hence any minimal normal subgroup is elementary abelian (see Section 2.4, Exercise 11). Such a subgroup is transitive and of degree p, so it must be the Sylow p-subgroup T of G, which is therefore normal and contained in every non-trivial normal subgroup N of G. The subgroup T is of order p and hence cyclic. Let τ be a generator; as a permutation this is a p-cycle. If $\sigma \in G$ commutes with τ, then σ is a permutation of p symbols which commutes with the p-cycle τ. But the only such permutations are the powers of τ, therefore T is its own centralizer.

(b) $\Rightarrow$ (c). By hypothesis T is of order p, hence cyclic. Let τ be a generator of T; then for any $\sigma \in G$ there exists $a \in \mathbf{Z}$ such that

$$\sigma^{-1}\tau\sigma = \tau^a. \tag{7.11.8}$$

Moreover, a is determined uniquely (mod p). We denote the unique residue class determined by a (mod p) by a_σ; then

$$\sigma \mapsto a_\sigma \tag{7.11.9}$$

is a homomorphism from G to $\mathbf{F}_p^\times$. If $a_\sigma = 1$, then $\sigma^{-1}\tau\sigma = \tau$ and so $\sigma \in T$ by hypothesis; conversely, if $\sigma \in T$, then $a_\sigma = 1$, so the kernel of (7.11.9) is T. Let us number the symbols permuted as $0, 1, \ldots, p-1$ in such a way that τ^ν maps 0 to ν. Then $x^\tau = x + 1$ and $x^{\tau\sigma} = x^{\sigma\tau^a}$, i.e. $(x+1)^\sigma = x^\sigma + a$. Taking $x = 0$, we find $1^\sigma = 0^\sigma + a$, hence by induction on x,

$$x^\sigma = ax + b, \quad \text{where } b = 0^\sigma. \tag{7.11.10}$$

(c) $\Rightarrow$ (a) is clear; the translations form a cyclic normal subgroup T in G with quotient isomorphic to a subgroup of $\mathbf{F}_p^\times$.

(c) $\Rightarrow$ (d). The affine transformation (7.11.7), where $\sigma \neq 1$, leaves at most one symbol fixed, because the congruence $ax + b \equiv x$ (mod p) does not hold identically and so has at most one solution.

(d) $\Rightarrow$ (b). By the orbit formula (see Section 2.1), the 'average number' of symbols fixed by a permutation is the number of orbits, which is 1. Since the identity fixes p

symbols, and the other permutations fix at most one symbol, there are exactly $p-1$ permutations fixing no symbol. Each is necessarily a cycle of length p; if one of them is denoted by τ, the others are powers of τ and the subgroup T generated by τ has order p. Since there are no other p-cycles in G, T is normal in G, and if $\sigma \in G$ satisfies $\sigma\tau = \tau\sigma$, then either σ fixes no symbol and so lies in T, or σ fixes a; then σ also fixes $a^\tau \neq a$, and so $\sigma = 1$. Thus T is its own centralizer in G.

Finally let G be as in (c) and let M be the subgroup of multiplications in G. Then $M \cap T = 1$ and $MT = G$, as is easily seen, and M is isomorphic to a subgroup of $\mathbf{F}_p^\times$ and therefore has order dividing $p-1$. Every non-trivial normal subgroup of G contains T and so these normal subgroups correspond to the subgroups of $G/T \cong M$. ∎

We remark that having prime degree is a strong condition on G. The group G is the whole affine group $\mathbf{Aff}_1(\mathbf{F}_p)$ precisely when G is doubly transitive (i.e. transitive on ordered pairs of symbols). To apply the above result, suppose that we have an irreducible equation of prime degree p. Its group is transitive of degree p, by Theorem 7.10.1, and condition (d) means that any permutation fixing two roots fixes all, so by the fundamental theorem (Theorem 7.6.2), all roots can be expressed rationally in terms of any two, whenever the equation is soluble. Thus we obtain from Proposition 7.11.6 Galois' criterion for the solubility of equations of prime degree:

Theorem 7.11.7. *Let*

$$f = 0 \tag{7.11.11}$$

be an irreducible equation of prime degree p over a field k of characteristic 0, and let E be a minimal splitting field of k. Then (7.11.11) *is soluble by radicals if and only if E can be generated over k by any two roots of* (7.11.11). *When this is so, its group is a group of affine transformations* mod *p including all translations, there is a cyclic extension F of k over which E has degree p, and every proper normal subextension of E/k is contained in F.* ∎

To apply the above results we construct some equations with group Sym_5 over $\mathbf{Q}$.

Example 1. The equation

$$x^5 + 10x^3 - 15 = 0$$

is irreducible by Eisenstein's criterion; it is $x^5 - 1 \pmod 2$, which is $x-1$ times an irreducible factor of degree 4, and it is $x^5 + x^3 \pmod 3$, which is x^3 times an irreducible factor of degree 2. By Theorem 7.11.2 its group must be Sym_5.

Example 2. Consider the equation

$$f(x) = x^5 - 5x + 1 = 0. \tag{7.11.12}$$

This is again irreducible by Eisenstein's criterion, as we see by putting $x = y - 1$. By

Descartes' rule (which tells us that an equation cannot have more positive roots than the number of sign changes in the sequence of coefficients (ignoring 0's)), (7.11.12) has at most three real roots, and in fact $f(-2)$, $f(-1)$, $f(1)$, $f(2)$ have signs $-, +, -, +$ respectively, hence there are exactly three real and two complex roots. If (7.11.12) were soluble, all its roots could be expressed rationally in terms of any two, by Theorem 7.11.7; choosing two real roots, we obtain a contradiction. Thus the group is Alt_5 or Sym_5 and the former can be ruled out because complex conjugation interchanges the complex roots and so is an odd permutation.

Exercises

1. Show that a real algebraic number admits quadrature iff the Galois group of its minimal equation over $\mathbf{Q}$ is a 2-group.
2. Carry out the construction of a regular n-gon for $n = 3, 4, 5, 6$.
3. Show that the binomial equation $x^n = a$ is soluble over $\mathbf{Q}$ but not necessarily abelian. Describe the group when n is prime, using Theorem 7.11.7.
4. Show that the equation $x^p - x - t = 0$ is not soluble by radicals over $\mathbf{F}_p(t)$, though its group is cyclic.
5. Show that if a separable equation over a field of prime characteristic p is soluble by radicals, then its group G is soluble (of order at most $p(p-1)$) and the converse holds when all prime factors of $|G|$ are less than p.
6. Show that a transitive permutation group G of prime degree p contains a cycle of length p. If p is odd and G also contains a transposition, then it must be the full symmetric group. (Hint. G contains an element of order p.)
7. Show that $x^4 - 5 = 0$ has the dihedral group of order 8 over $\mathbf{Q}$.
8. Let $f = 0$ be a cubic with roots α, β, γ over a field k containing a primitive cube root ω of 1. Verify that $\theta = (\alpha + \beta\omega + \gamma\omega^2)^3$ is of degree 1 or 2 over k, and show that $k \subseteq k(\theta) \subseteq k(\theta^{1/3}) = k(\alpha, \beta)$ is a root tower for the equation. What is the analogue for quartics? Show that for a quartic, $(\alpha - \beta + \gamma - \delta)^2$ has degree 1 or 3.
9. Let $f = 0$ be a quartic over a field k with a primitive sixth root of 1, and denote its roots in some splitting field by $\alpha_1, \ldots, \alpha_4$. Write $\gamma = \alpha_1 + \alpha_2$, $\beta = \alpha_1\alpha_2 + \alpha_3\alpha_4$ and let δ be the square root of the discriminant. Show that $k \subseteq k(\delta) \subseteq k(\delta, \beta) \subseteq k(\delta, \beta, \gamma) \subseteq k(\alpha_1, \alpha_2, \alpha_3)$ is a root tower.
10. Let $f = 0$ be a quintic with roots $\alpha_1, \ldots, \alpha_5$ and denote by ω a primitive fifth root of 1. Show that in general $\theta = (\alpha_1 + \alpha_2\omega + \ldots + \alpha_5\omega^4)^5$ has degree dividing 24, but when f is irreducible and soluble, then θ has degree dividing 4.
11. (L. Kronecker) Show that if f is an irreducible polynomial over $\mathbf{Q}$ of odd prime degree and soluble by radicals over $\mathbf{Q}$, then $f = 0$ either has just one real root or all its roots are real.
12. (E. Galois) Using Proposition 7.11.6 show that a soluble permutation group of degree 5 has order at most 20. Deduce that the general quintic is insoluble.

Further Exercises for Chapter 7

1. Let F_1 and F_2 be subfields of a field E. Show that if $F_1 \subseteq F_2$ and $[E : F_1] < \infty$, then $[E : F_1] \geq [E : F_2]$, with equality iff $F_1 = F_2$.
2. Show that if $k \subseteq F_i \subseteq E$ and F_i is finite over k $(i = 1, 2)$, then the subfield L generated by F_1 and F_2 is finite over k.
3. Show that $\mathbf{Q}(\sqrt{2}, \sqrt{-1})$ is a normal extension of $\mathbf{Q}$ and that $x^4 - 2x^2 + 9 = 0$ is a normal equation for it (i.e. the minimal equation of a primitive element; this is also called the *resolvent*).
4. Prove that $k[x]$ is a unique factorization domain by using splitting fields.
5. Let F/k be a finite extension and F be perfect. Show that k is perfect.
6. Let k be a field of prime characteristic p and $\alpha : x \mapsto x^p$ the Frobenius mapping. Show that there is a strictly descending tower of subfields $k \supseteq \operatorname{im} \alpha \supseteq \operatorname{im} \alpha^2 \supseteq \ldots$, whose intersection k_0 is a perfect subfield of k, and show that every perfect subfield of k is contained in k_0.
7. (König–Rados) Consider the equation

$$a_0 + a_1 x + \ldots + a_{q-2} x^{q-2} = 0$$

over $\mathbf{F}_q$, where q is a prime power. By examining the product of the Vandermonde matrix with rows $1, b, \ldots, b^{q-2}$ $(b \in \mathbf{F}_q^\times)$ and the circulant matrix whose rows are cyclic permutations of the coefficients $a_0, a_1, \ldots, a_{q-2}$, show that the number of non-zero roots of the above equation in $\mathbf{F}_q$ is $q - 1 - r$, where r is the rank of the circulant matrix.
8. Show that in a Galois extension of odd order the discriminant is a square. Conversely, show that in a cyclic Galois extension of even order the discriminant is not a square.
9. Let k be a finite field and $n \geq 1$. Show that there is an irreducible polynomial of degree n over k in which the coefficient of x is non-zero. (Hint. Consider the equation for x^{-1}.)
10. Let F/k be a finite extension and $|k| = q$. Show that $T(\alpha^q) = T(\alpha)$ for all $\alpha \in F$. Deduce that $T(\alpha) = 0$ iff the equation $x^q - x - \alpha = 0$ has a root in F.
11. (Kronecker) Let $f = 0$ be an irreducible equation over $\mathbf{Q}$ of odd prime degree p. Show that if f is soluble by radicals, then the number of its real roots is either 1 or p. Moreover, when $p \equiv 3 \pmod 4$, then the number of real roots is p or 1 according as the discriminant is positive or negative.
12. Let F/k be a Galois extension. If $1 \neq \sigma \in \operatorname{Gal}(F/k)$, find $a_i, b_i \in F$ $(i = 1, \ldots, n)$ such that $\sum a_i(b_i - b_i^\sigma) = 1$. Show that for $b_0 = 1$ and suitably chosen a_0,

$$\sum_0^n a_i b_i = 1, \sum_0^n a_i b_i^\sigma = 0.$$

By doing this for each $\sigma \neq 1$ and multiplying the results together, obtain $u_j, v_j \in F$ $(j = 1, \ldots, r)$ such that $\sum u_j v_j^\sigma = \delta_{j0}$. (This property can be used to characterize a notion of Galois extension of commutative rings, see Chase et al. (1965).)

13. Let p be a prime and n be an integer prime to p. Show that $\Phi_n(x^p) = \Phi_{pn}(x)\Phi_n(x)$. What happens if $p|n$?
14. (Vahlen–Capelli) Let F/k be a separable extension, say $F = k(\alpha)$, where α has the minimal polynomial f over k, and for any polynomial g over k take a complete factorization of $g(x) - \alpha$ over F: $g(x) - \alpha = \prod g_i(x)$. Using the formula $f(x) = N(x - \alpha)$, prove that

$$f(g(x)) = \prod N(g_i(x)),$$

where the norm is relative to the extension $F(x)/k(x)$. If $p(x)$ is an irreducible factor of $f(g(x))$ over k, show that $g_i(x)|p(x)$ for some i, and hence by passing to a splitting field of f over k, deduce that $N(g_i(x))|p(x)$. Use this fact to prove that the above equation is a complete factorization of $f(g(x))$ over k; in particular, show that $f(g(x))$ is irreducible over k iff $g(x) - \alpha$ is irreducible over F.
15. (Vahlen–Capelli) Show that $x^n = a$ (where $n > 1$, $a \neq 0$) is reducible over a field k iff either (i) $a = b^d$ for some $b \in k$ and $1 < d|n$, or (ii) $4|n$ and $a = -4c^4$. (For case (ii) use the identity $x^{4m} + 4c^4 = (x^{2m} - 2cx^m + 2c^2)(x^{2m} + 2cx^m + 2c^2)$. In the other direction use induction on n and Exercise 14; note also the identity $u^4 + v^4 + (u - v)^4 = 2[(u^3 + v^3)/(u + v)]^2$.)
16. (E. Artin) Let k be a field of characteristic 0 and K/k be a finite extension. Show that if K is algebraically closed then $[K : k] \leq 2$ and $K = k(\sqrt{-1})$. (The fact that char $k = 0$ can actually be proved from the other assumptions. Hint. Use Exercise 15.)
17. An equation is said to be *reciprocal* if with α, $1/\alpha$ is also a root. Show that if an irreducible equation over $\mathbf{Q}$ has a complex root of absolute value 1, then the equation is reciprocal of even degree.
18. Show that the group of a reciprocal equation of degree $2m$ or $2m + 1$ is of order at most $2^m . m!$. Determine the possible groups for a reciprocal quartic.
19. Let $f = 0$ be an irreducible cubic over $\mathbf{Q}$ with three real roots. Show that there is no root tower for $f = 0$ consisting entirely of real fields. (This is the 'casus irreducibilis', which expresses the fact that in this case none of the roots can be expressed in terms of real radicals alone, in contrast to the case of a cubic with only one real root.)
20. Let E/k be separable of degree r. If $E = k(\alpha)$ and $\beta \in E$, find a polynomial f over k of degree less than r such that $\beta = f(\alpha)$ (Hint. Use Lagrange interpolation, given that $f(\alpha^\sigma) = \beta^\sigma$.)
21. Let p be a prime and a be an integer. The *Legendre symbol* $\left(\frac{a}{p}\right)$ is defined as 0 if $p|a$, 1 if $x^2 \equiv a \pmod{p}$ has a solution and -1 otherwise; in the last two cases a is a *quadratic residue,* resp. *non-residue* mod p. Taking a to be an integer mod p, verify that the map of $\mathbf{F}_p^\times$ into $\mathbf{C}_2$ given by $a \mapsto \left(\frac{a}{p}\right)$ is a homomorphism. Let p be an odd prime and z be a generator of the multiplicative group $\mathbf{F}_p^\times$. Show that $z^{(p-1)/2} = -1$ and hence deduce *Euler's criterion*: for any a prime to p, $\left(\frac{a}{p}\right) \equiv a^{(p-1)/2} \pmod{p}$.

22. Let p, q be distinct odd primes and denote by w a primitive p-th root of 1 in an extension of $\mathbf{F}_q$. For any $a \in \mathbf{F}_p^\times$ define the *Gaussian sum* (in an extension of $\mathbf{F}_q$) as

$$\tau(a) = \sum_{x \in \mathbf{F}_p^\times} \left(\frac{x}{p}\right) w^{ax}.$$

Prove (i) $\tau(a) = \left(\frac{a}{p}\right)\tau(1)$, (ii) $\tau(1)^q = \tau(q)$, (iii) $\tau(1)^2 = (-1)^{(p-1)/2}p$ (Hint. Evaluate $\sum w^{ax}$ and use Euler's criterion.)

23. For any odd n, put $\varepsilon(n) \equiv (n-1)/2 \pmod 4$. Use Exercise 22 to show that $\left(\frac{q}{p}\right) = \tau(1)^{q-1}$. By evaluating $\tau(1)^{q-1} = [\tau(1)^2]^{(q-1)/2}$ in two ways, prove the *law of quadratic reciprocity*:

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

24. For any odd n, put $\omega(n) \equiv (n^2-1)/2 \pmod 8$. Let α be a primitive 8th root of 1 in an extension of $\mathbf{F}_p$ and put $\beta = \alpha + \alpha^{-1}$. Show that $\beta^2 = 2$, and using the Frobenius endomorphism $x \mapsto x^p$ and Euler's criterion to evaluate β^{p-1} in two ways, prove that $\left(\frac{2}{p}\right) = (-1)^{\omega(p)}$. (This formula, together with the law of quadratic reciprocity and Euler's criterion, enables one to evaluate Legendre symbols, e.g.

$$\left(\frac{29}{43}\right) = \left(\frac{43}{29}\right) = \left(\frac{14}{29}\right) = \left(\frac{2}{29}\right)\left(\frac{7}{29}\right) = -\left(\frac{7}{29}\right) = -\left(\frac{1}{7}\right) = -1,$$

thus $x^2 \equiv 29 \pmod{43}$ has no solution.)

25. Show that for $p \neq 13, 17$, one of 13, 17, 221 must be a quadratic residue mod p. Use the result to show that $x^4 - 15x^2 + 1$ is congruent mod p to a product of two quadratic polynomials with integer coefficients, for any prime p. (Hint. First factorize the quartic over $\mathbf{C}$.)

26. Prove the Euler summation formula relating the Euler and the Möbius function for $\alpha \leq 1$, $x \geq 2$:

$$\sum_{n \leq x} \frac{\varphi(n)}{n^\alpha} = \sum_{d \leq x} \frac{\mu(d)}{d^\alpha} \sum_{q \leq x/d} q^{1-\alpha} \approx \frac{x^{2-\alpha}}{2-\alpha} \sum_{d \leq x} \frac{\mu(d)}{d^2} + O\left(x^{1-\alpha} \sum_{d \leq x} \frac{\mu(d)}{d}\right)$$

$$= \frac{x^{2-\alpha}}{2-\alpha} \frac{1}{\zeta(2)} + O(x^{1-\alpha} \log x).$$

8 Quadratic Forms and Ordered Fields

Most readers will have met inner products before. Here we take up the subject in a more general form and look at the properties of quadratic forms over a general field and its group of isometries (Sections 8.1–8.3). With each quadratic form a certain algebra is associated, the Clifford algebra, and with the set of all forms on a field the Witt group is associated; these form the subject of Section 8.4 and Section 8.5 respectively, with a further development, the Witt ring of a field, in Section 8.9. In Section 8.10 we take a brief look at symplectic groups and in Section 8.11 we consider quadratic forms in characteristic 2. We also briefly discuss the related topic of ordered fields in Section 8.6, leading to a construction of the real numbers in Section 8.7 and formally real fields in Section 8.8. Some of the later topics are included for completeness, but do not really have a place in a basic account; thus at a first reading the later parts of Sections 8.7–8.11 can be omitted.

8.1 Inner Product Spaces

Let k be a field. By an *inner product space* we understand a pair (V, b) consisting of a finite-dimensional vector space V over k and a symmetric bilinear form b on V, i.e. a mapping

$$b : V \times V \to k,$$

which is *symmetric*: $b(x, y) = b(y, x)$, and *bilinear*:

$$b(\alpha x + \beta y, z) = \alpha b(x, z) + \beta b(y, z), \quad b(x, \alpha y + \beta z) = \alpha b(x, y) + \beta b(x, z).$$

There is a second way of defining inner product spaces, in terms of quadratic forms. By a *quadratic form* on a given vector space V over k we understand a mapping $q : V \to k$ such that

Q.1 $q(\alpha x) = \alpha^2 q(x)$ for all $x \in V$ and $\alpha \in k$,
Q.2 $q(x + y) - q(x) - q(y) = b_q(x, y)$ is a bilinear form on V.

The pair (V, q) is also called a *quadratic space*. By the *dimension* of q is meant the dimension of the underlying space; we speak of a *binary*, *ternary*, or *quaternary form* if the dimension is 2, 3 or 4.

Clearly the function b_q in **Q.2** is symmetric; it is called the *associated bilinear form*, and it can be used to define the inner product structure on V. If char $k \neq 2$, then every symmetric bilinear form b is associated with a quadratic form, namely $b = b_q$, where $q = \frac{1}{2}b(x, x)$. This follows because any bilinear form b on V satisfies

$$b(x+y, x+y) - b(x, x) - b(y, y) = b(x, y) + b(y, x).$$

When b is symmetric, this reduces to $2b(x, y)$, so in that case $b_q(x, y) = q(x+y) - q(x) - q(y) = b(x, y)$. So in characteristic $\neq 2$ there is complete equivalence between quadratic forms and bilinear forms, and we shall use both interchangeably. To be precise, we shall use $b(x, y)$ and $q(x) = b(x, x)$, so that

$$b(x, y) = \tfrac{1}{2}[q(x+y) - q(x) - q(y)]. \tag{8.1.1}$$

In characteristic 2 we can still associate with each quadratic form a bilinear form (as in **Q.2**) and each quadratic form can be expressed as $b(x, x)$ for some bilinear form b (see Exercise 3), but now b cannot in general be chosen to be symmetric. In what follows we shall concentrate on the case char $k \neq 2$ and briefly look at the case char $k = 2$ in Section 8.11.

Let V be an inner product space. Relative to the basis $e_1, \ldots, e_n$ of V, its form b is determined by the coefficients $a_{ij} = b(e_i, e_j)$, for if $x = \sum \xi_i e_i$, $y = \sum \eta_i e_i$, then by linearity we have $b(x, y) = \sum a_{ij}\xi_i\eta_j$. In matrix form this can be written as

$$b(x, y) = \xi A \eta^T, \tag{8.1.2}$$

where $\xi = (\xi_1, \ldots, \xi_n)$, $\eta = (\eta_1, \ldots, \eta_n)$ and T denotes transposition. Since b is symmetric, A is a symmetric matrix, i.e. $A^T = A$. Conversely, any symmetric matrix A defines a symmetric bilinear form on V by the formula (8.1.2).

If the basis of V is changed, let the new coordinates of x, y be the rows ξ', η', where $\xi = \xi' P$, $\eta = \eta' P$ for some invertible matrix P (describing the change of basis in V). Then $b(x, y) = \xi' A' \eta'^T = \xi' P A P^T \eta'^T$. Since this holds for all ξ', η', we conclude that

$$A' = PAP^T, \quad \text{where } P \text{ is invertible.} \tag{8.1.3}$$

Two matrices A, A' related as in (8.1.3) are said to be *congruent*; what we have said shows that matrices of the form b in different coordinate systems are congruent. Conversely, if b is represented by the matrix A in one coordinate system and A' is congruent to A, say (8.1.3) holds, then transformation by P will define a new basis, relative to which b has the matrix A'. This proves

Proposition 8.1.1. *Two matrices represent the same bilinear form in different coordinate systems if and only if they are congruent.* ■

The result holds for any bilinear forms, symmetric or not; for symmetric forms in characteristic not 2, the matrices can also be chosen to be symmetric. Moreover, given two matrices A, A', related as in (8.1.3), if one of them is symmetric, then clearly so is the other.

An *isometry* between two inner product spaces V, V' (or also between their forms) is an isomorphism $f : V \to V'$ which transforms the form of V into that of V'; thus if the forms are b, b' respectively, then

$$b'(xf, yf) = b(x, y) \quad \text{for all } x, y \in V.$$

Equivalently (in characteristic not 2), if the quadratic forms in V, V' are q, q' respectively, then $q'(xf) = q(x)$ for all $x \in V$. We shall then say that the spaces V, V' are *isometric*, and write $V \cong V'$. Relative to suitable bases the bilinear forms in spaces that are isometric are represented by the same matrix; hence two spaces with given bases are isometric iff the matrices of the forms relative to these bases are congruent.

Let (V, b) be an inner product space. Any subspace U of V is again an inner product space, the form being $b|U$, the restriction of the form b to U.

Two vectors $x, y \in V$ are said to be *orthogonal* if $b(x, y) = 0$; by the symmetry of b this is a symmetric relation. For any subset S of V we define its *orthogonal space* as

$$S^{\perp} = \{x \in V | b(x, y) = 0 \quad \text{for all } y \in S\}. \tag{8.1.4}$$

It is easily seen that $S^{\perp}$ is a subspace of V; this holds for any subset S of V, although we shall mainly use (8.1.4) when S is itself a subspace. In particular, the subspace $V^{\perp}$ is called the *radical* of V; it consists of all vectors orthogonal to all of V. If $V^{\perp} \neq 0$, then the form b (or also the space V) is called *singular*, otherwise V is *non-singular* or *regular*. An inner product space V over the real numbers is called *positive-definite* if its quadratic form q satisfies $q(x) > 0$ for all $x \neq 0$; such a space is just the familiar Euclidean space.

Let V be an inner product space whose form relative to a basis is given by (8.1.2). The radical of V is obtained by solving the equations $b(x, e_i) = 0$, i.e. $xA = 0$. These equations have a non-trivial solution precisely when A is singular, so we obtain

Proposition 8.1.2. *An inner product space is singular if and only if the matrix of its form (relative to any basis of the space) is singular.* ■

Let V be a regular space and U be any subspace. If $v_1, \ldots, v_n$ is a basis of V, chosen so that $v_1, \ldots, v_r$ is a basis of U, then a vector $x = \sum a_i v_i$ is orthogonal to U iff

$$0 = b(x, v_j) = \sum a_i (v_i, v_j), \quad j = 1, \ldots, r. \tag{8.1.5}$$

By hypothesis the $n \times n$ matrix $(b(v_i, v_j))$ is regular, hence the $n \times r$ matrix consisting of the first r columns has rank r. Therefore the system (8.1.5) has rank r and the space of solutions has dimension $n - r$. This proves the first part of

Proposition 8.1.3. *If V is a regular inner product space and U is any subspace of V, then*

$$\dim U + \dim U^{\perp} = \dim V \tag{8.1.6}$$

and

$$U^{\perp\perp} = U. \tag{8.1.7}$$

Proof. It only remains to prove (8.1.7). By (8.1.6) we have $\dim U + \dim U^\perp = \dim U^\perp + \dim U^{\perp\perp}$, hence $\dim U^{\perp\perp} = \dim U$. Since clearly $U \subseteq U^{\perp\perp}$, it follows that $U^{\perp\perp} = U$. ∎

We note that $U \cap U^\perp$ need not be 0; whether it is will depend on whether U is regular in the induced metric. We shall deal with this case in the next section.

For a regular space the determinant of the matrix, though not itself invariant, provides an invariant of the space. By (8.1.3) we have the relation

$$\det A' = \det A . (\det P)^2. \tag{8.1.8}$$

Here $(\det P)^2$ can assume the value of any non-zero square in the field k. If we denote this set of squares by $k^{\times 2}$, then by (8.1.8) each regular inner product space V determines a unique element of the factor group $k^\times / k^{\times 2}$, namely the residue class of $\det A$. This is called the *determinant* of the space V, or also of the form b.

Finally we remark that by fixing one of the arguments in an inner product, we obtain a linear functional on the space, i.e. an element of $V^* = \mathrm{Hom}_k(V, k)$. The inner product b of V thus defines a mapping

$$\varphi_b : V \to V^*, \quad \text{given by } x \mapsto b(x, -).$$

This mapping is linear, as is easily seen. It is injective iff b is regular; since V^* and V have the same dimension, φ_b is then an isomorphism.

Exercises

1. Let V be an inner product space. Show that for each $u \in V$ the mapping $\lambda_u : x \mapsto b(x, u)$ is an element of V^* and that the mapping $\varphi : u \mapsto \lambda_u$ is an isomorphism iff V is regular. Examine the case where the coefficient ring is not a field.
2. Let V be an inner product space and U_1, U_2 be subspaces. Show that (i) $U_1 \subseteq U_2 \Rightarrow U_1^\perp \supseteq U_2^\perp$, (ii) $U_1 \subseteq U_1^{\perp\perp}$, (iii) $U_1^\perp = U_1^{\perp\perp\perp}$.
3. Given a quadratic form q on a space V in characteristic 2, find a bilinear form b on V such that $q(x) = b(x, x)$. Find the conditions on q for which b can be chosen symmetric as well as bilinear. (Hint. See Section 8.11.)
4. Show that any quadratic form q satisfies the parallelogram law: $q(x+y) + q(x-y) = 2[q(x) + q(y)]$.
5. Let V be a regular inner product space and U_1, U_2 be subspaces. Show that $(U_1 + U_2)^\perp = U_1^\perp \cap U_2^\perp$, $(U_1 \cap U_2)^\perp = U_1^\perp + U_2^\perp$.

8.2 Orthogonal Sums and Diagonalization

Let V, V' be inner product spaces with quadratic forms q, q' respectively; their *orthogonal sum* is defined as the direct sum of the spaces V and V' with the quadratic form Q defined by

$$Q(x + x') = q(x) + q'(x'), \quad x \in V, x' \in V' \tag{8.2.1}$$

It is easily checked that Q is a quadratic form on $V \oplus V'$; the inner product space so defined is written $V \perp V'$. Clearly V' is orthogonal to V; moreover, if q, q' have the matrices A, A', then Q has the matrix

$$\begin{pmatrix} A & 0 \\ 0 & A' \end{pmatrix}$$

It follows that $\det Q = \det q . \det q'$.

The next result is useful in decomposing a space into an orthogonal sum. Note that the space itself is not required to be regular, only the subspace.

Lemma 8.2.1. *Let V be an inner product space and U be a subspace which is regular (under the inner product induced from V). Then*

$$V = U \perp U^{\perp}. \tag{8.2.2}$$

Proof. Let $e_1, \ldots, e_r$ be a basis of U; given $x \in V$, we write

$$x = \sum \xi_i e_i + y, \tag{8.2.3}$$

and try to solve the equations

$$\sum \xi_i b(e_i, e_j) = b(x, e_j), \tag{8.2.4}$$

which express that $b(y, e_j) = 0$ for all j, and so $y \in U^{\perp}$. Since U is regular, the matrix $(b(e_i, e_j))$ is invertible, hence the system (8.2.4) has a unique solution $\xi_1, \ldots, \xi_r$. Since x in (8.2.3) was arbitrary in V, this proves that $V = U + U^{\perp}$; the sum is direct because U is regular and so $U \cap U^{\perp} = 0$, and it is clearly orthogonal. ■

When we have a decomposition of V of the form (8.2.2), $U^{\perp}$ is called the *orthogonal complement* of U. Such a decomposition always exists in characteristic not 2:

Theorem 8.2.2. *Every inner product space over a field of characteristic not 2 is an orthogonal sum of one-dimensional spaces.*

Proof. Let V be the space, with quadratic form q. If $q = 0$, the result is clear; otherwise take $e_1 \in V$ such that $q(e_1) \neq 0$ and let U be the subspace spanned by e_1. Then U is regular and by Lemma 8.2.1, $V = U \perp U^{\perp}$. Now the result follows by induction on $\dim V$, because $\dim U^{\perp} < \dim V$. ■

In terms of matrices this means that every symmetric matrix in characteristic not 2 is congruent to a diagonal matrix. For example, for the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ we have

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix}.$$

We see that char $k \neq 2$ is essential for this example.

If we write V as an orthogonal sum of 1-dimensional spaces and take an adapted basis $e_1, \ldots, e_n$ this is called an *orthogonal basis* of V; relative to this basis the quadratic form on V has a diagonal matrix $\mathrm{diag}(a_1, \ldots, a_n)$. Let us write

$$\langle a_1, \ldots, a_n \rangle \tag{8.2.5}$$

for the inner product space with a matrix of this form; by Theorem 8.2.2 every inner product space (in char $\neq 2$) can be expressed in the form (8.2.5). In particular, $\langle a \rangle$ denotes a 1-dimensional space with the form ax^2; the n-dimensional space with the form $\sum ax_i^2$ is $\langle a, a, \ldots, a \rangle$. This will occasionally be written as $\langle a^n \rangle$; that notation will mainly be used for $a = 0$ or 1, so the risk of ambiguity is slight. We observe that for any $c_1, \ldots, c_n \in k^\times$, $\langle a_1, \ldots, a_n \rangle \cong \langle c_1^2 a_1, \ldots, c_n^2 a_n \rangle$.

From the proof of Theorem 8.2.2 we see that if $q(e_1) = a \neq 0$, then $V = \langle a \rangle \perp V'$, for some space V'. Let us say that the quadratic form q represents $a \in k$ if $q(x) = a$ for some $x \neq 0$. Then we can express the conclusion as follows:

Corollary 8.2.3. *Let V be an inner product space (in characteristic not 2) with quadratic form q. If q represents $a \neq 0$, then $V = \langle a \rangle \perp V'$, for some subspace V' of V.* ■

It is often easier to work with regular spaces; this can always be achieved by the following reduction.

Proposition 8.2.4. *Let V be an inner product space and let V_0 be a complement of $V^\perp$ in V (as vector space). Then*

$$V = V_0 \perp V^\perp. \tag{8.2.6}$$

The subspace V_0 is regular and is determined up to isometry by V.

Proof. Clearly V_0 is orthogonal to $V^\perp$, hence (8.2.6) follows. Moreover, $V_0^\perp = V^\perp$ and this meets V_0 in 0, so V_0 is regular. Since $b(x, y)$ vanishes for x or y in $V^\perp$, we can define the quotient $V/V^\perp$ with the natural homomorphism $x \mapsto \bar{x}$ from V to $V/V^\perp$ as an inner product space by putting $b(\bar{x}, \bar{y}) = b(x, y)$. With this definition it is clear that $V/V^\perp \cong V_0$, so the latter is unique up to isometry. ■

The space V_0 in Proposition 8.2.4 is called the *regular part* of V; its dimension is the *rank* of V or also of the quadratic form q, written $\mathrm{rk}\, q$. In view of this result we can mainly restrict ourselves to regular forms. We note the following criterion for isometry:

Corollary 8.2.5. *Let k be a field of characteristic not 2, in which every element is a square. Then two inner product spaces over k are isometric if and only if they have the same dimension and the same rank.*

Proof. The condition is clearly necessary; since $\langle a \rangle \cong \langle 1 \rangle$ for any $a \in k^\times$, any space of dimension n and rank r is isometric to $\langle 1^r, 0^{n-r} \rangle$, hence the condition is also sufficient. ■

In particular this solves the classification problem for any algebraically closed field of characteristic not 2.

The last result can be generalized as follows. A quadratic form is said to be *universal* if it represents every non-zero element of the field.

Lemma 8.2.6. *Let k be a field of characteristic not 2. If every quadratic form of rank v is universal (for some $v \geq 1$), then every inner product space of dimension n is isometric to $\langle \alpha_1, \ldots, \alpha_r, 1^s, 0^{n-r-s} \rangle$, where $\alpha_1, \ldots, \alpha_r$ are non-squares in k and $r < v$.*

Proof. Clearly $\langle c \rangle \cong \langle 1 \rangle$ for any non-zero square c in k, therefore any inner product space V of dimension n is isometric to $\langle \alpha_1, \ldots, \alpha_r, 1^s, 0^{n-r-s} \rangle$, where the α_i are non-squares. To show that $r < v$ we shall use induction on r. Suppose that $r \geq v$; then $\langle \alpha_1, \ldots, \alpha_r \rangle$ represents 1, by hypothesis, so by Corollary 8.2.3, we have $V \cong \langle \alpha'_2 \ldots, \alpha'_r, 1^{s+1}, 0^{n-r-s} \rangle$ and we can now apply induction on r to complete the proof. ∎

As an application let us take a finite field k of odd characteristic. In this case the hypothesis of Lemma 8.2.6 holds for $v = 2$. To establish this fact we must show that any binary quadratic form over k is universal; thus we have to solve the equation

$$ax^2 + by^2 = c \tag{8.2.7}$$

for any $a, b, c \in k^\times$. Let A be the set of all elements $ax^2 (x \in k)$ and B be the set of all elements $c - by^2 (y \in k)$. If the field k has q elements, it contains $(q+1)/2$ squares, because the group endomorphism $x \mapsto x^2$ of $k^\times$ has a kernel of order 2, so there are $(q-1)/2$ non-zero squares, which together with 0 give $(q+1)/2$ squares. Hence A, B have $(q+1)/2$ elements each and so $A \cap B \neq \varnothing$. This provides a solution for (8.2.7).

Applying Lemma 8.2.6, we see that every quadratic form of rank r over a finite field of characteristic not 2 has the form $\langle \alpha, 1^{r-1}, 0^s \rangle$, where $\alpha \in k^\times$. For a regular form this becomes $\langle \alpha, 1^{n-1} \rangle$. Now either α is a square, then we just have $\langle 1^n \rangle$; or α is not a square, then the previous form cannot be simplified. We note that the determinant of $\langle \alpha, 1^{n-1} \rangle$ is α, so we can decide to which case our form belongs by looking at the determinant. We also note that the ratio of any two non-zero non-squares is a square, because $(k^\times : k^{\times 2}) = 2$. The result may be summed up as

Theorem 8.2.7. *Let k be a finite field of odd prime characteristic. Then every quadratic form of rank at least 2 is universal. Further, two regular forms over k are isometric if and only if they have the same rank and determinant. More precisely, if λ is any non-square in k, then any regular form of rank n is isometric to $\langle 1^n \rangle$ or $\langle \lambda, 1^{n-1} \rangle$ according as the determinant is or is not a square.* ∎

Exercises

1. Verify that the function Q defined in (8.2.1) is a quadratic form on $V \oplus V'$.
2. Show that (in char $\neq 2$) a regular quadratic form which represents 0 is universal. (Hint. If $\sum a_i \alpha_i^2 = 0$, put $x_1 = \alpha_1(1+t)$, $x_i = \alpha_i(1-t)$, $i = 2, \ldots, n$.)

3. Show that a regular quadratic form $\langle a_1, \ldots, a_n \rangle$ represents $a \neq 0$ iff $\langle a_1, \ldots, a_n, -a \rangle$ represents 0.
4. Show that two binary quadratic forms in characteristic $\neq 2$ are isometric iff they have the same determinant and there exists $a \in k^\times$ which is represented by both.
5. Let V be a 3-dimensional inner product space of characteristic 2 whose bilinear form has the matrix

$$\begin{pmatrix} a & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

where $a \neq 0$. Find an orthogonal basis. (Hint. Take a basis including $(1,1,1)^T$.)
6. A bilinear form b on an inner product space V is said to be *alternating* if $b(x, x) = 0$ for all $x \in V$. Use the proof of Theorem 8.2.2 and Exercise 5 to show that in characteristic 2 every inner product space whose form is not alternating has an orthogonal basis.

8.3 The Orthogonal Group of a Space

Let V be an inner product space; the isometries of V with itself are called *orthogonal transformations*. Clearly they form a group, called the *orthogonal group* and denoted by $\mathbf{O}(V)$ or $\mathbf{O}(V, q)$ if the quadratic form q is to be emphasized. The elements of $\mathbf{O}(V)$ are the vector space automorphisms θ of V such that $q(x\theta) = q(x)$ for all $x \in V$. In terms of matrices, if q has the matrix A (relative to some fixed basis of V), then $\mathbf{O}(V)$ consists of all invertible matrices X such that

$$XAX^T = A. \tag{8.3.1}$$

When V is regular, we see from (8.3.1) by taking determinants that $(\det X)^2 = 1$, hence $\det X = \pm 1$. The *special orthogonal group* $\mathbf{SO}(V)$ is defined as the subgroup of $\mathbf{O}(V)$ consisting of all isometries of determinant 1. Its elements are the *proper* orthogonal transformations, also called *rotations*; the elements of $\mathbf{O}(V)$ not in $\mathbf{SO}(V)$ are called *improper*. By taking A diagonal in (8.3.1) we see that in char $\neq 2$ there are always improper orthogonal transformations, so that $\mathbf{SO}(V)$ is then a subgroup of index 2 in $\mathbf{O}(V)$.

In the special case of a space with the standard quadratic form $\sum x_i^2$ the condition (8.3.1) for a matrix X to be orthogonal reduces to

$$XX^T = I.$$

The set of all $n \times n$ matrices over k satisfying this condition is usually denoted by $\mathbf{O}_n(k)$ and is also called the *orthogonal group* of degree n over k, with the subgroup $\mathbf{SO}_n(k)$ of proper orthogonal matrices.

It is clear that isometric spaces have isomorphic orthogonal groups. Explicitly, if $\alpha : V \to V'$ is an isometry, then the map $\theta \mapsto \alpha^{-1}\theta\alpha$ is an isomorphism from $\mathbf{O}(V)$ to $\mathbf{O}(V')$.

A vector $x \neq 0$ in an inner product space V is called *isotropic* if $q(x) = 0$; otherwise x is *anisotropic*. A subspace U of V is *isotropic* if it contains an isotropic vector, and *anisotropic* otherwise. If every non-zero vector in U is isotropic, this means (in characteristic not 2) that the inner product restricted to U vanishes identically; in that case U is said to be *totally isotropic*.

Let us now take a closer look at the orthogonal group $\mathbf{O}(V)$ of an inner product space V. Throughout we shall suppose that char $k \neq 2$ and that V is regular.

Any $\alpha \in \mathbf{O}(V)$ such that $\alpha^2 = 1$ but $\alpha \neq 1$ is called an *involution*. Such transformations may be obtained as follows. Let $V = U \perp U'$ be an orthogonal sum decomposition; we define the *reflexion* with respect to (U, U') as the linear transformation of V which maps each $x \in U$ to $-x$ and leaves U' pointwise fixed. Clearly this is an orthogonal transformation $\neq I$ whose square is 1, so it is an involution (provided that $U \neq 0$). Conversely, let α be an involution on V and put

$$U_+ = \{x \in V | x\alpha = x\}, \quad U_- = \{x \in V | x\alpha = -x\}.$$

Then $U_+ \cap U_- = 0$, because char $k \neq 2$, and $V = U_+ + U_-$, since every $x \in V$ can be written as $x = u + v$, where $u = (x + x\alpha)/2$, $v = (x - x\alpha)/2$, and clearly $u \in U_+$, $v \in U_-$. Finally, if $x \in U_+$, $y \in U_-$ then $b(x, y) = b(x\alpha, y\alpha) = -b(x, y)$; hence $b(x, y) = 0$ and so $V = U_+ \perp U_-$. This shows that every involution is in fact a reflexion.

A reflexion with respect to (U, U') is said to be of type p ($p \in \mathbf{N}$), if $\dim U = p$. If we use a basis of V adapted to the decomposition $V = U_+ \perp U_-$, then α is represented by a diagonal matrix with $p - 1$'s and $n - p$ 1's, where $n = \dim V$. For example, a reflexion of type 2 in Euclidean 3-space is a rotation through an angle π.

The reflexions of type 1 are particularly important: they consist of reflexions in a hyperplane and the general reflexion can be expressed as a product of reflexions of type 1. To obtain an explicit formula for the reflexion with respect to $(u, u^\perp)$, where u is an anisotropic vector, consider the mapping

$$\sigma_u : x \mapsto x - \lambda(x)u,$$

where $\lambda = \lambda(x)$ is a linear function of x. This is a linear transformation of V and it will be orthogonal iff $q(x) = q(x - \lambda u) = q(x) - 2\lambda b(x, u) + q(u)\lambda^2$. Excluding the trivial case $\lambda = 0$, we see that the condition for an isometry is $\lambda q(u) - 2b(x, u) = 0$; hence

$$\sigma_u : x \mapsto x - \frac{2b(x, u)}{q(u)} u \tag{8.3.2}$$

is an orthogonal transformation. It is called the *symmetry* with respect to u and the vectors x, $x\sigma_u$ are called *symmetric* with respect to u. By using a basis adapted to the decomposition $V = (u) \perp u^\perp$ we see that it is the reflexion with respect to $(u, u^\perp)$: the vectors along u are reversed while the hyperplane of vectors orthogonal to u remains fixed. This makes it clear that the symmetries are just the reflexions of type 1. A reflexion of type p is described in a suitable orthogonal basis $e_1, \ldots, e_n$ by $e_i \mapsto -e_i$ $(i = 1, \ldots, p)$, $e_i \mapsto e_i$ $(i = p + 1, \ldots, n)$. But this can be written as $\sigma_{e_1} \ldots \sigma_{e_p}$, hence every reflexion can be written as a product of symmetries. The

reflexions in turn generate $\mathbf{O}(V)$; to establish this fact we first show that any anisotropic vector can be transformed into any other vector of the same length by a reflexion.

Lemma 8.3.1. *In an inner product space V (in characteristic $\neq 2$) let u, v be vectors such that $q(u) = q(v) \neq 0$. Then there is a reflexion of V which maps u to v.*

Proof. Since $q(u) = q(v)$, the vectors $x = (u+v)/2$, $y = (u-v)/2$ are orthogonal, as is easily checked, and they cannot both be isotropic, because $q(u) \neq 0$. Let X, Y be the spaces spanned by x, y respectively; if $q(x) \neq 0$, then $V = X \perp X^{\perp}$ and the reflexion with respect to $(X^{\perp}, X)$ maps $u = x + y$ to $v = x - y$; if $q(y) \neq 0$, then $V = Y \perp Y^{\perp}$ and the reflexion with respect to $(Y, Y^{\perp})$ maps u to v. ∎

Theorem 8.3.2. *Let V be a regular inner product space (in characteristic $\neq 2$) of dimension n. Then any orthogonal transformation in V can be written as a product of at most n reflexions.*

Proof. Let $\alpha \in \mathbf{O}(V)$ and choose an anisotropic vector $e_1 \in V$. Then $q(e_1) = q(e_1\alpha)$, hence by Lemma 8.3.1 there is a reflexion σ_1 such that $e_1\sigma_1 = e_1\alpha$, and so $\alpha\sigma_1$ leaves e_1 fixed. It follows that $\alpha\sigma_1$ also maps $V' = e^{\perp}$ into itself, but $\dim V' = n - 1$, so by induction on n we can write $\alpha\sigma_1|V' = \sigma'_n \ldots \sigma'_2$, where σ'_i is a reflexion in V'. Each σ'_i can be extended to a reflexion σ_i of V which leaves e_1 fixed. Then $\alpha\sigma_1\sigma_2 \ldots \sigma_n$ leaves e_1 fixed as well as every vector in V'. Hence it must be the identity, and so $\alpha = \sigma_n \ldots \sigma_1$, as required. ∎

Since every reflexion is a product of at most n symmetries, we have

Corollary 8.3.3. *Every orthogonal transformation on an n-dimensional regular space (in characteristic $\neq 2$) is a product of at most n^2 symmetries.* ∎

This bound can still be improved: if we examine the proof of Lemma 8.3.1 we see that the reflexion there can be taken to be of type 1 or 2, so the bound n^2 can be replaced by $2n$. But there is a more precise result: the Cartan–Dieudonné theorem states that every orthogonal transformation on an n-dimensional space can be written as a product of at most n symmetries (see e.g. Artin (1957)). For example, in three dimensions every orthogonal transformation is a product of at most three symmetries and a proper orthogonal transformation ($\neq 1$) is a product of two symmetries; each leaves a plane fixed and these two planes meet in a line, so the transformation must leave a line fixed, i.e. it is a rotation about that line.

Let α be any orthogonal transformation and u be an anisotropic vector in V. Then for any $x \in V$, x and $x\sigma_u$ are symmetric with respect to u, hence $x\alpha$ and $x\alpha\sigma_{u\alpha}$ are symmetric with respect to $u\alpha$, and so $(x\sigma_u)\alpha = x\alpha\sigma_{u\alpha}$. Thus we obtain

$$\sigma_{u\alpha} = \alpha^{-1}\sigma_u\alpha. \tag{8.3.3}$$

Since $\mathbf{O}(V)$ is generated by symmetries σ_u, its derived group $\mathbf{O}(V)'$ is generated by all commutators of symmetries $\sigma_u^{-1}\sigma_v^{-1}\sigma_u\sigma_v = (\sigma_u\sigma_v)^2$. For the subgroup H generated

by all $(\sigma_u\sigma_v)^2$ is normal in $\mathbf{O}(V)$, by (8.3.3), and any symmetries commute (mod H), hence $\mathbf{O}(V)/H$ is abelian, while clearly $H \subseteq \mathbf{O}(V)'$; therefore $H = \mathbf{O}(V)'$. We can also express $\mathbf{O}(V)'$ in terms of rotations as long as dim V is not too small:

Theorem 8.3.4. *Let V be a regular inner product space (in characteristic $\neq 2$) . Then the derived group $\mathbf{O}(V)'$ of the orthogonal group is generated by all $(\sigma_u\sigma_v)^2 (u, v \in V)$. If dim $V > 2$, then $\mathbf{O}(V)' = \mathbf{SO}(V)'$.*

Proof. The first part has been shown and it is clear that $\mathbf{SO}(V)' \subseteq \mathbf{O}(V)'$, so it only remains to show that $(\sigma_u\sigma_v)^2$ is a product of commutators of rotations. When dim V is odd, -1 is an improper orthogonal transformation and $(\sigma_u\sigma_v)^2 = [(-\sigma_u)(-\sigma_v)]^2$ is the required representation. There remains the case when dim V is even and so dim $V \geq 4$. We note that if w is a vector orthogonal to u, then $\sigma_u\sigma_w\sigma_u = \sigma_w$, hence σ_u commutes with σ_w. Thus if we can find an anisotropic vector w orthogonal to u and v, then $(\sigma_u\sigma_v)^2 = (\sigma_u\sigma_w.\sigma_v\sigma_w)^2$ and this is the desired expression of $(\sigma_u\sigma_v)^2$ as a commutator of two rotations. It remains to show that such a vector w can always be found. If this is not the case, then every vector orthogonal to u and v is isotropic. Let U be the subspace spanned by u and v; then $U^\perp$ is totally isotropic, hence $U^\perp \subseteq U^{\perp\perp} = U$, and dim $U^\perp = n - \dim U \geq n - 2 \geq 2$, so $U = U^\perp$ is totally isotropic, which contradicts the fact that $q(u) \neq 0$. ∎

We shall see later that when dim $V = 2$, then $\mathbf{SO}(V)$ is abelian, but $\mathbf{O}(V)$ need not be (see Exercise 4).

Exercises

1. Show that a product of r symmetries on an n-dimensional space leaves fixed a subspace of dimension at least $n - r$. Deduce that there are orthogonal transformations which cannot be written as a product of fewer than n symmetries.
2. Show that when dim $V = n$, then $\mathbf{O}(V)$ can be generated by all reflexions of type $n - 1$. What can be said about other types?
3. Show that the reflexion in Lemma 8.3.1 can be taken to be of type 1 or 2. (Hint. If $q(y) = 0$, take a reflexion with respect to $(H, H^\perp)$, where H is a 2-dimensional anisotropic subspace containing y and orthogonal to x.)
4. Find all fields k and dimensions n for which $\mathbf{O}_n(k)$ is commutative.
5. Show that any orthogonal transformation in the centre of $\mathbf{O}(V)$ leaves any anisotropic vector fixed or reverses it. Deduce that the centre of $\mathbf{O}(V)$ is $\{1, -1\}$.

8.4 The Clifford Algebra and the Spinor Norm

Let (V, q) be a quadratic space; then as we saw in Section 8.1,

$$B(x, y) = q(x + y) - q(x) - q(y) \tag{8.4.1}$$

is a symmetric bilinear form. Suppose that char $k \neq 2$, the case that will mainly concern us here. Then we define the bilinear form associated with the quadratic form q by

$$b(x, y) = \frac{1}{2} B(x, y)$$

With the help of the symmetric bilinear form b we can reach q by the equation

$$b(x, x) = q(x). \tag{8.4.2}$$

Thus we can pass from q to b and back, so in characteristic $\neq 2$ quadratic forms and bilinear forms are equivalent.

Given a quadratic space (V, q), suppose that we have a linear mapping $\lambda : V \to A$ into a k-algebra A. The mapping λ is said to be *admissible* and A is *compatible* with λ, if

$$(x\lambda)^2 = q(x) \quad \text{for all } x \in V. \tag{8.4.3}$$

On the left we have a product in A and on the right a scalar multiple of the unit element in A. From (8.4.3) we find by linearization

$$x\lambda . y\lambda + y\lambda . x\lambda = B(x, y), \tag{8.4.4}$$

where B is as in (8.4.1). In particular, if x and y are orthogonal vectors in V, then $x\lambda$ and $y\lambda$ anticommute while by (8.4.3) $x\lambda$ is invertible in A if x is anisotropic. Conversely, if $x\lambda$ has an inverse c in A, then $x\lambda . c = 1$, hence $q(x)c^2 = (x\lambda)^2 c^2 = 1$ and so $q(x) \neq 0$. Thus $x\lambda$ is invertible iff x is anisotropic. Such algebras exist for every quadratic form, in fact there is a distinguished one (first described by William Kingdon Clifford in 1878) which may be defined as follows.

Definition. Let (V, q) be a quadratic space over a field k. A *Clifford algebra* for (V, q) is a k-algebra C with a linear mapping $\mu : V \to C$ which is admissible, such that for any algebra A with an admissible map $\lambda : V \to A$ there exists a unique homomorphism $\lambda' : C \to A$ such that $\lambda = \mu\lambda'$. This is also expressed by saying that the pair C, μ is *universal* for this property.

We remark that the map $\mu : V \to C$ may be regarded as an initial object in the category of all admissible maps from V, and as such is unique up to isomorphism. It only remains to prove the existence of C. To do this we form the tensor algebra on V, as a mod 2 graded algebra: $\mathbf{T}(V) = T_0(V) \oplus T_1(V)$, where T_0, T_1 are the components of even and odd degree in $\mathbf{T}(V)$, respectively. Let I be the ideal of $\mathbf{T}(V)$ generated by the elements $x^2 - q(x) (x \in V)$. Since these elements are homogeneous of degree 0 (mod 2), I is a graded ideal and the quotient $C = \mathbf{T}(V)/I$ is again mod 2 graded, and it has the required universal property by construction: Given an admissible mapping $\lambda : V \to A$, we can extend λ to a homomorphism from $\mathbf{T}(V)$ to A because the former is free; this homomorphism maps I to 0, hence it can be factored via $\mathbf{T}(V)/I = C$, and it is unique because it is determined on a generating set of C, namely the image in C of V. Further, $C = C(V)$ is unique up to isomorphism, as universal object. Let us sum up our findings:

Proposition 8.4.1. *Every quadratic space (V, q) has a Clifford algebra $C(V)$, defined as the universal algebra compatible with a mapping $\mu_V : V \to C(V)$, and isometric spaces have isomorphic Clifford algebras.*

Proof. Only the last part still needs proof. If $f : V \to W$ is an isometry, then $f\mu_W : V \to W \to C(W)$ is admissible, hence there is a unique homomorphism $\varphi : C(V) \to C(W)$ such that $f\mu_W = \mu_V\varphi$. Since f is invertible, so is φ and it is the required isomorphism. ∎

As we have seen, the Clifford algebra C of V is mod 2 graded; we write $C = C_0 \oplus C_1$ and call the subalgebra C_0 the *even* Clifford algebra; the elements of C_0, C_1 are called *even* and *odd* respectively.

We shall soon see that the Clifford algebra of an n-dimensional space has dimension 2^n; for the moment we observe that when char $k \neq 2$,

$$\dim C \leq 2^n, \quad \text{where } \dim V = n. \tag{8.4.5}$$

For we can take an orthogonal basis $e_1, \ldots, e_n$ in V (Theorem 8.2.2). If the image of e_i in C is written u_i, then C is generated by $u_1, \ldots, u_n$ and by (8.4.3) and (8.4.4), $u_iu_j = -u_ju_i$ $(i \neq j)$, $u_i^2 = q(e_i)$. It follows that any product of u_i can be arranged in ascending order of suffixes by inserting a scalar coefficient. Thus C is spanned by the elements

$$u_{i_1} \ldots u_{i_r}, \quad i_1 < \ldots < i_r, \quad r = 0, 1, \ldots, n. \tag{8.4.6}$$

The number of these products is 2^n and (8.4.5) follows.

We pause to give some examples. As before, a space with the quadratic form $a_1x_1^2 + \ldots + a_nx_n^2$ (relative to an orthogonal basis) will be denoted by $\langle a_1, \ldots, a_n\rangle$.

Example 1. If V is 1-dimensional, say $V = \langle a\rangle$, then $C(V) = k[x]/(x^2 - a)$. Both C_0 and C_1 are one-dimensional and C has a basis 1, u with multiplication $u^2 = a$. If $a \neq 0$, C is either the direct sum of two copies of k or a quadratic extension field, namely $k(\sqrt{a})$, according as a is or is not a square in k.

Example 2. Let $V = \langle a, b\rangle$, where $a, b \in k^\times$. Then $C(V)$ is generated by two elements u, v such that $u^2 = a$, $v^2 = b$, $uv = -vu$; this is the quaternion algebra $(a, b; k)$, which we met in Section 5.4.

Example 3. If V has the basis u, v with $q(u) = q(v) = 0$, $b(u, v) = 1$, then $C(V)$ has the basis 1, u, v, uv with defining relations $u^2 = v^2 = 0$, $uv + vu = 1$. This is just the full matrix ring k_2 (see Section 4.4); we note that this applies even in characteristic 2.

Example 4. If the quadratic form on V is identically zero, then $C(V)$ reduces to the exterior algebra on V, see Section 6.4.

To study Clifford algebras we shall need the notion of a graded product. Given two graded algebras $A = \oplus A_i$, $B = \oplus B_i$, we define their graded tensor product $A \hat{\otimes} B$ as

the algebra whose underlying vector space is $A \otimes B$, again graded by total degree, with multiplication

$$(a_i \hat{\otimes} b_j)(c_r \hat{\otimes} d_s) = (-1)^{jr} a_i c_r \hat{\otimes} b_j d_s, \tag{8.4.7}$$

where the suffixes indicate the degree. In particular, this definition applies to mod 2 graded algebras. It is easily checked that the multiplication so defined is associative and distributive.

We can now describe the structure of Clifford algebras more precisely.

Theorem 8.4.2. *Let V be a quadratic space over a field k of characteristic not 2, with an orthogonal basis $e_1, \ldots, e_n$. Then the 2^n products*

$$e_{i_1} \ldots e_{i_r}, \quad i_1 < \ldots < i_r, \quad r = 0, 1, \ldots, n. \tag{8.4.8}$$

constitute a basis of $C(V)$; in particular, the admissible mapping $\mu : V \to C(V)$ is injective. Thus $\dim C(V) = 2^n$ *and if $n > 0$, then* $\dim C_0 = \dim C_1 = 2^{n-1}$.

If (V, q), (V', q') are two quadratic spaces and $(V \perp V', q + q')$ is their orthogonal sum, then

$$C(V \perp V') \cong C(V) \hat{\otimes} C(V'). \tag{8.4.9}$$

Proof. We begin with the last part. The inclusion $V \to V \perp V'$ induces a homomorphism $f : C(V) \to C(V \perp V')$; similarly we have a homomorphism $f' : C(V') \to C(V \perp V')$. The images of f and f' anticommute, for if $x \in V$, $x' \in V'$, then $x.x' = 0$ in $V \perp V'$, hence $xf.x'f' + x'f'.xf = 0$ in $C(V \perp V')$. Thus the generators anticommute, and by induction on the degree this holds for the subalgebras generated. We therefore have a homomorphism of graded algebras:

$$C(V) \hat{\otimes} C(V') \to C(V \perp V'). \tag{8.4.10}$$

To find the inverse, consider the mapping

$$(x, x') \mapsto x \hat{\otimes} 1 + 1 \hat{\otimes} x' \tag{8.4.11}$$

of $V \perp V'$ into $C(V) \hat{\otimes} C(V')$. We have

$$\begin{aligned}(x \hat{\otimes} 1 + 1 \hat{\otimes} x')^2 &= x^2 \hat{\otimes} 1 + (x \hat{\otimes} 1)(1 \hat{\otimes} x') + (1 \hat{\otimes} x')(x \hat{\otimes} 1) + 1 \hat{\otimes} x'^2 \\ &= q(x) + q(x') + x \hat{\otimes} x' - x \hat{\otimes} x' \\ &= q(x) + q(x').\end{aligned}$$

The right-hand side represents the quadratic form on $V \perp V'$, hence the algebra $C(V) \hat{\otimes} C(V')$ is compatible with the mapping (8.4.11), so by the universal property of $C(V \perp V')$ we obtain a homomorphism inverse to (8.4.10). So (8.4.9) is proved, as isomorphism of graded algebras.

Let (V, q) be any quadratic space of dimension n and write $V = \langle a_1, \ldots, a_n \rangle$. By what has been proved, we have an isomorphism

$$C(V) \cong C(\langle a_1 \rangle) \hat{\otimes} \ldots \hat{\otimes} C(\langle a_n \rangle).$$

Now dim $C\langle a\rangle = 2$ by the special case treated earlier, hence the right-hand side has dimension 2^n. As we saw, the 2^n elements (8.4.6) span $C(V)$, hence they form a basis, in particular $e_1, \ldots, e_n$ are linearly independent, so the admissible mapping μ is injective. The assertions about C_0, C_1 now follow because C_m is spanned by the products (8.4.8) with $r \equiv m \pmod 2$. ■

This result also provides another proof of Theorem 6.4.1 on exterior algebras, which is just the special case where the quadratic form is identically zero. In what follows we shall assume that V is a regular quadratic space, i.e. $V^\perp = 0$; the structure of $C(V)$ is then expressible in terms of quaternion algebras.

Let us write (a, b) for the quaternion algebra $(a, b; k)$ and (a) for the Clifford algebra of the space $\langle a\rangle$. Then we have the formulae

$$(a) \hat{\otimes} (b) = (a, b), \tag{8.4.12}$$

$$(a_1) \hat{\otimes} (a_2) \hat{\otimes} (a_3) = (-a_1a_2, -a_2a_3) \otimes (-a_1a_2a_3), \tag{8.4.13}$$

$$(a_1, a_2) \hat{\otimes} (a_3, a_4) = (-a_1a_2, -a_2a_3) \otimes (-a_1a_2a_3, a_4), \tag{8.4.14}$$

where on the right we can omit the $\hat{}$ because in each case the first factor is even. The formula (8.4.12) is a special case of (8.4.9). To prove (8.4.13), let (a_i) be spanned by $1, u_i$, where $u_i^2 = a_i$. The u_i anticommute, hence u_1u_2 and u_2u_3 anticommute and $(u_1u_2)^2 = -a_1a_2$, $(u_2u_3)^2 = -a_2a_3$, therefore u_1u_2 and u_2u_3 generate a quaternion algebra on the left of (8.4.13), while $u_1u_2u_3$ generates a subalgebra of the centre. Hence the left of (8.4.13) contains the right-hand side, so equality holds, since both sides are 8-dimensional. This proves (8.4.13); now using (8.4.12) and (8.4.13), we have

$$\begin{aligned}(a_1, a_2) \hat{\otimes} (a_3, a_4) &= (a_1) \hat{\otimes} (a_2) \hat{\otimes} (a_3) \hat{\otimes} (a_4)\\ &= (-a_1a_2, -a_2a_3) \hat{\otimes} (-a_1a_2a_3) \hat{\otimes} (a_4),\\ &= (-a_1a_2, -a_2a_3) \otimes (-a_1a_2a_3, a_4),\end{aligned}$$

i.e. (8.4.14). The formulae (8.4.12)–(8.4.14) can be used to describe the structure of $C(V)$, but we shall also need the determinant of V. In terms of a basis $u_1, \ldots, u_n$ of V this is $d = \det(b(u_i, u_j))$ and it is determined modulo a square (see Section 8.1). Sometimes we shall need the *discriminant* δ of V, defined as

$$\delta = (-1)^{n(n-1)/2} d = (-1)^m d, \quad \text{if } \dim V = n = 2m \text{ or } 2m + 1. \tag{8.4.15}$$

We note that the space is regular iff $d \neq 0$, and when $\dim V \equiv 0$ or $1 \pmod 4$, or when -1 is a square in k, then $\delta = d$; otherwise $\delta = -d$. For a regular quadratic form the Clifford algebra is close to being a simple algebra:

Theorem 8.4.3. *Let (V, q) be a regular quadratic space over a field k of characteristic not 2, and denote its discriminant by δ.*

(i) *If* dim V *is even then* $C(V)$ *is simple with centre* k*; if* dim V *is odd, then* $C(V)_0$ *is simple with centre* k.

(ii.a) *Suppose that* δ *is not a square in* k*. If* dim V *is odd, then* $C(V)$ *is simple with centre* $k(\sqrt{\delta})$*; if* dim V *is even, then* $C(V)_0$ *is simple with centre* $k(\sqrt{\delta})$.

(ii.*b*) *Suppose that* δ *is a square in* k*. If* dim V *is odd, then* $C(V)$ *is a direct product of two simple algebras with centre* k*; if* dim V *is even, then* $C(V)_0$ *is a direct product of two simple algebras with centre* k.

The relation between C, C_0 and their centres is easily remembered by bearing in mind that the dimension of a central simple algebra is a perfect square (Theorem 5.4.6). Thus if δ is not a square in k and $k(\sqrt{\delta}) = F$, then for dim $V = 2m$, $[C : k] = 2^{2m}$, $[C_0 : F] = 2^{2(m-1)}$, and when dim $V = 2m + 1$, then $[C : F] = 2^{2m}$, $[C : k] = 2^{2m}$.

Proof. Let $V = \langle a_1, \ldots, a_n \rangle$; using (8.4.13) we can step by step replace two-dimensional factors by quaternion algebras. At the first stage,

$$C(V) = (a_1) \hat{\otimes} \ldots \hat{\otimes} (a_n) = (-a_1a_2, -a_2a_3) \hat{\otimes} (-a_1a_2a_3) \hat{\otimes} (a_4) \hat{\otimes} \ldots \hat{\otimes} (a_n);$$

here the number of two-dimensional factors has been reduced by two, while the first factor is an even quaternion algebra. Suppose that $n = 2m$; then after m such steps we have a tensor product of quaternion algebras, hence $C(V)$ is then central simple over k. If $n = 2m + 1$, then after m steps we have a product of m quaternion algebras by $((-1)^m a_1 \ldots a_n) = (\delta)$, but as we have seen, this is $k[x]/(x^2 - \delta)$, which is $k(\sqrt{\delta})$ in case (ii.a) and $k \times k$ in case (ii.*b*). So we have either a simple algebra with centre $k(\sqrt{\delta})$ or a direct product of two central simple k-algebras.

Now consider $C_0 = C(V)_0$. When $n = 2m + 1$, C is the tensor product of m quaternion algebras, which is central simple over k; when $n = 2m$, we have after $m - 1$ steps a product of $m - 1$ even quaternion algebras by the algebra

$$((-1)^m a_1 \ldots a_{2m-1}) \hat{\otimes} (a_{2m}) \cong ((-1)^m a_1 \ldots a_{2m-1}, a_{2m}).$$

If the basis of this quaternion algebra is $1, u, v, uv$, then $(uv)^2 = (-1)^m a_1 \ldots a_{2m} = \delta$, so the even part of this algebra is again $k[x]/(x^2 - \delta)$, which is $k(\sqrt{\delta})$ in case (ii.a) and $k \times k$ in case (ii.*b*), so we again have either a central simple $k(\sqrt{\delta})$-algebra or a direct product of two central simple k-algebras. ∎

Sometimes it is useful to have an explicit expression for the centre of $C(V)$:

Proposition 8.4.4 *Let* (V, q) *be as in Theorem 8.4.3. If* dim $V = n$ *is odd, then the centre of* $C(V)$ *is two-dimensional with basis* $1, u_1 \ldots u_n$ *and no non-zero element anticommutes with all of* $C(V)$*. If* dim V *is even, then the centre of* $C(V)$ *is* k *and any anticommuting element is a scalar multiple of* $u_1 \ldots u_n$.

Proof. Let $u_1, \ldots, u_n$ be an orthogonal basis of V; we first note how an element changes under conjugation by u_i:

$$u_i^{-1} u_{i_1} \ldots u_{i_r} u_i = \lambda u_{i_1} \ldots u_{i_r},$$

where $\lambda = -1$ if an odd number of suffixes $i_1, \ldots, i_r$ differ from i and $\lambda = 1$ otherwise. For a given product both possibilities are realized unless $r = 0$ or n, and $u_1 \ldots u_n$ commutes with all elements of C for odd n and anticommutes for even n. The assertion is an immediate consequence. ■

It is possible to relate the orthogonal group of a quadratic space to its Clifford algebra; to carry this out we shall assume that char $k \neq 2$. We recall from Section 8.3 that for any anisotropic vector u we have the symmetry with respect to u:

$$\sigma_u : x \mapsto x - \frac{2b(x, u)}{q(u)} u, \tag{8.4.16}$$

which reverses vectors along u and preserves vectors orthogonal to u; thus it is the reflexion in the hyperplane orthogonal to u. We also recall from Corollary 8.3.3 that in an n-dimensional regular quadratic space any orthogonal transformation is a product of at most n^2 symmetries.

Let (V, q) be a regular quadratic space, $\mathbf{O}(V)$ be its orthogonal group and $C(V)$ be its Clifford algebra. Any $\theta \in \mathbf{O}(V)$ can be combined with the canonical inclusion $V \to C(V)$ and by the universal property extends to a unique automorphism of $C(V)$; thus we have a mapping

$$\mathbf{O}(V) \to \mathrm{Aut}(C(V)), \tag{8.4.17}$$

which is easily seen to be a group homomorphism. To analyse it further we shall require a result from FA, the Skolem–Noether theorem, which states that in a central simple k-algebra every automorphism is inner. Coupled with Theorem 8.4.3 this tells us that when dim V is even, every automorphism of $C(V)$ is inner. This suggests a representation of orthogonal transformations by inner automorphisms, but there are two complications. Firstly, the element inducing an inner automorphism is determined only up to a scalar factor, and secondly we have to take the grading into account. To meet the first point we shall start by taking an invertible element of $C(V)$ and use it to define an orthogonal transformation. To allow for the grading one could develop a theory of central simple *graded* algebras (as in Lam (1980) Chapter 4), but for our purpose the following development will be sufficient.

Let u be an anisotropic vector in V. Then u is invertible in $C(V)$ and for any $x \in V$ we have $xu + ux = 2b(x, u)$; hence

$$u^{-1}xu + x = 2b(x, u)u^{-1} = \frac{2b(x, u)}{q(u)} u.$$

Thus the symmetry (8.4.16) can be written as

$$x\sigma_u = -u^{-1}xu, \quad \text{where } x, u \in V, q(u) \neq 0. \tag{8.4.18}$$

The anisotropic vectors generate a group under multiplication in $C(V)$, because they are units; this group is denoted by $\Gamma(V)$ and is called the *Clifford group*; its elements are sometimes called *versors*. Let $z \in \Gamma(V)$; if $z = v_1 \ldots v_r$ say $(v_i \in V)$, then $\gamma_z = \sigma_{v_1} \ldots \sigma_{v_r}$ is an orthogonal transformation, which by (8.4.18) can be written

$x\gamma_z = (-1)^r z^{-1} x z$. If ν is the automorphism of $C(V)$ extending the orthogonal transformation $x \mapsto -x$ of V, then this may be written as

$$x\gamma_z = z^{-1} x z^\nu, \quad x \in V, z \in \Gamma(V). \tag{8.4.19}$$

It is clear that the mapping $z \mapsto \gamma_z$ is a homomorphism of $\Gamma(V)$ into $\mathbf{O}(V)$; since every orthogonal transformation is a product of symmetries (Corollary 8.3.3), this mapping is actually surjective. To find the kernel, assume that $\gamma_z = 1$. Then $xz^\nu = zx$ for all $x \in V$; writing $z = z_0 + z_1$, where $z_i \in C_i$, we have $x(z_0 - z_1) = (z_0 + z_1)x$, and on equating odd and even components we see that $xz_0 = z_0 x$, $xz_1 = -z_1 x$ for all $x \in V$. By Proposition 8.4.4, $z_1 = 0, z_0 \in k$, hence $\ker \gamma = k^\times$. We sum up our conclusions as

Theorem 8.4.5. *Let (V, q) be a regular quadratic space over k, where char $k \neq 2$. Then the mapping γ defined by* (8.4.19) *gives rise to an exact sequence*

$$1 \to k^\times \to \Gamma(V) \xrightarrow{\gamma} \mathbf{O}(V) \to 1. \qquad \blacksquare$$

This means that each orthogonal transformation determines a versor up to a scalar multiple, and it shows that (8.4.17) is an isomorphism, with inverse induced by γ, once the grading has been allowed for.

To make the correspondence more precise we restrict ourselves to versors of norm 1, which are obtained as follows. The mapping $x \mapsto -x$ of V can also be extended to a unique antiautomorphism $x \mapsto x^J$ of $C(V)$, called the *fundamental involution*, and the *norm* N on $C(V)$ is then defined as

$$N(x) = xx^J, \quad \text{for } x \in C(V). \tag{8.4.20}$$

It is clear that N defines a group homomorphism from $\Gamma(V)$ to $k^\times$ such that $N(x^\nu) = N(x)$. The versors x such that $N(x) = 1$ form a subgroup of $\Gamma(V)$ called the *reduced Clifford group* and denoted by $\mathrm{pin}(V)$, for reasons which will soon become clear. Its image under γ is called the *reduced orthogonal group* or *spinor kernel* and will be denoted by $\mathbf{O}'(V)$. To identify the cokernel of the map $\mathbf{O}'(V) \to \mathbf{O}(V)$ we define a homomorphism $\theta : \mathbf{O}(V) \to k^\times / k^{\times 2}$ by the following rule: any $\alpha \in \mathbf{O}(V)$ may be written as a product of symmetries:

$$\alpha = \sigma_{u_1} \dots \sigma_{u_r}. \tag{8.4.21}$$

When (8.4.21) holds, we define

$$\theta(\alpha) \equiv q(u_1) \dots q(u_r) \pmod{k^{\times 2}}. \tag{8.4.22}$$

To show that this is well-defined we have to verify that when $\sigma_{u_1} \dots \sigma_{u_r} = 1$, then $q(u_1) \dots q(u_r)$ is a square in k. In this case r is even, because 1 is a rotation, therefore the element $a = u_1 \dots u_r$ of $C(V)$ induces the identity on V and so lies in the centre of $C(V)$. Hence $a \in k$ and

$$q(u_1) \dots q(u_r) = u_1 \dots u_r u_r \dots u_1 = aa^J = a^2,$$

which proves the assertion. Thus $\theta(\alpha)$ is a well-defined residue class mod $k^{\times 2}$, and it

is clear that $\theta(\alpha\beta) = \theta(\alpha)\theta(\beta)$. The resulting homomorphism θ is called the *spinor norm.*

We remark that θ is not invariant under a scale change: if we replace q by λq, then $\theta(\sigma_u)$ changes by a factor λ. For this reason θ is usually restricted to $\mathbf{SO}(V)$; now a scale change $q \mapsto \lambda q$ replaces $\theta(\alpha)$ by $\lambda^r\theta(\alpha)$ and since r is even, this is a square. We also note that $\theta(-1)$ is the determinant of the quadratic form.

The relations between the various groups may be summarized in the following commutative diagram with exact rows and columns:

$$\begin{array}{ccccccccc} & & 1 & & 1 & & 1 & & \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 1 & \to & \{\pm 1\} & \longrightarrow & \mathrm{pin}(V) & \longrightarrow & \mathbf{O}'(V) & \to & 1 \\ & & \downarrow & & \downarrow & & \downarrow & & \\ 1 & \to & k^\times & \longrightarrow & \Gamma(V) & \xrightarrow{\gamma} & \mathbf{O}(V) & \to & 1 \\ & & \downarrow & & \downarrow N & & \downarrow \theta & & \\ 1 & \to & k^{\times 2} & \longrightarrow & k^\times & \longrightarrow & k^\times/k^{\times 2} & \to & 1 \\ & & \downarrow & & & & & & \\ & & 1 & & & & & & \end{array} \tag{8.4.23}$$

We note that the spinor norm is trivial whenever every norm is a square, e.g. for the complex numbers or a positive definite form over the real numbers.

If we restrict ourselves to proper orthogonal transformations and assume that every norm is a square, we obtain the following exact sequence, corresponding to the first row of the diagram (8.4.23) and revealing the (whimsical) origin of pin:

$$1 \to \{\pm 1\} \to \mathrm{Spin}(V) \to \mathbf{SO}(V) \to 1. \tag{8.4.24}$$

Here $\mathbf{SO}(V)$ is the special orthogonal group and $\mathrm{Spin}(V)$, called the *spin group*, is the inverse image under γ (restricted to $\mathrm{pin}(V)$) of $\mathbf{SO}(V)$. The exactness follows by definition.

As an example consider a regular two-dimensional space $V = \langle a, b\rangle$. Its Clifford algebra is the quaternion algebra (a, b). Specializing to the real quaternions we have $a = b = -1$; the algebra has the basis $1, i, j, ij = k$ and $N(x_0 + x_1 i + x_2 j + x_3 k) = x_0^2 + x_1^2 + x_2^2 + x_3^2$. Thus every norm is a square in $\mathbf{R}$; moreover, $\mathrm{Spin}(V)$ is the group of all matrices $P = \begin{pmatrix} x & -\bar{y} \\ y & \bar{x} \end{pmatrix}$, $x, y \in \mathbf{C}$. We have $\det P = x\bar{x} + y\bar{y} = 1$, which is the special unitary group $\mathbf{SU}_2(\mathbf{C})$. Now (8.4.24) takes the form

$$1 \to \{\pm 1\} \to \mathbf{SU}_2(\mathbf{C}) \to \mathbf{SO}_3(\mathbf{R}) \to 1.$$

This is the double covering of the rotation group in 3-space, also called the *spin representation.*

As a second example consider a two-dimensional quadratic space H with a basis u, v such that $q(u) = q(v) = 0$, $b(u, v) = 1$. This is known as a *hyperbolic plane* with *hyperbolic pair* u, v (see Section 8.5 below). On changing the basis to

$e = u + v, f = u - v$ we find that $b(e,f) = 0$, $q(e) = -q(f) = 1$; thus $H = \langle 1, -1\rangle$. Its Clifford algebra has basis 1, u, v, uv, and is easily seen to be the full matrix ring with matrix units $e_{12} = u$, $e_{21} = v$, $e_{11} = uv$, $e_{22} = vu$. More generally, we may consider a hyperbolic space, which is an orthogonal sum of hyperbolic planes. To find its Clifford algebra we note by (8.4.14) that $(-1,1)\hat{\otimes}(1,-1) = (1,1)\otimes(1,-1)$. Thus we can replace the graded by the ungraded tensor product and so find as the Clifford algebra of a $2r$-dimensional hyperbolic space $\mathfrak{M}_{2^r}(k)$.

Exercises

1. Show that the graded tensor product of anticommutative algebras is again anticommutative.
2. Prove Theorem 8.4.2 for infinite-dimensional quadratic spaces.
3. For any quadratic space V, find the relation between $V^\perp$ (the radical of V) and the Jacobson radical of $C(V)$.
4. Show that every Clifford algebra is a graded tensor product of a Clifford algebra on a regular quadratic space and an exterior algebra.
5. Let V be a quadratic space and u be an anisotropic vector in V. Verify that right multiplication by u defines an isomorphism $C(V)_0 \cong C(u^\perp)$.
6. Verify that $C\langle 1^{k+2}\rangle \cong C\langle(-1)^k\rangle \otimes C\langle 1^2\rangle$, $C\langle(-1)^{k+2}\rangle \cong C\langle 1^k\rangle \otimes C\langle(-1)^2\rangle$.
7. In any quadratic space (V,q) verify that $\theta(-1)$ is the determinant of q.
8. Show that $\theta : \mathbf{O}(V) \to k^\times/k^{\times 2}$ is the unique homomorphism such that $(\sigma_u)\theta \equiv q(u) \pmod{k^{\times 2}}$ for all anisotropic vectors u.
9. Show that a non-zero quaternion u is pure (i.e. of trace 0) iff $u \notin k$, $u^2 \in k$. Verify that every pure quaternion u of non-zero norm is invertible and satisfies $x\sigma_u = -uxu^{-1}$ for all pure quaternions x.

8.5 Witt's Cancellation Theorem and the Witt Group of a Field

So far we have studied a single quadratic form over a field. We now come to a construction which embodies information about the totality of all quadratic forms over a given field. Throughout this section k is a field of characteristic not 2.

We have seen in Theorem 8.2.2 that the matrix of a quadratic form can always be taken in diagonal form, for a suitable choice of basis. However, this is not always the most convenient form, particularly as the diagonal elements of the matrix are not invariants, and the relation between different diagonal matrices is not straightforward. We shall return to this point in Section 8.9, but for the moment consider a different way of decomposing an inner product space. This is the theory developed by Ernst Witt in 1937, which leads to an invariant of the space; moreover, these invariants form an abelian group, the Witt group associated with the field k. We begin with a cancellation theorem which is often useful.

Theorem 8.5.1 (Witt's cancellation theorem). *Let V be an inner product space (in characteristic not 2) and V_1, V_2 be regular subspaces. If $V_1 \cong V_2$, then $V_1^\perp \cong V_2^\perp$.*

Proof. Assume first that V_1, V_2 are one-dimensional, spanned by v_1, v_2 respectively, where $q(v_1) = q(v_2) \neq 0$. By Lemma 8.3.1 there is an orthogonal transformation θ mapping v_1 to v_2; hence θ also maps $V_1^\perp$ to $V_2^\perp$, so these spaces are isometric, as claimed.

In general we have $V = V_1 \perp V_1^\perp = V_2 \perp V_2^\perp$, by Lemma 8.2.1. Now we may take $V_1 \cong V_2 \cong \langle a_1, \ldots, a_r \rangle$, by Theorem 8.2.2, and applying the case just proved to $\langle a_1 \rangle$ we obtain

$$\langle a_2, \ldots, a_r \rangle \perp V_1^\perp \cong \langle a_2, \ldots, a_r \rangle \perp V_2^\perp.$$

The result now follows by induction on $\dim V_1$. ∎

We observe that the result may be stated in the form (for a regular subspace U)

$$U \perp W_1 \cong U \perp W_2 \Rightarrow W_1 \cong W_2.$$

This form accounts for the name 'cancellation theorem'. In characteristic 2 the result does not hold: let H be a two-dimensional inner product space in characteristic 2 with matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$; then $\langle 1 \rangle \perp H \cong \langle 1, 1, 1 \rangle \cong \langle 1 \rangle \perp \langle 1, 1 \rangle$, by Exercise 5 of Section 8.2, even though H is not isomorphic to $\langle 1, 1 \rangle$. However, the cancellation theorem does hold in characteristic 2 for alternating forms (see Further Exercise 5).

For any field k we define a monoid $M(k)$ as follows: the elements of $M(k)$ are the isometry classes of regular inner product spaces over k. If the isometry class of V is written $[V]$, we can define an addition by the formula

$$[U] + [V] = [U \perp V]. \tag{8.5.1}$$

If $U \cong U'$, $V \cong V'$, then $U \perp V \cong U' \perp V'$; this shows the right-hand side of (8.5.1) to depend only on $[U]$, $[V]$, not on U, V themselves; so the addition of classes is well-defined. The operation is clearly associative, with the class of the zero space as neutral element, so we obtain a monoid in this way. Now Theorem 8.5.1 shows that this monoid $M(k)$ satisfies cancellation.

Theorem 8.5.1 leads to a further reduction; to describe it we must first look at the spaces which are to be regarded as 'trivial'. They are the hyperbolic planes, as defined in Section 8.4, which may also be characterized as follows:

Proposition 8.5.2. *For any two-dimensional inner product space V in characteristic not 2, the following conditions are equivalent:*

(a) *V is a regular isotropic space;*
(b) *V has a basis u, v such that $b(u, v) = 1, q(u) = q(v) = 0$;*
(c) *$V \cong \langle 1, -1 \rangle$;*
(d) *V has determinant $-1 (\mathrm{mod}\ k^{\times 2})$.*

A basis u, v as in (b) is called a *hyperbolic pair* for V.

Proof (a) $\Rightarrow$ (b). Let u be an isotropic vector and complete it to a basis u, v' of V. Then $b(u, v') \neq 0$, because V is regular, and on multiplying v' by a suitable scalar we may assume that $b(u, v') = 1$. Now for any $\alpha \in k$, $b(v' - \alpha u, v' - \alpha u) = q(v') - 2\alpha$ and we can choose α so that the right-hand side is 0; then u and $v = v' - \alpha u$ form a hyperbolic pair for V.

(b) $\Rightarrow$ (c). If u, v is a hyperbolic pair, then $e = u + v/2$, $f = u - v/2$ is an orthogonal basis and $q(e) = -q(f) = 1$.

(c) $\Rightarrow$ (d) is clear. To prove (d) $\Rightarrow$ (a) we have by Theorem 8.2.2, $V \cong \langle a, b\rangle$, where $-ab$ is a square, say $-ab = c^2$. Then $-a^2 b = c^2 a$ and now $ax^2 + by^2 = 0$ for $x = c$, $y = a$, so V is isotropic. ■

From (b) or (c) it is clear that a hyperbolic plane is unique up to isometry. In studying the monoid $M(k)$ we shall regard orthogonal sums of hyperbolic planes as trivial. The reason for this is clear when one considers the Clifford algebra: hyperbolic planes are the spaces whose Clifford algebras are full matrix rings. Orthogonal sums of hyperbolic planes, called *hyperbolic spaces*, may be defined more simply as split spaces. An inner product space V is said to be *split* if it is regular and has a subspace U (the splitting space) such that $U^\perp = U$. It is clear that an orthogonal sum of hyperbolic planes is split: if

$$V \cong H_1 \perp \ldots \perp H_r, \tag{8.5.2}$$

where each H_i is a hyperbolic plane, with hyperbolic pair u_i, v_i say, write U for the subspace spanned by $u_1, \ldots, u_r$. Then it is clear that U is totally isotropic, hence $U \subseteq U^\perp$. If $x = \sum \alpha_i u_i + \sum \beta_i v_i \in U^\perp$, then $0 = b(x, u_i) = \beta_i$, hence $x = \sum \alpha_i u_i \in U$. This shows that $U^\perp \subseteq U$, so that U splits V.

The fact that conversely, every split space is of the form (8.5.2), follows from the next result, which is slightly more general.

Theorem 8.5.3. *Any regular inner product space V (in characteristic 2) can be written in the form*

$$V \cong H_1 \perp \ldots \perp H_r \perp U, \tag{8.5.3}$$

where $H_1, \ldots, H_r$ are hyperbolic planes and U is an anisotropic space. Here r is uniquely determined as the maximal dimension of a totally isotropic subspace of V and U is determined up to isometry.

The decomposition (8.5.3) is called the *Witt decomposition* of V, and the integer r is called the *Witt index* of V, while U is called the *anisotropic part* of V. Thus the form $\langle 1^r, (-1)^s\rangle$ has Witt index $\min(r, s)$.

Proof. Let W be any totally isotropic subspace of V, with basis $u_1, \ldots, u_s$ and choose $v_1 \in V$ such that $b(u_1, v_1) \neq 0$, $b(u_i, v_1) = 0$ for $i > 1$; since V is regular, this is always possible. The subspace H spanned by u_1, v_1 is regular and it contains the isotropic vector u_1, hence it is a hyperbolic plane, and by Lemma 8.2.1 we have the decomposition

$$V = H_1 \perp V',$$

where $V' = H_1^{\perp}$. Now V' contains $u_2, \ldots, u_s$; hence by induction on s there exist $v_2, \ldots, v_s \in V'$ such that u_i, v_i span a hyperbolic plane H_i and we have a decomposition

$$V \cong H_1 \perp \ldots \perp H_s \perp U. \tag{8.5.4}$$

If U is isotropic, we can split off more hyperbolic planes, and the process ends when we reach a component U which is anisotropic. Thus we always have a decomposition (8.5.3) where U is anisotropic, and by construction $r \geq s$. It is clear that V given by (8.5.3) contains a totally isotropic subspace of dimension r, hence r is uniquely determined as the maximal dimension of a totally isotropic subspace of V. Finally, U is unique up to isometry by the Witt cancellation theorem (Theorem 8.5.1). ■

If V itself is split, by a subspace W of dimension r, and we take a decomposition (8.5.3) of V, then since $W^{\perp} = W$, we have $U = 0$ and hence we obtain

Corollary 8.5.4. *Any split space in characteristic $\neq 2$ is an orthogonal sum of hyperbolic planes.* ■

For any field k of characteristic not 2 we now define the *Witt group* $W(k)$ as follows. In the monoid $M(k)$ let us identify any two classes $[V], [V']$ whose spaces have anisotropic parts that are isometric. Writing again $[V]$ for the class of V, we now have

$$[V] = [V'] \Leftrightarrow V \perp mH \cong V' \perp nH,$$

where nH stands for the orthogonal sum of n hyperbolic planes. We thus obtain a monoid $W(k)$, which is in fact a group. For if V is an n-dimensional space with regular form q, and V' is a space with form $-q$, then we can by Theorem 8.2.2 write $V \cong \langle a_1, \ldots, a_n \rangle$, $V' \cong \langle -a_1, \ldots, -a_n \rangle$; therefore $V \perp V' \cong \langle a_1, -a_1, \ldots, a_n, -a_n \rangle$. Now $\langle a_i, -a_i \rangle$ is regular isotropic, hence a hyperbolic plane, and so $V \perp V' \cong nH$. It follows that $[V]$ has the inverse $-[V] = [V']$, so $W(k)$ is indeed a group, the *Witt group*. Clearly it is an invariant of the field k. For example, for an algebraically closed field the Witt group is of order 2; for the real field $\mathbf{R}$, quadratic forms can be classified by $n - 2\nu$, where ν is the index, hence we have $\mathrm{W}(\mathbf{R}) \cong \mathbf{Z}$. In Section 8.9 we shall see that a ring structure can be defined on $W(k)$, giving the *Witt ring* of k.

By examining the proof of Theorem 8.5.3 we obtain a useful result on extending partial isometries:

Theorem 8.5.5 (Witt's extension theorem). *Let V be a regular inner product space (in characteristic $\neq 2$) and W be a subspace of V. Then any isometric mapping $\theta : W \to V$ can be extended to an orthogonal transformation of V.*

Proof. Let W_0 be the radical of W and write $W = W_0 \perp X$. Here X is regular, so by Lemma 8.2.1, $V = X \perp X^{\perp}$. Now $X^{\perp} \supseteq W_0$ and as in the proof of Theorem 8.5.3 we can write $X^{\perp} = Z \perp Y$, where Z is a regular subspace split by W_0. Since $X \cong X^{\theta}$, we have, by Theorem 8.5.1, $X^{\perp} \cong (X^{\theta})^{\perp}$; but $(X^{\theta})^{\perp} \supseteq W_0^{\theta}$ because W_0 is orthogonal to

X. Therefore we can write $(X^\theta)^\perp = Z' \perp Y'$, where Z' is a regular subspace split by W_0^θ. Let $u_1, \ldots, u_s$ be a basis of W_0 and write $Z = H_1 \perp \ldots \perp H_s$, $Z' = H_1' \perp \ldots \perp H_s'$, where H_i has the hyperbolic pair (u_i, v_i) and H_i' has the hyperbolic pair (u_i^θ, v_i'). The mapping $u_i \mapsto u_i^\theta$, $v_i \mapsto v_i'$ is an isometry $Z \to Z'$ whose restriction to W_0 agrees with θ; thus we have extended θ to the regular space $U = X \perp Z$. Again we have $U \cong U^\theta$, so $U^\perp \cong (U^\theta)^\perp$ and this latter isomorphism gives the desired extension of θ to all of $V = U \perp U^\perp$. ■

Exercises

1. Show that a space V is split iff it is regular and has a linear mapping α into itself such that $q(x) + q(x\alpha) = 0$ for all $x \in V$.
2. Show that an orthogonal transformation of a split space which is the identity on a maximal totally isotropic subspace is a rotation.
3. Show that a hyperbolic plane has exactly two isotropic lines (i.e. every isotropic vector is proportional to one of exactly two vectors). Show also that each vector of a hyperbolic pair uniquely determines the other.
4. Show that an orthogonal transformation of a hyperbolic plane is a symmetry iff it interchanges the two isotropic lines.
5. For any hyperbolic plane H show that $\mathbf{SO}(H) \cong k^\times$. Describe the orthogonal group of H. What is the orthogonal group of $H \perp \langle 0 \rangle$?
6. Prove Theorem 8.5.5 in the special case where W (as well as V) is regular. When can the orthogonal transformation in Theorem 8.5.5 be chosen to be a rotation?
7. Give a direct proof (not using Theorem 8.5.3) that any two split spaces of the same dimension are isomorphic.
8. Give a proof of Theorem 8.5.1 using Theorem 8.5.5.

8.6 Ordered Fields

In the further theory of quadratic forms reality conditions play an important role, and it is useful to develop these ideas in a more general form for ordered fields.

By an *ordered ring* we understand a non-trivial ring R (not necessarily commutative) with a strict total ordering '$>$' compatible with the ring operations, in the sense that the following rules hold:

O.1 $x > x', y > y' \Rightarrow x + y > x' + y'$,
O.2 $x > 0, y > 0 \Rightarrow xy > 0$.

If in **O.2** we replace x, y by $x - x'$, $y - y'$ and rearrange the result, using **O.1**, we obtain the more general rule

O.2′ $x > x', y > y' \Rightarrow xy + x'y' > xy' + x'y$.

In discussing the ordering on R we shall, as usual, write '$x \geq y$' to mean '$x > y$ or $x = y$' and use $\leq, <$ for the opposite ordering; x is called *positive* if $x > 0$ and *negative* if $x < 0$.

For any element a of an ordered ring we define the *absolute value* by

$$|a| = \begin{cases} a & \text{if } a \geq 0, \\ -a & \text{if } a \leq 0. \end{cases}$$

Clearly $|a| \geq 0$ with equality iff $a = 0$, and we have the *triangle inequality*

$$|a + b| \leq |a| + |b|, \tag{8.6.1}$$

as well as the product rule

$$|ab| = |a|.|b|. \tag{8.6.2}$$

Of these, (8.6.2) is an immediate consequence of the formula $|a| = \sqrt{a^2}$. To prove (8.6.1) we remark that this clearly holds (with equality) when a, b have the same sign. If $a \geq 0 \geq b$ say, then $a + b \leq a \leq a - b = |a| + |b|$ and $-a - b \leq -b \leq a - b = |a| + |b|$, hence (8.6.1) follows in this case. By symmetry (8.6.1) also holds when $b \geq 0 \geq a$.

An ordered ring can also be described by its order set; this is a subset P containing 1 but not 0 and containing with any x, y also $x + y$, xy. For example, in an ordered ring R, the set consisting of all strictly positive elements

$$P = \{x \in R | x > 0\}$$

is an order set, called the *positive order set* of R. Moreover, since R is totally ordered, the ordering defines a trichotomy on R:

$$\text{each } x \in R \text{ lies in just one of the sets } \{0\}, P, -P = \{-x | x \in P\}. \tag{8.6.3}$$

Conversely, every order set satisfying (8.6.3) defines an ordering on R: we put $x > y$ iff $x - y \in P$. We note that any ordered ring, and more generally any ring with an order set, is non-trivial, i.e. $1 \neq 0$.

Proposition 8.6.1. *Any ordered ring is an integral domain, and the square of any non-zero element is positive.*

Proof. Let R be an ordered ring; then $1 \neq 0$ by definition. Given $x, y \in R$, if $x, y > 0$, then $xy > 0$. Similarly, if $x, y < 0$, then $-x, -y > 0$ and so $xy = (-x)(-y) > 0$. There remains the case when x, y have opposite signs, say $x > 0 > y$. Then $x, -y > 0$, hence $-xy = x(-y) > 0$ and so $xy < 0$; similarly if $x < 0 < y$, then $xy < 0$. In each case $xy \neq 0$, hence R is an integral domain. In particular, when $x = y$, the last two cases cannot occur and $x^2 > 0$. ∎

For the rest of this section all our rings will be commutative. Thus an ordered commutative ring is an integral domain, and hence has a field of fractions K. We claim that there is just one way of ordering K so as to extend the ordering of R, namely by the rule:

$$\text{for any } a, b \in R, \text{ where } b \neq 0, a/b > 0 \text{ in } K \text{ iff } ab > 0. \tag{8.6.4}$$

In the first place, since $a/b.b^2 = ab$ and $b^2 > 0$, any ordering of K which extends that on R must satisfy (8.6.4). It remains to show that (8.6.4) defines an ordering. It is

well-defined, for if $a/b = a'/b'$, then $ab' = ba'$ and hence $abb'^2 = a'b^2b'$. It follows that $ab > 0 \Leftrightarrow abb'^2 > 0 \Leftrightarrow a'b^2b' > 0 \Leftrightarrow a'b' > 0$. Further, it clearly defines a trichotomy on K as in (8.6.3), and if $a_1b_1 > 0$, $a_2b_2 > 0$, then $(a_1b_2 + a_2b_1)b_1b_2 = a_1b_1b_2^2 + a_2b_2b_1^2 > 0$, $a_1a_2b_1b_2 = a_1b_1a_2b_2 > 0$. This shows that if the fractions $f_i = a_i/b_i$ satisfy $f_i > 0$ $(i = 1, 2)$, then $f_1 + f_2 > 0, f_1f_2 > 0$.

We also note that $1, 1 + 1, \ldots$ are all > 0, hence R and with it K must be of characteristic 0. The result may be summed up as

Theorem 8.6.2. *Any ordered commutative ring is an integral domain of characteristic 0. The ordering can be extended to an ordering of its field of fractions in just one way.* ■

As an example consider the integers **Z**; in any ordering $1, 1 + 1, \ldots$ must all be positive, and since these numbers and their negatives, with 0, exhaust **Z**, there is only one way of ordering **Z**, namely the familiar 'natural' ordering of the integers. By Theorem 8.6.2 the same applies to **Q** and we have

Corollary 8.6.3. *There is only one way of making* **Q** into an ordered field, namely by the usual ordering. ■

The real numbers **R** also admit only one ordering, but for a different reason: every square must be positive (or 0), and since the squares, their negatives and 0 exhaust **R**, no other ordering is possible apart from the usual one. We shall return to this point in Section 8.8. An example of a field with different orderings is $\mathbf{Q}(\alpha)$, where α is a root of $x^2 = 2$. There are two ways of embedding this field into **R**, mapping α to $\sqrt{2}$ or to $-\sqrt{2}$, and correspondingly two ways of ordering it.

Two rings R, R' are said to be *order-isomorphic* if there is a ring isomorphism $x \mapsto x'$ from R to R' which preserves the ordering, i.e. $x > 0 \Leftrightarrow x' > 0$; clearly it is enough to require that $x > 0 \Rightarrow x' > 0$. For example, the prime subfield of any ordered field K is order-isomorphic to **Q**; we also say that there is an *order-embedding* of **Q** in K.

Exercises

1. Show that in any ordered field, $a > b > 0$ implies $b^{-1} > a^{-1}$.
2. Let K be an ordered field. Given $a, b \in K$, show that if $a < b$, then there exists $c \in K$ such that $a < c < b$. (This property, that between any two elements of K a third can be found, is expressed by saying that K is *dense in itself.*)
3. Let K be an ordered field. Show that the polynomial ring $K[x]$ can be ordered by taking as positive order set the set of all polynomials with positive highest coefficient; hence obtain an ordering of the rational function field $K(x)$. Compare this with the ordering of $K(x)$ obtained by taking the polynomials with positive coefficient of the lowest term as positive order set.
4. Show that the mapping $\mathbf{Q}(x) \to \mathbf{R}$ defined by $x \mapsto \pi$ provides an ordering of $\mathbf{Q}(x)$; compare this with the orderings obtained in Exercise 3.

5. Show that a skew field D can be ordered iff $D^\times$ has a subgroup of index 2 closed under addition.

8.7 The Field of Real Numbers

There are essentially two ways of constructing the real numbers: (a) Dedekind's method of completion by cuts and (b) Cantor's method of sequences. Of these (a) is logically the simpler, while (b) corresponds more closely to the practical way of approximating real numbers. For us (b) has the further advantage of making it easier to define the algebraic operations; we shall therefore concentrate on (b). Later on, when we come to study valuations (in Chapter 9), we shall find that the same construction can be used for the p-adic numbers.

Let K be an ordered field. By a *null sequence* in K we understand a sequence $\{a_n\}(a_n \in K)$ such that for any $\varepsilon > 0$ in K there exists $n_0 \in \mathbf{N}$ such that $|a_n| < \varepsilon$ for all $n > n_0$. As usual in analysis we also say 'a_n converges to 0 as n tends to ∞' and write '$a_n \to 0$' or 'lim $a_n = 0$'. A sequence $\{a_n\}$ is said to *converge* to $a \in K$, $a_n \to a$, if $a_n - a \to 0$, and we shall call a the *limit* of the sequence; it is clear that the limit, if it exists, is unique.

The following is a familiar necessary condition for the convergence of a sequence, which does not mention the value of the limit:

Cauchy's Condition. *If a sequence $\{a_n\}$ converges to a limit, then the double sequence $\{c_{mn}\}$, where $c_{mn} = a_m - a_n$, converges to 0.*

In detail this means: given $\varepsilon > 0$, there exists n_0 such that $|a_m - a_n| < \varepsilon$ for all $m, n > n_0$. This condition follows easily from the triangle inequality.

We shall call $\{a_n\}$ a *Cauchy sequence* if $|a_m - a_n| \to 0$; thus Cauchy's condition states that every convergent sequence is a Cauchy sequence. The converse need not hold, e.g. the rational numbers 1, 1.4, 1.41, 1.414, ... obtained in calculating successive approximations to $\sqrt{2}$ form a Cauchy sequence, but they are not convergent within **Q**. The case where the converse holds is an important property of ordered fields and we make the

Definition. An ordered field is said to be *complete* if every Cauchy sequence of its elements is convergent.

Cauchy's Convergence Principle just states that **R** with the usual ordering is complete. Although an ordered field need not be complete, it always has a completion, constructed in a canonical fashion, as in the process that leads from **Q** to **R**. If K is any ordered field, a subfield K' is called *dense* in K if between any two elements of K there is an element of K'. An embedding as a dense subfield is called a *dense embedding*.

Theorem 8.7.1. *Let K be an ordered field. Then there exists a complete ordered field $\tilde{K}$ and a dense order-embedding $\lambda : K \to \tilde{K}$ such that to each order-embedding $f : K \to L$*

into a complete ordered field L there corresponds a unique order-embedding $f' : \tilde{K} \to L$ such that $f = \lambda f'$.

This is just the universal property of completions; to prove it in detail is somewhat lengthy (although not difficult). We therefore outline the main steps and leave some of the verifications to the reader.

Let R be the set of all Cauchy sequences in K; this is a subset of $K^{\mathbf{N}}$, the set of all sequences in K. Now $K^{\mathbf{N}}$ is a ring, the direct power of K, and R is clearly a subring, the operations being carried out componentwise: $\{a_n\} \pm \{b_n\} = \{a_n \pm b_n\}$, $\{a_n\}\{b_n\} = \{a_n b_n\}$. Since every constant sequence is clearly a Cauchy sequence, we can regard K as a subfield of R by mapping $c \in K$ to the constant sequence $c, c, \ldots$. Consider the set $\mathfrak{n}$ of all null sequences in R; we claim that $\mathfrak{n}$ is an ideal in R. Let us show e.g. if $\{a_n\} \in R$, $\{b_n\} \in \mathfrak{n}$, then $\{a_n b_n\} \in \mathfrak{n}$. Since $\{a_n\}$ is a Cauchy sequence, there exists n_0 such that $|a_m - a_n| < 1$ for all $m, n > n_0$, hence $|a_m| < |a_{n_0}| + 1$ for $m > n_0$. It follows that $|a_m| < M$, where

$$M = \max\{|a_1|, |a_2|, \ldots, |a_{n_0 - 1}|, |a_{n_0}| + 1\};$$

thus every Cauchy sequence is bounded. Now $b_n \to 0$, hence for any $\varepsilon > 0$, $|b_n| < \varepsilon/M$ for $n > n_1$, so $|a_n b_n| < \varepsilon$ for $n > n_1$, i.e. $a_n b_n \to 0$. This proves that $\{a_n b_n\} \in \mathfrak{n}$, and the remaining properties of $\mathfrak{n}$ are established similarly.

We claim that $\mathfrak{n}$ is a maximal ideal in R. For $\mathfrak{n}$ is proper, e.g. the constant sequence $1, 1, \ldots$ is a Cauchy sequence but not a null sequence. Moreover, if $\{a_n\}$ is a Cauchy sequence which is not null, then it has an inverse (mod $\mathfrak{n}$). To prove this fact, let $\{a_n\}$ be not null; then by definition there exists $p \in K$, $p > 0$, such that to each n there corresponds $n' > n$ with $|a_{n'}| \geq p$. Since $\{a_n\}$ is Cauchy, there exists n_0 such that $|a_m - a_n| < p/2$ for all $m, n > n_0$. Choose any $n > n_0$ and take $n' > n$ as before; then

$$|a_n| \geq |a_{n'}| - |a_n - a_{n'}| > p/2.$$

Therefore the sequence a_n^{-1} is bounded for $n > n_0$. If we define

$$b_n = \begin{cases} 1 & \text{for } n \leq n_0, \\ a_n^{-1} & \text{for } n > n_0, \end{cases}$$

then $\{b_n\}$ is a Cauchy sequence and $\{a_n b_n\}$ converges to 1, i.e. $\{a_n\}.\{b_n\} \equiv 1$ (mod $\mathfrak{n}$). This shows $\mathfrak{n}$ to be maximal.

Hence $R/\mathfrak{n}$ is a field, which we denote by $\tilde{K}$. We write λ for the natural homomorphism $K \to R \to R/\mathfrak{n}$ obtained by mapping $a \in K$ to the residue class of the constant sequence $a, a, \ldots$. Like every homomorphism between fields, this is an embedding, and we shall identify K with its image in $\tilde{K}$.

It is clear how to extend the ordering to $\tilde{K}$: given $\alpha \in \tilde{K}$, represented by a Cauchy sequence $\{a_n\}$, either this is a null sequence or there exists $\varepsilon > 0$ and n_0 such that $a_n > \varepsilon$ for $n > n_0$, or there exists n_1 such that $a_n < -\varepsilon$ for $n > n_1$. Moreover, all Cauchy sequences representing α have the same property and accordingly we set $\alpha = 0$, $\alpha > 0$ or $\alpha < 0$. We leave the reader to verify that K is dense in $\tilde{K}$. Now given a Cauchy sequence $\{\alpha_n\}$ in $\tilde{K}$, either α_n is constant from some n onwards,

then it is a Cauchy sequence in K and so it converges to a limit in $\tilde{K}$; or $\{\alpha_n\}$ contains infinitely many distinct terms, in that case, omitting repetitions, we may assume all the α_n to be distinct. For each n we can choose $a_n \in K$ to lie between α_n and α_{n+1}; the resulting sequence $\{a_n\}$ is easily seen to be a Cauchy sequence in K which has a limit α in $\tilde{K}$. Clearly $\lim \alpha_n = \lim a_n = \alpha$, and this shows $\tilde{K}$ to be complete.

Finally let $f : K \to L$ be an order-embedding in a complete field L. Any element α of $\tilde{K}$ is obtained as the limit of a Cauchy sequence $\{a_n\}$ in K; it is easily seen that $\{a_n f\}$ is a Cauchy sequence in L and so has a limit b, say. Any other Cauchy sequence tending to α differs from $\{a_n\}$ by a null sequence, hence its image in L again tends to b; therefore b depends only on α and we may put $\alpha f' = b$. Now the reader may verify without difficulty that f' is an order-embedding such that $f = \lambda f'$ and that it is the only mapping satisfying this equation. ∎

The field $\tilde{K}$ whose existence is proved in Theorem 8.7.1 is called the *completion* of K; since it is obtained as the solution of a universal problem, it is determined up to order-isomorphism by the properties listed in Theorem 8.7.1. For example, if we take $K = \mathbf{Q}$, then $\tilde{K} = \mathbf{R}$.

The real numbers, as ordered field, have other important properties which can also be used to characterize them. Below we briefly look at two of them.

An ordered field K is said to be *Archimedean* if for any $a \in K$ there exists $n \in \mathbf{N}$ such that $n > a$. We note that this is so precisely when $\mathbf{Q}$ is dense in K. For if $\mathbf{Q}$ is dense in K and $a \leq 0$ in K, then $1 > a$. If $a > 0$ in K, then $0 < (2a)^{-1} < a^{-1}$, hence there exists $\alpha \in \mathbf{Q}$ such that $(2a)^{-1} < \alpha < a^{-1}$, so if $\alpha = m/n$, then $a < n/m < n$; this shows K to be Archimedean. Conversely, assume that K is Archimedean and let $0 < a < b$ in K. Then $(b-a)^{-1} < n$ for some $n \in \mathbf{N}$, hence $0 < 1/n < b - a$; further, $na < m$ for some $m \in \mathbf{N}$. With the least such m we have $(m-1)/n \leq a$, hence $m/n \leq a + 1/n < a + (b-a) = b$, and so we have $a < m/n < b$. If $a < b < 0$, we can find $\alpha \in \mathbf{Q}$ between $-b$ and $-a$ and so $a < -\alpha < b$, while for $a < 0 < b$ we can take $\alpha = 0$. Thus we obtain

Proposition 8.7.2. *An ordered field K is Archimedean if and only if the prime subfield* $\mathbf{Q}$ is dense in K. ∎

In particular this result shows $\mathbf{R}$ to be Archimedean. If K is an ordered subfield of $\mathbf{R}$, then $\mathbf{Q}$ is again dense in K, so K is Archimedean. Conversely, if K is any Archimedean ordered field, then $\mathbf{Q}$ is dense in K, by Proposition 8.7.2, and now the proof of Theorem 8.7.1 shows that there is an order-embedding of K in $\mathbf{R}$. Thus we have proved

Theorem 8.7.3. *Any ordered subfield of* $\mathbf{R}$ *is Archimedean, and conversely, any Archimedean ordered field is order-isomorphic to a subfield of* $\mathbf{R}$. ∎

Clearly no proper subfield of $\mathbf{R}$ is complete; we therefore have

Corollary 8.7.4. *Any complete Archimedean ordered field is isomorphic to* **R**. ■

A second important property of the real numbers is the upper bound property. In any partially ordered set S the supremum of any subset X, when it exists, is unique. It may or may not be a member of X, e.g. $\{0, -1, -2, \dots\}$ and $\{-1, -1/2, -1/3, \dots\}$ both have the supremum 0, which belongs to the first set but not the second. A set is said to be *bounded above* if it has an upper bound.

In terms of upper bounds we can characterize the real numbers by the

Upper Bound Property. *Every non-empty subset that is bounded above has a least upper bound.*

We observe that any field possessing the upper bound property is Archimedean. For if $\alpha \in K$ but $n \leq \alpha$ for all $n \in \mathbf{N}$, then the set $\mathbf{N} = \{1, 2, \dots\}$ is bounded above and so has a supremum γ, say. It follows that $\gamma - 1 < \gamma$ hence $\gamma - 1 < n$ for some $n \in \mathbf{N}$ and so $\gamma < n + 1$, which contradicts the definition of γ. Thus $n > \alpha$ for some $n \in \mathbf{N}$ and so K is Archimedean, as asserted.

We can now give a characterization of the real numbers in terms of the upper bound property.

Theorem 8.7.5. *The field* **R** *of real numbers possesses the upper bound property, and any ordered field with the upper bound property is order-isomorphic to* **R**.

Proof. Let X be any set of real numbers, bounded above. If X has a greatest element α, then $\alpha = \sup X$. Otherwise let Y be the set of all its upper bounds and X' be the complement of Y in **R**. Then $X \subseteq X'$ and the members of X' may be characterized as the numbers that are exceeded by some member of X. Moreover, every number in X' is less than every number in Y. Since they are complementary, we can, for each $n = 1, 2, \dots$ find $a_{2n-1} \in X'$, $a_{2n} \in Y$ such that

$$|a_{2n-1} - a_{2n}| < 1/n, \quad a_{2n-3} < a_{2n-1} < a_{2n} < a_{2n-2}.$$

Clearly $\{a_n\}$ is a Cauchy sequence; let α be its limit. If $n \in \mathbf{N}$, then $\alpha + 1/n$ is an upper bound for X, for it exceeds some a_{2r} and so is in Y, while $\alpha - 1/n$ is less than some a_{2r-1} and so lies in X'. It follows that $\alpha = \sup X$, so the upper bound property holds in **R**.

Now let K be any ordered field possessing the upper bound property. By the remark preceding the theorem we see that K is Archimedean, and hence, by Theorem 8.7.3, order-isomorphic to a subfield of **R**. We may therefore identify K with a subfield of **R**; clearly K contains **Q** as a subfield. But every element α of **R** may be expressed as a supremum of some subset A of **Q**, and A also has a supremum α' say in K, by the upper bound property. Clearly $\alpha \leq \alpha'$; if the inequality were strict, we could find $a \in \mathbf{Q}$ such that $\alpha < a < \alpha'$, and this would contradict the fact that $\alpha' = \sup A$ in K. Hence $\alpha' = \alpha$, and it follows that $K = \mathbf{R}$. ■

As a consequence we have the *intermediate value property* of real numbers:

Proposition 8.7.6. *Given a polynomial f in $\mathbf{R}[x]$ and $\alpha, \beta \in \mathbf{R}$ such that $\alpha < \beta$, if $f(\alpha)f(\beta) < 0$, then there exists $\gamma \in \mathbf{R}$ such that $\alpha < \gamma < \beta$ and $f(\gamma) = 0$.*

Proof. By hypothesis $f(\alpha)$, $f(\beta)$ have opposite signs, and we may assume that $f(\alpha) < 0 < f(\beta)$, replacing f by $-f$ if necessary. Let A be the set of real numbers t such that $t < \beta$ and $f(t) < 0$. This set contains α, is bounded above (by β) and so has a supremum γ say, where $\alpha \leq \gamma < \beta$. This means that for $\gamma < t < \beta$, $f(t) \geq 0$, but if $t < \gamma$, we can find t_1 such that $t < t_1 < \gamma$ and $f(t_1) < 0$. Now express $f(\gamma + x)$ as a polynomial in x:

$$f(\gamma + x) = a_0 x^n + \ldots + a_n, \quad \text{where } a_n = f(\gamma).$$

If $a_n \neq 0$, then the right-hand side has the same sign as a_n for sufficiently small x. But we have seen that $f(t)$ changes sign in each interval about γ; hence $a_n = 0$, i.e. $f(\gamma) = 0$. ■

The proof shows incidentally that f is continuous; this can of course also be established directly.

Exercises

1. Show that an Archimedean ordered field has no order-preserving automorphism other than the identity.
2. Let K be an ordered field, algebraic over a subfield E. Show that if E is Archimedean, then so is K.
3. Fill in the details omitted in the proof of Theorem 8.7.1.
4. Adapt the proof of Theorem 8.7.3 to prove that every Archimedean ordered skew field is commutative. (Hint. Conjugation is an order-preserving automorphism fixing $\mathbf{Q}$.)
5. Show that if an ordered field K has an Archimedean subfield dense in K, then K is itself Archimedean.
6. Show that the ordered fields constructed in Exercise 3 of Section 8.6 are not Archimedean. Use the method of Exercise 4 of Section 8.6 to find uncountably many distinct Archimedean orderings on $\mathbf{Q}(x)$.

8.8 Formally Real Fields

After a brief look at general ordered fields in Section 8.6 we developed the theory of real fields in Section 8.7, but there is an algebraic theory of ordered fields that puts the special role of real fields clearly in evidence. This theory goes back to a classic series of papers by Emil Artin and Otto Schreier in the 1920s, but there was further development in the later 20th century, which helped to simplify the basic theory.

We recall from Section 8.6 that an ordered ring may be described by its positive order set P; let us define a *cone* as a subset containing all squares but not -1 and

closed under sums and products. If R is an ordered ring with order set P_0, then $P = P_0 \cup \{0\}$ satisfies

$$R = P \cup -P, P \cap -P = \{0\}. \tag{8.8.1}$$

In an integral domain R, any cone P satisfying (8.8.1), with 0 removed, is the positive order set of an ordering on R. For a field the second condition in (8.8.1) is redundant, for if $P \cap -P$ contains a non-zero element c, then $c, -c \in P$, hence $-1 = c.(-c).(c^{-1})^2 \in P$.

To study the possible orderings of a field it is helpful to consider the elements that are positive under every ordering. For any field K let us define its *core* as the set C of sums of squares

$$C = C(K) = \left\{ \sum a_i^2 | a_i \in K \right\}.$$

We observe that C is closed under addition, multiplication and inversion. For addition this is clear; next if $a = \sum a_i^2$, $b = \sum b_j^2$, then $ab = \sum_{ij} (a_i b_j)^2$ and $a^{-1} = a(a^{-1})^2$. By definition C contains every square in K, so it is a cone precisely when $-1 \notin C$.

If K is an ordered field, then clearly all non-zero elements of $C(K)$ are positive, in particular, $-1 \notin C(K)$. A field is said to be *formally real* if its core does not contain -1, and so it is a cone. By what has been said, every ordered field is formally real. Further, every formally real field is of characteristic 0, for C contains $1, 1+1, \ldots$ and these must all be different from -1. The following rephrasing of the definition is often useful:

Proposition 8.8.1. *A field is formally real if and only if -1 cannot be written as a sum of squares.* ■

Our first task is to show that every formally real field can be ordered. It is no harder to prove a relative version, using a relative notion of core.

Let K be any field with a subfield F and suppose that F is ordered. Then the *F-core* of K is the set $\Gamma_F(K)$ of all sums of squares with positive coefficients from F:

$$\Gamma_F(K) = \left\{ \sum a_i \lambda_i^2 | \lambda_i \in K, a_i \in F, a_i > 0 \right\}. \tag{8.8.2}$$

If K is any field of characteristic 0, it contains $\mathbf{Q}$ as prime subfield and this is ordered; now the $\mathbf{Q}$-core of K is just the core defined earlier. As in that case we can show that $\Gamma_F(K)$ is closed under addition, multiplication and inversion and contains all squares. Hence $\Gamma_F(K)$ is a cone iff it does not contain -1. If K has an ordering extending that of F, then $\Gamma_F(K)$ clearly does not contain -1 and so is a cone. We shall show conversely that if $\Gamma_F(K)$ is a cone, then the ordering of F can be extended to K. We single out two properties of cones:

Lemma 8.8.2. *Let K be any field.*

(i) *If P is a cone on K and $a, b \in K$ are such that $ab \in P$, then either $P + aP$ or $P - bP$ is a cone.*

(ii) *Every cone on K is contained in a maximal cone and any maximal cone P satisfies $P \cup -P = K$.*

Proof. (i) Clearly $P + aP$, $P - bP$ contain P and they admit addition, multiplication and inversion. For example, if $x = x_1 - bx_2$, $y = y_1 - by_2$, then $xy = x_1y_1 + b^2x_2y_2 - b(x_1y_2 + x_2y_1)$, and similarly for the other cases. Hence if neither $P + aP$, $P - bP$ is a cone, then both contain -1, say $x + au = -1 = y - bv$; it follows that $-abuv = 1 + x + y + xy$ and so $-1 = x + y + xy + abuv \in P$, a contradiction.

(ii) The existence of a maximal cone containing a given cone follows by Zorn's lemma. Now let P be a maximal cone and take $a \in K^\times$. Then $a^2 \in P$, hence by (i), $P + aP$ or $P - aP$ is a cone, say $P + aP$; but $P + aP \supseteq P$, so by maximality $P = P + aP$ and thus $a \in P$. ∎

Now it is an easy matter to establish a criterion for extendibility:

Theorem 8.8.3. *Let K be a field with an ordered subfield F. Then the ordering of F can be extended to an ordering of K if and only if the F-core $\Gamma_F(K)$ is a cone, i.e. $-1 \notin \Gamma_F(K)$.*

Proof. We have seen the necessity of this condition. Conversely, when it holds, then by Lemma 8.8.2(ii), $\Gamma_F(K)$ is contained in a maximal cone P and $P \cup -P = K$. It follows that P defines an ordering on K and this ordering extends the ordering on F because P contains $\Gamma_F(K)$, which by construction contains the positive cone of F. ∎

If we apply this result to a field K of characteristic 0, taking F to be $\mathbf{Q}$, we see that K can be ordered iff it is formally real. More precisely we have

Corollary 8.8.4. *Let K be a formally real field and P be any cone on K. Then P is the intersection of all maximal cones containing P. In particular, K can be ordered. Thus a field can be ordered if and only if it is formally real.*

Proof. It is clear that the intersection of maximal cones containing P is again a cone containing P, so to prove the first part we have to find, for any $a \notin P$, a maximal cone containing P but not a. We note that $P - aP$ is a cone, for if not, then $-1 = x - ay$ for some $x, y \in P$, and so $ay = x + 1$, hence $a = (x + 1)y^{-1} \in P$, which is a contradiction. By Lemma 8.8.2 there is a maximal cone Q containing $P - aP$; hence $Q \supseteq P$, but $a \notin Q$ because $-a \in Q$. Now it follows that the intersection of all maximal cones containing P contains no element not in P and so equals P. The second part follows because a formally real field always has a cone, namely its core. ∎

From Corollary 8.8.4 we see that in any formally real field K the core is just the intersection of all maximal cones. In other words, an element of K is a sum of squares

iff it is positive under every ordering of K. Such an element is also called *totally positive.*

We next investigate which algebraic extensions of ordered fields allow an extension of the ordering. A field is said to be *real closed* if it is formally real but has no algebraic extensions that are formally real; thus a real closed field is a maximal formally real extension field in its algebraic closure. By the *real closure* of a formally real field K we understand a maximal formally real extension of K in its algebraic closure. For example, the field of all real algebraic numbers is real closed and it is the real closure of $\mathbf{Q}$. It is easy to see that such a real closure always exists.

Theorem 8.8.5. *Every formally real field has a real closure.*

Proof. Let K be formally real; by Corollary 8.8.4 we may take it to be ordered. Consider the family $\mathscr{F}$ of all ordered extensions of K in its algebraic closure; $\mathscr{F}$ itself is partially ordered by order-preserving inclusions, thus for $E_1, E_2 \in \mathscr{F}$ we write $E_1 \leq E_2$ if E_2 is an extension of E_1 and the ordering of E_1 is induced by that of E_2. Clearly $\mathscr{F}$ is inductive, and so by Zorn's lemma it has a maximal member E. This is the required real closure of K. ■

Real closed fields can be described more explicitly; to do so we need to look at two particular cases of algebraic extensions.

Lemma 8.8.6. *Let K be an ordered field and L be a finite algebraic extension of K. Then the ordering of K can be extended to L provided that either $[L : K]$ is odd, or $L = K(\sqrt{a})$, where $a \in K, a > 0$.*

Proof. Suppose first that L/K is an extension of odd degree n. We shall use induction on n. Since $\operatorname{char} K = 0$, the extension is simple, by the theorem of the primitive element (Theorem 7.9.2), say $L = K(\alpha)$, where the minimal polynomial p of α over K has degree n. For $n = 1$ the result clearly holds, so assume that $n > 1$. By Theorem 8.8.3 we have to show that $\Gamma_K(L)$, the K-core of L, is a cone. If this is not so, it must contain -1, so we have

$$-1 = \sum \lambda_i f_i(\alpha)^2, \tag{8.8.3}$$

where $\lambda_i > 0$ in K and each f_i is a polynomial over K of degree at most $n-1$. Since p is the minimal polynomial of α, we have the identity

$$-1 = \sum \lambda_i f_i(x)^2 + p(x)q(x), \tag{8.8.4}$$

for some polynomial q. Now $\sum \lambda_i f_i(x)^2$ has even degree, because the highest term is a sum of squares with positive coefficients and so cannot vanish. Moreover, this degree is positive, for otherwise (8.8.3) would give a contradiction, and it cannot exceed $2n-2$, because $\deg f_i \leq n-1$. Thus pq has even degree $\leq 2n-2$, hence q has odd degree $\leq n-2$. It follows that q has an irreducible factor of odd degree $< n$, and if β is a zero, then by induction on n the ordering extends to $K(\beta)$, but on putting $x = \beta$ in (8.8.4) we find $-1 = \sum \lambda_i f_i(\beta)^2$, a contradiction. Hence $\Gamma_K(L)$ is a cone and this shows that the ordering of K can be extended to L.

Next let $L = K(\sqrt{a})$ and suppose again that $\Gamma_K(L)$ is not a cone. Then

$$-1 = \sum \lambda_i (x_i + y_i\sqrt{a})^2.$$

Multiplying out and observing that $1, \sqrt{a}$ are linearly independent over K, we obtain the relation

$$-1 = \sum \lambda_i (x_i^2 + y_i^2 a) > 0,$$

which is a contradiction. Hence $\Gamma_K(L)$ is a cone and the result follows. ■

It is useful at this point to introduce a special type of ordered field. A field K is called *Euclidean* if for each $a \in K^\times$ exactly one of $a, -a$ is a square; this just means that the set of all squares in K is the positive cone of an ordering. Such a field can always be ordered by writing $a \geq 0$ iff a is a square; clearly this is the only possible ordering of K. Our next result, essentially due to Emil Artin, in a presentation following Winfried Scharlau (1985), shows how Euclidean fields are related to real closed fields.

Theorem 8.8.7. *For any field K the following conditions are equivalent:*

(a) *K is real closed,*
(b) *K is Euclidean and every equation over K of odd degree has a root in K,*
(c) *K is not algebraically closed but $K(\sqrt{-1})$ is algebraically closed.*

Proof. (a) $\Rightarrow$ (b). If K is real closed it is formally real and so can be ordered, by Corollary 8.8.4. By Lemma 8.8.6 this ordering can be extended to $K(\sqrt{a})$ for any $a > 0$, and since $K(\sqrt{a}) = K$ by the maximality of K as formally real field, it follows that a is a square in K. Of course every square is ≥ 0, so K is indeed Euclidean, and no equation of odd degree > 1 can be irreducible, because it would lead to a larger ordered field, by Lemma 8.8.6. Hence every equation of odd degree > 1 is reducible over K, and by induction on the degree it has a root in K, thus (b) holds.

(b) $\Rightarrow$ (c). Assume (b); then -1 is not a square in K, so K is not algebraically closed. Moreover, since K is Euclidean, it can be ordered by writing $a \geq 0$ iff a is a square. Now consider $K(\sqrt{-1})$ and write $i = \sqrt{-1}$; we have to show that $K(i)$ is algebraically closed. In the first place, every element is a square, for if $c \in K$, either $c \geq 0$ and $\sqrt{c} \in K$ or $c < 0$ and $\sqrt{-c} \in K$. Next if $c \in K(i)$, say $c = a + bi$, where $a, b \in K$ and $b \neq 0$, then we have to solve

$$a + bi = (u + vi)^2 = u^2 - v^2 + 2uvi. \tag{8.8.5}$$

By (8.8.5) we have $a^2 + b^2 = (u^2 + v^2)^2$; now $a^2 + b^2 > 0$ in K, hence it is a square in K, say $a^2 + b^2 = d^2$, and it remains to solve

$$(u^2 + v^2)^2 = d^2, 2uv = b. \tag{8.8.6}$$

Now d is only determined up to sign and we may assume that $d \geq 0$. Then d is a square, say $d = e^2$ and the first equation (8.8.6) can be replaced by $u^2 + v^2 = e^2$.

Either b or $-b$ is a square; if b is a square, then $e^2 + b > 0$, hence $e^2 + b = h^2$, and if we now solve for u, v from $u + v = h$, $u - v = a/h$, we find that (8.8.5) holds with these values. Similarly, if $-b$ is a square, then $e^2 - b > 0$, hence $e^2 - b = k^2$ and we have instead to solve $u - v = k$, $u + v = a/k$. Thus every element of $K(i)$ has a square root, and the usual formula for quadratic equations shows that every quadratic equation over $K(i)$ has a root.

Suppose now that $K(i)$ is not algebraically closed. Then any finite extension of $K(i)$ is contained in a Galois extension L/K (recall that char $K = 0$, so every finite extension is separable). Let $G = \mathrm{Gal}(L/K)$, take a Sylow 2-subgroup S of G and denote by H the fixed field of S. Then $[H : K] = (G : S)$ is odd, by the definition of S, hence any element α of H has odd degree over K, but by (b) all such elements belong to K, so $H = K$ and $G = S$ is a 2-group. It follows that the Galois extension $L/K(i)$ also has a 2-group T, say. By Corollary 2.1.7, every non-trivial 2-group has a subgroup of index 2. A subgroup of index 2 in T corresponds to an extension of $K(i)$ of degree 2, but this contradicts the fact that every quadratic equation over $K(i)$ has a root in $K(i)$. Hence $T = 1$ and $L = K(i)$, i.e. $K(i)$ has no proper algebraic extensions, and is therefore algebraically closed.

(c) $\Rightarrow$ (a). It is enough to show that K is formally real; this will follow if we show that $a^2 + b^2$ is a square for all $a, b \in K$. Over $K(i)$, $a + bi$ has a square root, thus $a + bi = (\alpha + \beta i)^2$ say, and hence $a^2 + b^2 = (\alpha^2 + \beta^2)^2$. ∎

As we saw, every Euclidean field can be ordered in just one way; hence we have

Corollary 8.8.8. *Every real closed field can be ordered in just one way; for this ordering every positive element has a square root, and any automorphism necessarily preserves the ordering.* ∎

With the help of Theorem 8.8.7 we can also deduce the fundamental theorem of algebra (see Section 7.3). We shall need the intermediate value property established in Proposition 8.7.6, which depended on the continuity of polynomials over **R**.

Proposition 8.8.9. *An ordered field is real closed if and only if it has the intermediate value property for polynomials.*

Proof. Let K be a real closed field, f be any polynomial over K and $\alpha, \beta \in K$, say $\alpha < \beta$, such that $f(\alpha)f(\beta) < 0$. Since its algebraic closure is of degree 2 over K (by Theorem 8.8.7), we can write f as a product of linear and quadratic factors, irreducible over K. Each irreducible quadratic factor has the form

$$x^2 + px + q = (x + p/2)^2 + (q - p^2/4),$$

and here $q - p^2/4 > 0$, by irreducibility. Hence this factor is positive for all values of x, so f can change sign only when a linear factor changes sign, and this happens only at a zero of f, so there must be a zero lying between α and β.

Conversely, when the intermediate value property holds for K, then we shall show that K is real closed by verifying Theorem 8.8.7(b). Let $a > 0$ in K; the polynomial $f = x^2 - a$ is positive for $x = 1 + a$ and negative for $x = 0$, hence it has a zero, and

this is the required square root. Similarly if f is a polynomial of odd degree, then f has the same sign as its leading coefficient for large positive x and the opposite sign for large negative x, hence it changes sign and so has a zero in K. Thus Theorem 8.8.7(b) holds and it follows that K is real closed. ■

We have seen in Proposition 8.7.6 that $\mathbf{R}$ has the intermediate value property, hence we obtain

Corollary 8.8.10. *The field* $\mathbf{R}$ *of real numbers is real closed and the field of complex numbers, defined as* $\mathbf{C} = \mathbf{R}(\sqrt{-1})$ *is algebraically closed.* ■

This proof of the fundamental theorem of algebra (and the part of Theorem 8.8.7 on which it depends) is due to E. Artin.

In Theorem 8.8.5 we saw that every formally real field has a real closure and we now ask to what extent this real closure is unique. The answer is at first disappointingly vague, but it can be made precise once the problem is reformulated: by Theorem 8.8.5 and Corollary 8.8.8 every formally real field can be ordered. If we now take an ordered field K as our starting point, we shall find that its real closure is unique, up to a unique K-isomorphism. One method of proof is by Sturm's theorem, which determines the exact number of real roots of an equation over $\mathbf{R}$ in any interval (see Exercise 12). We shall instead use the proof of Knebusch 1972, as simplified by Becker and Spitzlay 1975, see Scharlau (1985).

Let K be any ordered field. A quadratic form q over K is called *positive-definite* if $q(x) > 0$ for all $x \neq 0$; if this holds for $-q$, q is said to be *negative-definite*. Given any quadratic space (V, q) over K with a regular form q, we can write V as an orthogonal sum

$$V = V^+ \perp V^- \tag{8.8.7}$$

of two subspaces, V^+ where q is positive-definite and V^- where q is negative-definite. These subspaces are not unique, but their dimensions are unique, and the difference

$$\operatorname{sgn}(q) = \dim(V^+) - \dim(V^-) \tag{8.8.8}$$

is called the *signature* of q. For the case $K = \mathbf{R}$ this just reduces to Sylvester's law of inertia and the uniqueness proof is the same as in that case: if $V = W^+ \perp W^-$ is another such decomposition, then $W^+ \cap V^- = 0$, hence $\dim(W^+) \leq \dim(V) - \dim(V^-) = \dim(V^+)$; by symmetry we find that $\dim(W^+) = \dim(V^+)$, hence also $\dim(W^-) = \dim(V^-)$.

The definition (8.8.8) can still be used for singular spaces, by taking the signature of the regular part.

Consider a finite-dimensional K-algebra A. We have the trace form on A, given by the trace of the regular representation. If $\rho_a : x \mapsto xa$, then the *trace* of a, $\operatorname{tr}(a)$, is defined as the trace of the linear mapping ρ_a. This provides us with a quadratic form $\operatorname{tr}(x^2)$ on A, which is regular whenever A is semisimple. For consider the set

$$\mathfrak{N} = \{y \in A | \operatorname{tr}(xy) = 0 \quad \text{for all } x \in A\}. \tag{8.8.9}$$

This set $\mathfrak{N}$ is an ideal in A, because $\mathrm{tr}(x(ay)) = \mathrm{tr}((xa)y)$, $\mathrm{tr}(x(ya)) = \mathrm{tr}((ax)y)$. By definition (8.8.9), for any $a \in \mathfrak{N}$, all powers of ρ_a have zero trace, hence ρ_a is nilpotent, and so a is nilpotent, because ρ is faithful. Hence $\mathfrak{N}$ is a nilpotent ideal and so $\mathfrak{N} = 0$, because A is semisimple. This shows the quadratic form $\mathrm{tr}(x^2)$ on A to be regular.

We shall need the following lemma, which goes back to Sylvester.

Lemma 8.8.11. *Let K be an ordered field, with real closure $\overline{K}$. If f is a polynomial in x over K, $A = K[x]/(f)$ and q is the quadratic form on A defined by $q(x) = \mathrm{tr}(x^2)$, then* $\mathrm{sgn}(q)$ *equals the number of distinct zeros of f in $\overline{K}$. Hence the number of these zeros depends only on the ordering of K, not on $\overline{K}$.*

Proof. Over $\overline{K}$ the polynomial f splits into factors that are either linear or quadratic:

$$f(x) = p_1(x)^{\alpha_1} \dots p_r(x)^{\alpha_r},$$

hence

$$A_{\overline{K}} = \overline{K}[x]/(f) = E_1 \times \dots \times E_r, \quad \text{where } E_i = \overline{K}[x]/(p_i^{\alpha_i}). \tag{8.8.10}$$

It is clear that the signature of q is the same over K and over $\overline{K}$, and to compute it over $\overline{K}$ we may use a basis adapted to the decomposition (8.8.10). The polynomial p_i has degree 1 or 2; we consider these cases separately; in $\overline{K}[x]/((x-a)^\alpha)$ we can take as our basis $1, x-a, \dots, (x-a)^{\alpha-1}$. The quadratic form has rank 1 and $\mathrm{tr}(1) = 1$, $\mathrm{tr}((x-a)^i) = 0$ for $i > 0$, hence the signature is 1 in this case.

If $E = \overline{K}[x]/(p^\alpha)$, where $p = x^2 + \lambda x + \mu$, and $4\mu > \lambda^2$ (because p is irreducible over $\overline{K}$), then on writing $y = (2x+\lambda)(4\mu - \lambda^2)^{-1/2}$, we can express this as $\overline{K}[y]/((y^2+1)^\alpha)$. Taking as our basis $1, y, y^2+1, y(y^2+1), (y^2+1)^2, \dots$, we find that our quadratic form now has rank 2 and $\mathrm{tr}(y^i(y^2+1)^j) = 0$ if $j \geq 1$, $\mathrm{tr}(1) = 2$, $\mathrm{tr}(y) = 0$, $\mathrm{tr}(y^2) = -2$; therefore the signature is 0. Thus in each case the signature equals the number of distinct zeros of f in $\overline{K}$, as claimed. ∎

We shall use this result to examine the extensions of the ordering of K to a simple extension:

Proposition 8.8.12. *Let K be an ordered field and $L = K(\alpha)$ be a simple algebraic extension, with quadratic form defined by the trace on L: $q(x) = \mathrm{tr}_{L/K}(x^2)$. If* $\mathrm{sgn}(q) = r > 0$, *then L has an ordering extending that of K and there are at most r distinct orderings of L extending the ordering on K, obtained by taking r K-homomorphisms of L into a real closure of K.*

Proof. Denote by $\overline{K}$ a real closure of K; by Lemma 8.8.11, r is the number of zeros of the minimal polynomial of α in $\overline{K}$, and each zero α_i gives rise to an embedding σ_i of L in $\overline{K}$ by mapping $\alpha \mapsto \alpha_i$. The ordering induced in L in this way clearly extends the ordering of K. If there is an ordering $>_o$ of L distinct from these r, then for $i = 1, \dots, r$ there exists $c_i \in L$ such that $c_i >_o 0$, but $c_i^{\sigma_i} < 0$ in $\overline{K}$. By repeated application of Lemma 8.8.6 the ordering $>_o$ extends to

$$L(\sqrt{c_1}, \dots, \sqrt{c_r}) = K(\gamma), \tag{8.8.11}$$

for a primitive element γ over K with minimal polynomial g. Now $K(\gamma)$ has a real closure M, which is a real closed algebraic extension of K, and so is a real closure for K. Since g has a zero in M, g also has a zero in $\overline{K}$, by Lemma 8.8.11. Hence there is an embedding $\tau : K(\gamma) \to \overline{K}$; the restriction of τ to K must equal one of the σ_i, say σ_1. But then $c_1^{\sigma_1} = c_1^{\tau} = ((\sqrt{c_1})^{\tau})^2 > 0$, which contradicts the fact that $c_1^{\sigma_1} < 0$. ∎

If L is an algebraic extension of an ordered field K, then the ordering cannot necessarily be extended to L; but when it can, then the bound given in Proposition 8.8.12 is always attained. This can be proved at the same time as the uniqueness of the real closure.

Theorem 8.8.13. *Let K be an ordered field, L be a finite algebraic extension, say $L = K(\gamma)$, and let $r = \operatorname{sgn}(q)$, where $q(x) = \operatorname{tr}_{L/K}(x^2)$. If the ordering of K can be extended to L, there are exactly r such extensions and for each such ordering L has an order-preserving embedding in the real closure $\overline{K}$ of K.*

Further, any two real closures of K are isomorphic by a unique isomorphism and this isomorphism preserves the ordering.

Proof. We first show that any ordered algebraic extension L of K has an order-embedding in $\overline{K}$. For by Zorn's lemma there is a maximal subfield L_0 of L with an order-embedding in $\overline{K}$. If $L_0 \neq L$, say $\alpha \in L \backslash L_0$, then $L_0(\alpha)$ is an ordered algebraic extension of L_0 and by Proposition 8.8.12, $L_0(\alpha)$ has an order-embedding in $\overline{K}$; this contradicts the maximality of L_0, hence $L_0 = L$ and so L has an order-embedding in $\overline{K}$.

Next we prove the last part. Suppose that L_1, L_2 are real closures of K; by what has been shown there exists a K-homomorphism $\sigma : L_1 \to L_2$. Since the image L_1^{σ} is real closed, we have $L_1^{\sigma} = L_2$ and so σ is an isomorphism; since $(\alpha^2)^{\sigma} = (\alpha^{\sigma})^2$ for any $\alpha \in L_1$, we have $\alpha > 0$ in L_1 iff $\alpha^{\sigma} > 0$ in L_2, thus σ preserves the ordering. Let $\tau : L_1 \to L_2$ be another isomorphism; we have to show that $\alpha^{\sigma} = \alpha^{\tau}$ for any $\alpha \in L_1$. Denote by f the minimal polynomial of α over K and by $\alpha_1, \ldots, \alpha_r$ its conjugates in L_1, where $\alpha_1 < \ldots < \alpha_r$ say. Since σ, τ are order-preserving, we have

$$\alpha_1^{\sigma} < \ldots < \alpha_r^{\sigma}, \quad \alpha_1^{\tau} < \ldots < \alpha_r^{\tau}.$$

Now $\alpha_1^{\sigma}, \ldots, \alpha_r{}^{\sigma}$ are the zeros of f in L_2 and likewise $\alpha_1^{\tau}, \ldots, \alpha_r{}^{\tau}$, hence $\alpha_i{}^{\sigma} = \alpha_i{}^{\tau}$ for $i = 1, \ldots, r$, and so $\alpha^{\sigma} = \alpha^{\tau}$, as required.

Finally let $L = K(\gamma)$ be as stated; then there are r embeddings $\sigma_1, \ldots, \sigma_r$ in $\overline{K}$ and we have to show that they define distinct orderings of L. If σ_1, σ_2 define the same ordering on L, take a real closure M of L for this ordering. By the first part, σ_1 and σ_2 can be extended to K-isomorphisms $M \to \overline{K}$ and we now have distinct K-isomorphisms between the real closed fields M, $\overline{K}$; but we have seen that there is only one such isomorphism. Hence there are r distinct orderings on L. ∎

Several ways are known of estimating the number of roots of an equation, such as the Harriot–Descartes rule, which dates from 1637. However, this leaves the question of finding the exact number of real roots open, and it was not until nearly 200 years

later that this problem was solved, by Jacques-Charles-François Sturm (his solution, announced in 1829, was first published in 1835). We shall prove Sturm's theorem here using only the intermediate value property of **R** established in Proposition 8.7.6. It holds for any real closed field and can be used to give another proof of the uniqueness of the real closure of a formally real field, but we have preferred the above more direct method. The result, though of independent interest, will not be needed later and can be omitted without loss of continuity.

Our object is to determine the exact number of real roots of a given equation over **R**. More precisely, given real numbers $\alpha < \beta$, we shall describe a method of calculating the number of roots γ satisfying $\alpha < \gamma < \beta$.

Let f be a polynomial over **R**; by a *Sturm sequence* of polynomials for f in the interval $[\alpha, \beta]$ we understand a sequence of polynomials over **R**:

$$f_0 = f, f_1, \ldots, f_r \tag{8.8.12}$$

such that

S.1 f_r has no zeros in $[\alpha, \beta]$;
S.2 $f_0(\alpha), f_0(\beta) \neq 0$;
S.3 for $i = 1, \ldots, r-1$, if $f_i(\gamma) = 0$, where $\alpha < \gamma < \beta$, then $f_{i-1}(\gamma) f_{i+1}(\gamma) < 0$;
S.4 if $f(\gamma) = 0$, where $\alpha < \gamma < \beta$, then $f_0(x) f_1(x)$ is an increasing function of x at $x = \gamma$, thus $f_0 f_1$ changes from negative to positive values in a small enough interval about γ.

We shall soon find how to construct such sequences; for the moment we show how a Sturm sequence can be used to answer the question raised above. In any sequence of real numbers $\lambda_1, \ldots, \lambda_n$ we shall define the number of sign changes as the number of times there is a change from a positive to a negative value or vice versa, ignoring zeros; e.g. $2, 0, 3, -1, 0, 0, 4$ has two changes.

Theorem 8.8.14. *Let f be a real polynomial of degree n, with Sturm sequence* (8.8.12) *in the interval $[\alpha, \beta]$, and for any $\gamma \in \mathbf{R}$ denote by $\delta(\gamma)$ the number of sign changes in the sequence*

$$f_0(\gamma), f_1(\gamma), \ldots, f_r(\gamma). \tag{8.8.13}$$

Then the number of distinct zeros of f in $[\alpha, \beta]$ is $\delta(\alpha) - \delta(\beta)$.

Proof. Consider how the sequence (8.8.13) changes as γ varies from α to β. The only sign changes happen when γ passes through a zero of some f_i (by the intermediate value property for **R**). We first take the case $i > 0$. By **S.1** we also have $i < r$, thus $f_i(\gamma) = 0$ and by **S.3**, $f_{i-1}(\gamma) f_{i+1}(\gamma) < 0$, say $f_{i-1}(\gamma) < 0 < f_{i+1}(\gamma)$. Then in the section f_{i-1}, f_i, f_{i+1} of (8.8.13) there is one change of sign. We still have $f_{i-1}(\gamma') < 0 < f_{i+1}(\gamma')$ for any γ' near γ; for such γ' the number of sign changes in the sequence f_{i-1}, f_i, f_{i+1} is still 1, whatever the value of $f_i(\gamma')$. Hence the number of sign changes in (8.8.13) remains constant as the argument passes through a zero of f_i $(i > 0)$. Next take the case when the argument passes through a zero γ of f_0; then $f_1(\gamma) \neq 0$ by **S.3** for $i = 1$. By taking a sufficiently small interval about Γ we may assume that γ is the only zero of f_0 in this interval and that f_1 has no zeros there.

Assume first that $f_0(\gamma') < 0$ for $\gamma' < \gamma$; then $f_1(\gamma') > 0$ by **S.4**, and either $f_0(\gamma'') > 0$, $f_1(\gamma'') > 0$ for $\gamma'' > \gamma$, or $f_0(\gamma'') < 0$, $f_1(\gamma'') < 0$ for $\gamma'' > \gamma$. In the second case f_1 changes from positive to negative values and hence vanishes, against the hypothesis. Thus the first alternative must hold, and as x increases from $\gamma' < \gamma$ to $\gamma'' > \gamma$, the sequence $f_0(x)$, $f_1(x)$ changes from $-$, $+$ to $+$, $+$, so one sign change is lost as x passes through a zero of f. The same argument applies if $f_0(\gamma') > 0$ for $\gamma' < \gamma$. By addition we find that the number of sign changes in $[\alpha, \beta]$ is $\delta(\alpha) - \delta(\beta)$, as asserted. ■

Now let f be any polynomial and define the *standard sequence* for f as follows: $f_0 = f$, $f_1 = f'$ (the derivative) and for $i \geq 1$, f_i is defined as the remainder in the division algorithm, with the sign changed:

$$f_{i-1} = f_i q_i - f_{i+1}, \quad \deg f_{i+1} < \deg f_i. \tag{8.8.14}$$

Let f_r be the last non-zero remainder; by the Euclidean algorithm, f_r is the highest common factor of f and f'. Thus if f has no repeated zeros, f_r is a non-zero constant. We claim that in this case $f_0, f_1, \ldots, f_r$ is a Sturm sequence for f, for any interval $[\alpha, \beta]$ such that $f(\alpha)f(\beta) \neq 0$. **S.1** holds since f_r is constant, and **S.2** holds by assumption; further no two successive f_i can vanish for the same value of x, for if they did, then by (8.8.14) all succeeding f_i must vanish and this contradicts **S.1**. Now if $f_i(\gamma) = 0$, then by (8.8.14), $f_{i-1}(\gamma)f_{i+1}(\gamma) < 0$, so **S.3** holds. Finally **S.4** holds because if $f(\gamma) = 0$, then f^2 has a minimum at $x = \gamma$ and so $(f^2)' = 2f_0f_1$ is increasing at $x = \gamma$.

Suppose now that f is an arbitrary polynomial and let d be the highest common factor of f and f', say $f = dg_0$, $f' = dg_1$. Clearly d is a factor of each f_i, so writing $f_i = dg_i$, we have

$$g_{i-1} = g_i q_i - g_{i+1}, \quad \deg g_{i+1} < \deg g_i. \tag{8.8.15}$$

We claim that $g_0, g_1, \ldots, g_r$ is a Sturm sequence for g_0. **S.1–S.3** follow as before; to establish **S.4** we note that when $g_0(\gamma) = 0$, then $f(\gamma) = 0$, f^2 has a minimum and so $(f^2)' = 2d^2g_0g_1$ is increasing at $x = \gamma$. Clearly $2d^2$ is non-negative, so g_0g_1 is increasing at $x = \gamma$.

We thus find that

$$g_0, g_1, \ldots, g_r \tag{8.8.16}$$

is a Sturm sequence for g_0. Now the sequence

$$f_0, f_1, \ldots, f_r \tag{8.6.17}$$

is obtained by multiplying each term of (8.8.16) by d; hence the numbers of sign changes of (8.8.16) and (8.8.17) are the same at any point Γ not a zero of d. We can therefore use (8.8.17) to determine the zeros of g_0. But they are just the distinct zeros of f, so we have proved

Theorem 8.8.15 (Sturm's theorem). *Let f be a polynomial over* **R** and let $f_0 = f, f_1 = f', f_2, \ldots, f_r$ *be the standard sequence for f. Given $\alpha, \beta \in \mathbf{R}$, not zeros of f, such that $\alpha < \beta$, the number of distinct zeros of f (not counting multiplicities) in $[\alpha, \beta]$ is $\delta(\alpha) - \delta(\beta)$.* ■

We note that by taking α, β sufficiently large negative and positive numbers respectively, we obtain a formula for the total number of real zeros of f. It may be shown that if $f = x^n + a_1 x^{n-1} + \ldots + a_n$ and $M = \max\{|a_1|, \ldots, |a_n|\} + 1$, then all the real zeros of f lie in the interval $[-M, M]$.

Exercises

1. Verify that any cone satisfying (8.8.1) defines an ordering.
2. Let R be any ring and P be a maximal cone in R. Show that $P \cup -P = R$ and $P \cap -P$ is a prime ideal in R.
3. Let K be a field of characteristic not 2 which is not formally real. Show that the core of K is the whole of K. Deduce that every element of K is a sum of squares.
4. Show that there is a least Euclidean subfield E of $\mathbf{R}$ and verify that E is not real closed.
5. Show that there is a least subfield F of $\mathbf{R}$ over which every equation of odd degree has a root in F; verify that F is not Euclidean.
6. Show that a field is real closed iff its algebraic closure is of finite degree > 1 over it.
7. Let $f = 0$ be an equation with real coefficients, without repeated roots. Show that if the number of pairs of conjugate complex roots of $f = 0$ is s, then the discriminant of f has sign $(-1)^s$.
8. In Theorem 8.8.7 show that (b) $\Rightarrow$ (c) without using Sylow theory by forming, for any equation $f = 0$ with roots $\alpha_1, \ldots, \alpha_n$ the equation with roots $\alpha_i + \alpha_j + \lambda\alpha_i\alpha_j$ $(i < j)$, where λ is chosen to make all these values distinct, using induction on the 2-power dividing deg f.
9. Let K be a field such that $x^2 + 1$ is irreducible over K, while $K(\sqrt{-1})$ is algebraically closed. Show that any polynomial over K splits into linear and quadratic factors, and that the sum of two squares in K is a square. Deduce that K is formally real and in fact real closed. (Hint. See the proof (c) $\Rightarrow$ (a) of Theorem 8.8.7.)
10. Apply Sturm's theorem to determine the roots of $x^4 - 3x^3 - x^2 + 8x - 4 = 0$ to the nearest integer.
11. The Legendre polynomials, defined as

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n,$$

$(n = 0, 1, \ldots)$ satisfy the relations

$$nP_n - (2n-1)xP_{n-1} + (n-1)P_{n-2} = 0,$$

$$(1 - x^2)P_n' + nxP_n - nP_{n-1} = 0 \quad (n = 1, 2, \ldots).$$

Show that for any $n \geq 0$, $P_n, P_{n-1}, \ldots, P_0$ is a Sturm sequence for P_n in $[-1, 1]$. Deduce that P_n has n distinct zeros in $[-1, 1]$ which are separated by those of P_{n-1}.

12. Prove the last part of Theorem 8.8.13 by using Theorem 8.8.15.

8.9 The Witt Ring of a Field

In Section 8.5 we associated with any field k the monoid $M(k)$ whose elements are equivalence classes (under isometry) of regular quadratic spaces. By the Witt cancellation theorem (Theorem 8.5.1) this is a cancellation monoid, hence $M(k)$ is embedded in its universal group $\hat{W}(k)$ (Proposition 2.1.8). Formally the elements of $\hat{W}(k)$ can be written as differences of isometry classes: $[U] - [V]$, and we have $[U] - [V] = [U'] - [V']$ iff $U \perp V' \cong U' \perp V$. Our next object is to define a multiplication on $\hat{W}(k)$ which will make it into a ring.

Let V, V' be regular quadratic spaces with associated bilinear forms b, b'; we can define a bilinear form on the tensor product $V \otimes V'$ by the formula

$$F(x \otimes x', y \otimes y') = b(x, y)b'(x', y') \quad (x, y \in V, x', y' \in V').$$

By linearity this determines F completely; we observe that in terms of the corresponding quadratic forms q, q', P we have

$$P(x \otimes x') = q(x)q'(x'),$$

but this formula by itself is not enough to determine P, because P is not linear and we need to define it for sums $\sum x_i \otimes x_i'$, not merely products $x \otimes x'$. In terms of matrices, if the forms on V, V' have the matrices A, A' respectively (relative to any bases), then the matrix for $V \otimes V'$ is the Kronecker product $A \otimes A'$. From the properties of the tensor product it is clear that this operation induces a multiplication on $\hat{W}(k)$:

$$[U][V] = [U \otimes V].$$

Clearly this multiplication is associative and commutative, and on the basis of the formula

$$U \otimes (V \perp V') \cong U \otimes V \perp U \otimes V'$$

it follows easily that the distributive law holds in $\hat{W}(k)$. Further, we have $U \otimes \langle 1 \rangle \cong U$ and if Z is the zero-dimensional space, then $U \otimes Z \cong Z$. Hence $\hat{W}(k)$ is a commutative ring with the zero space as 0 and $\langle 1 \rangle$ as unit element; it is called the *Witt–Grothendieck ring* of quadratic forms over k. In terms of a diagonal representation we have

$$\langle a_1, \ldots, a_r \rangle \otimes \langle b_1, \ldots, b_s \rangle \cong \langle a_1 b_1, a_1 b_2, \ldots, a_1 b_s, a_2 b_1, \ldots, a_r b_s \rangle.$$

For any $a, b \in k$, if $a, b, a + b \neq 0$, we have the *Witt relation*

$$\langle a, b \rangle \cong \langle a + b, (a + b)ab \rangle. \tag{8.9.1}$$

For if u, v is an orthogonal basis for $\langle a, b \rangle$, then $u + v$, $bu - av$ is again an orthogonal basis and $q(u + v) = a + b$, $q(bu - av) = b^2 a + a^2 b = (a + b)ab$, so we obtain the right-hand side of (8.9.1). Thus in $\hat{W}(k)$, $\langle a \rangle + \langle b \rangle = \langle a + b \rangle + \langle (a + b)ab \rangle$.

We next try to find a presentation for $\hat{W}(k)$. To this end let us define an abstract ring $A(k)$ with generators $[a]$, $a \in k^\times$ and defining relations

R.1 $[1] = 1$;
R.2 $[a][b] = [ab]$ for all $a, b \in k^\times$;
R.3 $[a] + [b] = [a+b] + [(a+b)ab]$ for all $a, b \in k^\times$ such that $a \neq -b$.

By **R.2**, $A(k)$ is commutative; we also note the following consequence of **R.1–R.3**:

$$[ab^2] = [a]. \tag{8.9.2}$$

To see this we put $a = b$ in **R.3**: $2[b] = [2b] + [2b^3]$. Multiplying by $[a/2b]$ we get $2[a/2] = [a] + [ab^2]$, and since the left-hand side is independent of b, it follows that $[ab^2] = [a1^2] = [a]$, i.e. (8.9.2). As a relation in $\hat{W}(k)$, (8.9.2) is obvious, since it states that $\langle ab^2\rangle \cong \langle a\rangle$, which follows by a change of variable.

We have seen that **R.1–R.3** hold for the generators of $\hat{W}(k)$; therefore there is a unique ring homomorphism $\varphi : A(k) \to \hat{W}(k)$ defined by $[a] \mapsto \langle a\rangle (a \in k^\times)$; we claim that φ is an isomorphism. Since $\hat{W}(k)$ is additively generated by the $\langle a\rangle$, φ is surjective. Now any element of $A(k)$ can by **R.2** be written in the form

$$[a_1] + \ldots + [a_n] - [b_1] - \ldots - [b_m] \quad (a_i, b_j \in k^\times); \tag{8.9.3}$$

if $\ker \varphi \neq 0$, we may choose a non-zero element (8.9.3) of $\ker \varphi$ so as to minimize n. This means that $\langle a_1, \ldots, a_n\rangle \cong \langle b_1, \ldots, b_m\rangle$ and hence $m = n$, by the invariance of the rank. Here we must have $n > 1$, for otherwise $a_1 = b_1 x^2$ and so $[a_1] = [b_1]$ by (8.9.2). Now since b_1 is represented by $\langle a_1, \ldots, a_n\rangle$, we have an expression

$$b_1 = a_1 x_1^2 + \ldots + a_p x_p^2, \quad \text{where } p \leq n. \tag{8.9.4}$$

Suppose that the a_i in (8.9.3) are chosen so as to minimize p in (8.9.4). If $p > 1$, then $d = a_1 x_1^2 + a_2 x_2^2$ is non-zero and by **R.3** and (8.9.2),

$$[a_1] + [a_2] = [a_1 x_1^2] + [a_2 x_2^2] = [d] + [d.a_1 x_1^2 a_2 x_2^2] = [d] + [da_1 a_2].$$

Therefore $[a_1] + \ldots + [a_n] = [d] + [da_1 a_2] + [a_3] + \ldots + [a_n]$ and now $b_1 = d + a_3 x_3^2 + \ldots + a_p x_p^2$, which contradicts the choice of p. Thus $p = 1$, $b_1 = a_1 x_1^2$, hence $[b_1] = [a_1]$ and by Witt's cancellation theorem (Theorem 8.5.1), $\ker \varphi$ contains

$$[a_2] + \ldots + [a_n] - [b_2] - \ldots - [b_n].$$

This contradicts the minimality of n, hence $\ker \varphi = 0$ as claimed.

Thus we have proved

Theorem 8.9.1. *Let k be a field of characteristic not 2. Then the Witt–Grothendieck ring $\hat{W}(k)$ has a presentation with generators $\langle a\rangle, a \in k^\times$ and defining relations*

W.1 $\langle 1\rangle = 1$;
W.2 $\langle a\rangle\langle b\rangle = \langle ab\rangle$ *for* $a, b \in k^\times$;
W.3 $\langle a\rangle + \langle b\rangle = \langle a+b\rangle + \langle (a+b)ab\rangle$ *for* $a, b \in k^\times$ *such that* $a \neq -b$. ∎

A closer examination reveals a further consequence. Let us call two diagonal forms $\langle a_1, \ldots, a_n\rangle$ and $\langle b_1, \ldots, b_n\rangle$ *simply related* if there exist i, j such that

$\langle a_i, a_j\rangle \cong \langle b_i, b_j\rangle$ and $a_k = b_k$ for all $k \neq i, j$. Here i may equal j, in which case we interpret $\langle a_i, a_i\rangle$ as $\langle a_i\rangle$. Now the proof of Theorem 8.9.1 shows the truth of

Theorem 8.9.2. (Witt's chain equivalence theorem). *Let q, q' be two regular diagonal quadratic forms (over a field of characteristic not 2) which are equivalent. Then there is a sequence of diagonal forms $q_1 = q, q_2, \ldots, q_r = q'$ such that for $i = 2, \ldots, r$, q_{i-1} is simply related to q_i.* ∎

Every invariant of quadratic forms gives rise to a function on $\hat{W}(k)$. Consider for example the rank; clearly isometric forms have the same rank, hence the rank function induces a mapping

$$\mathrm{rk} : \hat{W}(k) \to \mathbf{Z}, \tag{8.9.5}$$

and since $\mathrm{rk}(U \perp V) = \mathrm{rk}\, U + \mathrm{rk}\, V$, $\mathrm{rk}(U \otimes V) = \mathrm{rk}\, U . \mathrm{rk}\, V$, $\mathrm{rk}\langle 1\rangle = 1$, it follows that (8.9.5) is a ring homomorphism; the kernel $\hat{I}(k)$ is called the *augmentation ideal* of $\hat{W}(k)$. Since (8.9.5) is split by the natural mapping $\mathbf{Z} \to \hat{W}(k)$, we have a direct decomposition

$$\hat{W}(k) \cong \mathbf{Z} \oplus \hat{I}(k).$$

A second important ideal in $\hat{W}(k)$ consists of the split spaces. We recall from Section 8.5 that a *split* quadratic space is an orthogonal sum of hyperbolic planes. Hence a space is split iff its dimension is even, say $n = 2r$, and relative to a suitable basis its matrix is $\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$. A quadratic space is split whenever it is regular of even dimension $2r$ and has a totally isotropic r-dimensional subspace; for then its matrix relative to a suitable basis has the form

$$\begin{pmatrix} 0 & P \\ P^T & Q \end{pmatrix}$$

where P, Q are $r \times r$ and P is invertible. Now we have

$$\begin{pmatrix} I & 0 \\ A & B \end{pmatrix}\begin{pmatrix} 0 & P \\ P^T & Q \end{pmatrix}\begin{pmatrix} I & A^T \\ 0 & B^T \end{pmatrix} = \begin{pmatrix} 0 & PB^T \\ BP^T & BP^TA^T + APB^T + BQB^T \end{pmatrix},$$

and this reduces to $\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$ when $B = (P^T)^{-1}$, $A = -(1/2)BQB^T$.

We also recall Theorem 8.5.3, which states that in characteristic not 2 any regular quadratic space V can be written as

$$V = L \perp U, \tag{8.9.6}$$

where L is split and U is anisotropic. If $\dim L = 2r$, then r, and hence the isometry type of L, is uniquely determined as the maximal dimension of a totally isotropic subspace of V, while U is unique up to isometry.

The decomposition (8.9.6) is called the *Witt decomposition* of V, r is the *Witt index* and U is the *anisotropic part* of V.

If L is any split space, then so is $L \otimes U$, for any regular space U. For if $\dim L = 2r$, $\dim U = n$, then L has a totally isotropic subspace L_0 of dimension r, hence $L_0 \otimes U$ is a totally isotropic subspace of $L \otimes U$ of dimension rn. This means that the additive group spanned by the split spaces in $\hat{W}(k)$ is already an ideal. We shall often write H for the element of $\hat{W}(k)$ representing the hyperbolic plane, thus $H = [\langle 1, -1\rangle]$; the ideal of split spaces may then be written $\mathbf{Z}H$. The residue class ring $W(k) = \hat{W}(k)/\mathbf{Z}H$ is called the *Witt ring* of k. A more direct description of $W(k)$ is given in

Proposition 8.9.3. *Let k be any field of characteristic not 2. The correspondence*

$$V \mapsto [V] (\mathrm{mod}\ \mathbf{Z}H) \tag{8.9.7}$$

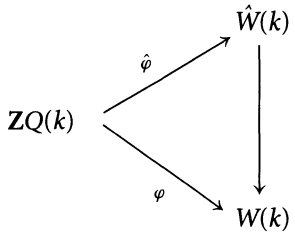
induces a bijection between the set of isometry classes of anisotropic spaces and the Witt ring $W(k)$; the negative of (V, q) is $(V, -q)$. Given two anisotropic spaces U, V, neither $U \perp V$ nor $U \otimes V$ need be anisotropic, but if we take their anisotropic parts provided by the Witt decomposition we obtain the sum and product in $W(k)$.

Proof. Let us call two spaces (or also their forms) *similar* if their anisotropic parts (obtained from the Witt decomposition) are isometric. It is clear that similar spaces have the same image in $W(k)$. Since $H = [\langle a, -a\rangle]$ for any $a \in k^\times$, it follows that $-[\langle a\rangle] = [\langle -a\rangle]$ in $W(k)$; hence every element of $W(k)$ is represented by a space V. If we take a Witt decomposition $V = V_h \perp V_a$, where V_h is split and V_a is anisotropic, then V is similar to V_a and so $[V] = [V_a]$ and it follows that (8.9.7) is surjective. Now let $[U] = [V]$; then in $\hat{W}(k)$ we have $[U] = [V] + nH$, where $n \in \mathbf{Z}$. If $n \geq 0$, say, then $U \cong V \perp nH$, hence U and V are similar. This shows that (8.9.7) is injective, as claimed. Now the rest is clear. ∎

There is another presentation of $W(k)$, related more closely to k, which is often useful. Let us write $Q(k)$ for the multiplicative group of classes mod squares of k:

$$Q(k) = k^\times / k^{\times 2}. \tag{8.9.8}$$

We shall use multiplicative notation for $Q(k)$, writing $\langle a\rangle$ for the class of $a \in k^\times$. Each element of $Q(k)$ corresponds to a one-dimensional quadratic form and the product in $Q(k)$ corresponds to the tensor product of forms. Hence we have a homomorphism φ of $Q(k)$ into the unit group $U(W(k))$, which extends to a ring homomorphism $\mathbf{Z}Q(k) \to W(k)$, again denoted by φ. Likewise there is a homomorphism $\hat{\varphi} : Q(k) \to U(\hat{W}(k))$, which can again be extended to a homomorphism $\hat{\varphi} : \mathbf{Z}Q(k) \to \hat{W}(k)$. Since every quadratic space has an orthogonal basis, it follows that φ and $\hat{\varphi}$ are surjective.



Theorem 8.9.4. *Let k be a field of characteristic not 2 and define $Q(k)$ as in* (8.9.8) *above. Then the Witt–Grothendieck ring and the Witt ring can be described in terms of the group ring $\mathbf{Z}Q(k)$ by*

$$\hat{W}(k) \cong \mathbf{Z}Q(k)/\mathfrak{a}, \quad W(k) \cong \mathbf{Z}Q(k)/\mathfrak{b}, \tag{8.9.9}$$

where $\mathfrak{a}$ is the ideal generated by $(1 + \langle a\rangle)(1 - \langle 1 + a\rangle)$, $a \neq 0, -1$, and $\mathfrak{b}$ is the ideal generated by $\mathfrak{a}$ together with $1 + \langle -1\rangle$.

Proof. It is clear that (8.9.9) holds with $\mathfrak{a} = \ker \hat{\varphi}$, $\mathfrak{b} = \ker \varphi$ and it remains to determine these ideals. Now the proof of Theorem 8.9.1 shows that $\mathfrak{a}$ is generated by $\langle a\rangle + \langle b\rangle - \langle c\rangle - \langle d\rangle$, where $\langle a, b\rangle \cong \langle c, d\rangle$, and clearly also by $1 + \langle b\rangle - \langle c\rangle - \langle d\rangle$, where $\langle 1, b\rangle \cong \langle c, d\rangle$. Here $\langle b\rangle = \langle cd\rangle$, since the determinants are equal, so also $\langle d\rangle = \langle bc\rangle$, and hence

$$1 + \langle b\rangle - \langle c\rangle - \langle d\rangle = (1 + \langle b\rangle)(1 - \langle c\rangle). \tag{8.9.10}$$

Now $c = x^2 + y^2 b$; if $x = 0$, then $\langle c\rangle = \langle b\rangle$ and (8.9.10) reduces to 0. Otherwise we have $\langle c\rangle = \langle 1 + a\rangle$, where $a = x^{-2}y^2 b$ and $\langle a\rangle = \langle b\rangle$, so the right-hand side of (8.9.10) reduces to $(1 + \langle a\rangle)(1 - \langle 1 + a\rangle)$. This confirms the description of $\mathfrak{a}$; now $W(k) \cong \hat{W}(k)/\mathbf{Z}(1 + \langle -1\rangle)$, so $\mathfrak{b}$ is generated by $\mathfrak{a}$ and $1 + \langle -1\rangle$ as claimed. ■

The two ideals $\hat{I}(k)$ and $\mathbf{Z}H$ of $\hat{W}(k)$ lead to the following commutative diagram with exact rows and columns:

$$\begin{array}{ccccccccc}
 & & & & 0 & & 0 & & \\
 & & & & \downarrow & & \downarrow & & \\
 & & 0 & \longrightarrow & \mathbf{Z}H & \longrightarrow & 2\mathbf{Z} & \to & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \to & \hat{I}(k) & \to & \hat{W}(k) & \xrightarrow{rk} & \mathbf{Z} & \to & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \to & I(k) & \to & W(k) & \xrightarrow{e} & \mathbf{Z}/2 & \to & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
 & & 0 & & 0 & & 0 & &
\end{array}$$

From the diagram we see that the rank mapping (8.9.5) induces a homomorphism $e : W(k) \to \mathbf{Z}/2$, called the *dimension index*. Its kernel is $I(k) = \hat{I}(k)$, the set representing forms of even rank.

By contrast the determinant function on $W(k)$ cannot in general be factored via $\hat{W}(k)$, because $\det(H) \neq 1$ in general (i.e. when -1 is not a square). The discriminant $\mathrm{dis}(q)$, defined as in (8.4.15), has the advantage that $\mathrm{dis}(H) = 1$, but we now have $\mathrm{dis}(q \perp q') \neq \mathrm{dis}(q).\mathrm{dis}(q')$ in general. In order to restore multiplicativity we can proceed as follows. Consider again the group $k^\times/k^{\times 2}$, now written additively, and define the set $E(k) = \mathbf{Z}/2\mathbf{Z} \times k^\times/k^{\times 2}$ as a group by the rule

$$(e, d) + (e', d') = (e + e', d + d' + ee'\{-1\}),$$

where $\{-1\}$ is the class of -1. Clearly this is an abelian group with $k^\times/k^{\times 2}$ as a subgroup of index 2. We can also define a multiplication on $E(k)$ by setting

$$(e, d)(e', d') = (ee', ed' + e'd),$$

and with this definition it is easily verified that $E(k)$ is a commutative ring. Now it can be shown that the mapping

$$q \mapsto (e(q), \{\mathrm{dis}(q)\}), \tag{8.9.11}$$

where $e(q)$ is the rank of q (mod 2), defines a surjective homomorphism from $W(k)$ to $E(k)$ with kernel $I(k)^2 = \{\sum a_i b_i | a_i, b_i \in I(k)\}$. This shows that $I(k)^2$ consists of all classes of even-dimensional forms of discriminant 1.

Returning for a moment to Clifford algebras, we recall that for any quadratic form q, either C or C_0 is central simple (depending on whether the rank is even or odd). Thus each quadratic form gives rise to an element of the Brauer group $\mathbf{B}_k$ (see Section 5.4), called the *Witt invariant* of q. By taking the graded structure of C into account, one obtains a larger group, the *Brauer–Wall group* $\mathbf{BW}_k$. We shall not enter into details here but confine ourselves to the remark that the quotient $\mathbf{BW}_k/\mathbf{B}_k$ is isomorphic to the additive group of $E(k)$ (see Lam (1980); Scharlau (1985)).

Another invariant of q, equivalent to the Witt invariant, but a little easier to handle, is obtained by writing q in diagonal form: $q = \langle a_1, \ldots, a_n\rangle$ and taking the element of $\mathbf{B}_k$ corresponding to

$$\prod_{i<j} (a_i, a_j; k),$$

where $(a_i, a_j; k)$ is the quaternion algebra defined in Section 5.4. This is called the *Hasse invariant*; it is not hard to verify (using Theorem 8.9.2) that this is in fact an invariant of q. It can be shown that two forms of the same dimension ≤ 3 are isometric iff they have the same determinant and the same Hasse invariant (see Exercise 11).

The ideal structure of $W(k)$ is closely related to the ways of ordering k, and to end this section we describe the connexion.

Let k be an ordered field and V be a quadratic space over k; it is clear that the signature $\sigma(V)$ depends only on the class of V in $W(k)$, so that we have a mapping

$$\sigma : W(k) \to \mathbf{Z}. \tag{8.9.12}$$

Moreover, since $\sigma(U \perp V) = \sigma(U) + \sigma(V)$, $\sigma(\langle ab\rangle) = \sigma(\langle a\rangle)\sigma(\langle b\rangle)$, $\sigma(\langle 1\rangle) = 1$, it follows that (8.9.12) is a ring homomorphism. Its kernel is generated by the Witt classes of all forms $\langle 1, -a\rangle = 1 - \langle a\rangle$ such that $a > 0$. In particular, if k is Euclidean, then every positive element is a square, so all such forms split and σ is an isomorphism. Thus we have

Proposition 8.9.5. *For any Euclidean field k the signature mapping* (8.9.12) *is an isomorphism; in particular this holds for every real closed field.* ∎

From Theorem 8.8.5 we know that any ordered field k has a real closure K and it is clear that the inclusion $i : k \mapsto K$ induces a homomorphism $i^* : W(k) \to W(K)$. But $W(K) \cong \mathbf{Z}$, by Proposition 8.9.5, hence $W(k)/\ker i^* \cong \mathbf{Z}$. Thus each ordering of k determines a prime ideal of $W(k)$ with residue class ring $\mathbf{Z}$. The full connexion between orderings and prime ideals is described in Theorem 8.9.7 below, but first we need a lemma.

Lemma 8.9.6. *Let $\mathfrak{p}$ be a prime ideal in $W(k)$, for any field k. Then $W(k)/\mathfrak{p}$ is isomorphic to $\mathbf{Z}$ or $\mathbf{Z}/p$, for some prime number p.*

Proof. $A = W(k)/\mathfrak{p}$ is an integral domain by definition of $\mathfrak{p}$, and for any $a \in k^\times$, $(\langle a\rangle + 1)(\langle a\rangle - 1) = \langle a^2\rangle - 1 = 0$, hence $\langle a\rangle \equiv \pm 1 \pmod{\mathfrak{p}}$. Since $W(k)$ is additively generated by $\langle a\rangle$, $a \in k^\times$, the additive group of A is cyclic and hence is a homomorphic image of $\mathbf{Z}$, as claimed. ∎

A prime ideal $\mathfrak{p}$ in $W(k)$ is said to be of *characteristic* 0 or p according as $W(k)/\mathfrak{p}$ is isomorphic to $\mathbf{Z}$ or $\mathbf{Z}/p$.

Theorem 8.9.7 (Harrison–Lorenz–Leicht). *Let k be any field and $W(k)$ be its Witt ring with augmentation ideal $I(k)$. Then $I(k)$ is the unique prime ideal of characteristic 2 in $W(k)$. For an odd prime p there is a natural bijection between* (i) *the orderings of k,* (ii) *the prime ideals of characteristic 0 in $W(k)$ and* (iii) *the prime ideals of characteristic p in $W(k)$.*

Proof. Let $\mathfrak{p}$ be a prime ideal of characteristic 2 in $W(k)$. Then $\langle a^2\rangle \equiv 1 \pmod{\mathfrak{p}}$, hence $\langle a\rangle \equiv 1 \pmod{\mathfrak{p}}$ for all $a \in k^\times$, because $\mathfrak{p}$ is prime, and so $I(k) \subseteq \mathfrak{p}$. Since $I(k)$ is maximal, it follows that $I(k) = \mathfrak{p}$, so $I(k)$ is the unique prime ideal of characteristic 2 in $W(k)$.

In the remark after Proposition 8.9.5 we have seen how to associate with each ordering P of k a prime ideal $\mathfrak{p}$ of characteristic 0 in $W(k)$. Since $W(k)/\mathfrak{p} \cong \mathbf{Z}$, the ideal $p\mathbf{Z}$ in $\mathbf{Z}$ corresponds to a prime ideal $\bar{\mathfrak{p}}$ of characteristic p in $W(k)$. Now let $\mathfrak{p}$ be any prime ideal of odd prime characteristic p in $W(k)$ and define

$$P = \{a \in k | \langle a\rangle \equiv 1 (\mathrm{mod}\ \mathfrak{p})\}.$$

We claim that P is the positive order set of an ordering on k. It is clear that $0 \notin P$ and that P is closed under multiplication. Further, if $a, b \in P$ and $c = a + b$, then $c \neq 0$, for otherwise $\langle b\rangle = \langle -a\rangle \equiv -1 \pmod{\mathfrak{p}}$, i.e. $2 \equiv 0 \pmod{\mathfrak{p}}$, which contradicts the fact that p is odd. Now $\langle a\rangle + \langle b\rangle = \langle c\rangle(1 + \langle ab\rangle)$ by **W.3**, hence $2\langle c\rangle = 2$, and so $\langle c\rangle \equiv 1 \pmod{\mathfrak{p}}$; this shows that P is closed under addition. Moreover, any $a \in k^\times$ satisfies $a \equiv \pm 1 \pmod{\mathfrak{p}}$, so either a or $-a$ is in P, but not both and therefore P defines an ordering on k.

It only remains to show that the mappings $P \mapsto \mathfrak{p}$, $\mathfrak{p} \mapsto \bar{\mathfrak{p}}$, $\bar{\mathfrak{p}} \mapsto P$ when carried out in succession lead back to the same ordering, resp. prime ideal; this is straightforward and may be left to the reader to check. ∎

Let k be a field such that $W(k) \cong \mathbf{Z}$; then k can be ordered and the signature σ is an isomorphism, hence $\langle a \rangle = 1$ for all $a > 0$. Thus every positive element of k is a square, i.e. k is Euclidean. Together with Proposition 8.9.5 this proves

Corollary 8.9.8. *A field k is Euclidean if and only if $W(k) \cong \mathbf{Z}$.* ∎

Exercises

1. Verify the homomorphism (8.9.11) of $W(k)$ into $E(k)$ and show that the kernel is $I(k)^2$. (Hint. Define a map $E(k) \to W(k)/I^2$ by $(0, \{a\}) \mapsto \langle 1, -a \rangle + I^2$, $(1, \{a\}) \mapsto \langle a \rangle + I^2$.)
2. Show that $E(k)$ splits over $Q(k)$ (see (8.9.8)) iff -1 is a square in k.
3. Show that the inverse image of $Q(k)$ in the homomorphism (8.9.11) is $I = I(k)$, hence that I/I^2 is a $W(k)/I$-module. Deduce that $W(k)$ is Noetherian iff $Q(k)$ is finite. (Hint. Recall that $W(k)$ is additively generated by the $\langle a \rangle$, $a \in k^\times$.)
4. Show that the Witt ring $W(k)$ has no idempotents apart from 0 and 1.
5. Show that if k is real closed then $\hat{W}(k) \cong \mathbf{Z} \oplus \mathbf{Z}$, and determine the multiplication. Determine $\hat{W}(k)$ when (i) k is finite, (ii) every element of k is a square.
6. Show that k is not formally real iff $W(k)$ is a local ring, and the unique maximal ideal is then $I(k)$.
7. Show that k is formally real iff $W(k)$ is not a torsion group.
8. A field is said to be *Pythagorean* if every sum of squares is a square. Show that a formally real field k is Pythagorean iff $W(k)$ is torsion-free.
9. Let A be an anisotropic space. Show that if U and $U \perp A$ are both split, then $A = 0$. Deduce that V corresponds to the 0 of $W(k)$ iff V is split.
10. Show that the Witt ring of the p-adic number field $\mathbf{Q}_p$ (see Section 9.3 below) is given by $W(\mathbf{Q}_p) \cong W(\mathbf{F}_p) \oplus W(\mathbf{F}_p)$ for $p \neq 2$, $W(\mathbf{Q}_2) \cong \mathbf{C}_8 \oplus \mathbf{C}_2 \oplus \mathbf{C}_2$.
11. Show that any regular ternary (i.e. 3-dimensional) quadratic form of determinant -1 can be written $\langle a, b, -ab \rangle$. Verify that its Hasse invariant is $(a, b; k)$ and deduce that two ternary forms of the same determinant and the same Hasse invariant are isometric.
12. Show that a field is Euclidean iff it is formally real and $|Q(k)| = 2$.

8.10 The Symplectic Group

Let k be any field. By a *symplectic space* V over k we understand a vector space V over k with a regular bilinear form b which is *alternating*, i.e. such that

$$b(x, x) = 0 \quad \text{for all } x \in V. \tag{8.10.1}$$

By linearization we find that b is antisymmetric: $b(y, x) = -b(x, y)$, and relative to a basis of V, B is described by an *alternating* or *skew-symmetric* matrix, i.e. a matrix A such that $A^T = -A$. If $\dim V = n$, we find by taking determinants, $\det A = \det A^T = \det(-A) = (-1)^n \det A$. Since the space is regular, $\det A \neq 0$. When $\operatorname{char} k \neq 2$, it follows that n must be even, but this will follow generally (even for $\operatorname{char} k = 2$) from Theorem 8.10.1.

Symplectic spaces are of some importance in the Hamiltonian theory of dynamics. Their theory is considerably simpler than that of inner product spaces, for as we saw, every symplectic space is even-dimensional, say $n = 2m$, and as in the proof of Proposition 8.5.2 we can find a basis of the form $u_1, \ldots, u_m, v_1, \ldots, v_m$ such that $b(u_i, v_j) = \delta_{ij}$, $b(u_i, u_j) = b(v_i, v_j) = 0$; such a basis is said to be *symplectic*. Any pair of vectors u, v satisfying $b(u, v) = 1$ is called *hyperbolic*; thus a symplectic basis consists of mutually orthogonal hyperbolic pairs. Relative to a symplectic basis the matrix of a symplectic form becomes

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{8.10.2}$$

Let us again derive the existence of a symplectic basis in a slightly more general form. Any 2-dimensional symplectic space clearly has a basis consisting of a hyperbolic pair; we shall call it again a *hyperbolic plane.*

Theorem 8.10.1. *Let V be a symplectic space, U be any subspace and U_0 be a maximal totally isotropic subspace of U. Then* $\dim U \leq 2\dim U_0$ *and any basis $u_1, \ldots, u_r$ of U_0 can (after suitable renumbering) be completed to a basis $u_1, \ldots, u_r, v_1, \ldots, v_s$ of U such that (u_i, v_i) $(i = 1, \ldots, s \leq r)$ are mutually orthogonal hyperbolic pairs. This basis of U can be completed to a symplectic basis of V; in particular, V is an orthogonal sum of hyperbolic planes and so* $\dim V$ *is even.*

Proof. Let U, U_0 be as stated and let $u_1, \ldots, u_r$ be a basis of U_0. If $U_0 \neq U$, then any vector of $U \backslash U_0$ cannot be orthogonal to all of $u_1, \ldots, u_r$, so we can choose $v_1 \in U$ such that $b(u_i, v_1) = 0$ for all i except one, say (by renumbering the u_i), $b(u_1, v_1) \neq 0$. On multiplying v_1 by a suitable scalar we may assume that $b(u_1, v_1) = 1$. If $\langle U_0, v_1 \rangle \neq U$, we can repeat the process and after a finite number of steps we reach the required basis of U. By the maximality of U_0 we have $s \leq r$, hence $\dim U = r + s \leq 2r$.

Now if $s < r$ we can in V find v_r such that $b(u_i, v_r) = \delta_{ir}$, because V is regular, and continuing in this way we obtain $u_1, \ldots, u_r, v_1, \ldots, v_r$, a symplectic basis for a subspace W of V. If $W \neq V$, we have $V = W \perp W^\perp$ and by induction on $\dim V$ we can find a symplectic basis of $W^\perp$ which together with the basis of W already found is a symplectic basis of V. As sum of hyperbolic planes V has even dimension. ∎

The isometries of V are also called *symplectic mappings*; they form a group $\mathbf{Sp}_{2m}(k)$, called the *symplectic group*. If we refer all our mappings to a fixed symplectic basis, we see that the symplectic mappings are described by matrices P such that

$$PJP^T = J, \quad \text{where } J \text{ is as in (8.10.2).} \tag{8.10.3}$$

A matrix P satisfying (8.10.3) is said to be *symplectic.*

In a special case the symplectic group is easily identified:

Proposition 8.10.2. *For any field k,* $\mathbf{Sp}_2(k) \cong \mathbf{SL}_2(k)$.

Proof. Let u, v be a symplectic basis. Any linear transformation has the form

$$u \mapsto u' = au + bv,$$

$$v \mapsto v' = cu + dv,$$

and this will be symplectic iff $b(u', v') = 1$, i.e. $ad - bc = 1$. Hence $\mathbf{Sp}_2(k)$ consists precisely of the matrices with determinant 1, as claimed. ■

As for linear groups we can define the *projective* symplectic group $\mathbf{PSp}_{2m}(k) = \mathbf{Sp}_{2m}(k)/C$, where C is the centre; it can be shown that $C = \{\pm 1\}$ (see FA Section 3.6).

Let us consider some examples of symplectic matrices. We have

$$S_P = \begin{pmatrix} P & 0 \\ 0 & (P^T)^{-1} \end{pmatrix}, \quad P \in \mathbf{GL}_m(k), \tag{8.10.4}$$

and

$$R_Q = \begin{pmatrix} I & Q \\ 0 & I \end{pmatrix}, \quad \text{where } Q \in k_m, Q^T = Q. \tag{8.10.5}$$

It is easily verified that S_P, R_Q, R_Q^T all lie in $\mathbf{Sp}_{2m}(k)$. In fact these matrices form a generating set:

Theorem 8.10.3. *The symplectic group* $\mathbf{Sp}_{2m}(k)$ *is generated by the matrices* S_P $(P \in \mathbf{GL}_m(k))$ *and* R_Q, R_Q^T $(Q \in k_m, Q^T = Q)$.

Proof. Consider the general matrix in $\mathbf{Sp}_{2m}(k)$:

$$F = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \quad A, B, C, D \in k_m. \tag{8.10.6}$$

The condition $F \in \mathbf{Sp}_{2m}(k)$ means that $FJF^T = J$. On writing this out, we obtain

$$AB^T = BA^T, CD^T = DC^T, AD^T - BC^T = I, \tag{8.10.7}$$

so if A and C are regular, $A^{-1}B$ and $C^{-1}D$ must be symmetric. Then it is easily verified that

$$F = R_{CA^{-1}}^T R_{BA^T} S_A. \tag{8.10.8}$$

In the general case we use induction on the nullity of A, i.e. $m - r$, where r is the rank of A. By left and right multiplication by suitable matrices S_P we can reduce F to the form

$$F_1 = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix}, \quad \text{where } A_1 = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}.$$

The condition that $B_1 A_1^T$ is symmetric means that B_1 has the form

$$B_1 = \begin{pmatrix} G & H \\ 0 & K \end{pmatrix}, \quad \text{where } G \text{ is symmetric.}$$

Since F_1 is non-singular, it follows that $K \neq 0$. If we multiply F_1 on the right by R_Q^T, where $Q = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$, partitioned in conformity with A_1, we obtain

$$F_2 = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix}, \quad \text{where } A_2 = \begin{pmatrix} I & H \\ 0 & K \end{pmatrix},$$

and this has smaller nullity than A. Now the result follows by induction. ■

Exercises

1. Show that $\mathbf{Sp}_{2m}(k)$ is generated by S_P ($P \in \mathbf{GL}_m(k)$), R_Q ($Q \in k_m$, $Q^T = Q$) and J.
2. Show that $\mathbf{PSp}_2(\mathbf{F}_2) \cong \mathrm{Sym}_3$, $\mathbf{PSp}_2(\mathbf{F}_3) \cong \mathrm{Alt}_4$.
3. Let V be a 4-dimensional symplectic space over $\mathbf{F}_2$ and define a *pentagon* as a 5-tuple of vectors $u_1, \ldots, u_5$ satisfying $b(u_i, u_j) = 1$ for $i \neq j$. Show that (i) every pentagon contains 20 hyperbolic pairs, and (ii) every hyperbolic pair occurs in exactly one pentagon. Deduce that there are exactly 6 pentagons. Use the results to show that $\mathbf{Sp}_4(\mathbf{F}_2) \cong \mathrm{Sym}_6$.
4. Verify directly that if $F \in \mathbf{Sp}_{2m}(k)$, where F is as in (8.10.6), and A is invertible, then CA^{-1} is symmetric (this is implicit in the formula (8.10.8) for F).
5. Show that $\mathbf{Sp}_{2m}(k)$ contains a subgroup isomorphic to $\mathbf{GL}_m(k)$. (Hint. Consider a maximal totally isotropic subspace.)

8.11 Quadratic Forms in Characteristic Two

In the treatment of quadratic forms we have at various points used the possibility of dividing by 2; we shall now consider the case of characteristic 2, where this is no longer possible. In particular, there is now no longer an equivalence between quadratic forms and symmetric bilinear forms.

Let us first suppose that we have a space V over a field of characteristic 2, with a symmetric bilinear form B. If B is *alternating*, i.e. $\mathrm{B}(x, x) = 0$ for all $x \in V$, then the rank of B is even, equal to $2m$ say, and in a suitable basis the form has the matrix

$$\begin{pmatrix} 0 & I_m & 0 \\ I_m & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}; \tag{8.11.1}$$

by Theorem 8.10.1. Since the property of being alternating is clearly independent of the choice of basis, it follows that we cannot transform B to diagonal form, unless

it is identically zero. But for any symmetric bilinear form that is non-alternating, such a transformation exists:

Proposition 8.11.1. *Let V be a vector space over a field k of characteristic 2, with a symmetric bilinear form B which is not alternating. Then V has an orthogonal basis.*

Proof. By hypothesis there exists $u_1 \in V$ such that $B(u_1, u_1) \neq 0$. We have $V = \langle u_1 \rangle \perp u_1^\perp$ and we can use induction on dim V unless B is alternating on $u_1^\perp = U$. In that case we can take a basis in U so that the matrix takes the form (8.11.1). Consider the subspace of V spanned by u_1 and a hyperbolic plane in U (which will exist unless the form is zero on U). Its matrix is

$$\begin{pmatrix} c & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

where $c = B(u_1, u_1) \neq 0$. Let u_1, u_2, u_3 be the corresponding basis and put $v_1 = u_1 + u_2 + u_3$; then $B(v_1, v_1) = c \neq 0$ and $v_1^\perp$ has the basis $v_2 = u_1 + cu_2$, $v_3 = u_1 + cu_3$. The matrix in this basis is

$$\begin{pmatrix} c & 0 & 0 \\ 0 & c & c + c^2 \\ 0 & c + c^2 & c \end{pmatrix}$$

and the earlier argument shows that this can be diagonalized; now the diagonalization can be completed by induction on dim V. ∎

This result, together with the earlier remark on alternating forms deals with the case of bilinear forms, and we now turn to the case of a quadratic form q. From it we obtain as before a symmetric bilinear form B, given by

$$B(x, y) = q(x + y) - q(x) - q(y), \tag{8.11.2}$$

but in characteristic 2 this is always alternating: $B(x, x) = 0$, so it cannot be used to reconstruct q. Nevertheless it is possible to define $q(x)$ in terms of a bilinear form $b(x, y)$, but this form cannot always be chosen to be symmetric.

Given a quadratic form q and its derived bilinear form B given by (8.11.2), we define a bilinear form b in terms of a basis $u_1, \ldots, u_n$ of V by

$$b(u_i, u_j) = \begin{cases} B(u_i, u_j) & \text{if } i < j, \\ q(u_i) & \text{if } i = j, \\ 0 & \text{if } i > j. \end{cases} \tag{8.11.3}$$

Together with bilinearity this defines b on V and we have

$$b(x, x) = q(x) \quad \text{for all } x \in V, \tag{8.11.4}$$

as is easily checked. We define orthogonality for q in terms of the alternating form B given by (8.11.2). If this form B is regular, the form q is said to be *non-defective.*

In that case the dimension is even: $n = 2m$, and we can find a *symplectic* basis $u_1, \ldots, u_m, v_1, \ldots, v_m$; thus $B(u_i, v_i) = B(v_i, u_i) = 1 (i = 1, \ldots, m)$, while all other products are zero.

If the subspace spanned by the pair u_i, v_i is denoted by V_i, then we have

$$V = V_1 \perp \ldots \perp V_m,$$

and the form b defined by (8.11.3) has the matrix $P = P_1 \oplus \ldots \oplus P_m$, where

$$P_i = \begin{pmatrix} a_i & 1 \\ 0 & b_i \end{pmatrix}, \quad a_i = q(u_i), b_i = q(v_i).$$

We remark that the matrix P is not uniquely determined by the given basis, since $P + N$, for any alternating matrix N, determines the same form. However, P is the unique upper triangular matrix for this basis.

Since the matrix of a quadratic form is only determined up to a summand which is alternating, its determinant is not an invariant, but there is an invariant to take its place, which was introduced by Cahit Arf in 1940 and is known as the Arf invariant. To define it we remark that for any field k of characteristic 2 the set $\{x + x^2 | x \in k\}$ is a subgroup of its additive group, denoted by $\wp(k)$ (see Section 11.10 below).

Theorem 8.11.2 (C. Arf). *Let V be a vector space over a field k of characteristic 2, with a non-defective quadratic form q. Then for any symplectic basis $u_1, \ldots, u_m, v_1, \ldots, v_m$ of V the expression*

$$\delta(q) = \sum q(u_i) q(v_i) \tag{8.11.5}$$

is unique modulo $\wp(k)$.

The element of $k/\wp(k)$ defined by $\delta(q)$ is called the *Arf invariant* of the form q.

Proof. The matrix of q relative to the given basis is $\begin{pmatrix} A & I \\ 0 & B \end{pmatrix}$, where A, B are diagonal matrices, and $\delta(q) = \mathrm{tr}(AB)$. If we add an alternating matrix, A, B are replaced by $A' = A + M$, $B' = B + N$, where M, N are alternating. Now M, N have zeros on the main diagonal; so do AN, MB and $\mathrm{tr}(MN) = 0$, hence $\mathrm{tr}(AB) = \mathrm{tr}(A'B')$.

Thus $\delta(q)$ is independent of the matrix chosen for q in the given basis and it remains to show that it is unchanged mod $\wp(k)$ by a change of symplectic basis. It is enough to verify this for the generating set of $\mathbf{Sp}_{2m}(k)$ given in Theorem 8.10.3. Consider first S_P; we have

$$\begin{pmatrix} P & 0 \\ 0 & P^{-T} \end{pmatrix} \begin{pmatrix} A & I \\ 0 & B \end{pmatrix} \begin{pmatrix} P^T & 0 \\ 0 & P^{-1} \end{pmatrix} = \begin{pmatrix} PAP^T & I \\ 0 & P^{-T}BP^{-1} \end{pmatrix},$$

and $\mathrm{tr}(PAP^T . P^{-T}BP^{-1}) = \mathrm{tr}(AB)$. For R_Q we have

$$\begin{pmatrix} I & Q \\ 0 & I \end{pmatrix} \begin{pmatrix} A & I \\ 0 & B \end{pmatrix} \begin{pmatrix} I & 0 \\ Q & I \end{pmatrix} = \begin{pmatrix} A + Q + QBQ & I + QB \\ BQ & B \end{pmatrix}.$$

Bearing in mind the relation $\mathrm{tr}(C^2) = \mathrm{tr}(C)^2$ (in characteristic 2), we find that $\mathrm{tr}((A + Q + QBQ)B) = \mathrm{tr}(AB + QB + (QB)^2) = \mathrm{tr}(AB) + \mathrm{tr}(QB) + \mathrm{tr}(QB)^2$. Hence

R_Q does not change the residue class of $\delta(q)$, and a similar argument applies to R_Q^T. Since $\mathbf{Sp}_{2m}(k)$ is generated by all S_P, R_Q, R_Q^T, this shows $\delta(q)$ to be invariant. ■

Exercises

In these exercises the ground field is assumed to have characteristic 2.

1. Show that over a perfect field any diagonal quadratic form of dimension > 1 is singular.
2. Let J be the matrix (8.11.1) for a regular form (of dimension $2m$). Show that any isometry is given by a matrix P satisfying $PJP^T = J$.
3. Let V be an inner product space. Show that the set $I = \{x \in V | q(x) = 0\}$ is a subspace containing the radical and give examples where these two spaces are distinct. Prove Theorem 8.2.2 in the case $I = V^{\perp}$.
4. Let q be a 2-dimensional quadratic form which is non-defective. If there exists $u \neq 0$ such that $q(u) = 0$, show by choosing a symplectic basis including u, that q is equivalent to the form x_1x_2.
5. Verify that $\delta(q_1 \perp q_2) = \delta(q_1) + \delta(q_2)$.
6. Show that a two-dimensional form q is isotropic iff $\delta(q) = 0$.
7. Show that a two-dimensional form is determined up to isometry by its Arf invariant and a single value, i.e. two forms q_1, q_2 are isometric iff $\delta(q_1) = \delta(q_2)$ and $q_1(x_1) = q_2(x_2)$ for some vectors $x_1, x_2 \neq 0$.

8. Show that the Clifford algebra of an even-dimensional regular quadratic space over a finite field k is a full matrix ring over k. What happens in odd dimensions?
9. Show that the Clifford algebra $C(V)$ forms an algebra for the multiplication

$$a \wedge b = ab - (-1)^{mn}ba, \quad \text{where } m = \deg a, n = \deg b,$$

and that the subalgebra generated by V is just the exterior algebra on V.
10. Show that an orthogonal transformation of a split space which is the identity on a maximally isotropic subspace is a rotation.
11. Let L/K be a field extension. Show that an ordering on K has an extension to L iff the signature homomorphism on $W(K)$ maps the kernel of the natural homomorphism $W(K) \to W(L)$ to 0. Deduce another proof of Lemma 8.8.6.
12. Let $L = K(\alpha)$ be a quadratic extension, where $\alpha^2 = a \in K$. If V is an anisotropic K-space such that the extended space V_L is isotropic, show that V splits off a subspace isometric to $\langle c, -ac\rangle$, for some $c \in K$. (Hint. Take an isotropic vector $x + \alpha y$ and consider the space spanned by x, y.)
13. (A. Pfister) Let $L = K(\alpha)$ be a quadratic extension, where $\alpha^2 = a \in K$. Show that the kernel of the natural homomorphism $W(K) \to W(L)$ is the ideal generated by $\langle 1, -a\rangle$.
14. Show that for extensions of odd degree, the natural map $W(K) \to W(L)$ is injective.
15. Show that in a field k of characteristic not 2 every element is a square iff $W(k)$ is of order 2.
16. Prove the uniqueness of the limit of a convergent sequence from the definitions.
17. In any ordered field K the subset $A = \{a \in K \mid |a| \leq n \text{ for some } n \in \mathbf{N}\}$ is a subring whose non-units form an ideal $\mathfrak{m}$, the set of inverses of elements of $K \setminus A$. Show that $A/\mathfrak{m}$ has a natural ordering and is isomorphic to a subfield of $\mathbf{R}$.
18. Given a polynomial f over $\mathbf{R}$ and $a < b$ in $\mathbf{R}$, show that there exists $c \in \mathbf{R}$, $a < c < b$, such that $f(b) - f(a) = (b - a)f'(c)$.
19. Let I be the augmentation ideal in $W(K)$. Show that there is a homomorphism from I^2 to the Sylow 2-subgroup of the Brauer group $\mathbf{B}_K$ (the *symbol homomorphism*; Merkuryev showed in 1981 that I^2/I^3 is isomorphic to the Sylow 2-subgroup of $\mathbf{B}_K$).
20. Let A be a symmetric matrix over $\mathbf{R}$ and denote by $f_i(x)$ the principal minor of order i (i.e. formed from the first i rows and columns) of $xI - A$. Show that $f_i g_i - h_i^2 = f_{i+1} f_{i-1}$, where g_i, h_i are certain polynomials in x. (Hint. Use Further Exercise 16 of Chapter 6.)
21. Show that the number of real roots of a polynomial equation $f = 0$ can be determined from the leading terms of a Sturm sequence. Show that the number of non-real roots is the number of sign changes of $1, s_0, \begin{vmatrix} s_0 & s_1 \\ s_1 & s_2 \end{vmatrix}, \ldots$, where the s_i are the power sums of the roots.
22. Let $r = r(x)$ be a rational function of x. At a pole ξ of r define the *index* of r to be 1 if $r(x)$ changes from negative to positive values as x increases through ξ, -1 if $-r$ has index 1 at ξ and 0 otherwise. Write $r = f_1/f_0$, where f_0, f_1 are coprime polynomials and let $f_0, f_1, f_2, \ldots$ be the sequence of negative remainders in the Euclidean algorithm for f_0, f_1. If $\alpha < \beta$ are not poles of r, show that the sum

of the indices of $r(x)$ in $[\alpha, \beta]$ is $\delta(\alpha) - \delta(\beta)$, where δ is defined as in Theorem 8.8.14. Deduce Theorem 8.8.15 as a corollary.

23. Show that $(a^2 + b^2 + c^2 + d^2)(c^2 + d^2) = (ac + bd)^2 + (ad - bc)^2 + (c^2 + d^2)^2$. Deduce that if -1 is a sum of three squares, then it is also a sum of two squares.
24. Show that in a field of characteristic $\neq 2$ every element is a square iff the Witt ring $W(k)$ has just two elements.
25. Let L/K be a finite extension of degree d. Show that there is an injective homomorphism $\mathbf{Sp}_{2m}(L) \to \mathbf{Sp}_{2dm}(K)$.
26. Show that over a perfect field of characteristic 2, a non-defective quadratic form of dimension $2m$ has Witt index at least $m - 1$ and is determined up to isometry by its dimension and its Arf invariant.

9 Valuation Theory

Valuation theory may be described as the study of divisibility (in commutative rings) in its purest form, but that is only one aspect. The general formulation leads to the introduction of topological concepts like completion, which provides a powerful tool. It also emphasizes the parallel with the absolute value on the real and complex numbers. After the initial definitions (in Section 9.1) we shall prove the essential uniqueness of the absolute value on **R** and **C** (in Section 9.2) and go on to describe the p-adic numbers in Section 9.3 and integral elements in Section 9.4, before looking at simple cases of the extension problem in Section 9.5.

9.1 Divisibility and Valuations

Let R be a (commutative) integral domain with field of fractions K; our object will be to study the divisibility on R. We note that the divisibility can be defined on K as well by writing for any $a, b \in K^\times$, $a|b$ (*a divides b*) whenever $b = ac$ for some $c \in R$. This relation is reflexive and transitive, thus it is a preordering of $K^\times$. Moreover, this preordering is preserved by multiplication:

$$a|b \Rightarrow ad|bd \quad \text{for all } a, b, d \in \mathrm{K}^\times,$$

and so K may be regarded as a preordered group. It is not generally ordered, because $a|b$, $b|a$ need not imply $a = b$; but if we define

$$a \sim b \text{ whenever } a|b \text{ and } b|a,$$

then we obtain an equivalence on $K^\times$. Its classes are just the classes of associated elements of $K^\times$, and these classes form a partially ordered group Γ. If $v(a)$ denotes the class of $a \in K$, then $a \mapsto v(a)$ is a homomorphism from K to Γ. We shall write the operation in Γ as addition and put $v(a) \leq v(b)$ or $v(b) \geq v(a)$ to indicate that $a|b$. Then the mapping v satisfies the conditions:

V.1 $v(a), v(b) > \gamma \Rightarrow v(a + b) > \gamma$, for all $\gamma \in \Gamma$;
V.2 $v(ab) = v(a) + v(b)$;
V.3 $a \in R \Leftrightarrow v(a) \geq 0$.

It is convenient to define v on the whole of K, including 0. To achieve this we shall introduce a symbol ∞ satisfying $\infty + \gamma = \infty$, while $\gamma < \infty$ for all $\gamma \in \Gamma$. If we now put $v(0) = \infty$, then **V.1–V.3** hold for all $a, b \in K$ and $\gamma \in \Gamma \cup \{\infty\}$. Of course this is just a formal device, designed to ensure that our formulae hold without exception; in defining v we can ignore $v(0)$ since its value is prescribed.

Given a field K, it is clear that any function v from K to a partially ordered group Γ, satisfying **V.1**, **V.2**, determines the subset $R = \{x \in K | v(x \geq 0\}$, which is easily seen to be a subring, and **V.3** holds for this R. So here we have another way of describing R, which stresses the divisibility in R. In general this offers no advantage over the usual description, but it suggests looking at rings for which Γ has a particularly simple form.

An important simplification is obtained by assuming Γ to be totally ordered. Then **V.1** can be replaced by

V.1′ $v(a + b) \geq \min\{v(a), v(b)\}$.

In this case v is called a *valuation* on K and K is *valuated*. Thus a valuation on K is a mapping $v : K^\times \to \Gamma$ to a totally ordered group Γ, together with $v(0) = \infty$, satisfying **V.1′** and **V.2**. The ring R defined by **V.3** is then called the *valuation ring*, or the ring of *valuation integers*. It is easy to characterize valuation rings directly:

Proposition 9.1.1. *Let K be a field. Then a subring V of K is a valuation ring in K if and only if for any $x \in K^\times$, either $x \in V$ or $x^{-1} \in V$.*

Proof. Let v be a valuation on K and $V = \{x \in K | v(x) \geq 0\}$ be the ring of integers. Then for any $x \in K$, $v(x) + v(x^{-1}) = v(xx^{-1}) = v(1) = 0$; hence either $v(x) \geq 0$ or $v(x^{-1}) \geq 0$, so V satisfies the given conditions. Conversely, let V be a subring of K satisfying this condition and denote by Γ the group of classes of associated elements, defined as above. Then for any value $\gamma \in \Gamma$, $\gamma = v(x)$ for some $x \in K^\times$. Now either $x \in V$ or $x^{-1} \in V$ and accordingly $\gamma \geq 0$ or $-\gamma \geq 0$; this means that Γ is totally ordered. ∎

Frequently the group Γ will be still further restricted. Thus if $\Gamma = \mathbf{R}$, we speak of a *real-valued valuation*. When $\Gamma = \mathbf{Z}$, the simplest non-trivial case, we speak of a *principal valuation*; this is sometimes called a *discrete rank one valuation*.

Every field has the *trivial* valuation, given by

$$v(x) = \begin{cases} 0 & \text{if } x \neq 0 \\ \infty & \text{if } x = 0. \end{cases}$$

Here are other examples, important in what follows:

Example 1. The p-adic valuation on $\mathbf{Q}$. Every rational number can, up to sign, be written uniquely as a product of powers of different primes. Fix a prime p; for any $a \in \mathbf{Q}^\times$ let p^ν (where $\nu \in \mathbf{Z}$) be the exact power of p occurring in a and define $v(a) = \nu$. This is easily seen to be a valuation, called the *p-adic valuation*. For example, taking $p = 3$, we have $v(111) = 1$, $v(10/9) = -2$. We remark that

this valuation is completely determined by its values on **Z**. This is a general property: if K is the field of fractions of an integral domain A, then any valuation defined on A has a unique extension to K (see Exercise 2).

Example 2. On $k(x)$, the field of rational functions in x over a field k, a valuation may be defined by writing any element φ of $k(x)$ in reduced form: $\varphi = f/g$, where f, g are polynomials in x over k which are coprime, and setting $v(\varphi) = \deg g - \deg f$. The integers in this valuation are the rational functions remaining finite at ∞. Other valuations of $k(x)$ are obtained by singling out an irreducible polynomial p over k and defining the value of φ as the exponent of p in a complete factorization of φ. In particular, for $k = \mathbf{C}$ (or any algebraically closed field) any irreducible polynomial has the form $x - \alpha$, for some $\alpha \in \mathbf{C}$ (up to a constant factor) and the integers for the corresponding valuation are the rational functions that remain finite at $x = \alpha$. Thus the valuations we have found correspond to the different places on the Riemann sphere.

In what follows we shall mainly be concerned with principal valuations; although some of the results will hold for the general case, we shall not always mention this fact explicitly.

If v is a principal valuation on a field K, then the image $v(K^\times)$ is a subgroup of **Z**. It is either 0, when v is trivial, or it is of the form $v(K^\times) = r\mathbf{Z}$ for some $r > 0$, and hence isomorphic to **Z**. In that case $(1/r)v(x)$ is again a valuation, this time with value group exactly **Z**. We call the valuation *normalized* if the value group is the whole of **Z** and express the preceding statement by saying: every non-trivial (principal) valuation can be normalized. Nevertheless, it is not always expedient to normalize our valuations; e.g. a normalized valuation on K may be non-trivial on a subfield F and yet not normalized on F.

We now turn to the consequences of the definition. As in every homomorphism of groups, the neutral element is preserved, i.e. $v(1) = 0$. It follows that

$$v(-1) = 0 \quad \text{and } v(-x) = v(x) \quad \text{for all } x, \tag{9.1.1}$$

for $v(-1) + v(-1) = v((-1)^2) = v(1) = 0$, hence $v(-1) = 0$ and now $v(-x) = v(-1) + v(x) = v(x)$.
The following relation will frequently be needed:

Lemma 9.1.2. *For any x, y in a field with a valuation v,*

$$v(x-y) \geq \min\{v(x), v(y)\}, \tag{9.1.2}$$

with equality unless $v(x) = v(y)$.

Proof. From **V.1′** we have $v(x-y) \geq \min\{v(x), v(-y)\} = \min\{v(x), v(y)\}$, by (9.1.1), and this proves (9.1.2). If $v(x) \neq v(y)$, say $v(x) > v(y)$, then $v(y) \geq \min\{v(x), v(x-y)\}$ and $v(x-y) \geq v(y)$ by (9.1.2), hence $v(x-y) = v(y)$. Similarly if $v(y) > v(x)$. ∎

By an easy induction we find that $v\left(\sum_1^n a_i\right) \geq \min\{v(a_1), \ldots, v(a_n)\}$, and hence we obtain, as in Lemma 9.1.2,

$$\text{if } \sum_1^n a_i = 0, \text{ then } v(a_i) = v(a_j) \text{ for some } i \neq j. \tag{9.1.3}$$

If we think of $v(x)$ as indicating the degree of divisibility of x by a certain prime, property (9.1.2) is obvious. Later (in Section 9.2) we shall meet a less obvious interpretation.

We have already seen that associated with every valuation there is a ring

$$V = \{x \in K | v(x) \geq 0\},$$

the ring of *valuation integers*. The set of all non-units in this ring is

$$\mathfrak{p} = \{x \in K | v(x) > 0\};$$

clearly this is an ideal in V, by **V.1′**, **V.2**, and it is maximal, since a proper ideal cannot contain any units. Moreover, it is the unique maximal ideal of V. Hence the quotient ring $V/\mathfrak{p}$ is a field $\overline{K}$, called the *residue class field* of the valuation v.

If v is the trivial valuation, then $V = K$, $\mathfrak{p} = 0$ and the residue class field is just K itself. Leaving this case aside, we may characterize the valuation rings of principal valuations as follows:

Proposition 9.1.3. *Let K be a field with a valuation v and valuation ring V. Then v is a principal valuation if and only if V is a principal ideal domain, and in that case K contains an element p such that*

$$K^\times = \langle p \rangle \times U,$$

where $\langle p \rangle$ is the (multiplicative) cyclic group on p and U is the group of units of V. Thus every element a of $K^\times$ has the form $a = p^n u$, where $n \in \mathbf{Z}$ and $u \in U$, and if v is normalized, then $v(a) = n$; in any case $a \in V$ if and only if $n \geq 0$.

The element p is called a *uniformizer* or also a *prime element* of v; it is determined up to a unit factor.

Proof. Let v be a principal valuation and take an element p of least positive value. By taking v to be normalized we may assume that $v(p) = 1$; hence for any $a \in K^\times$, if $v(a) = n \in \mathbf{Z}$, then $v(ap^{-n}) = v(a) - nv(p) = 0$, so $ap^{-n} = u \in U$, and we can therefore write $a = p^n u$, where $n = v(a)$, $u \in U$. It is clear that the representation is unique once p has been chosen. Moreover, V is principal; all its ideals have the form 0 or (p^n), $n \geq 0$.

Conversely, assume that V is principal; then its maximal ideal can be written in the form (p). We claim that $\cap(p^n) = 0$. For if not, suppose that $\cap(p^n) = (q)$; then $q = a_n p^n (a_n \in V)$ for all $n \geq 0$, hence $qp^{-1} = a_n p^{n-1}$ for all $n \geq 1$, so $qp^{-1} \in \cap(p^n) = (q)$, say $qp^{-1} = qb (b \in V)$. Thus $q(1 - pb) = 0$, but $q \neq 0$, hence $pb = 1$, a contradiction, and this shows that $\cap(p^n) = 0$. Now take $a \in K^\times$; by what has been proved, $a \notin (p^{n+1})$ for some n. Choose the least such n; then $a \in (p^n)$, so

$a = p^n u$, $u \in V$, and here u is a unit, for if $u \in (p)$, we would have $a \in (p^{n+1})$. Thus we have expressed a in the form $a = p^n u$ ($u \in U$), and this form is unique, for if $p^n u = p^m v$, where $u, v \in U$ and $n \geq m$, say, then $p^{n-m} = vu^{-1} \in U$, hence $n = m$, $v = u$. It follows that $v(a) = nv(p)$ for $a \in K$, and v is normalized iff $v(p) = 1$. ■

A valuation ring which is also a principal ideal domain but not a field is called a *principal valuation ring*. This term (due to Mumford) seems preferable to the usual term for such rings: 'discrete rank 1 valuation ring'.

Two valuations v_1, v_2 on a field K are said to be *equivalent* if there is an order-isomorphism θ between their value groups such that $v_2(x) = v_1(x)^\theta$ for all $x \in K$. It is clear from this definition that two valuations have the same valuation ring iff they are equivalent:

Theorem 9.1.4. *On any field K there is a bijection between the equivalence classes of valuations on K and valuation rings in K. In this correspondence principal valuation rings correspond to principal valuations.* ■

We have already briefly mentioned some examples of valuations. Generally, in any unique factorization domain R with field of fractions K, we can write each element of $K^\times$ in the form

$$a = up_1^{\alpha_1} \dots p_r^{\alpha_r} \quad (\alpha_i \in \mathbf{Z}),$$

where $p_1, \dots, p_r$ are pairwise non-associated prime elements of R. We single out a prime p and write $a = p^v a'$, where a' is prime to p (i.e. the exponent of p in the factorization of a' is 0). If we now put $v(a) = v$, this provides a valuation on K associated with the prime p, called again the *p-adic valuation* on K. Here the valuation ring V is the set of all elements $p^n a'$, with $n \geq 0$, i.e. the set of elements with denominator prime to p, while the maximal ideal in V is the set of all elements with numerator divisible by p.

In the particular case $R = \mathbf{Z}$ we get a p-adic valuation for each prime number p (introduced by Kurt Hensel in 1908). The elements of V are then called *p-adic integers*, e.g. 5/24 is a 7-adic integer, but not a 3-adic integer. No other valuations exist on $\mathbf{Q}$, by

Proposition 9.1.5. *The only non-trivial valuations on $\mathbf{Q}$ are the p-adic valuations.*

Proof. By Theorem 9.1.4 it is enough to determine all valuation rings on $\mathbf{Q}$. Let V be a valuation ring on $\mathbf{Q}$, with maximal ideal $\mathfrak{p}$, say. Since $1 \in V$, it follows that $\mathbf{Z} \subseteq V$. Now $\mathbf{Z} \cap \mathfrak{p}$ is a prime ideal in $\mathbf{Z}$, and either $\mathbf{Z} \cap \mathfrak{p} = 0$; then every non-zero element of $\mathbf{Z}$ is a unit in V, so $V = \mathbf{Q}$ and the valuation is trivial. Or $\mathbf{Z} \cap \mathfrak{p} = p\mathbf{Z}$ for some prime number p; then every $n \in \mathbf{Z}$ is either divisible by p or prime to p, and hence a unit in V. This means that we have the p-adic valuation on $\mathbf{Q}$. ■

If we examine what makes this proof work, we find that it depends essentially on the fact that $\mathbf{Z}$ is a principal ideal domain; e.g. the proof also applies to the

polynomial ring $k[x]$ over a field k. Here the valuations of the field of fractions $k(x)$ which are trivial on k are the p-adic valuations for the different irreducible polynomials p in $k[x]$, and one extra valuation is associated with the degree. For let V be a valuation ring in $k(x)$ containing k and assume first that $x \in V$. Then $k[x] \subseteq V$, and the same argument as in the proof of Proposition 9.1.5 shows that the valuation is associated with some irreducible polynomial, because every maximal ideal of $k[x]$ has such a polynomial as generator. If $x \notin V$, then $y = x^{-1} \in V$ and the same conclusion holds with x replaced by y. Moreover, y is a non-unit in V, so if we are given $f = a_0 + a_1 x + \ldots + a_n x^n (a_n \neq 0)$, then $f = (a_0 y^n + \ldots + a_n) y^{-n}$, and so $v(f) = -n$. Thus we get an extra valuation $v(f) = -\deg f$. In particular, when k is algebraically closed, then the valuations correspond to the elements of k, together with a 'point at infinity', as we have seen in the case $k = \mathbf{C}$. The residue class field is k at each point, while the valuation indicates the order of the function at the given place: if $f = (x - \alpha)^n u$ or $f = x^{-n} u$, n indicates the order to which f vanishes at α or at ∞.

Exercises

1. Show that a finite field has only the trivial valuation.
2. Let v be a mapping from a commutative ring R to a totally ordered group Γ (with $v(0) = \infty$) such that **V.1′** and **V.2** hold. Show that R is an integral domain and there is exactly one extension of v to its field of fractions, defined by $v(a/b) = v(a) - v(b)$.
3. Show that the relation $a|b$ in an integral domain R is an ordering iff 1 is the only unit in R. Give examples of such rings.
4. Show that a principal ideal domain with a single maximal ideal is either a principal valuation ring or a field.
5. Show that every automorphism of the additive group of $\mathbf{R}$ which preserves the natural order is of the form $x \mapsto \lambda x$, where $\lambda > 0$. Deduce that two real-valued valuations v_1, v_2 are equivalent iff $v_2(x) = \lambda v_1(x)$ for all $x \in K$ and a fixed $\lambda > 0$.

9.2 Absolute Values

It is possible to interpret a valuation as a distance function. This is a fruitful approach, which enables us to introduce metric notions such as completion. We begin with a quite general definition.

Let R be a commutative ring. An *absolute value* on R is a real-valued function $x \mapsto |x|$ on R such that

A.1 $|x + y| \leq |x| + |y|$ (triangle inequality),
A.2 $|xy| = x|.|y|$,
A.3 $|x| \geq 0$ with equality iff $x = 0$.

Any non-trivial ring R with an absolute value, briefly a *valued* ring, is an integral domain. For $1 \neq 0$ by definition, and if $x, y \neq 0$, then $|xy| = |x|.|y| \neq 0$, hence $xy \neq 0$. Conversely, any integral domain has at least one absolute value: we put

$$|x| = \begin{cases} 1 & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases} \tag{9.2.1}$$

This is called the *trivial* absolute value, all others are *non-trivial*. A field with the trivial absolute value is said to be *discretely* valued, or *discrete*. If K is a field with a real-valued valuation v, then we can define an absolute value on K by putting

$$|x| = 2^{-v(x)}.$$

A.2, **A.3** are immediate, while **A.1** holds in the stronger form:

A.1′ $|x + y| \leq \max\{|x|, |y|\}$ (ultrametric inequality).

As in Lemma 9.1.2 we derive from **A.1′** the consequence

$$|x - y| \leq \max\{|x|, |y|\}, \quad \text{with equality unless } |x| = |y|. \tag{9.2.2}$$

Geometrically this may be expressed by saying that 'all triangles are isosceles'.

An absolute value is called *non-Archimedean* if it satisfies **A.1′**, *Archimedean* otherwise. If $|\ |$ is non-Archimedean, then by setting $v(x) = -\log_2 |x|$ we obtain a valuation; hence the non-Archimedean absolute values just correspond to the real-valued valuations, with the trivial absolute value corresponding to the trivial valuation. An example of an Archimedean absolute value is the usual absolute value on $\mathbf{Q}$, or more generally, the absolute value defined on any Archimedean ordered field, as in Section 8.7.

The following criterion for a non-Archimedean absolute value is often useful. In the proof we shall need an elementary limit: for any real positive α,

$$\lim_{n \to \infty} (1 + n\alpha)^{1/n} = 1. \tag{9.2.3}$$

To prove (9.2.3) we note that $(1 + n\alpha)^{1/n} \geq 1$ for all $n \geq 1$; further, for any $\delta > 0$ we have $(1 + \delta)^n \geq 1 + n\alpha$ for all sufficiently large n (by the binomial theorem), and so $(1 + n\alpha)^{1/n} \leq 1 + \delta$. Since δ was arbitrary, (9.2.3) follows.

Proposition 9.2.1. *For any absolute value $|\ |$ on a field K the following conditions are equivalent:*

(a) $|\ |$ *is non-Archimedean,*
(b) $|n.1| \leq 1$ *for all* $n \in \mathbf{Z}$,
(c) $|n.1|$ *is bounded for all* $n \in \mathbf{Z}$,
(d) $|z| \leq 1 \Rightarrow |1 + z| \leq 1$.

Proof. (a) $\Rightarrow$ (b). If $|\ |$ is non-Archimedean, then $|n.1| \leq \max\{|1|, \ldots, |1|\} = 1$, so (b) holds.

(b) $\Rightarrow$ (c) is clear, and to prove (c) $\Rightarrow$ (d), suppose that $|n.1| \leq M$ for all $n \in \mathbf{N}$. Then for any $z \in K$ such that $|z| \leq 1$, we have

$$|1+z|^n = |(1+z)^n| = \left|\sum \binom{n}{i} z^i\right| \leq (n+1)M.$$

Taking n-th roots, we find that

$$|1+z| \leq (1+n)^{1/n} M^{1/n};$$

letting n tend to ∞ and remembering (9.2.3), we obtain (d).

Finally to prove (d) $\Rightarrow$ (a) take $x, y \in K$. Without loss of generality we may suppose that $x, y \neq 0$. Suppose that $|x| \geq |y|$; then $|y/x| \leq 1$, hence by (d), $|1 + y/x| \leq 1$ and multiplying by $|x|$, we obtain $|x+y| \leq |x| = \max\{|x|, |y|\}$, which is (a). ∎

Since Proposition 9.2.1(c) clearly holds in finite characteristic, we deduce

Corollary 9.2.2. *A field with an Archimedean absolute value must be of characteristic 0.* ∎

If R is a ring with an absolute value $|\ |$, we can define a metric on R by putting $d(x, y) = |x - y|$. This makes R into a metric space; the ring operations are continuous by **A.1**, **A.2**, so that we have a topological ring. We shall indicate how to form the completion of such a ring, rather in the fashion in which $\mathbf{R}$ was formed from $\mathbf{Q}$ in Section 8.7. The same method applies, because the construction in Section 8.7 depended not on the ordering of $\mathbf{Q}$ but only on the ordering of the absolute values. We shall sketch the method, using the fact that $\mathbf{R}$ is already known to be complete.

As in Section 8.7, we shall say that the sequence $\{c_\nu\}$ of elements of R converges to $c \in R$ if $|c_\nu - c| \to 0$ as $\nu \to \infty$. By a *Cauchy sequence* we understand a sequence $\{c_\nu\}$ such that $|c_\mu - c_\nu| \to 0$ as $\mu, \nu \to \infty$. As before we see that every convergent sequence is a Cauchy sequence; if the converse holds, R is said to be *complete* with respect to the absolute value. When R is not complete, we can form the completion by taking the set C of all Cauchy sequences over R and verifying that this is a ring under componentwise addition and multiplication, with the constant sequences ($a_\nu = a$ for all ν) as a subring isomorphic to R. By identifying $a \in R$ with the constant sequence $\{a, a, \ldots\}$, we can embed R in C. The sequences convergent to 0, the *null sequences*, again form an ideal $\mathfrak{n}$ in C; here we use the fact that for every Cauchy sequence $\{c_\nu\}$ the sequence $|c_\nu|$ is bounded. Clearly $R \cap \mathfrak{n} = 0$, so if we set $\overline{R} = C/\mathfrak{n}$, we have a mapping $R \to C \to C/\mathfrak{n} = \overline{R}$, which is an embedding, because the kernel is $R \cap \mathfrak{n} = 0$, by the second isomorphism theorem.

We extend the absolute value to $\overline{R}$ by putting $|\{c_\nu\}| = \lim |c_\nu|$; by the completeness of $\mathbf{R}$ this defines an absolute value on $\overline{R}$, which extends the given absolute value on R. Moreover, R is dense in $\overline{R}$, i.e. every element of $\overline{R}$ is a limit of elements of R. Suppose further that R is a field; then so is $\overline{R}$. For if $c \in \overline{R}$, say $c = \lim c_\nu$, and $c \neq 0$, then as in Section 8.7 we can show that c_ν is bounded for all $\nu > \nu_o$ and $c_\mu^{-1} - c_\nu^{-1} = c_\mu^{-1}(c_\nu - c_\mu)c_\nu^{-1} \to 0$ as $\mu, \nu \to \infty$, so $\{c_\nu^{-1}\}$ $(\nu > \nu_o)$ is again a Cauchy sequence,

convergent to the element c^{-1} of $\overline{R}$. It only remains to show that $\overline{R}$ is complete and is determined up to isomorphism by R; this follows as before. Summing up, we have

Theorem 9.2.3. *Let R be a commutative absolute-valued ring. Then there exists a complete absolute-valued ring $\overline{R}$ and an embedding $R \to \overline{R}$ preserving absolute values, such that the image of R is dense in $\overline{R}$, and $\overline{R}$ is uniquely determined by R, up to (metric) isomorphism. If R is a field, then so is $\overline{R}$.* ∎

The ring $\overline{R}$ is again called the *completion* of R with respect to the absolute value. For example, consider $\mathbf{Q}$ with the p-adic valuation v_p, for a given prime p. The completion is called the *field of p-adic numbers* and is denoted by $\mathbf{Q}_p$. The elements x of $\mathbf{Q}_p$ such that $v_p(x) \geq 0$ form a subring $\mathbf{Z}_p$, the *ring of p-adic integers*. Each element of $\mathbf{Q}_p$ may be written in the form of a series

$$a = \sum_{i=-k}^{\infty} a_i p^i \quad (0 \leq a_i < p). \tag{9.2.4}$$

If $a_i = 0$ for $i < 0$, we have an element of $\mathbf{Z}_p$. The finite series form a subring of $\mathbf{Z}_p$ isomorphic to $\mathbf{Z}$; here (9.2.4) reduces to the expression of an ordinary integer in the base of p.

We note that if the absolute value on a field K corresponds to a principal valuation, then the value groups for K and its completion $\overline{K}$ are the same; for $v(K)$ is a discrete subgroup of $\mathbf{R}$, hence closed, and $v(a) = \lim v(a_\nu)$ whenever $a_\nu \to a \in \overline{K}$. The residue class field likewise is unchanged by completion: let V, V' be the valuation rings and $\mathfrak{p}, \mathfrak{p}'$ be their maximal ideals in $K, \overline{K}$ respectively. Then $V \subseteq V'$, $\mathfrak{p} \subseteq \mathfrak{p}'$; since $V \cap \mathfrak{p}' \supseteq \mathfrak{p}$ and $V \cap \mathfrak{p}' \neq V$, we have $V \cap \mathfrak{p}' = \mathfrak{p}$ by the maximality of the latter, and $V + \mathfrak{p}' = V'$, because any $c \in V'$ can be written $c = c_0 + c_1$, where $c_0 \in V$, $c_1 \in \mathfrak{p}'$. Hence $V'/\mathfrak{p}' = (V + \mathfrak{p}')/\mathfrak{p}' \cong V/(V \cap \mathfrak{p}') = V/\mathfrak{p}$.

We have seen that an absolute value defines a topology. Two absolute values on a field K are said to be *equivalent* if they induce the same topology on K. Any absolute value on a field inducing the discrete topology must be trivial: if $|\ |$ is not trivial, then for some $a \in K$, $|a| \neq 0, 1$, hence $a \neq 0$ and $x = a$ or $x = a^{-1}$ satisfies $|x| < 1$; it follows that $x^n \to 0$. The relation between equivalent absolute values can be described quite explicitly as follows:

Proposition 9.2.4. *Let $|\ |_1, |\ |_2$ be any two non-trivial absolute values on a field K. Then the following conditions are equivalent:*

(a) $|\ |_1$ *is equivalent to* $|\ |_2$,
(b) $|x|_1 < 1 \Rightarrow |x|_2 < 1$ *for all* $x \in K$,
(c) $|x|_1 = |x|_2^{\gamma}$ *for all* $x \in K$ *and some real constant* γ.

The third condition shows that the notion defined here agrees with the definition of equivalence of valuations given in Section 9.1. We also note that not every value of γ in (c) will give an absolute value. Given an absolute value $|\ |$, the function $|x|^{\gamma}$ is again an absolute value for $0 < \gamma \leq 1$, and in certain cases (e.g. if $|\ |$ is non-Archimedean) γ may be taken > 1, but not always.

Proof. (a) $\Rightarrow$ (b). If (a) holds and $|x|_1 < 1$, then $x \to 0$ in the $||_1$-metric, hence also in the $||_2$-metric, and so $|x|_2 < 1$.

(b) $\Rightarrow$ (c). We first show that (b) is in fact symmetric, i.e.

$$|x|_1 < 1 \Leftrightarrow |x|_2 < 1. \tag{9.2.5}$$

If this were not so, then for some $a \in K$, $|a|_2 < 1$ and $|a|_1 \geq 1$, or on writing $b = a^{-1}$, $|b|_1 \leq 1$ and $|b|_2 > 1$. Choose $c \in K$ such that $0 < |c|_1 < 1$; then by (b), $0 < |c|_2 < 1$, hence $|b^n c|_1 < 1$ for all $n \geq 0$, so $|b^n c|_2 < 1$, i.e. $|b|_2^n < |c^{-1}|_2$ for all n, which is possible only if $|b|_2 \leq 1$. This contradiction shows that (9.2.5) holds.

Let us write $f_i(x) = -\log |x|_i$ $(i = 1, 2)$; then (9.2.5) takes the form

$$f_1(x) > 0 \Leftrightarrow f_2(x) > 0, \tag{9.2.6}$$

and we must show that $f_2(x) = cf_1(x)$ for some $c > 0$, bearing in mind that $f_i(xy) = f_i(x) + f_i(y)$, by the definition and **A.2**. Take $a \in K^\times$ such that $f_1(a) > 0$; then for any $x \in K$ and non-zero integers m, n we have

$$\begin{aligned} f_1(x) > (m/n)f_1(a) &\Leftrightarrow f_1(x^n a^{-m}) > 0 \\ &\Leftrightarrow f_2(x^n a^{-m}) > 0 \\ &\Leftrightarrow f_2(x) > (m/n)f_2(a). \end{aligned}$$

Thus for all rational r we have

$$f_1(x) > rf_1(a) \Leftrightarrow f_2(x) > rf_2(a).$$

Letting r tend to $f_1(x)/f_1(a)$ from below, we find that

$$\frac{f_2(x)}{f_2(a)} \geq \frac{f_1(x)}{f_1(a)};$$

by symmetry we have equality here, thus

$$\frac{f_2(x)}{f_1(x)} = \frac{f_2(a)}{f_1(a)} = c,$$

for all $x \in K$, which shows that (c) holds. Finally, to prove (c) $\Rightarrow$ (a), when (c) holds for any $\gamma > 0$, we clearly have $|a_\nu|_1 \to 0 \Leftrightarrow |a_\nu|_2 \to 0$, i.e. (a). ■

Just as this result shows that equivalent absolute values are very much alike, so the next result shows that inequivalent ones are very different.

Theorem 9.2.5 (Approximation-theorem, Artin–Whaples, 1945). *Let $||_1, \ldots, ||_n$ be non-trivial absolute values on a field K which are pairwise inequivalent. Then for any $a_1, \ldots, a_n \in K$ and any $\varepsilon > 0$, there exists $\alpha \in K$ such that*

$$|\alpha - a_i|_i < \varepsilon \quad \text{for } i = 1, \ldots, n.$$

Proof. First we observe that for any absolute value $|\ |$, as $r \to \infty$,

$$\lim \left| \frac{a^r}{1+a^r} \right| = \begin{cases} 0 & \text{if } |a| < 1, \\ 1 & \text{if } |a| > 1. \end{cases} \tag{9.2.7}$$

Our next objective is to find $c \in K$ such that

$$|c|_1 > 1, \quad |c|_i < 1 \quad \text{for } i = 2, \ldots, n. \tag{9.2.8}$$

Since $||_1$ and $||_2$ are inequivalent, we can by Proposition 9.2.4 find $a \in K$ such that $|a|_1 > 1 \geq |a|_2$ and $b \in K$ such that $|b|_1 \leq 1 < |b|_2$. Now $c = ab^{-1}$ satisfies $|c|_1 > 1 > |c|_2$. This is (9.2.8) in case $n = 2$; we may therefore take $n > 2$ and use induction on n. By hypothesis there exists $a \in K$ such that $|a|_1 > 1 > |a|_i$ $(i = 3, \ldots, n)$ and there exists $b \in K$ such that $|b|_1 > 1 > |b|_2$. Either $|a|_2 \leq 1$; then we put $c_r = a^r b$; now $|c_r|_1 > 1 > |c_r|_2$ for all r, and if r is large enough, then $|c_r|_i < 1$ for $i = 3, \ldots, n$ also. Or $|a|_2 > 1$; then we put $c_r = a^r b/(1 + a^r)$. By (9.2.7), as $r \to \infty$, $|c_r|_1 \to |b|_1 > 1$, $|c_r|_2 \to |b|_2 < 1$, $|c_r|_i \to 0$ $(i = 3, \ldots, n)$, hence (9.2.8) is satisfied by $c = c_r$ when r is large enough.

Take c satisfying (9.2.8); from (9.2.7) it follows that the sequence $c^r/(1 + c^r)$ tends to 1 at $||_1$ and to 0 at $||_i$ for $i = 2, \ldots, n$. Thus, given $\delta > 0$ and $1 \leq i \leq n$, we can find $u_i \in K$ such that $|u_i - 1| < \delta$, $|u_i|_j < \delta$ for $j \neq i$. We take such u_i for $i = 1, \ldots, n$ and put $\alpha = \sum a_i u_i$. Then

$$|\alpha - a_i|_i \leq |a_i(u_i - 1)|_i + \sum_{j \neq i} |a_j u_j|_i < \delta n M,$$

where $M = \max_{ij} \{|a_j|_i\}$. So by choosing $\delta < \varepsilon/nM$ we obtain the required element α. ∎

Corollary 9.2.6. *If $||_1, \ldots, ||_r$ are inequivalent non-trivial absolute values, then there is no relation*

$$|x|_1^{r_1} \ldots |x|_n^{r_n} = 1 \quad \text{for all } x \in K,$$

except the trivial one, where $r_1 = \ldots = r_n = 0$.

Proof. If $r_1 \neq 0$, say, choose $x \in K$ so that $|x|_1$ is small and $|x - 1|_i$ is small for $i = 2, \ldots, n$. Then $|x|_i^{r_i}$ is near 1 for $i > 1$ and the given relation cannot hold. ∎

By contrast, for infinitely many absolute values there is a product formula of this form (see Exercise 6).

Later we shall study methods of extending absolute values to extension fields, and it is convenient to treat the existence and uniqueness separately. The uniqueness can be proved under quite general conditions, namely for any finite-dimensional space over a complete valued field. We digress briefly to introduce the necessary definitions.

Let K be a field with an absolute value $|\ |$. A *normed vector space* over K is a vector space V over K with a function $x \mapsto ||x||$ taking values in $\mathbf{R}$ and satisfying the following conditions:

N.1 $||x+y|| \leq ||x|| + ||y||$ for all $x, y \in V$;
N.2 $||\alpha x|| = |\alpha|.||x||$ for $x \in V$, $\alpha \in K$;
N.3 $||x|| \geq 0$, with equality if and only if $x = 0$.

Clearly any field extension, with an absolute value extending that of K, is a normed vector space. Moreover, any finite-dimensional vector space over K has at least one norm, the *cubical norm*, defined as follows. Pick a basis $e_1, \ldots, e_n$ in V and for $x = \sum \alpha_i e_i$ define $||x|| = \max\{|\alpha_1|, \ldots, |\alpha_n|\}$. It is straightforward to verify that this is a norm. We can define convergence, Cauchy sequences and completeness in V as in the case of fields (see Section 8.7), using the norm in place of the absolute value. Then we have

Proposition 9.2.7. *Let V be a normed space of finite dimension over a complete absolute-valued field K. Then V is complete and its topology is induced by any cubical norm on V; thus the topology on V is uniquely determined.*

Proof. We first show that V is complete in the cubical norm. Let $e_1, \ldots, e_n$ be a basis of V and let $x_\nu = \sum \alpha_{i\nu} e_i$ be a Cauchy sequence: $||x_\mu - x_\nu|| \to 0$, hence $|\alpha_{i\mu} - \alpha_{i\nu}| \to 0$ for $i = 1, \ldots, n$. By the completeness of K, $\lim \alpha_{i\nu} = \alpha_i$ exists in K. Put $x = \sum \alpha_i e_i$; then $|\alpha_i - \alpha_{i\nu}| \to 0$ as $\nu \to \infty$ $(i = 1, \ldots, n)$, hence $||x - x_\nu|| \to 0$, i.e. $x_\nu \to x$, and so V is indeed complete in the cubical norm defined by the e_i.

Now let $N(x)$ be any norm on V; we must show that this defines the same topology as the cubical norm $||\ ||$. We use induction on $\dim V$, which is finite by hypothesis. For $\dim V = 0$ there is nothing to prove, so let $\dim V > 0$. In the first place, if $x = \sum \alpha_i e_i$, then

$$N(x) \leq \sum N(\alpha_i e_i) = \sum |\alpha_i|.N(e_i) \leq ||x||.\Big(\sum N(e_i)\Big).$$

Thus $N(x) \leq c||x||$ for a fixed c, hence convergence in the cubical topology entails convergence in the N-topology (i.e. the cubical topology is finer than the N-topology). Conversely, let $\{x_\nu\}$ be a sequence such that $N(x_\nu) \to 0$ and suppose that $||x_\nu||$ does not tend to 0. Write $x_\nu = \sum \alpha_{i\nu} e_i$, so that $||x_\nu|| = \max_i \{|\alpha_{i\nu}|\}$. If $|\alpha_{i\nu}| \to 0$ for each i, we would have $||x_\nu|| \to 0$. This is not so, hence for some i, say $i = 1$, $|\alpha_{i\nu}|$ does not tend to 0. Going over to a subsequence, we may assume that $|\alpha_{1\nu}| \geq \varepsilon$ for some $\varepsilon > 0$ and all ν. We write $y_\nu = (\alpha_{1\nu})^{-1} x_\nu = \sum \beta_{i\nu} e_i$, where $\beta_{1\nu} = 1$ by construction. Then

$$N(y_\nu) = |\alpha_{1\nu}|^{-1} N(x_\nu) \leq \varepsilon^{-1} N(x_\nu),$$

hence again $N(y_\nu) \to 0$, and so $\sum_2^n \beta_{i\nu} e_i \to -e_1$ in the N-topology. But $e_2, \ldots, e_n$ span an $(n-1)$-dimensional subspace; by induction on n this has a unique topology and is complete, hence closed. Thus e_1 belongs to the subspace spanned by $e_2, \ldots, e_n$, a contradiction. Therefore $||x_\nu|| \to 0$ and the result follows. ∎

Any discrete space is necessarily complete; hence we have

Corollary 9.2.8. *Any finite-dimensional normed space over a discrete field is itself discrete.* ∎

As a further consequence we have the uniqueness property of extensions.

Theorem 9.2.9. *Let K be a complete valued field and L be a finite algebraic extension field of K. Then the absolute value on K has at most one extension to L, and L is complete.*

Proof. Let $|\,|$ be the absolute value on K and let $|\,|_1, |\,|_2$ be two absolute values on L extending $|\,|$. By Proposition 9.2.7 both induce the same topology on L, and L is complete in this topology. If $|\,|$ is trivial, then K is discrete, hence so is L and $|\,|_1, |\,|_2$ are both trivial. Otherwise we have $|x|_1 = |x|_2^\gamma$ for all $x \in L$ and some γ, by Proposition 9.2.4, and there exists $a \in K$ such that $|a| \neq 0, 1$. Taking $x = a$ we see that $\gamma = 1$, hence $|\,|_1 = |\,|_2$. ∎

Any absolute value on a finite field must be trivial, since the topology is discrete. More directly, in a field of q elements, every $x \neq 0$ satisfies $x^{q-1} = 1$, hence $|x|^{q-1} = 1$ and so $|x| = 1$. More generally, this clearly holds for any algebraic extension of a finite field.

We end this section by determining all complete valued fields with an Archimedean absolute value, following Ostrowski (with simplifications by E. Artin). It turns out that besides the two well-known examples of $\mathbf{R}$ and $\mathbf{C}$ there are no others.

Theorem 9.2.10 (Ostrowski's first theorem). *Any non-trivial absolute value on $\mathbf{Q}$ is equivalent either to the usual absolute value or to a p-adic valuation.*

Proof. Let f be any absolute value on $\mathbf{Q}$. If f is non-Archimedean, then the corresponding valuation must be p-adic for some prime p, as we have seen in Proposition 9.1.5; so we may assume that f is Archimedean. We observe that for any $n \in \mathbf{N}$, $f(n) \leq f(1) + \ldots + f(1) = n$, hence

$$f(n) \leq n \quad \text{for all } n \in \mathbf{Z}. \tag{9.2.9}$$

Given any integers $m, n > 1$, we express m in the base of n:

$$m = a_0 + a_1 n + \ldots + a_\nu n^\nu, \quad \text{where } 0 \leq a_i < n, a_\nu \neq 0. \tag{9.2.10}$$

In particular, $m \geq n^\nu$ and so

$$\nu \leq \frac{\log m}{\log n}. \tag{9.2.11}$$

By (9.2.10), $f(m) \leq f(a_0) + f(a_1)f(n) + \ldots + f(a_n)f(n)^\nu$, but $f(a_i) \leq a_i < n$, therefore

$$f(m) \leq n[1 + f(n) + \ldots + f(n)^\nu].$$

According as $f(n) \leq 1$ or > 1, we replace all terms in the brackets by the first or last term and so obtain

$$f(m) \leq n(1 + \nu) \max\{1, f(n)^\nu\}.$$

By (9.2.11) we can rewrite this as

$$f(m) \leq n\left(1+\frac{\log m}{\log n}\right).\max\{1, f(n)^{\log m/\log n}\}.$$

In this formula replace m by m^r and take r-th roots:

$$f(m) \leq n^{1/r}\left(1+\frac{r.\log m}{\log n}\right)^{1/r}.\max\{1, f(n)^{\log m/\log n}\}.$$

Letting $r \to \infty$ and remembering (9.2.3), we obtain

$$f(m) \leq \max\{1, f(n)^{\log m/\log n}\}. \tag{9.2.12}$$

Since f is Archimedean, we know that $f(n_0) > 1$ for some n_0, by Proposition 9.2.1. Taking $m = n_0$, we see that $f(n) > 1$ for all $n > 1$, so we can rewrite (9.2.12) as

$$f(m) \leq f(n)^{\log m/\log n}$$

i.e.

$$\frac{\log f(m)}{\log m} \leq \frac{\log f(n)}{\log n}.$$

By symmetry we have equality, say both sides equal γ. Then $\log f(m) = \gamma \log m$ for all $m \geq 1$, i.e. $f(m) = m^\gamma$. It follows that $f(-m) = f(m) = m^\gamma$ and $f(m/n) = (m/n)^\gamma$, hence f is equivalent to the usual absolute value on $\mathbf{Q}$, as claimed. ■

It remains to determine the Archimedean absolute values on arbitrary fields. As a complete proof from first principles is rather long (and the result is not needed elsewhere in this book), we shall merely indicate how it follows from standard results in analysis.

Theorem 9.2.11 (Ostrowski's second theorem). *Let K be a field with an Archimedean absolute value for which K is complete. Then $K \cong \mathbf{R}$ or $K \cong \mathbf{C}$, and the absolute value is equivalent to the usual absolute value.*

Proof. Let f be the given absolute value on K. By Corollary 9.2.2, K has characteristic 0 and so contains $\mathbf{Q}$ as a subfield. Moreover, since f is Archimedean, $f(n) > 1$ for some n, so f is non-trivial on $\mathbf{Q}$ and by Theorem 9.2.10 there exists $\alpha \in \mathbf{R}$ such that

$$|x| = f(x)^\alpha \text{ for all } x \in \mathbf{Q}. \tag{9.2.13}$$

We shall use (9.2.13) to define $|\ |$ over the whole of K; this extends the usual absolute value and, moreover, it satisfies the triangle inequality on K. For we have

$$f(x+y) \leq f(x)+f(y) \leq 2\max\{f(x), f(y)\},$$

hence by (9.2.13),

$$|x+y| \leq 2^\alpha.\max\{|x|, |y|\}.$$

By repeated application we have for any $x_i \in K$, $1 \leq i \leq m = 2^r$,

$$|x_1 + \ldots + x_m| \leq (2^r)^\alpha . \max\{|x_1|, \ldots, |x_m|\}. \tag{9.2.14}$$

Given $x_1, \ldots, x_n \in K$, we can choose r so that $2^{r-1} \leq n < 2^r$. If we apply (9.2.14) with this value of r, we get

$$|x_1 + \ldots + x_n| \leq (2n)^\alpha . \max\{|x_1|, \ldots, |x_n|\}.$$

In particular,

$$\begin{aligned} |a+b|^n &= \left|\sum \binom{n}{i} a^i b^{n-i}\right| \\ &\leq [2(n+1)]^\alpha . \max\left\{|a|^n, \binom{n}{1}|a|^{n-1}|b|, \ldots, |b|^n\right\} \\ &\leq [2(n+1)]^\alpha (|a| + |b|)^n. \end{aligned}$$

Taking n-th roots, we have

$$|a+b| \leq [2(n+1)]^{\alpha/n}(|a| + |b|),$$

and letting $n \to \infty$, we obtain by (9.2.3) and the fact that $\lim 2^{\alpha/n} = 1$,

$$|a+b| \leq |a| + |b| \quad \text{for all } a, b \in K.$$

Thus we have an extension of $||$ to K satisfying the triangle inequality as well as being multiplicative. Since K is complete, it must contain $\mathbf{R}$. If $x^2 + 1 = 0$ has a root in K, we adjoin this root to $\mathbf{R}$ and find that $\mathbf{C} \subseteq K$, and we have to show that $K = \mathbf{C}$ or $K(i) = \mathbf{C}$ respectively, where i is a root of $x^2 + 1 = 0$. In either case we have a complete normed space over $\mathbf{C}$, where in the second case the norm is given by

$$||x + iy|| = (|x|^2 + |y|^2)^{1/2}.$$

It is easily seen that this is indeed a norm on $K(i)$. By Proposition 9.2.7 $K(i)$ is then a complete space and it only remains to show that a field K which is a complete normed space over $\mathbf{C}$ is $\mathbf{C}$ itself (Gelfand–Mazur theorem). We shall outline the proof, referring e.g. to Rudin (1966) for details.

Suppose that $c \in K \backslash \mathbf{C}$; then $c - \alpha \neq 0$ for all $\alpha \in \mathbf{C}$ and so $(c-\alpha)^{-1}$ is an element of K for all $\alpha \in \mathbf{C}$. Consider linear functionals on K, i.e. $\mathbf{C}$-linear mappings from K to $\mathbf{C}$. Such a functional f is bounded if $|f(x)| \leq \gamma ||x||$ for some constant γ depending on f. When K is finite-dimensional over $\mathbf{C}$, it is algebraic over $\mathbf{C}$ and hence $\mathrm{K} = \mathbf{C}$, because $\mathbf{C}$ is algebraically closed. In the infinite-dimensional case f is bounded iff it is continuous, or equivalently, if the hyperplane $\ker f$ is closed. Given $b \in K^\times$, we can find f such that $f(b) \neq 0$, by constructing a closed hyperplane not containing b, and this can be done by Zorn's lemma (this is the assertion of the Hahn–Banach theorem).

Now return to $(c-\alpha)^{-1}$ and consider $f((c-\alpha)^{-1})$, for a bounded linear functional f. This is an analytic function of α for all $\alpha \in \mathbf{C}$ and it is bounded as $\alpha \to \infty$, hence by Liouville's theorem it is a constant, so for any $\alpha, \beta \in \mathbf{C}$,

$$f((c-\alpha)^{-1} - (c-\beta)^{-1}) = f((c-\alpha)^{-1}) - f((c-\beta)^{-1}) = 0.$$

Since this holds for all f, we must have $(c-\alpha)^{-1} - (c-\beta)^{-1} = 0$, which is a contradiction when $\alpha \neq \beta$. Hence $K = \mathbf{C}$ and the theorem follows. ■

Exercises

1. Show that an absolute value $N(x)$ on a field satisfies the triangle inequality whenever $N(x) \leq 1 \Rightarrow N(1+x) \leq 2$, and the ultrametric inequality if $N(x) \leq 1 \Rightarrow N(1+x) \leq 1$.
2. In a non-Archimedean metric show that two open balls either are disjoint or one is contained in the other. Does this hold for closed balls?
3. Let $||$ be a non-Archimedean absolute value. Show that if an infinite series $\sum a_n$ is convergent and $|a_n| < |a_1|$ for all $n > 1$, then $|\sum a_n| = |a_1|$ (this is called the *principle of domination*).
4. Show that on a field with a non-Archimedean absolute value, an infinite series is convergent iff the n-th term tends to 0 (sometimes called 'the Freshman's dream').
5. Show that the trivial valuation on a field has no non-trivial extension to an algebraic extension field.
6. Show that the absolute values on $\mathbf{Q}$ can be normed so that $\prod_i |a|_i = 1$ for all $a \neq 0$. Do the same for $k(x)$, where k is a finite field.
7. Show that every non-trivial subgroup of $\mathbf{R}$ is either of the form $\alpha\mathbf{Z}$, where $\alpha > 0$, or is dense in $\mathbf{R}$. Deduce that every discrete $\mathbf{R}$-valued valuation is principal.
8. Let F be a complete field for an Archimedean absolute value. Show directly (without using Ostrowski's theorem) that $1 + a$ is a square whenever $4|a| < |4|$.

9.3 The *p*-adic Numbers

The definition of the p-adic field $\mathbf{Q}_p$ as a completion of $\mathbf{Q}$ in Section 9.2 suggests that the methods of analysis may be applicable. This is indeed the case, as was first observed by Kurt Hensel, and in this section we shall look at ways of solving equations over $\mathbf{Q}_p$. Throughout this section, p will be a fixed prime number, arbitrary except when otherwise specified. The p-adic valuation on $\mathbf{Q}_p$ will be denoted by v or v_p.

Every natural number a has a p-adic expansion, i.e. we can write it in the base p:

$$a = a_0 + a_1 p + a_2 p^2 + \ldots + a_r p^r \quad (0 \leq a_i < p). \tag{9.3.1}$$

In terms of the valuation v_p we can think of a as being obtained by successive

approximations: $a_0, a_0 + a_1 p, \ldots, a$. Similarly the general element of $\mathbf{Z}_p$ can be written as an infinite series

$$b = b_0 + b_1 p + b_2 p^2 + \ldots; \tag{9.3.2}$$

it is the limit of the sequence of integers formed by its partial sums $b_0, b_0 + b_1 p, \ldots$. For the elements of $\mathbf{Q}_p$ we also have to allow a finite number of negative powers of p:

$$c = c_{-k} p^{-k} + c_{1-k} p^{1-k} + \ldots + c_{-1} p^{-1} + c_0 + c_1 p + \ldots. \tag{9.3.3}$$

For example, when $p = 7$, we have

$$-1 = 6 + 6.7 + 6.7^2 + \ldots. \tag{9.3.4}$$

Rational numbers are in $\mathbf{Z}_7$ as long as their denominators are prime to 7. Thus to find the 7-adic expansion of 1/2 we have $1/2 = (-6+7)/2 = -3 + (1/2).7 = -3 - 3.7 - 3.7^2 - \ldots$, hence

$$\frac{1}{2} = \frac{8-7}{2} = 4 - \frac{1}{2}.7 = 4 + 3.7 + 3.7^2 + \ldots. \tag{9.3.5}$$

With the help of this expansion we can solve $x^2 = 2$ in $\mathbf{Z}_7$, using the binomial theorem. We have $x^2 = 2$ precisely when $(2x)^2 = 8 = 1 + 7$, hence $2x = (1+7)^{1/2}$, and

$$x = \frac{1}{2}(1+7)^{1/2} = \sum \frac{1}{2}\binom{1/2}{n}.7^n = 4 + 5.7 + 4.7^2 + 5.7^4 + \ldots.$$

The latter is an element of $\mathbf{Z}_7$, as we see by using (9.3.5). We shall soon meet a systematic way of solving such equations.

Clearly $\mathbf{Z}_p$ is a principal valuation ring, the ideal of non-units being generated by p. If we adjoin an inverse for p, we obtain $\mathbf{Q}_p = \mathbf{Z}_p[p^{-1}]$. Let us write U for the group of units of $\mathbf{Z}_p$; then every $x \in \mathbf{Q}_p$ can be uniquely written in the form

$$x = p^v u, \quad \text{where } v = v(x) \text{ and } u \in U,$$

and of course $x \in \mathbf{Z}_p$ iff $v \geq 0$.

If in (9.3.2) we ignore all terms after p^{n-1}, we obtain an integer mod p^n; thus we have a natural homomorphism $\varepsilon_n : \mathbf{Z}_p \to \mathbf{Z}/p^n$ which consists in mapping b, given by (9.3.2), to the residue class (mod p^n) of $b_0 + b_1 p + \ldots + b_{n-1} p^{n-1}$. Clearly ε_n is surjective, with kernel $p^n \mathbf{Z}_p$, so we have an exact sequence

$$0 \to \mathbf{Z}_p \xrightarrow{p^n} \mathbf{Z}_p \xrightarrow{\varepsilon_n} \mathbf{Z}/p^n \to 0, \tag{9.3.6}$$

where p^n indicates multiplication by p^n. In particular this shows that

$$\mathbf{Z}_p/p^n\mathbf{Z}_p \cong \mathbf{Z}/p^n. \tag{9.3.7}$$

Let us consider a polynomial f with coefficients in $\mathbf{Q}_p$. On multiplying by a suitable power of p we may assume the coefficients of f to lie in $\mathbf{Z}_p$. Moreover, we can always arrange for f to be primitive; in the present case this means that at least one of the coefficients is a unit. If we apply ε_n to the coefficients of f, we obtain a polynomial

with coefficients in $\mathbf{Z}/p^n$ which we shall denote by $\varepsilon_n f$. Suppose that the equation $\varepsilon_n f = 0$ has a solution in $\mathbf{Z}/p^n$ for all n; we wish to show that $f = 0$ has a solution in $\mathbf{Z}_p$.

Let

$$a_n = a_{n0} + a_{n1}p + \ldots + a_{nn-1}p^{n-1} \quad (0 \leq a_{ij} < p)$$

be a root of $\varepsilon_n f = 0$. Then we have an infinite sequence of integers

$$\begin{aligned} a_1 &= a_{10}, \\ a_2 &= a_{20} + a_{21}p, \\ a_3 &= a_{30} + a_{31}p + a_{32}p^2, \\ &\ldots \end{aligned} \tag{9.3.8}$$

Consider the numbers in the first column on the right: $a_{10}, a_{20}, \ldots$; they can assume only finitely many values $0, 1, \ldots, p-1$, so at least one of these values must occur infinitely often. By choosing an appropriate subsequence of the equations (9.3.8) we may therefore assume that $a_{10} = a_{20} = \ldots = b_0$ say. Next consider the sequence of coefficients of p in the equations that remain: $a_{21}, a_{31}, \ldots$; again we can pass to an infinite subsequence in which all coefficients of p are the same, say b_1. Continuing in this way, we obtain a p-adic integer

$$b = b_0 + b_1 p + \ldots,$$

with the property that for each n_0 there is an $n > n_0$ such that $\varepsilon_n b = a_n$ and hence $\varepsilon_n f(b) = 0$. Thus $f(b) \equiv 0 \pmod{p^n}$ for arbitrarily large n, and it follows that $f(b) = 0$; the same conclusion holds for more than one variable. This establishes

Proposition 9.3.1. *Let $f \in \mathbf{Z}_p[x]$, where p is any prime number. If the equation $\varepsilon_n f = 0$ has a root in $\mathbf{Z}/p^n$ for all n, then $f = 0$ has a root in $\mathbf{Z}_p$. More generally, this applies for polynomials in several variables.* ■

For the benefit of readers who have met inverse limits we remark that instead of forming $\mathbf{Z}_p$ as a completion we can also construct it as an inverse limit of the rings $\mathbf{Z}/p^n$. Each of the latter, being finite, is compact and the diagonal argument in the proof of Proposition 9.3.1 can also be used to deduce the compactness of $\mathbf{Z}_p$ from this fact.

In practice it is not necessary to solve each of the equations $\varepsilon_n f = 0$; it suffices to solve one or two and then improve the approximation by a method analogous to the Newton–Fourier rule for calculating roots. We state the essential step separately in the next lemma, where f' as usual denotes the derivative of f.

Lemma 9.3.2. *Given $f \in \mathbf{Z}_p[x]$, suppose that $\alpha \in \mathbf{Z}_p$ satisfies $f(\alpha) \equiv 0 \pmod{p^n}$ and $v(f'(\alpha)) = r$, where $0 \leq 2r < n$. Then $\beta = \alpha - f(\alpha)/f'(\alpha)$ is in $\mathbf{Z}_p$,*

$$f(\beta) \equiv 0 \pmod{p^{n+1}}, \; v(f'(\beta)) = r,$$

and

$$\beta \equiv \alpha \pmod{p^{n-r}}.$$

Proof. We write $\beta = \alpha + \gamma p^{n-r}$ and try to determine $\gamma \in \mathbf{Z}_p$. To this end consider $f(\alpha + h)$; if we regard this as a polynomial in h and expand it in powers of h, we obtain

$$f(\alpha + h) = f(\alpha) + hf'(\alpha) + h^2 g(h), \quad \text{where } g \in \mathbf{Z}_p[x]. \tag{9.3.9}$$

The form of the coefficient of h follows by Taylor's theorem, but we note that a straightforward application of Taylor's theorem to derive (9.3.9) would not make it clear that g has integral coefficients.

If we now put $h = \gamma p^{n-r}$ in (9.3.9), we get

$$f(\beta) = f(\alpha) + p^{n-r}\gamma f'(\alpha) + p^{2n-2r}c,$$

for some $c \in \mathbf{Z}_p$. By hypothesis, $f(\alpha) = ap^n$, $f'(\alpha) = up^r$, where $a \in \mathbf{Z}_p$ and $u \in U$, the unit group of $\mathbf{Z}_p$. Hence $f(\beta) = [a + \gamma u + p^{n-2r}c]p^n$; we can solve $a + \gamma u \equiv 0 \pmod{p}$ for γ, because u is a unit, and with this value for γ we have $f(\beta) \equiv 0 \pmod{p^{n+1}}$ and $\beta \equiv \alpha \pmod{p^{n-r}}$. The last congruence shows that $f'(\beta) \equiv f'(\alpha) \pmod{p^{n-r}}$, hence $f'(\beta) \equiv p^r u \pmod{p^{n-r}}$, and since $n - r > r$, we see that $v(f'(\beta)) = r$. It is easily seen that $\gamma = -a/u$, and this gives the stated value for β. ■

Explicitly, if α is an approximate zero of f (and $v(f'(\alpha))$ is not too high), then $\alpha - f(\alpha)/f'(\alpha)$ is a better approximation. Now the general result follows by iteration:

Theorem 9.3.3. *Let $f \in \mathbf{Z}_p[x_1, \ldots, x_m]$, $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbf{Z}_p^m$ be such that*

$$f(\alpha) \equiv 0 \pmod{p^n}.$$

Suppose that for some j, r $(1 \le j \le m, 0 \le 2r < n)$, $v(D_j f(\alpha)) = r$, where $D_j f$ is the partial derivative with respect to x_j. Then there is a zero $\beta = (\beta_1, \ldots, \beta_m)$ of f in $\mathbf{Z}_p^m$ and $\beta_i \equiv \alpha_i \pmod{p^{n-r}}$.

Proof. Assume first that $m = 1$. We apply the lemma to $\alpha^{(0)} = \alpha$ and obtain $\alpha^{(1)} \in \mathbf{Z}_p$ such that $\alpha^{(1)} \equiv \alpha^{(0)} \pmod{p^{n-r}}$ and $f(\alpha^{(1)}) \equiv 0 \pmod{p^{n+1}}$, $v(f'(\alpha^{(1)})) = r$. By induction on k we obtain a sequence $\{\alpha^{(k)}\}$ such that

$$\alpha^{(k+1)} \equiv \alpha^{(k)} \pmod{p^{n+k-r}}, \tag{9.3.10}$$

and further, $f(\alpha^{(k)}) \equiv 0 \pmod{p^{n+k}}$, $v(f'(\alpha^{(k)})) = r$. The congruence (9.3.10) shows that $\{\alpha^{(k)}\}$ is a Cauchy sequence; its limit β satisfies $f(\beta) = 0$, $\beta \equiv \alpha \pmod{p^{n-r}}$, and this is what we had to prove.

Now the general case is easily deduced by writing $f_1(x) = f(\alpha_1, \ldots, \alpha_{j-1}, x, \alpha_{j+1}, \ldots, \alpha_m)$ and applying the result just proved to f_1. ■

When $n = 1$, $r = 0$, we obtain the usual form of Newton's rule:

Corollary 9.3.4. *Let $f \in \mathbf{Z}_p[x]$ and let $\bar{f}$ be the polynomial obtained by reducing the coefficients* mod p, *thus $\bar{f} = \varepsilon_1 f$. Then every simple zero of $\bar{f}$ in* $\mathbf{Z}/p$ can be lifted to a zero in $\mathbf{Z}_p$. ∎

For multiple zeros this breaks down, e.g. $x^2 + 1 = 0$ has a root (mod 2), but no root in $\mathbf{Z}_2$. A look at Theorem 9.3.3 suggests that we reduce mod 4, not mod 2, and indeed, $x^2 + 1 \equiv 0 \pmod 4$ has no solutions. In the case of a double zero of f, Theorem 9.3.3 tells us that we need to start from a root α of $f(x) \equiv 0 \pmod{p^3}$ when $p \neq 2$, and for $p = 2$ find a root of $f(x) \equiv 0 \pmod 4$.

We now examine the structure of $\mathbf{Q}_p$ more closely. We know already that $\mathbf{Q}_p^\times \cong \mathbf{Z} \times U$, by Proposition 9.1.3, and it remains to describe the group of units U. Each element of U is of the form

$$u = a_0 + a_1 p + a_2 p^2 + \ldots, \quad 1 \leq a_0 < p, 0 \leq a_i < p \ (i > 0).$$

Consider the natural homomorphism $U \to \mathbf{Z}/p^n$ obtained by ignoring powers of p higher than p^{n-1}. The kernel is the set $U_n = 1 + p^n \mathbf{Z}_p$ and we have the chain

$$U = U_0 \supset U_1 \supset U_2 \supset \ldots, \quad \cap U_n = 1. \tag{9.3.11}$$

U_1 is called the group of 1-units ('Einseinheiten'); they are the p-adic units that are congruent to 1 (mod p). We note that the first factor group in the series (9.3.11) is isomorphic to the multiplicative group of $\mathbf{F}_p$ while the remaining factors are isomorphic to the additive group of $\mathbf{F}_p$:

$$U_0/U_1 \cong \mathbf{F}_p^\times, \quad U_n/U_{n+1} \cong \mathbf{F}_p \quad (n > 0).$$

For when $n > 0$, we have $(1 + p^n x)(1 + p^n y) \equiv 1 + p^n(x + y) \pmod{p^{n+1}}$.

Next consider the equation

$$x^{p-1} = 1. \tag{9.3.12}$$

It has $p - 1$ roots in $\mathbf{F}_p$, all distinct, and so each can be lifted uniquely to a solution in $\mathbf{Z}_p$. Thus (9.3.12) has $p - 1$ roots in $\mathbf{Z}_p$; these roots form a subgroup L of U, called the *set of multiplicative representatives* of $\mathbf{F}_p^\times$ in $\mathbf{Z}_p$. We assert that

$$U = L \times U_1. \tag{9.3.13}$$

For any $x \in U$ can be written $c(1 + a_1 p + \ldots) \in LU_1$, and this representation is unique, because the only element of L with constant term 1 is 1, so that $L \cap U_1 = 1$. This establishes (9.3.13).

It remains to find the structure of U_1; here we need a homomorphism from the additive to the multiplicative group of $\mathbf{Q}_p$. In the case of $\mathbf{R}$ such a mapping is provided by the exponential function. Now $\exp x$ can still be defined as a power series over $\mathbf{Q}_p$, but it will no longer converge for all x, because the coefficients $1/n!$ do not tend to 0 as $n \to \infty$. Instead we shall use the binomial series.

Lemma 9.3.5. *For any $z \in \mathbf{Z}_p$ and $n \in \mathbf{N}$,* $\binom{z}{n} \in \mathbf{Z}_p$.

Proof. Given any integer N, we can write $z = z_N + p^N z'$, where $z_N \in \mathbf{Z}$ and $z' \in \mathbf{Z}_p$. Now for any polynomial f, $f(x) - f(y) = (x-y)g(x,y)$, where g is a polynomial in x and y which only depends on f. Taking $x = z, y = z_N, f(x) = \binom{x}{n}$, we find that

$$\binom{z}{n} - \binom{z_N}{n} = p^N z' g(z, z_N). \tag{9.3.14}$$

Here the coefficients of g lie in $\mathbf{Q}_p$ and depend only on n, not on N, so for large enough N the right-hand side of (9.3.14) lies in $\mathbf{Z}_p$. Since clearly $\binom{z_N}{n} \in \mathbf{Z}_p$, we have $\binom{z}{n} \in \mathbf{Z}_p$, as claimed. ∎

Now consider the binomial series in $\mathbf{Z}_p$:

$$(1+p)^x = \sum \binom{x}{n} p^n. \tag{9.3.15}$$

By the lemma $\binom{x}{n} \in \mathbf{Z}_p$, hence the series converges for all $x \in \mathbf{Z}_p$, and as in elementary analysis one proves that

$$(1+p)^x (1+p)^y = (1+p)^{x+y}.$$

So we have a homomorphism $\varphi : \mathbf{Z}_p \to U_1$ given by $x \mapsto (1+p)^x$. To find the kernel, we write $x = ap^r + \ldots$, where a is prime to p. Then

$$(1+p)^x = 1 + xp + \binom{x}{2} p^2 + \ldots \equiv 1 + ap^{r+1} \pmod{p^{r+2}},$$

at least for $p \neq 2$. This is impossible if $(1+p)^x = 1$, hence the mapping φ is injective for $p \neq 2$. When $p = 2$, suppose that $(1+p)^x = 1$; then $1 = (1+2)^{2x} = (1+2^3)^x$, and now we can argue as before: if $x = a2^r + \ldots$, where a is odd, then $(1+2^3)^x \equiv 1 + a2^{r+3} \pmod{2^{r+4}}$, which is again a contradiction. Thus we have an injection even for $p = 2$.

The mapping $\varphi : \mathbf{Z}_p \to U_1$ induces homomorphisms

$$\varphi_n : \mathbf{Z}_p \to U_1/U_{n+1} \quad (n = 0, 1, \ldots),$$

and it is clear from the previous argument that $\ker \varphi_n = p^n \mathbf{Z}_p$ for $p \neq 2$. Hence the induced mapping

$$\overline{\varphi}_n : \mathbf{Z}_p/p^n\mathbf{Z}_p \to U_1/U_{n+1}$$

is an injection for $p \neq 2$. Here both sides have the same finite order p^n, hence $\overline{\varphi}$ is an isomorphism. In particular, this shows $\overline{\varphi}_n$ and with it φ_n to be surjective. It follows from the construction of U_1 that $\varphi : \mathbf{Z}_p \to U_1$ is surjective, hence it is an isomorphism. This proves the case $p \neq 2$ of

Theorem 9.3.6. *For any odd prime p, $\mathbf{Z}_p \cong U_1$, while for $p = 2$, $\mathbf{Z}_2 \cong U_2$ and $U_1 \cong \langle -1 \rangle \times U_2$.*

Proof. It only remains to consider the case $p = 2$. The formula (9.3.15) still applies and it gives an injection $\mathbf{Z}_2 \to U_1$, but this is no longer surjective. We note that now U_1/U_3 is not cyclic: $(1+2)^2 \equiv 1 \pmod{2^3}$, so U_1/U_3 is isomorphic to the Klein 4-group. Instead of φ we shall use the mapping $\psi : \mathbf{Z}_2 \to U_2$ given by

$$\psi(x) = (1+4)^x = \sum \binom{x}{n} 2^{2n}.$$

Clearly this is injective, and as before we get an isomorphism

$$\psi_n : \mathbf{Z}_2/2^n\mathbf{Z}_2 \to U_2/U_{n+2}.$$

Hence ψ is an isomorphism, and clearly $U_1 = U_2 \times \langle -1 \rangle$. ■

If we combine the expression for U_1 obtained in this theorem with earlier results, we obtain

Corollary 9.3.7. *The multiplicative structure of* $\mathbf{Q}_p$ is given by

$$\mathbf{Q}_p^\times \cong \mathbf{Z} \times \mathbf{Z}_p \times \mathbf{C}_{p-1} \quad \text{for } p \neq 2,$$

$$\mathbf{Q}_2^\times \cong \mathbf{Z} \times \mathbf{Z}_2 \times \mathbf{C}_2. \quad \blacksquare$$

The study of functions on $\mathbf{Q}_p$ shows many features quite different from the familiar case of $\mathbf{R}$. We shall not go into detail, but content ourselves with giving a criterion for the continuity of functions on $\mathbf{Z}_p$ (see Mahler (1981)).

Let f be any function on $\mathbf{Z}_p$ with values in $\mathbf{Q}_p$. With f we associate a series of coefficients $\{a_n\}$, defined as

$$a_n = \sum (-1)^k \binom{n}{k} f(n-k). \tag{9.3.16}$$

These are just the iterated differences of f at 0; we define the translation operator T by

$$Tf(x) = f(x+1),$$

and the difference operator D as $T - 1$:

$$Df(x) = (T-1)f(x) = f(x+1) - f(x).$$

Then (9.3.16) is equivalent to the formula

$$a_n = D^n f(0).$$

We remark that the definition of a_n uses only the values of f on $\mathbf{N}$; moreover, we regain f by the formula

$$f(n) = \sum_k \binom{n}{k} a_k. \tag{9.3.17}$$

For on substituting the values of a_k we find

$$\sum \binom{n}{k} D^k f(0) = (1+D)^n f(0) = T^n f(0) = f(n).$$

Now the continuity criterion states that f is continuous iff the a_n tend to 0:

Theorem 9.3.8 *A function f from $\mathbf{N}$ to $\mathbf{Z}_p$ is continuous if and only if its coefficients a_n given by* (9.3.16) *tend to 0 as $n \to \infty$. Moreover, when this is so, then f has a unique continuous extension f^* to $\mathbf{Z}_p$, given by*

$$f^*(x) = \sum \binom{x}{n} a_n. \tag{9.3.18}$$

Proof. Let us state the condition for f to be continuous; since $\mathbf{Z}_p$ is compact, continuity and uniform continuity mean the same thing on $\mathbf{Z}_p$:

$$\begin{aligned} &f \text{ is continuous on } \mathbf{Z}_p \text{ iff, given } s \in \mathbf{N}, \text{ there exists } t \in \mathbf{N} \text{ such that} \\ &v(f(x)-f(y)) \geq s \text{ for all } x, y \in \mathbf{Z}_p \text{ such that } v(x-y) \geq t, \end{aligned} \tag{9.3.19}$$

and of course the same condition applies if f is only defined on $\mathbf{Z}$. In particular, if f is periodic with period p^t, then $f(x) = f(y)$ whenever $v(x-y) \geq t$, and it follows that such a function is always continuous. From another point of view such a function could also be described as a *step function.*

We begin by proving the result for periodic functions; thus we show that for a periodic function f, with period $q = p^t$ say, the coefficients $a_n = D^n f(0)$ tend to 0. The translation operator T, restricted to the space of functions with period q, satisfies $T^q = 1$, hence by the binomial theorem,

$$D^q = (T-1)^q = T^q + pG - 1 = pG,$$

where G is an operator with integral coefficients. It follows that $D^{qs} = p^s G^s$ and for any $n > qs$,

$$a_n = D^n f(0) = p^s (G^s D^{n-qs} f)(0),$$

because f has period q. Since D, G have integer coefficients, it follows that $v(a_n) \geq s$ for $n \geq p^t s$, and so $a_n \to 0$, as claimed.

Next we show that any continuous function can be approximated by periodic functions with sufficiently small period. Let f be continuous, so that (9.3.19) holds, and fix $s \geq 1$. For each $x \in \mathbf{Z}_p$ there is a unique rational integer $g(x)$ in the range

$$0 \leq g(x) < p^s \tag{9.3.20}$$

such that

$$v(f(x) - g(x)) \geq s. \tag{9.3.21}$$

Since f is continuous, we can choose t so that (9.3.19) holds; if $x, y \in \mathbf{Z}_p$ are such that $v(x-y) \geq t$, then by (9.3.19) and the definition of g, we have

$$v(g(x)-g(y)) \geq \min\{v(f(x)-g(x)), v(f(x)-f(y)), v(f(y)-g(y))\} \geq s.$$

Thus $g(x) \equiv g(y) \pmod{p^s}$, and by (9.3.20) it follows that $g(x) = g(y)$. On writing $y = x+q$ we have $g(x+q) = g(x)$, so g has period q, and from (9.3.21) we see that it approximates f.

If f, g have coefficients a_n, b_n respectively, then by (9.3.21) and Lemma 9.3.5,

$$v(a_n - b_n) = v\left(\sum (-1)^k \binom{n}{k} [f(n-k) - g(n-k)]\right) \geq s.$$

Since g has period q, $b_n \to 0$, so for large enough n, $v(a_n) \geq s$. But s was arbitrary, so this shows that $a_n \to 0$, as claimed.

Conversely, if $\{a_n\}$ are the coefficients for f and $a_n \to 0$, then the series (9.3.18) for f^* converges for all $x \in \mathbf{Z}_p$, because $\binom{x}{n} \in \mathbf{Z}_p$ by Lemma 9.3.5, and it is continuous, for, given $s \in \mathbf{N}$, we can choose n_0 such that $v(a_n) \geq s$ for $n > n_0$ and then choose $v(x-y)$ so large that $v\left(\binom{x}{n} - \binom{y}{n}\right) \geq s$ for $n = 1, 2, \ldots, n_0$. Comparing (9.3.17) and (9.3.18), we see that f^* agrees with f on $\mathbf{N}$. It follows that f is continuous and f^* is its extension to $\mathbf{Z}_p$, unique because $\mathbf{N}$ is dense in $\mathbf{Z}_p$. ■

Exercises

1. Solve $x^3 = 4$ in $\mathbf{Z}_5$, $\mathbf{Z}_2$, $\mathbf{Z}_3$ when possible.
2. How can the positive expression (9.3.5) be reconciled with the negative expression for 1/2 in the line above? (Hint. Remember (9.3.4).)
3. Show that $\exp x$ converges for $v(x) > 1/(p-1)$. Use exp instead of the binomial function to prove Theorem 9.3.6. (Hint. Show first that $v(n!) = [n/p] + [n/p^2] + \ldots$, where $[\xi]$ is the greatest integer $\leq \xi$.)
4. Show that a p-adic number $\sum a_i p^i$ is rational iff the a_i from some index onwards are periodic. Describe the set of all p-adic numbers represented by a finite series $\sum a_i p^i$.
5. Show that the equation $x^2 + 1 = 0$ is irreducible over $\mathbf{Q}_2$. (Hint. Try $x = a + 2b$.) Do the same for $x^2 - 2 = 0$.
6. Let p be an odd prime. Show that if α, β are units in $\mathbf{Z}_p$, then $\alpha x^2 + \beta y^2 = 1$ has a solution in $\mathbf{Z}_p$. Deduce that two regular quadratic forms f, g of the same rank over $\mathbf{Q}_p$ are equivalent iff $\det f$ and $\det g$ define the same residue class in $\mathbf{Q}_p^\times / \mathbf{Q}_p^{\times 2}$.
7. Let $x \in \mathbf{Q}_p^\times$ have the form $x = p^n u$ ($n \in \mathbf{Z}$, $u \in U$). Show that x is a square iff n is even and u is a quadratic residue mod p when p is odd, and $u \equiv 1 \pmod 8$ when $p = 2$. Hence find $\mathbf{Q}_p^\times / \mathbf{Q}_p^{\times 2}$.
8. Let $p \neq 2$ and $v(a_1) = v(a_2) = v(a_3)$. Show that $a_1 x_1^2 + a_2 x_2^2 + a_3 x_3^2$ is isotropic over $\mathbf{Q}_p$. Deduce that any quadratic form in at least five variables over $\mathbf{Q}_p$ is isotropic. Conclude that $\mathbf{Q}_p$ cannot be ordered.

9. Let f be a p-adic function with coefficients a_n. Show that $\sum f(n)x^n = (1+t)\sum a_n t^n$, where $t = x(1-x)^{-1}$. Deduce that the function g defined by $g(n) = (-1)^n a_n$ has coefficients $b_n = (-1)^n f(n)$.

9.4 Integral Elements

Let S be a commutative ring and R be a subring of S. An element of S is said to be *integral over* R if it satisfies a monic equation with coefficients in R:

$$x^n + a_1 x^{n-1} + \ldots + a_n = 0 \quad (a_i \in R). \tag{9.4.1}$$

This generalizes the definition of an algebraic integer given in Section 5.1, which was the special case $R = \mathbf{Z}$. Even in that case we have, besides the obvious instances such as $\sqrt{2}$, also integers like $\frac{1}{2}(-1+\sqrt{-3})$, which arises as root of $x^3 = 1$.

There are several equivalent ways of expressing the definition, which are often useful:

Proposition 9.4.1. *Let $R \subseteq S$ be rings. For any $c \in S$ the following conditions are equivalent:*

(a) *c is integral over R;*
(b) *$R[c]$ is finitely generated as R-module;*
(c) *there is a finitely generated R-submodule M of S with zero annihilator in $R[c]$, such that $cM \subseteq M$.*

When c is a unit in S, all are equivalent to

(d) $c \in R[c^{-1}]$.

Proof. (a) $\Rightarrow$ (b). Let c satisfy (9.4.1); we claim that $R[c]$ is generated as R-module by $1, c, c^2, \ldots, c^{n-1}$. For it is clearly generated by all the powers c^ν, and for $\nu \geq n$ we have, on multiplying (9.4.1) for $x = c$ by $c^{\nu-n}$ and rearranging:

$$c^\nu = -(a_1 c^{\nu-1} + \ldots + a_n c^{\nu-n}).$$

By induction on n this expresses all powers of c in terms of $1, c, \ldots, c^{n-1}$, and (b) follows. We note that when c is a unit in S, we find in the same way

$$c = -(a_1 + a_2 c^{-1} + \ldots + a_n c^{1-n}). \tag{9.4.2}$$

Thus when a unit c of S is integral over R, then $c \in R[c^{-1}]$. Conversely, if $c \in R[c^{-1}]$, we have an equation of the form (9.4.2) with $a_i \in R$, and this can be rearranged as (9.4.1) (with $x = c$); hence c is then integral over R. This shows that (a) $\Leftrightarrow$ (d) when c is a unit in S.

(b) $\Rightarrow$ (c) is clear, since $R[c]$ is a submodule of the required sort, and to prove (c) $\Rightarrow$ (a), let M be generated by $u_1, \ldots, u_n$. Then

$$cu_i = \sum a_{ij} u_j, \quad \text{where } a_{ij} \in R.$$

Hence $\sum_j (c\delta_{ij} - a_{ij})u_j = 0$; if we write $A = (a_{ij})$ and multiply by $adj(cI - A)$, we obtain

$$\det(cI - A)u_i = 0, \quad i = 1, \ldots, n.$$

By hypothesis, the annihilator of all the u_i in $R[c]$ is 0, and it follows that $\det(cI - A) = 0$, which on expansion gives a monic equation for c over R. ■

Given a ring S with a subring R, S is said to be *integral* over R if every element of S is integral over R. We note that if $c_1, \ldots, c_n \in S$ are integral over R then $R[c_1, \ldots, c_n]$ is finitely generated as R-module. For $R[c_1]$ is a finitely generated R-module and $c_2, \ldots, c_n$ are integral over $R[c_1]$, therefore $R[c_1, c_2, \ldots, c_n]$ is a finitely generated R-module, by induction on n. This enables us to prove the transitivity of integrality:

Corollary 9.4.2. *Let R be a ring, S be an integral extension of R and T be an integral extension of S. Then T is integral over R.*

Proof. We have to prove that every element of T is integral over R. Any $c \in T$ satisfies an equation $c^n + a_1 c^{n-1} + \ldots + a_n = 0$, where $a_i \in S$. Each a_i is integral over R, so by the above remark, $R[a_1, \ldots, a_n]$ is a finitely generated R-module, say

$$R[a_1, \ldots, a_n] = \sum_{j=1}^{r} Ru_j.$$

Likewise $S[c]$ is a finitely generated $R[a_1, \ldots, a_n]$-module, so that

$$S[c] = \sum_{h=1}^{s} R[a_1, \ldots, a_n]v_h$$

$$= \sum_{h=1}^{s} \sum_{j=1}^{r} Ru_j v_h.$$

Thus $S[c]$ is a finitely generated R-module and this proves that c is integral over R. ■

The set $\overline{R}$ of all elements of S integral over R is easily seen to be a subring of S. It is called the *integral closure* of R in S, and R is said to be *integrally closed* or an *order* in S if $\overline{R} = R$. Generally an integral domain is called 'integrally closed' (without qualification) if it is integrally closed in its field of fractions. For example, $\mathbf{Z}[\sqrt{-1}]$ is integrally closed, as is easily verified, but $\mathbf{Z}[\sqrt{-3}]$ is not, for its integral closure contains the element $\frac{1}{2}(-1 + \sqrt{-3})$.

There is a close connexion between the integral closure and general valuation rings; to describe it we shall need an existence lemma for valuation rings, which is also useful elsewhere. On any field K we consider pairs $P = (R, \mathfrak{a})$ consisting of a subring R of K and a proper ideal $\mathfrak{a}$ in R. Given two such pairs $P = (R, \mathfrak{a})$, $P' = (R', \mathfrak{a}')$, we shall say that P' *dominates* P, in symbols $P' \geq P$, if $R' \supseteq R$ and $\mathfrak{a}' \supseteq \mathfrak{a}$, and use $\leq, <, >$ as usual.

Lemma 9.4.3 (Chevalley). *Let K be a field, R be a subring of K and $\mathfrak{a}$ be a non-zero proper ideal in R. Then there is a subring V with an ideal $\mathfrak{p}$ such that $(V, \mathfrak{p})$ is maximal among pairs dominating $(R, \mathfrak{a})$. Further, any such maximal pair $(V, \mathfrak{p})$ consists of a valuation ring $V \neq K$ and its maximal ideal $\mathfrak{p}$.*

Proof. The family of pairs dominating $(R, \mathfrak{a})$ is clearly inductive and so, by Zorn's lemma it has a maximal element $(V, \mathfrak{p})$. It remains to show that V is a valuation ring in K and $\mathfrak{p}$ is its maximal ideal. Let $c \in K$; we must show that $c \in V$ or $c^{-1} \in V$. Assume the contrary; then since $c \notin V$, we have $V \subset V[c]$, and if the ideal $\mathfrak{p}'$ generated by $\mathfrak{p}$ in $V[c]$ is proper, then $(V[c], \mathfrak{p}') > (V, \mathfrak{p})$, which contradicts the maximality. Hence $\mathfrak{p}' = V[c]$, i.e. we have an equation

$$1 = a_0 + a_1 c + \ldots + a_m c^m, \ a_i \in \mathfrak{p}. \tag{9.4.3}$$

Similarly, since $c^{-1} \notin V$, we have an equation

$$1 = b_0 + b_1 c^{-1} + \ldots + b_n c^{-n}, \ b_j \in \mathfrak{p}. \tag{9.4.4}$$

We may assume that m, n are chosen as small as possible, and by symmetry we may take $m \geq n$. Multiplying (9.4.4) by c^m we get

$$(1 - b_0)c^m = b_1 c^{m-1} + \ldots + b_n c^{m-n}. \tag{9.4.5}$$

If we now multiply (9.4.3) by $1 - b_0$ and substitute for $(1 - b_0)c^m$ from (9.4.5), we obtain an equation of the form (9.4.3), but with a lower value for m. This is a contradiction, and it shows that either c or c^{-1} lies in V, i.e. V is a valuation ring in K. Further, $\mathfrak{p}$ is a maximal ideal of V, by the maximality of the pair $(V, \mathfrak{p})$. Finally, $V \neq K$, for if $V = K$, then its maximal ideal would be 0 and so could not contain $\mathfrak{a}$. ∎

In this lemma we assumed that $\mathfrak{a} \neq 0$. If $\mathfrak{a} = 0$, we can still apply the lemma, replacing $\mathfrak{a}$ by any non-zero proper ideal, as long as R is not a field. Even when R is a field, if K is transcendental over R, we can enlarge R to a subring $R[x]$ not a field and proceed as before. However, if R is a field and K is algebraic over R, then any subring between R and K is a field; in that case any pair dominating $(R, 0)$ is of the form $(L, 0)$, where L is a field between R and K.

We can now characterize the integral closure of a subring as follows.

Theorem 9.4.4. *Let K be a field and A be a subring. Then the integral closure $\overline{A}$ of A in K is the intersection of all general valuation rings of K containing A.*

Proof. Let $\{V_\lambda\}$ be the family of all valuation rings in K that contain A, and denote by $\mathfrak{p}_\lambda$ the maximal ideal of V_λ. If c is integral over A, then either $c = 0$ or its monic equation over A may be written

$$1 + a_1 c^{-1} + \ldots + a_n c^{-n} = 0. \tag{9.4.6}$$

Suppose that $c \notin V_\lambda$; then $c^{-1} \in \mathfrak{p}_\lambda$ and now (9.4.6) shows that $1 \in \mathfrak{p}_\lambda$, a contradiction; hence $c \in \cap V_\lambda$. Conversely, if c is not integral over A, then $c \neq 0$ and by

Proposition 9.4.1, $c \notin A[c^{-1}]$, hence (c^{-1}) is a proper ideal in $A[c^{-1}]$. By Lemma 9.4.3 there exists a valuation ring $V \supseteq A[c^{-1}]$ with maximal ideal containing (c^{-1}). It follows that $c \notin V$, so $c \notin \cap V_\lambda$; therefore we have $\overline{A} = \cap V_\lambda$ as claimed. ∎

We observe that Theorem 9.4.4 provides another proof that the elements of a field integral over a given subring themselves form a ring. The theorem also shows that a subring of a field K is integrally closed iff it is the intersection of all the valuation rings containing it. Here it is of course important to include all valuations, not merely the principal ones.

We note that a UFD may also be described as the intersection of the valuation rings associated with the various atoms. Hence we obtain

Corollary 9.4.5. *A unique factorization domain is integrally closed (in its field of fractions).* ∎

For some applications it is useful to have a characterization of real-valued valuations.

Theorem 9.4.6. *Let K be a field with a subring V. Then V is a valuation ring of a non-trivial real-valued valuation on K if and only if V is a maximal subring of K which is not a field.*

Proof. Let V be the valuation ring of a non-trivial real-valued valuation v on K. Any ring strictly containing V contains an element c such that $v(c) < 0$. Now for any $x \in K$ there exists $n \in \mathbf{N}$ such that $nv(c) < v(x)$, hence $v(xc^{-n}) > 0$, so $a = xc^{-n} \in V$ and therefore $x = c^n a \in V[c]$. Hence $V[c] = K$ and this shows V to be a maximal proper subring of K.

Conversely, suppose that V is a maximal proper subring of K which is not a field. By Lemma 9.4.3, V is a valuation ring. We complete the proof by showing that its value group Γ is Archimedean ordered, for then it can be embedded in $\mathbf{R}$ (see Theorem 8.7.3). Let $a, b \in V$, $a^{-1} \notin V$; then $V[a^{-1}] = K$, hence $b^{-1} = ca^{-n}$ for some $c \in V$ and some $n \geq 0$, and so $v(a^n b^{-1}) = v(c) \geq 0$, i.e. $nv(a) \geq v(b)$. Thus Γ is indeed Archimedean ordered and we can embed Γ in $\mathbf{R}$ by fixing $\alpha \in \Gamma$, $\alpha > 0$, and defining

$$\varphi(\beta) = \inf\{m/n | m\alpha \geq n\beta\}.$$

Hence v is equivalent to a real-valued valuation. ∎

Another important consequence of Lemma 9.4.3 is the extension theorem; to describe it we need yet another way of looking at valuations.

Let K be a field. By a *place* of K in a field k we understand a mapping

$$\varphi : K \to k \cup \{\infty\},$$

such that $k\varphi^{-1} = V$ is a subring of K and the restriction $\varphi|V$ is a ring homomorphism. Thus with the usual operations on ∞ we have

$$(x - y)\varphi = x\varphi - y\varphi, \quad 1\varphi = 1, \quad (xy)\varphi = x\varphi . y\varphi,$$

whenever the right-hand side is defined (i.e. different from $0.\infty$ and $\infty - \infty$). The set $k\varphi^{-1}$ is called the set of elements where φ is *finite*.

For example, a valuation on K with valuation ring V and maximal ideal $\mathfrak{p}$ defines a place on the residue class field $V/\mathfrak{p}$, which is finite on V. Conversely, if φ is a place of K in k and V is the set where φ is finite, then if $x \notin V$, we have $x^{-1} \in V$, for otherwise we should have $1 = x\varphi.x^{-1}\varphi = \infty.\infty = \infty$, a contradiction. Hence V is a valuation ring and $\varphi|V : V \to V/\mathfrak{p}$ is the residue class homomorphism. It is clear that two places $\varphi_i : K \to k_i$ $(i = 1, 2)$ correspond to the same valuation iff there is an isomorphism $\alpha : k_1 \to k_2$ such that $\varphi_1\alpha = \varphi_2$. In this case the places are said to be *equivalent*. The situation is summed up in

Proposition 9.4.7. *The valuations on a field K correspond to the places of K, with equivalent valuations corresponding to equivalent places.* ■

For the next result we need a special case of localization, which forms the subject of Section 10.3. We shall describe this briefly here and refer the reader who wants to know more to Section 10.3.

Let R be a ring and S be a subset of R. A ring homomorphism $f : R \to R'$ is called *S-inverting* if it maps the elements of S to invertible elements in R'. It is easily seen that the set of all elements inverted by a given homomorphism is *multiplicative*, i.e. it contains 1 and is closed under multiplication. If R is an integral domain with field of fractions K, and S is a multiplicative subset of R, then the set

$$R_S = \{x \in K | x = a/u, \text{ where } a \in R, u \in S\}$$

is easily seen to be a subring of K containing R; it is called the *ring of fractions with denominators in S*. In particular, when S is the complement of a prime ideal $\mathfrak{p}$, then R_S is a *local ring*, i.e. the set of all non-units forms an ideal. In this case one usually writes (somewhat illogically) $R_\mathfrak{p}$ to mean R_S; the risk of confusion is small, since the multiplicative set S does not usually contain 0, whereas $\mathfrak{p}$ always does.

Theorem 9.4.8 (Extension theorem). *Let K be a field containing a subring A and let $f : A \to E$ be a homomorphism of A into a field E. Then f can be extended to a place of K in an extension F of E.*

Proof. Write $\mathfrak{p} = \ker f$ and $S = A\backslash\mathfrak{p}$; then f is S-inverting and so can be extended to a homomorphism of the local ring $A_\mathfrak{p}$ into E. If $A_\mathfrak{p}$ is a field and K is algebraic over it, then we can extend f further to K, by Lemma 7.5.3. Otherwise there exists by Lemma 9.4.3 (and the remark following it) a valuation ring V with maximal ideal $\mathfrak{m}$ dominating $A_\mathfrak{p}$ and its maximal ideal $\mathfrak{p}_1$, say. Thus $\mathfrak{m} \cap A_\mathfrak{p} \supseteq \mathfrak{p}_1$ and in fact equality holds here, because every element of $A_\mathfrak{p}$ not in $\mathfrak{p}_1$ is a unit. Thus we have the diagram:

$$\begin{array}{ccccc} & & V/\mathfrak{m} & & \\ & & \uparrow & & \\ A/\mathfrak{p} & \longrightarrow & A_\mathfrak{p}/\mathfrak{p}_1 & \longrightarrow & E \end{array}$$

here $A/\mathfrak{p}$ is a subfield of $\mathbf{V}/\mathfrak{m}$, hence the embedding can be extended to one of $V/\mathfrak{m}$ in a field $F \supseteq E$, and this is a place of K in F. ■

A closer analysis shows that the field F can be taken within an algebraic closure of E, but this fact will not be needed here (see Exercise 8).

Exercises

1. Verify that $\mathbf{Z}[\sqrt{-1}]$ is integrally closed.
2. Let $R \subseteq S$ be rings. Show that an element c integral over R is a non-zerodivisor iff the constant term in the monic equation of least degree for c is a non-zerodivisor in S.
3. Let $R \subset R'$ be integral domains such that R' is integral over R. Show that for every prime ideal $\mathfrak{p}$ of R there is a prime ideal $\mathfrak{p}'$ of R' such that $\mathfrak{p}' \cap R = \mathfrak{p}$.
4. Let K/k be a finitely generated field extension. Show that there exists a valuation trivial on k but non-trivial on K iff K/k is not algebraic.
5. In Lemma 9.4.3 show that if $\mathfrak{a}$ is prime, then $\mathfrak{p}$ can be chosen so that $\mathfrak{p} \cap R = \mathfrak{a}$. (Hint. See the proof of Theorem 9.4.8.)
6. Let A be an integral domain, integrally closed in its field of fractions K. Show that for any monic polynomials f, g over K, if $fg \in A[x]$, then $f, g \in A[x]$. Give an example to show that the condition (on integral closure) cannot be omitted.
7. Let A be an integrally closed domain and L be any field containing A. If $a_1, \ldots, a_n \in L$ are such that for each $i = 1, \ldots, n$ there exists $m_i \in \mathbf{N}$ such that $a_i^{m_i} = f_i(a_1, \ldots, a_n)$, where f_i is a polynomial of degree $< m_i$ with coefficients in A, then $a_1, \ldots, a_n$ are integral over A. Deduce that any extension of A finitely generated as A-module is integral over A.
8. Show that if the field F in the proof of Theorem 9.4.8 is transcendental over E, it can be replaced by a place of F in an extension of E. Deduce that F can always be chosen to be algebraic over E.

9.5 Extension of Valuations

We now turn to ways of extending a valuation given on a field to a larger field. There are essentially two cases to consider, the algebraic and the transcendental extensions. It is possible to treat both at once, but we shall take these cases separately, since it is then possible to treat the algebraic case more simply, while the transcendental case can be treated more explicitly.

Let L/K be any field extension, and let v be a valuation on K, with an extension w to L. We shall write V, $\mathfrak{p}$, Γ for the valuation ring, maximal ideal and value group of v, and W, $\mathfrak{P}$, Δ for the corresponding entities for w. Clearly we have $\Gamma = v(K) \subseteq w(L) = \Delta$, hence Γ is a subgroup of Δ. The index

$$(\Delta : \Gamma) = e \tag{9.5.1}$$

is always denoted by e and is called the *ramification index*. Next we have $\mathfrak{P} \cap V = \mathfrak{p}$, hence the natural homomorphism

$$V/\mathfrak{p} \to W/\mathfrak{P} \tag{9.5.2}$$

is an embedding, so denoting the residue class fields by $\overline{K}, \overline{L}$ respectively, we see that $\overline{L}$ is an extension of $\overline{K}$. The degree

$$[\overline{L} : \overline{K}] = f \tag{9.5.3}$$

is called the *residue degree*. From the properties of the index and the degree it is clear that both the ramification index and the residue degree are multiplicative under extensions.

For example, let $K = \mathbf{Q}, L = \mathbf{Q}(\sqrt{2})$. The 2-adic valuation v_2 has an extension to L for which $e = 2, f = 1$ (2 is 'ramified' in L). If p is an odd prime, either 2 is a quadratic residue mod p (i.e. $x^2 \equiv 2 \pmod{p}$ has a solution), then 2 is a square in $\mathbf{Q}_p$ and v_p has two extensions to L and $e = f = 1$; or 2 is a quadratic non-residue mod p, then 2 is not a square in $\mathbf{Q}_p$ (p is 'inert' in L), now v has just one extension to L, $e = 1, f = 2$ and the residue field is enlarged from $\mathbf{F}_p$ to $\mathbf{F}_p(\sqrt{2})$. This illustrates the general situation described in Theorem 9.5.7 below.

In comparing the values of v and w the following relation is often useful: let $u_1, \ldots, u_r \in W$ be such that their residues mod $\mathfrak{P}$, $\overline{u}_1, \ldots, \overline{u}_r$ say, are linearly independent over $\overline{K} = V/\mathfrak{p}$. Then we claim that for any $\alpha_1, \ldots, \alpha_r \in K$, we have

$$w(\alpha_1 u_1 + \ldots + \alpha_r u_r) = \min\{v(\alpha_1), \ldots, v(\alpha_r)\}. \tag{9.5.4}$$

To prove this assertion we may assume, after appropriate renumbering, that $v(\alpha_1) = \ldots = v(\alpha_k) < v(\alpha_j) (j > k)$, so that $\alpha_i/\alpha_1 \in V$ for all i. Since $\overline{u}_i \neq 0$, it follows that $w(u_i) = 0$, so that the left-hand side of (9.5.4) has a value $\geq v(\alpha_1)$. If this value is $> v(\alpha_1)$, then after dividing by α_1 and reducing mod $\mathfrak{P}$ we get

$$\overline{u}_1 + (\overline{\alpha}_2/\overline{\alpha}_1)\overline{u}_2 + \ldots + (\overline{\alpha}_r/\overline{\alpha}_1)\overline{u}_r = 0.$$

This contradicts the linear independence of the $\overline{u}_i$ over $\overline{K}$ and it establishes equality in (9.5.4).

An important relation between e and f is given by

Theorem 9.5.1. *Let L/K be a finite field extension of degree n, and let v be a valuation on K with an extension w to L. Then the ramification index e and residue degree f satisfy the relation*

$$ef \leq n. \tag{9.5.5}$$

If v is real-valued, or principal, or trivial, then so is w.

If moreover K is complete and v is principal, then equality holds in (9.5.5).

Proof. Denote the valuation rings of K, L by V, W respectively, so that $W \cap K = V$, and denote by $\mathfrak{p}, \mathfrak{P}$ the corresponding maximal ideals. Let $u_1, \ldots, u_r \in W$ be such

that their residues mod $\mathfrak{P}$ are linearly independent over $V/\mathfrak{p}$, and take $\pi_1, \ldots, \pi_s \in L$ such that the $w(\pi_j)$ are incongruent mod Γ, the value group of v. We assert that the rs elements $u_i\pi_j$ are linearly independent over K. For if there is a relation

$$\sum \alpha_{ij} u_i \pi_j = 0, \quad \text{where } \alpha_{ij} \in K, \tag{9.5.6}$$

let us write $a_j = \sum_i u_i \alpha_{ij}$, so that (9.5.6) reads $\sum a_j \pi_j = 0$. If the a_j are not all 0, it follows that $w(a_h\pi_h) = w(a_k\pi_k)$ for some h, k, $1 \leq h < k \leq s$. This means that $w(\pi_h/\pi_k) = w(\pi_h) - w(\pi_k) = w(a_k) - w(a_h)$. Here the right-hand side is in Γ, by (9.5.4), but the left-hand side is not, which is a contradiction. Hence all the a_i vanish, and so, again by (9.5.4), the α_{ij} must also vanish, which shows the $u_i\pi_j$ to be linearly independent. It follows that $rs \leq n$; in particular, e and f must be finite and taking $s = e, r = f$, we obtain (9.5.5).

From the definition of e it follows that $e\Delta \subseteq \Gamma$, where Δ is the value group of w; thus if Γ is embedded in $\mathbf{R}$, so is Δ, and if $\Gamma \subseteq \mathbf{Z}$, then by renormalizing we can ensure that $\Gamma \subseteq e\mathbf{Z}$; it follows that $e\Delta \subseteq e\mathbf{Z}$, i.e. $\Delta \subseteq \mathbf{Z}$. If v is trivial, then $\Gamma = 0$, hence $e\Delta = 0$ and so w is also trivial.

Now assume that v is principal and K is complete. Then so is L, as a finite-dimensional K-space, by Proposition 9.2.7. Let π, Π be uniformizers for v, w respectively; then every element of L can be written as $\sum \alpha_i \Pi^i$, where α_i belongs to a transversal of the residue field $\overline{L}$ in W. Instead of the powers Π^i we can also use $\pi^i, \Pi\pi^i, \ldots, \Pi^{e-1}\pi^i$, and instead of a transversal of $\overline{L}$ in W we can take elements $u_1, \ldots, u_f$ of W such that $\overline{u}_1, \ldots, \overline{u}_f$ form a basis of $\overline{L}$ over $\overline{K}$ and use $\sum \alpha_i u_i$, where the α_i run over a transversal of $\overline{K}$ in V. Then every element of L may be written

$$x = \sum_{i=-N}^{\infty} \sum_{\mu\nu} a_{\mu\nu i} u_\mu \Pi^\nu \pi^i, \quad \text{where } a_{\mu\nu i} \in K.$$

This shows that L is spanned by the ef elements $u_\mu \Pi^\nu$ over K. Hence $[L : K] \leq ef$ and it follows that we have equality in (9.5.5). ∎

We begin by looking at transcendental extensions; here we need not postulate completeness, nor even assume that the valuation is principal.

Proposition 9.5.2. *Let K be a field with a valuation v and residue class field $\overline{K}$. If $K(t)$ is a purely transcendental extension of K, then there is just one extension of v to $K(t)$ such that t remains transcendental over $\overline{K}$. This valuation on $K(t)$ has the same value group as v and has the residue class field $\overline{K}(t)$.*

The valuation on $K(t)$ determined in this way is called the *Gaussian extension* of v to $K(t)$.

Proof. If there is such an extension w of v, then on $K[t]$ it is given by

$$w(a_0 + a_1 t + \ldots + a_n t^n) = \min\{v(a_0), \ldots, v(a_n)\}, \tag{9.5.7}$$

by (9.5.4). This is already enough to determine w on the field of fractions $K(t)$. Thus there can be at most one such extension, and there is one, since w given by (9.5.7)

clearly satisfies all the conditions. It is clear from (9.5.7) that v and w have the same value group. Further, the residue class field clearly contains $\overline{K}(t)$, and for any element f/g such that $w(f/g) \geq 0$ we can find $u \in K(t)$, $\overline{u} \in \overline{K}(t)$ such that $w(f/g - u) \geq 0$; this is most easily seen by writing f/g as a Laurent series in t. This shows $\overline{K}(t)$ to be the residue class field. ∎

Next we prove that extensions always exist in the algebraic case.

Proposition 9.5.3. *Let L/K be a field extension of finite degree and let v be a valuation on K. Then v has an extension to L; moreover, any such extension is real-valued, principal or trivial whenever v is.*

If v is real-valued and K is complete, then there is a unique extension w of v to L, given in terms of the norm by

$$w(\gamma) = [L : K]^{-1} v(N_{L/K}(\gamma)) \quad \text{for } \gamma \in L. \tag{9.5.8}$$

Proof. We already know from Theorem 9.5.1 that if v is real-valued, principal or trivial, then so is any extension to L. It only remains to prove the existence.

Let V be the valuation ring of v on K. Then $V \subseteq L$ and by the extension theorem (Theorem 9.4.8) the natural homomorphism $V \to V/\mathfrak{p} = \overline{K}$ can be extended to a place of L in an extension field of $\overline{K}$. This gives the required valuation on L. When K is complete (for a real-valued valuation v), then the extension w to L is unique, by Theorem 9.2.9. Now take any $\gamma \in L$, suppose that f is its minimal polynomial over K, of degree n say, and let E be a splitting field of f over L; w has a unique extension to E, still denoted by w. Over E we can write $f = \prod (x - \gamma_i)$, where $\gamma_1 = \gamma$ say. Since f is irreducible over K, we have $K(\gamma_i) \cong K(\gamma)$ for $i = 1, \ldots, n$, hence by uniqueness, $w(\gamma_i) = w(\gamma)$ and so $v(N_{K(\gamma)/K}(\gamma)) = \sum w(\gamma_i) = nw(\gamma)$. If $[L : K(\gamma)] = r$, then $[L : K] = rn$ and so $v(N_{L/K}(\gamma)) = rv(N_{K(\gamma)/K}(\gamma)) = rnw(\gamma)$; on division by rn this yields (9.5.8). ∎

We observe that this result has an analogue in the Archimedean case: then K must be $\mathbf{R}$ or $\mathbf{C}$, by Ostrowski's second theorem (Theorem 9.2.11) and $[L : K]$ is 1 or 2. The only non-trivial case is that where $K = \mathbf{R}$, $L = \mathbf{C}$ and the absolute value on $\mathbf{C}$ is given by

$$||x|| = |x\overline{x}|^{1/2},$$

which is the multiplicative analogue of (9.5.8).

The following necessary condition for irreducibility over a complete field is an easy consequence of Proposition 9.5.3.

Corollary 9.5.4. *Let K be a field, complete under a real-valued valuation v. If $f = a_0 x^n + a_1 x^{n-1} + \ldots + a_n$ is an irreducible polynomial over K, then*

$$v(a_i) \geq \min\{v(a_0), v(a_n)\}. \tag{9.5.9}$$

Proof. Suppose first that $v(a_0) \leq v(a_n)$, so that $v(a_n/a_0) \geq 0$, and consider the monic

polynomial $a_0^{-1}f$. If its zeros in a splitting field are $\alpha_1, \ldots, \alpha_n$, and w denotes the unique extension to E, then

$$w(\alpha_i) = \frac{1}{n} v(a_n/a_0) \geq 0.$$

Hence the elementary symmetric functions of the α_i are integers, and so $v(a_i/a_0) \geq 0$, i.e. $v(a_i) \geq v(a_0)$, and so (9.5.9) is proved in this case. When $v(a_0) > v(a_i)$, we can make a reduction to the previous case by the substitution $y = 1/x$, hence (9.5.9) holds generally. ■

For complete fields there is a useful method for lifting factorizations over the residue class field, known as Hensel's lemma. It is usually proved by a somewhat lengthy verification, but it can be obtained more directly from Theorem 9.4.8 and Theorem 9.2.9 (asserting the existence and uniqueness of extensions). We single out the essential step as a lemma.

We note that a polynomial f over a valuation ring V is primitive (i.e. its coefficients have no common factor) precisely when its image in the residue class ring $V/\mathfrak{p}$ is non-zero.

Lemma 9.5.5. *Let K be a complete field under a real-valued valuation v, with valuation ring V, maximal ideal $\mathfrak{p}$ and residue class field $V/\mathfrak{p} = \overline{K}$. Write $\alpha \mapsto \overline{\alpha}$ for the residue class map $V \to V/\mathfrak{p}$. If*

$$f = a_0 x^n + a_1 x^{n-1} + \ldots + a_n \tag{9.5.10}$$

is any irreducible polynomial over V, then one of the following three cases will arise:

(i) *a_0, a_n are both units and $\overline{a}_0^{-1} f$ is a power of an irreducible polynomial,*
(ii) *a_0 is a unit and $\overline{f} = \overline{a}_0 x^n$,*
(iii) *a_n is a unit and $\overline{f} = \overline{a}_n$.*

Proof. Suppose first that a_0 is a unit. Let E be a splitting field of f over K and w be the unique extension of v to E. We write W for the valuation ring of w in E and $\mathfrak{P}$ for its maximal ideal, so that its residue class field is $W/\mathfrak{P} = \overline{E}$. If σ is an automorphism of E over K and $\gamma \in E$, then $w(\gamma^\sigma) = w(\gamma)$, because γ^σ and γ are conjugate over K. This shows that σ maps W into itself and likewise $\mathfrak{P}$, so it induces an automorphism $\overline{\sigma}$ of $W/\mathfrak{P} = \overline{E}$ over $\overline{K}$. Now f is irreducible over V, hence it is also irreducible over K, by inertia (Theorem 7.7.2), and if its zeros are $\alpha_1, \ldots, \alpha_n$, then

$$w(\alpha_i) = \frac{1}{n} v(a_n) \geq 0,$$

by Proposition 9.5.3. This shows that $\alpha_i \in W$ and we can therefore factorize the image $\overline{f}$ of f over $\overline{E}$ as $\overline{f} = \overline{a}_0 \prod (x - \overline{\alpha}_i)$. For any i, j in the range $1, \ldots, n$ there is an automorphism σ of E/K such that $\alpha_i^\sigma = \alpha_j$, hence $\overline{\alpha}_i^\sigma = \overline{\alpha}_j$, and so the zeros of $\overline{f}$ are permuted transitively by the automorphisms of $\overline{E}/\overline{K}$. It follows that $\overline{f}$ is a power of an irreducible polynomial. If a_n is a unit, we are in case (i), while for $a_n \in \mathfrak{p}$, $\overline{f}$ has the factor x and so it must be a power of x, and we have case (ii).

Now suppose that a_0 is a non-unit. Since f is primitive, some a_i is a unit, hence a_n is a unit by (9.5.9). Consider $g(y) = y^n f(y^{-1})$; this is again irreducible and case (ii) applies because the highest coefficient is a unit, while the constant term is not, so $\bar{g} = \bar{a}_n y^n$; therefore $\bar{f} = \bar{a}_n$ and we have case (iii). ∎

Theorem 9.5.6 (Hensel's lemma). *Let K be a complete field for a real-valued valuation v, with valuation ring V, maximal ideal $\mathfrak{p}$ and residue class field $V/\mathfrak{p} = \overline{K}$. Given a primitive polynomial f over V, suppose that $g_0, h_0 \in V[x]$ are such that*

$$f \equiv g_0 h_0 (\mathrm{mod}\ \mathfrak{p}), \tag{9.5.11}$$

and that g_0, h_0 are relatively prime (mod $\mathfrak{p}$) *and g_0 is monic. Then there exist unique polynomials $g, h \in V[x]$ such that g is monic and g, h satisfy*

$$f = gh,\ g \equiv g_0,\ h \equiv h_0 (\mathrm{mod}\ \mathfrak{p}). \tag{9.5.12}$$

Moreover, $\deg g = \deg g_0$.

Proof. We factorize f over $V[x]$ as $f = \prod_1^m p_i^{r_i}$, where the p_i are distinct and irreducible over V, hence by inertia also irreducible over K. Since f is primitive, each p_i is primitive, and by Lemma 9.5.5, either $\bar{p}_i = \pi_i^{s_i}$ for some irreducible polynomial π_i over $\overline{K}$ and $s_i \geq 1$ (cases (i), (ii)), or $\bar{p}_i = \pi_i \in \overline{K}$ (case (iii)).

Suppose first that the highest coefficient of f is a unit. On dividing by it we may take f monic, and all the p_i (and hence the π_i) may also be taken monic. We have $\bar{g}_o \bar{h}_0 = \prod_1^m \pi_i^{r_i s_i}$ and since $\bar{g}_0$, $\bar{h}_0$ are relatively prime we can number the p_i so that $\bar{g}_0 = \prod_1^t \pi_i^{r_i s_i}$, $\bar{h}_0 = \prod_{t+1}^m \pi_i^{r_i s_i}$; here the π_i need not be distinct, but if $\pi_i = \pi_j$ then $i, j \leq t$ or $i, j > t$. Hence there is just one way to satisfy (9.5.12), namely by putting

$$g = \prod_1^t p_i^{r_i},\ h = \prod_{t+1}^m p_i^{r_i}.$$

If the highest coefficient of f is a non-unit, then case (iii) will occur, i.e. $\bar{p}_i \in \overline{K}$ for some i. Now we can number the p_i so that $p_1, \ldots, p_u$ are monic, while $p_{u+1}, \ldots, p_m$ have a non-unit as highest coefficient, and of course $u < m$. We can further renumber the $p_1, \ldots, p_u$ so that $g_0 = \prod_1^t \pi_i^{r_i s_i}$, $h_0 = \prod_{t+1}^m \pi_i^{r_i s_i}$, where $t \leq u$. Now there are several ways of satisfying (9.5.12), by taking one or more of the factors $p_{u+1}, \ldots, p_m$ to g, but if g is to be monic, we must have $g = \prod_1^t p_i^{r_i}$, $h = \prod_{t+1}^m p_i^{r_i}$. This then is the unique solution; since $\bar{g} = \bar{g}_0$ and both g, g_0 are monic, they have the same degree. ∎

We remark that in the proof of Hensel's lemma we assumed the field K to be complete. This is a 'topological' condition, involving limiting processes; thus the completion of a field generally has a higher cardinal than the field itself. Sometimes it is preferable, instead of completeness, to assume the conclusion of Hensel's lemma. A field with this property is called *Henselian*, and every field with a real-valued valuation has a least Henselian extension, its 'Henselization'. This obtained by a purely algebraic process, analogous to forming the algebraic closure. We also note that a

field is Henselian iff its valuation has a unique extension to any finite field extension (see Endler (1972) and Exercise 9 below).

We next consider algebraic extensions of incomplete fields. In Section 11.7 we shall show that for two fields E, F over a common subfield k, such that F/k is separable of degree n, we have

$$E \otimes_k F \cong K_1 \times \ldots \times K_r, \tag{9.5.13}$$

where $K_1, \ldots, K_r$ are fields containing a copy of E and of F and generated by them. Let $\alpha \in F$ and write f, g_i for the characteristic polynomial of α over k, and of its image in K_i over E respectively. Then $f = g_1 \ldots g_r$; for by definition, $f(x) = \det(xI - A)$, where $A = (a_{ij})$ and $\alpha v_i = \sum a_{ij} v_j$ for a basis $v_1, \ldots, v_n$ of F/k. Clearly f is also the characteristic polynomial of α as an element of $E \otimes F$ over E, because the v_i will be a basis for $E \otimes F$ over E. We now change the basis in $E \otimes F$ by choosing a basis adapted to the decomposition on the right of (9.5.13). Each K_i is mapped into itself by α, hence we have $f = g_1 \ldots g_r$. In particular, comparing last coefficients and second coefficients in this equation we find

$$N_{F/K}(\alpha) = \prod N_{K_i/E}(\alpha), \tag{9.5.14}$$

$$T_{F/K}(\alpha) = \sum T_{K_i/E}(\alpha). \tag{9.5.15}$$

With these preparations we can describe the relation between the ramification indices and residue degrees which takes the place of (9.5.5) in the incomplete case.

Theorem 9.5.7. *Let k be a field with a principal valuation v, and let K/k be a separable extension of degree n. Then there are r extensions $w_1, \ldots, w_r$ of v to K, where $1 \leq r \leq n$, and if w_i has ramification index e_i and residue degree f_i, then*

$$\sum e_i f_i = n. \tag{9.5.16}$$

Moreover, if k, K have completions $\tilde{k}, K_i$ under v, w_i respectively, then

$$\tilde{k} \otimes_k K \cong K_1 \times \ldots \times K_r. \tag{9.5.17}$$

Proof. Since we are dealing with principal valuations whose restrictions to k are all equal, two extensions of v to K are equivalent iff they are actually equal.

If $\tilde{K}$ is a completion of K with respect to a valuation w extending v, then $\tilde{k}$ is isomorphic to a subfield of $\tilde{K}$. We shall identify it with its image; then $\tilde{k}K$ is a dense subset of $\tilde{K}$, but since K/k is finite, it is a complete $\tilde{k}$-space by Proposition 9.2.7, so $\tilde{K} = \tilde{k}K$. Conversely, given a field of the form $\hat{K} = \tilde{k}K$, there is by Proposition 9.5.3 a unique valuation w on $\hat{K}$ extending the valuation v defined on $\tilde{k}$, and $\hat{K}$ is complete by Proposition 9.2.7. Since k is dense in $\tilde{k}$, K is dense in $\hat{K}$, so $\hat{K}$ is the completion of K with respect to the restriction of w to K. Thus we have a bijection between the set of valuations of K extending v and the set of fields generated by $\tilde{k}$ and K.

By (9.5.13) we have a decomposition of the form (9.5.17), where the K_i are fields generated by $\tilde{k}$ and K, and hence are the completions of K with respect to the different valuations extending v.

Finally let $[K_i : \tilde{k}] = n_i$. For a principal valuation the residue class field and value group are not affected by passing to completions, hence by Theorem 9.5.1, $n_i = e_i f_i$ and by (9.5.17), $n = [K : k] = [\tilde{k} \otimes K : \tilde{k}] = \sum n_i = \sum e_i f_i$. ∎

We again observe that this result has an analogue for Archimedean absolute values. There $\tilde{k} = \mathbf{R}$ or $\mathbf{C}$ and if e.g. $\mathbf{R} \otimes K = K_1 \times \ldots \times K_r$, we then have $[K_i : \mathbf{R}] = 1$ or 2 according as K_i is real or complex.

For Galois extensions the statement of Theorem 9.5.7 can be simplified:

Corollary 9.5.8. *If in Theorem 9.5.7 K/k is Galois, then the automorphisms of $\tilde{k} \otimes K$ induced by $\mathrm{Gal}(K/k)$ permute the K_i transitively. Hence $e_i = e, f_i = f$ independently of i, and*

$$[K : k] = efr,$$

where r is the number of extensions of v to K.

Proof. The result will follow if we can show that for some $\sigma \in \mathrm{Gal}(K/k)$, $w_1 = \sigma w_2$, i.e. $w_1(x) = w_2(x^\sigma)$ for all $x \in K$. To this end consider the valuations of K defined by

$$x \mapsto w_1(x^\sigma), \quad x \mapsto w_2(x^\sigma),$$

as σ ranges over $\mathrm{Gal}(K/k) = G$. If these two sets of valuations are disjoint, then by the approximation theorem there exists $a \in K$ such that $w_1(a^\sigma) > 0$, $w_2(a^\sigma - 1) > 0$ for all $\sigma \in G$. Hence $w_2(a^\sigma) = 0$ and $v(N(a)) = \sum w_2(a^\sigma) = 0$, but also $v(N(a)) = \sum w_1(a^\sigma) > 0$, which is a contradiction. Hence the sets are not disjoint, say $w_1(x^\sigma) = w_2(x^\tau)$ for some $\sigma, \tau \in G$, and so $w_1 = \sigma^{-1}.\tau w_2$ as desired. This shows that the K_i are permuted transitively and the rest is clear. ∎

We remark that the factors r, f, e correspond to the three kinds of extension: decomposition, residue class field extension and ramification respectively.

Exercises

1. Let K be a field with a principal valuation. Show that the residue class field has the same characteristic as K iff the valuation is trivial on the prime subfield of K.
2. Prove the first part of Proposition 9.5.3 for real-valued valuations by finding a maximal proper subring containing the valuation ring of v and using Theorem 9.4.6.
3. Use Proposition 9.5.2 and Theorem 9.4.4 to show that if an integral domain R is integrally closed, then so is the polynomial ring $R[x]$.
4. Let K be a complete valued field with residue class field k of characteristic 0. Show that K contains a subfield k_0 isomorphic to k and mapped to k by the residue class mapping. (Hint. Use Lemma 9.3.2.)

5. Let K be a field with a valuation v and consider the rational function field $K(t)$. If x is another generator, say $x = (at + b)(ct + d)^{-1}$, find the conditions under which x and t define the same Gaussian extension of v.
6. Let K be a field with a principal valuation v and let $K(t)$ be the rational function field. With a positive real number λ define a function w by $w(a_0 + a_1 t + \ldots + a_n t^n) = \min_i \{v(a_i) + i\lambda\}$. Show that w is a valuation on $K[t]$ which can be extended to $K(t)$. Determine its value group and residue class field.
7. Let E/K be a field extension of degree n and suppose that E is complete under a principal valuation v. Denote the valuation rings in K, E by V, W and their residue class fields by $\overline{K}$, $\overline{E}$. Show that if $\overline{E}/\overline{K}$ is a separable extension then W is a free V-module with basis $1, \alpha, \ldots, \alpha^{n-1}$ for a suitable $\alpha \in W$. (Hint. Use Theorem 9.5.1 and the fact that a separable extension is simple.)
8. Use Hensel's lemma to find a root of $x^5 = 2$ in $\mathbf{Q}_7$.
9. Let K be a field with a real-valued valuation v which has a unique extension to any finite extension of K. Show that formula (9.5.8) holds, and that Corollary 9.5.4, Lemma 9.5.5 and Theorem 9.5.6 all still hold. (This exercise suggests that we can obtain the Henselization of K by taking each irreducible polynomial f over K, primitive and with coefficients in V, say, and if $\overline{f}$ can be factorized into relatively prime factors over $\overline{K}$, say $\overline{f} = \overline{g}\overline{h}$, adjoining enough elements from an algebraic closure of K to K to enable us to lift $\overline{g}$, $\overline{h}$ to K; see Endler (1972).)
10. Let K be a field with a real-valued valuation v. Show that v has a unique extension to any purely inseparable extension of K.

Further Exercises for Chapter 9

1. Let Γ be any totally ordered abelian group. Verify that the group algebra $k\Gamma$ (for any field k) is an integral domain. (If Γ is written additively, it is convenient to write the elements of $k\Gamma$ as formal 'polynomials' $\sum c_\alpha x^\alpha$, where $c_\alpha \in k$, $\alpha \in \Gamma$, with the rule $x^\alpha x^\beta = x^{\alpha+\beta}$.) Show that the field of fractions K of $k\Gamma$ has a valuation with value group Γ. Determine the valuation ring and maximal ideal of this valuation. Examine the particular case $\Gamma = \mathbf{Z}$.
2. Show that Theorem 9.2.3 still holds for non-commutative rings.
3. Show that a function f from $\mathbf{Z}_p$ to $\mathbf{Q}_p$ takes integer values on $\mathbf{Z}$ iff all its coefficients $D^n f(0)$ are integers.
4. Show that $\mathbf{Q}_p$ and $\mathbf{Q}_q$ for distinct primes p, q are not isomorphic. (Hint. Look for solutions of $x^2 = p$.)
5. Let E, F, G be fields, let α be a place of E in F and β be a place of F in G. Show that $\alpha\beta$ (suitably defined) is a place of E in G.
6. Let $k \subset K$ be fields and V be a minimal valuation ring in K containing k. Using Exercise 5, show that the corresponding place is algebraic over k. Use Zorn's lemma to show that such minimal valuation rings always exist. Deduce that if A is a subring of a field K, then any homomorphism of A into an algebraically closed field E can be extended to a place of K in E.
7. Let A be an integral domain and K be its field of fractions; A is called *completely integrally closed* in K if for any $x \in K$ for which there exists $d \in K^\times$ such that

$dx^n \in A$ for all $n > 0$, it follows that $x \in A$. Verify that a completely integrally closed ring is integrally closed, and show that the intersection of any set of principal valuation rings is completely integrally closed.

8. Show that a valuation ring is completely integrally closed iff its value group is Archimedean ordered (see Exercise 7 and Section 8.7).
9. A finite extension E/K of complete valued fields is said to be *totally ramified* if $[E : K] = e$ is the ramification index. Given an extension E/K with a principal valuation v and with uniformizers p, π in K, E respectively, show that E/K is totally ramified iff π satisfies an equation $x^e + a_1 p x^{e-1} + \ldots + a_{e-1} p x + a_e p = 0$, where $v(a_i) \geq 0$ for $i = 1, \ldots, e-1$, $v(a_e) = 0$, and $E = K(\pi)$. (Such an equation is called an *Eisenstein equation* and the corresponding polynomial is an *Eisenstein polynomial.*)
10. (a) Show that any polynomial over $\mathbf{Q}$ close (in the usual absolute value) to a polynomial with n simple real zeros itself has n simple real zeros. (b) Show that a polynomial over $\mathbf{Q}$ close (in the p-adic valuation) to an Eisenstein polynomial for the prime p is itself Eisenstein and hence irreducible. (c) For any integers $0 \leq n \leq m$ construct an irreducible polynomial over $\mathbf{Q}$ of degree m with exactly n real zeros.
11. (Krasner's lemma) Let L be a complete field under a non-Archimedean absolute value and K be a complete subfield. Given $\alpha, \beta \in L$, where α is algebraic over K and β is separable algebraic over $K(\alpha)$, show that if α is closer to β than are any of the conjugates of β over K, then β is fixed under all automorphisms of $K(\alpha, \beta)/K(\alpha)$ and deduce that $\beta \in K(\alpha)$.
12. Let K be a complete valued field. Show that if f is monic, irreducible and separable over K, then any polynomial g sufficiently close to f is also irreducible, and to any zero α of f there corresponds a zero β of g such that $K(\alpha) = K(\beta)$.
13. Let K be an algebraically closed field with a real-valued valuation. Show that its completion $\tilde{K}$ is again algebraically closed. (Hint. First show that the zeros of a polynomial are continuous functions of the coefficients. Now any polynomial f over $\tilde{K}$ may be approximated by polynomials f_v over K, and if f is separable of degree d, then the zeros of these polynomials f_v can be arranged in d Cauchy sequences.)
14. Let K be a complete valued field and F be its algebraic closure. Show that the valuation has a unique extension to F, which is real-valued if the original valuation was so, and in that case the completion of F is again algebraically closed.
15. In the algebraic closure F of $\mathbf{Q}_p$ take any sequence (a_n) such that a_r is a power of a_s for $r \leq s$ (e.g. a_r could be taken a primitive $(r!)$-th root of 1); denote the degree of a_r over $\mathbf{Q}_p$ by n_r. Show that for any $n < n_r$ and any $h \in \mathbf{N}$ there exists $k > h$ such that a_r satisfies no congruence of degree n (mod p^k). Deduce the existence of a sequence of integers $k_1 < k_2 < \ldots$ such that a_r satisfies no congruence of degree $< n_r$ (mod p^{k_r}), and hence show that the series $\sum a_r p^{k_r}$ converges to an element transcendental over $\mathbf{Q}_p$. Deduce that F is not complete (its completion is algebraically closed, by Exercise 13).

10 Commutative Rings

Commutative ring theory has its origins in number theory and algebraic geometry in the 19th century. Today it is of particular importance in algebraic geometry, and there has been an interesting interaction of algebraic geometry and number theory, using the methods of commutative algebra. Here we can do no more than describe the basic techniques and take the first steps in the subject. In Section 10.1 we define the various operations on ideals and use them in Section 10.2 to study unique factorization. In Section 10.3 we give an account of fractions and examine the effect of chain conditions in Section 10.4. Many rings of algebraic numbers fail to have unique factorization of elements, but instead have unique factorization of ideals, and the consequences are studied in Sections 10.5 and 10.6. Sections 10.7–10.10 deal with the properties of rings used in algebraic geometry (but also of importance in commutative ring theory): equations (Section 10.7), decomposition of ideals (Section 10.8), dimension (Section 10.9) and the relation between ideals and algebraic varieties (Section 10.10).

Throughout this chapter all rings will be commutative, unless otherwise stated; for this reason there is no need to distinguish between left and right modules.

10.1 Operations on Ideals

Historically the first ring to be studied was the ring $\mathbf{Z}$ of integers; the term 'ring' was first used by David Hilbert (1897) in his *Zahlbericht* for a ring of algebraic integers (Zahlring). In $\mathbf{Z}$ every ideal is principal; in fact ideals were first introduced (by Ernst-Eduard Kummer) as 'ideal numbers' in rings of algebraic integers which lacked unique factorization. In $\mathbf{Z}$ we can from any two numbers a, b form their highest common factor (HCF; also called greatest common divisor) (a, b), their product ab and their least common multiple (LCM) $[a, b]$. These operations correspond to operations on ideals in any ring.

We shall denote ideals in a ring R by small gothic letters $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}, \ldots$. If $\mathfrak{a}$ is an ideal in a commutative ring R, generated by elements $a_1, \ldots, a_n$, we write $\mathfrak{a} = (a_1, \ldots, a_n)$; then the elements of $\mathfrak{a}$ are all the linear combinations $\sum r_i a_i$ $(r_i \in R)$. In particular, when $n = 1$, $\mathfrak{a}$ is principal; thus (a) is the ideal consisting of all elements ra $(r \in R)$.

Given ideals $\mathfrak{a}$, $\mathfrak{b}$ in R, we define their *sum* as

$$\mathfrak{a} + \mathfrak{b} = \{x + y | x \in \mathfrak{a}, y \in \mathfrak{b}\}.$$

It is easily seen that this is an ideal, the least ideal containing $\mathfrak{a}$ and $\mathfrak{b}$. If $\mathfrak{a}$, $\mathfrak{b}$ are principal, say $\mathfrak{a} = (a)$, $\mathfrak{b} = (b)$, then their sum, if principal, has the form (d), where d is an HCF of a and b. For example in $\mathbf{Z}$, $(36) + (10) = (2)$; on the other hand in $\mathbf{Z}[x]$, $(x) + (2) = (x, 2)$ and this cannot be simplified.

Similarly the product is defined as

$$\mathfrak{a}\mathfrak{b} = \left\{\sum x_i y_i | x_i \in \mathfrak{a},\ y_i \in \mathfrak{b}\right\}.$$

This operation is again associative and commutative, so that ideals form a commutative monoid under multiplication, with $(1) = R$ as neutral element. If $\mathfrak{b} = \mathfrak{a}$, the product $\mathfrak{a}\mathfrak{a}$ is written $\mathfrak{a}^2$ and generally $\mathfrak{a}^n$ is defined recursively by $\mathfrak{a}^n = \mathfrak{a}^{n-1}\mathfrak{a}$. A third operation is the intersection $\mathfrak{a} \cap \mathfrak{b}$, which corresponds to the LCM when all ideals are principal.

Let us note the form these definitions take for finitely generated ideals. If $\mathfrak{a} = (a_1, \ldots, a_r)$, $\mathfrak{b} = (b_1, \ldots, b_s)$, then $\mathfrak{a} + \mathfrak{b} = (a_1, \ldots, a_r, b_1, \ldots, b_s)$, while $\mathfrak{a}\mathfrak{b} = (a_1 b_1, a_1 b_2, \ldots, a_1 b_s, a_2 b_1, \ldots, a_r b_s)$. This shows in particular that the sum and product of finitely generated ideals are again finitely generated. By contrast there is no simple expression for $\mathfrak{a} \cap \mathfrak{b}$, and it need not be finitely generated, even when both $\mathfrak{a}$ and $\mathfrak{b}$ are, but we note that there is a short exact sequence relating $\mathfrak{a} \cap \mathfrak{b}$ to $\mathfrak{a} + \mathfrak{b}$:

$$0 \to \mathfrak{a} \cap \mathfrak{b} \xrightarrow{\lambda} \mathfrak{a} \oplus \mathfrak{b} \xrightarrow{\mu} \mathfrak{a} + \mathfrak{b} \to 0. \tag{10.1.1}$$

Here μ is defined as $(x, y)\mu = x - y$ $(x \in \mathfrak{a}, y \in \mathfrak{b})$ and $\ker \mu \cong \mathfrak{a} \cap \mathfrak{b}$; this is easily verified and will be left to the reader to do.

There is also a form of division for ideals. Given ideals $\mathfrak{a}$, $\mathfrak{b}$, we define

$$\mathfrak{a} : \mathfrak{b} = \{x \in R | x\mathfrak{b} \subseteq \mathfrak{a}\}.$$

The reader will have no difficulty in verifying that this is an ideal. This is so even if $\mathfrak{b}$ is an arbitrary subset of R, not necessarily an ideal. Thus for any subset S of R we have

$$\mathfrak{a} : S = \{x \in R | xS \subseteq \mathfrak{a}\}.$$

The addition, multiplication and intersection of ideals are associative and commutative and satisfy the distributive law: $\mathfrak{a}(\mathfrak{b} + \mathfrak{c}) = \mathfrak{a}\mathfrak{b} + \mathfrak{a}\mathfrak{c}$, as is easily checked. Division is related to the other operations by the formulae:

$$(\mathfrak{a} : \mathfrak{b})\mathfrak{b} \subseteq \mathfrak{a} \subseteq \mathfrak{a} : \mathfrak{b}, \tag{10.1.2}$$

$$(\cap \mathfrak{a}_i) : \mathfrak{b} = \cap(\mathfrak{a}_i : \mathfrak{b}), \quad \mathfrak{a} : \left(\sum \mathfrak{b}_i\right) = \cap(\mathfrak{a} : \mathfrak{b}_i), \tag{10.1.3}$$

$$(\mathfrak{a} : \mathfrak{b}) : \mathfrak{c} = \mathfrak{a} : \mathfrak{b}\mathfrak{c}. \tag{10.1.4}$$

For example, to prove (10.1.4), we have, for any $x \in R$, $x \in (\mathfrak{a} : \mathfrak{b}) : \mathfrak{c} \Leftrightarrow x\mathfrak{c} \subseteq \mathfrak{a} : \mathfrak{b} \Leftrightarrow x\mathfrak{b}\mathfrak{c} \subseteq \mathfrak{a} \Leftrightarrow x \in \mathfrak{a} : \mathfrak{b}\mathfrak{c}$.

The other rules are established similarly.

Exercises

1. Prove the formulae (10.1.2)–(10.1.4).
2. Show that $(\mathfrak{a} \cap \mathfrak{b})(\mathfrak{a} + \mathfrak{b}) \subseteq \mathfrak{a}\mathfrak{b}$ for any ideals $\mathfrak{a}$, $\mathfrak{b}$ in a ring R, and give an example in $\mathbf{Z}[x]$ where the inequality is strict.
3. Show that $(\mathfrak{a} + \mathfrak{c})(\mathfrak{b} + \mathfrak{c}) \subseteq \mathfrak{a}\mathfrak{b} + \mathfrak{c}$ for any ideals $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$ in a ring R, and give examples where the inequality is strict.
4. Let $\mathfrak{a}$, $\mathfrak{b}$ be finitely generated ideals in an integral domain. Show that if $\mathfrak{a} + \mathfrak{b}$ is principal, then $\mathfrak{a} \cap \mathfrak{b}$ is finitely generated. (Hint. Use the exact sequence (10.1.1).)
5. Let $K = k(x, y, z_1, z_2, \ldots)$, where k is a field and x, y, z_i $(i = 1, 2, \ldots)$ are indeterminates. Show that in the k-subalgebra of K generated by $x, y, z_i, xy^{-1}z_i$ $(i = 1, 2, \ldots)$ neither $(x) \cap (y)$ nor $(x) : (y)$ is finitely generated.

10.2 Prime Ideals and Factorization

In any integral domain R an element is said to be *irreducible* or an *atom* if it is a non-unit, which cannot be expressed as a product of two non-units. For example, in $\mathbf{Z}$ the atoms are just the prime numbers; $\mathbf{Z}$ has the further important property that every integer can be written as a product of prime numbers in essentially only one way. Let us define more generally a *unique factorization domain* (UFD) as an integral domain in which every non-unit has a *complete factorization*, i.e. it can be written as a product of atoms, and in any two complete factorizations of an element the factors are the same except for the order and unit factors. In order to study UFDs it is convenient to single out another property of atoms. By a *prime* we shall understand an element p in an integral domain R, not zero or a unit such that for any $a, b \in R$, $p|ab \Rightarrow p|a$ or $p|b$. Any prime is necessarily an atom, for if p is a prime and $p = ab$ $(a, b \neq 0)$, then by definition, $p|a$ or $p|b$, say $p|a$. Hence $a = pc = abc$; it follows that $bc = 1$, so b is a unit and p is an atom, *associated* to a (i.e. differing by a unit factor). We note that the converse, with a finiteness condition, characterizes UFDs:

Theorem 10.2.1. *Let R be an integral domain. Then R is a unique factorization domain if and only if*

(i) *every element of R not zero or a unit has a complete factorization and*
(ii) *every atom is a prime.*

Proof. If R is a UFD, then every element not zero or a unit has a complete factorization. Now let p be an atom and suppose that $p|ab$, say $ab = pc$. Take complete factorizations $a = p_1 \ldots p_r$, $b = q_1 \ldots q_s$, $c = z_1 \ldots z_t$. Then $pz_1 \ldots z_t =$

$p_1 \dots p_r q_1 \dots q_s$ and by unique factorization p must be associated to some p_i or q_j and accordingly $p|a$ or $p|b$.

Conversely, assume (i), (ii) and take two complete factorizations of an element (which exist by (i)):

$$c = p_1 \dots p_r = q_1 \dots q_s. \tag{10.2.1}$$

We have to show that $r = s$ and after suitable renumbering each q_i is associated to p_i. We shall use induction on r; for $r = 1$ there is nothing to prove, for then s is also 1. Now let $r > 1$; p_1 is an atom, hence prime and $p_1|q_1 \dots q_s$, therefore $p_1|q_j$ for some j, say $j = 1$ (by renumbering the q_j). Since q_1 is also an atom, it must be associated to p_1, say $p_1 = q_1 u$ for a unit u. Dividing (10.2.1) by q_1, we obtain

$$up_2 \dots p_r = q_2 \dots q_s.$$

By induction, $r - 1 = s - 1$, hence $r = s$; moreover, we can renumber $q_2, \dots, q_r$ so that p_i is associated to q_i. This also holds for $i = 1$ and it proves the uniqueness of the factorization (10.2.1), so R is a UFD. ∎

Thus a UFD may be characterized as an integral domain in which every element not zero or a unit can be written as a product of primes. An integral domain satisfying (i) will be called *atomic*.

It is easily seen that every Noetherian domain is atomic. For let R be Noetherian and take $a \in R$. If a is not 0 or a unit, we can split off a non-unit factor b_1 and continue in this way: $a = b_1 c_1 = b_1 b_2 c_2 = \dots$. Since the b_i are non-units, we have an ascending chain

$$aR \subset c_1 R \subset c_2 R \subset \dots;$$

this chain must break off, so some c_i must be an atom. We now repeat the process, taking each b_i to be an atom (as we may, by what has just been shown) and so obtain a factorization of a as a product of atoms. Thus we have

Corollary 10.2.2. *Any Noetherian domain is a unique factorization domain if and only if every atom is a prime.* ∎

Corollary 10.2.3. *Every principal ideal domain is a unique factorization domain.*

Proof. Let R be a PID; then R is Noetherian, so we need only prove that every atom is prime. Let a be an atom and suppose that a divides bc but not b, say $bc = ah$. Let $(a, b) = (d)$; then d is a common factor of a and b, and so must be a unit, which may be taken to be 1, i.e. $au + bv = 1$. Hence $c = auc + bcv = a(uc + hv)$, so $a|c$, as required. ∎

For example, the well-known unique factorization of the integers follows from Corollary 10.2.3 because $\mathbf{Z}$ is a PID. This is most easily proved by the division algorithm, which states that given $a, b \in \mathbf{Z}$, where $b \neq 0$, there exist $q, r \in \mathbf{Z}$ such that

$$a = bq + r, \quad |r| < |b|. \tag{10.2.2}$$

More briefly, there exists $q \in \mathbf{Z}$ such that

$$|a - bq| < |b|.$$

More generally, a commutative integral domain R is said to be *Euclidean* if with each element $a \in R$ a non-negative integer $|a|$ is associated such that $|ab| \geq |a|$ for all $a, b \neq 0$ in R and q, r exist to satisfy (10.2.2). Such a ring is always a PID, for if $\mathfrak{a}$ is any ideal in R, if $\mathfrak{a} \neq 0$, let $c \in \mathfrak{a}$ be such that $|c|$ is least. Given $a \in \mathfrak{a}$, we can write $a = cq + r$, where $|r| < |c|$; clearly $r = a - cq \in \mathfrak{a}$, so by minimality of $|c|$ we have $r = 0$, therefore $c \in (a)$ and this shows $\mathfrak{a}$ to be principal. Thus we have

Corollary 10.2.4. *Every Euclidean domain is a principal ideal domain and hence also a unique factorization domain.* ■

In the study of algebraic number theory it was found that rings of algebraic integers always satisfy (i) but not (ii) of Theorem 10.2.1, so they may not be UFDs; certain pairs of elements a, b fail to have an HCF, so that (a, b) is non-principal. It was this fact that led Kummer and Dedekind to develop ideal theory. Here the analogue of an atom is a maximal ideal, i.e. an ideal which is maximal among proper ideals (see Section 3.2). Even more important is the notion of prime ideal. In any commutative ring R a *prime ideal* is a proper ideal $\mathfrak{p}$ such that $xy \in \mathfrak{p}$ implies $x \in \mathfrak{p}$ or $y \in \mathfrak{p}$; note that R itself is not a prime ideal. To illustrate the definition, an element p is prime iff (p) is a non-zero prime ideal. An illustration of non-principal prime ideals is the following chain of prime ideals in the polynomial ring $k[x_1, \ldots, x_n]$, where k is a field:

$$0 \subset (x_1) \subset (x_1, x_2) \subset \ldots \subset (x_1, x_2, \ldots, x_n).$$

We note that an ideal $\mathfrak{p}$ in a ring R is prime iff $R/\mathfrak{p}$ is an integral domain. Since $\mathfrak{p}$ is maximal iff $R/\mathfrak{p}$ is a field, we deduce

Proposition 10.2.5. *In a commutative ring R every maximal ideal is prime. In particular, every non-trivial commutative ring has prime ideals.*

Proof. The first part is clear by what has been said. The second part follows because by Krull's theorem (Theorem 4.2.6), every non-trivial ring has maximal ideals. ■

Krull's theorem has a useful generalization. Let us call a subset S of R *multiplicative* if $1 \in S$ and $x, y \in S$ implies $xy \in S$.

Theorem 10.2.6. *Let R be a commutative ring, S be a multiplicative subset and $\mathfrak{a}$ be an ideal of R disjoint from S. Then there exists an ideal $\mathfrak{m}$ in R which contains $\mathfrak{a}$, is disjoint from S and is maximal subject to these conditions. Any such ideal $\mathfrak{m}$ is prime.*

Proof. Let $\mathscr{A}$ be the set of all ideals $\mathfrak{a}'$ with the properties $\mathfrak{a}' \supseteq \mathfrak{a}$, $\mathfrak{a}' \cap S = \varnothing$. $\mathscr{A}$ is inductive, for if $\{\mathfrak{c}_\lambda\}$ is a chain of ideals in $\mathscr{A}$, their union $\mathfrak{c}$ is an ideal containing $\mathfrak{a}$; of course if the chain is empty, then $\mathfrak{c} = \mathfrak{a}$. If $\mathfrak{c} \cap S \neq \varnothing$, take $x \in \mathfrak{c} \cap S$; then $x \in \mathfrak{c}_\lambda$

for some λ and so $\mathfrak{c}_\lambda \cap S \neq \varnothing$, a contradiction. Hence $\mathscr{A}$ is inductive; by Zorn's lemma it has a maximal member $\mathfrak{m}$ and this is an ideal with the required properties.

Now let $\mathfrak{m}$ be as stated; since $1 \in S$, $1 \notin \mathfrak{m}$, so $\mathfrak{m}$ is proper. If $b, c \notin \mathfrak{m}$, $bc \in \mathfrak{m}$, then by the maximality of $\mathfrak{m}$, $(\mathfrak{m} + (b)) \cap S \neq \varnothing$, say $s = bu + x \in S$, where $u \in R$, $x \in \mathfrak{m}$, and similarly $t = cv + y \in S$, where $v \in R$, $y \in \mathfrak{m}$. Then S contains

$$st = (bu + x)(cv + y) = bcuv + x(cv + y) + buy \in \mathfrak{m},$$

and this contradicts the fact that $\mathfrak{m} \cap S = \varnothing$; it follows that $\mathfrak{m}$ is prime. ■

If S is a multiplicative subset of R such that $0 \notin S$, then we can apply Theorem 10.2.6 with $\mathfrak{a} = 0$ and obtain

Corollary 10.2.7. *Given a multiplicative set S not containing 0 in a commutative ring R, there exist ideals in R that are maximal subject to being disjoint from S.* ■

As examples of multiplicative sets we mention (i) the multiplicative set generated by an element f of R, viz. $\{1, f, f^2, \dots\}$, (ii) the complement of a prime ideal and (iii) the set $1 + \mathfrak{a} = \{1 + a | a \in \mathfrak{a}\}$, where $\mathfrak{a}$ is an ideal.

A multiplicative set S is said to be *saturated* if $ab \in S$ implies $a \in S$. Since $ab = ba$, it then also follows that $b \in S$. Clearly the complement of any prime ideal is multiplicative and saturated. In the opposite direction we have

Proposition 10.2.8. *Let R be a commutative ring and S be a subset. Then S is multiplicative and saturated if and only if its complement $R\backslash S$ is a union of prime ideals.*

Proof. Clearly the complement of any union of prime ideals is multiplicative and saturated. Conversely, if S is multiplicative and saturated, let $a \notin S$; then $ab \notin S$ for all $b \in R$, hence $(a) \cap S = \varnothing$, so by Theorem 10.2.6 there is a prime ideal $\mathfrak{p}$ containing a and disjoint from S. It follows that the union of all prime ideals disjoint from S is precisely $R\backslash S$. ■

We can also use Theorem 10.2.6 to describe the intersection of all prime ideals in a ring.

Proposition 10.2.9. *In a commutative ring R, the set $\mathfrak{N}$ of all nilpotent elements is an ideal, equal to the intersection of all prime ideals.*

Proof. We have to show that

$$\mathfrak{N} = \cap \mathfrak{p}, \tag{10.2.3}$$

where the intersection is over all prime ideals of R. If $a \in \mathfrak{N}$, then $a^n = 0$ for some $n \geq 1$. Hence for any prime ideal $\mathfrak{p}$, $a^n \in \mathfrak{p}$ and so $a \in \mathfrak{p}$; thus $\mathfrak{N} \subseteq \cap \mathfrak{p}$. Now let $a \notin \mathfrak{N}$; then $a^n \neq 0$ for all $n \geq 1$, hence $0 \notin \{1, a, a^2, \dots\}$ and by Theorem 10.2.6 there is a prime ideal disjoint from $\{1, a, a^2, \dots\}$. Hence the right-hand side of (10.2.3) does not contain a and it follows that we have equality. ■

The ideal $\mathfrak{N}$ consisting of all nilpotent elements of R is called the *nilradical* of R. Given any ideal $\mathfrak{a}$ of R, we can define the *radical* of $\mathfrak{a}$ in R as

$$\sqrt{\mathfrak{a}} = \{x \in R | x^n \in \mathfrak{a} \text{ for some } n \geq 1\}.$$

Thus $\sqrt{\mathfrak{a}}$ is the inverse image, under the natural homomorphism $R \to R/\mathfrak{a}$, of the nilradical of $R/\mathfrak{a}$. Applying Proposition 10.2.9 and bearing in mind that the prime ideals of $R/\mathfrak{a}$ are the images of prime ideals of R containing $\mathfrak{a}$, by the third isomorphism theorem, we see that $\sqrt{\mathfrak{a}}$ is the intersection of all prime ideals containing $\mathfrak{a}$. An ideal coinciding with its radical is called a *radical ideal.*

In terms of prime ideals there is another criterion for unique factorization.

Theorem 10.2.10. *An integral domain is a unique factorization domain if and only if every non-zero prime ideal contains a prime element.*

Proof. In any domain R let S be the set of all products of prime elements and units. Clearly S is multiplicative; it is also saturated, for if $ab \in S$, either ab is a unit; then a, b are both units; or $ab = p_1 \dots p_r u$ ($r \geq 1$, p_i prime, u a unit), then $p_1 | ab$, hence $p_1 | a$ or $p_1 | b$, say the former, $a = p_1 a_1$. Then $a_1 b = p_2 \dots p_r u$; now by an induction on r we see that $a_1, b \in S$, hence also $a = p_1 a_1 \in S$, so S is indeed saturated. By Proposition 10.2.8 its complement is a union of prime ideals. But any prime element belongs to S, so the prime ideals disjoint from S contain no prime elements. Therefore if R satisfies the conditions of the theorem, then S consists of all non-zero elements and so R is then a UFD.

Conversely, let R be a UFD and $\mathfrak{p}$ be a non-zero prime ideal. Take $0 \neq a \in \mathfrak{p}$; by unique factorization, $a = p_1 \dots p_r$, where p_i is prime, $r \geq 1$ and $p_1 \dots p_r \in \mathfrak{p}$, hence $p_i \in \mathfrak{p}$ for some i, because $\mathfrak{p}$ is a prime ideal. This shows the condition to be necessary. ∎

In a UFD every minimal non-zero prime ideal is principal. For let $\mathfrak{p}$ be a minimal non-zero prime ideal and p be a prime element in $\mathfrak{p}$; then $0 \subset (p) \subseteq \mathfrak{p}$, hence $\mathfrak{p} = (p)$, by minimality, and so $\mathfrak{p}$ is principal. In Noetherian rings the converse holds: a Noetherian domain in which each minimal non-zero prime ideal is principal is a UFD (see e.g. Nagata (1962) Theorem 13.1).

Exercises

1. Show that $\sqrt{\sqrt{\mathfrak{a}}} = \sqrt{\mathfrak{a}}$ and $\sqrt{\mathfrak{a}^n} = \sqrt{\mathfrak{a}}$, for any $n \geq 1$.
2. Show that $\sqrt{(\mathfrak{ab})} = \sqrt{(\mathfrak{a} \cap \mathfrak{b})} = \sqrt{\mathfrak{a}} \cap \sqrt{\mathfrak{b}}$ and $\sqrt{(\mathfrak{a} + \mathfrak{b})} = \sqrt{(\sqrt{\mathfrak{a}} + \sqrt{\mathfrak{b}})}$.
3. Show that in an Artinian ring every prime ideal is maximal.
4. Show that the intersection (and union) of any chain of prime ideals is again prime. Deduce the existence of minimal prime ideals in any non-trivial ring.
5. Show that if $a^m = b^n = 0$, then $(a + b)^{m+n-1} = 0$. Hence give a direct proof that the set of all nilpotent elements in a (commutative) ring is an ideal.
6. Show that an integral domain is Euclidean provided it has a norm function $|a|$ such that $|ab| = |a| . |b|$ and for any a, b with $|a| \geq |b|$ there exists c such that $|a - bc| < |a|$.

7. Show that an ideal $\mathfrak{a}$ in a ring R is prime iff it is proper and for any ideals $\mathfrak{b}$, $\mathfrak{c}$ in R, if $\mathfrak{bc} \subseteq \mathfrak{a}$, then $\mathfrak{b}$ and $\mathfrak{c}$ cannot both contain $\mathfrak{a}$ properly.
8. Show that for any ideal $\mathfrak{a}$ in a Noetherian ring, $(\sqrt{\mathfrak{a}})^n \subseteq \mathfrak{a}$, for some n depending on $\mathfrak{a}$. Give an example to show that this may fail without the Noetherian condition.
9. Let R be a ring and $\mathfrak{N}$ be its nilradical. Show that the following conditions are equivalent: (a) $R/\mathfrak{N}$ is an integral domain, (b) $\mathfrak{ab} = 0 \Rightarrow \mathfrak{a}^2 = 0$ or $\mathfrak{b}^2 = 0$, (c) $xy = 0 \Rightarrow x$ or y is nilpotent.
10. Show that $A = \{f \in \mathbf{Q}[x] | f(0) \in \mathbf{Z}\}$ is a ring in which every irreducible element is prime. Is it a UFD?

10.3 Localization

In the construction of the field of fractions of an integral domain the essential property of the set of denominators is its closure under multiplication. Let us now take any commutative ring R, not necessarily an integral domain, and let S be a multiplicative subset of R. Then we can construct fractions a/s with denominators in S as follows. We define an equivalence relation on the product set $R \times S$ by setting

$$(a, s) \sim (a', s') \Leftrightarrow (as' - sa')t = 0 \quad \text{for some } t \in S. \tag{10.3.1}$$

This relation is clearly reflexive and symmetric; to prove that it is transitive, let $(a, s) \sim (a', s')$ and $(a', s') \sim (a'', s'')$, say $(as' - sa')t = 0$, $(a's'' - s'a'')t' = 0$, where $t, t' \in S$. Then

$$\begin{aligned}(as'' - sa'')s'tt' &= as's''tt' - sa's''tt' + sa's''tt' - sa''s'tt' \\ &= (as' - sa')t.s''t' + (a's'' - s'a'')t'.st \\ &= 0.\end{aligned}$$

Since $s'tt' \in S$, this proves that $(a, s) \sim (a'', s'')$, and it follows that (10.3.1) is indeed an equivalence relation. The proof shows why we had to introduce t in definition (10.3.1). If S contains only non-zerodivisors we can replace the condition in (10.3.1) by $as' - sa' = 0$.

Let us write $\frac{a}{s}$ or a/s for the equivalence class containing (a, s) and define addition and multiplication by the rules

$$\frac{a}{s} + \frac{a'}{s'} = \frac{as' + sa'}{ss'}, \quad \frac{a}{s} \cdot \frac{a'}{s'} = \frac{aa'}{ss'}. \tag{10.3.2}$$

It is routine to verify that these operations are well-defined and that the set R_S of all equivalence classes forms a ring under the operations (10.3.2), with 0/1 as zero and 1/1 as unit element. The natural mapping $\lambda : R \to R_S$ given by

$$\lambda : x \mapsto x/1 \tag{10.3.3}$$

is clearly a homomorphism which maps every element of S to a unit in R_S, for if $s \in S$, then $(s/1)(1/s) = s/s = 1$. This ring R_S is called the *ring of fractions* with denominators in S.

Given any ring R and a subset S of R, we say that a homomorphism $f : R \to R'$ is *S-inverting* if it maps the elements of S to invertible elements of R'. As we have just seen, the homomorphism (10.3.3) is S-inverting, but we can say more than that:

Theorem 10.3.1. *Let R be a commutative ring and S be a multiplicative subset of R. Then there exists a ring R_S and a homomorphism $\lambda : R \to R_S$ which is universal S-inverting, i.e. it is S-inverting and for every S-inverting homomorphism $f : R \to R'$ there is a unique homomorphism $f' : R_S \to R'$ such that $f = \lambda f'$. Moreover, this property determines R_S up to isomorphism.*

The elements of R_S can be written as fractions a/s ($a \in R$, $s \in S$), where $a/s = a'/s'$ if and only if $(as' - sa')t = 0$ for some $t \in S$; the addition and multiplication in R_S are defined by (10.3.2) *and λ is given by* (10.3.3), *with kernel*

$$\ker \lambda = \{a \in R | at = 0 \text{ for some } t \in S\}.$$

Proof. We saw that λ as defined in (10.3.3) is S-inverting. Now let $f : R \to R'$ be any S-inverting homomorphism and define a mapping $f_1 : R \times S \to R'$ by

$$(a, s)f_1 = (af)(sf)^{-1}.$$

This is possible because f is S-inverting, and f_1 takes the same value on equivalent pairs: if $(as' - sa')t = 0$, then $(af.s'f - sf.a'f)tf = 0$ and hence $af.(sf)^{-1} = a'f.(s'f)^{-1}$. Thus we obtain a well-defined mapping $f' : R_S \to R'$ by putting $(a/s)f' = af.(sf)^{-1}$. This mapping is easily seen to be a homomorphism, with the help of (10.3.3), and it has the property

$$(a/1)f' = af, \tag{10.3.4}$$

i.e. $\lambda f' = f$. Moreover, it is the only such mapping, for (10.3.4) determines the values of f' on the elements $a/1$ and its value on $1/s$ must then be the inverse of its value on $s/1$. The uniqueness of R_S follows as usual by universality. Finally take $a \in \ker \lambda$; by (10.3.3) this means that $a/1 = 0/1$, i.e. $at = 0$ for some $t \in S$. ∎

We see in particular that the mapping $\lambda : R \to R_S$ is injective precisely when S consists of non-zerodivisors. For example, when R is an integral domain and $S = R^\times$, then R_S is just the field of fractions of R. At the other extreme, if $0 \in S$, then $a/s = 0$ for all $a \in R$, $s \in S$, because $(a.1 - s.0)0 = 0$; hence $R_S = 0$. Leaving this trivial case aside, we may suppose that $0 \notin S$.

An important application is to the case where S is the complement of a prime ideal $\mathfrak{p}$. Here one often writes $R_\mathfrak{p}$ in place of R_S, as we have done in Section 9.4. The ring $R_\mathfrak{p}$ just constructed is a local ring; in the canonical homomorphism $R \to R_\mathfrak{p}$ the prime ideal $\mathfrak{p}$ corresponds to the unique maximal ideal of $R_\mathfrak{p}$ consisting of all the non-units. This ring $R_\mathfrak{p}$ is also called the *local ring* of R at $\mathfrak{p}$, and the process of forming R is called *localization*.

It is essential to be clear about the distinction between $R/\mathfrak{p}$ and $R_\mathfrak{p}$. In rough terms we may think of $R/\mathfrak{p}$ as being formed from R by 'putting the elements in $\mathfrak{p}$ equal to 0', while $R_\mathfrak{p}$ is formed by 'making the elements outside $\mathfrak{p}$ invertible'. These two rings arise whenever we have a homomorphism from a ring into a field, $f : R \to K$ say, such that K is generated, as field, by the image of f. The kernel of this homomorphism is a prime ideal $\mathfrak{p}$ say, and there are two ways of analysing f, which form a commutative diagram:

$$\begin{array}{ccc} R & \longrightarrow & R_\mathfrak{p} \\ \downarrow & & \downarrow \\ R/\mathfrak{p} & \longrightarrow & K \end{array}$$

We can either take $R/\mathfrak{p}$, an integral domain, and form its field of fractions, or we can take the local ring $R_\mathfrak{p}$ and form the quotient by its maximal ideal, called the *residue class field* of $R_\mathfrak{p}$. Each time we get the same field K. Although one habitually uses the first route, it turns out that the second route (via $R_\mathfrak{p}$) is easier to generalize to non-commutative rings (see Cohn (1985) Chapter 7).

It is important to know the relation between ideals in R and in R_S. With every ideal $\mathfrak{a}$ in R we associate an *expanded ideal* $\mathfrak{a}_S$ or $\mathfrak{a}^e$ of R generated by the image $\mathfrak{a}\lambda$:

$$\mathfrak{a}^e = \{a/s | a \in \mathfrak{a},\ s \in S\}.$$

To verify that $\mathfrak{a}^e$ is an ideal, we note that if a/s, $a'/s' \in \mathfrak{a}^e$, then $a/s + a'/s' = (as' + sa')/ss' \in \mathfrak{a}^e$ and for any $b/t \in R_S$, $a/s.b/t = ab/st \in \mathfrak{a}^e$. Now take an ideal $\mathfrak{A}$ in R_S and define the corresponding *contracted ideal* in R by

$$\mathfrak{A}^c = \mathfrak{A}\lambda^{-1} = \{x \in R | x\lambda \in \mathfrak{A}\}.$$

Clearly this is an ideal, and the next result shows when these operations are inverse:

Proposition 10.3.2. *Let R be a commutative ring, S be a multiplicative subset and R_S be the corresponding ring of fractions, with the natural homomorphism $\lambda : R \to R_S$. Then*

(i) *for any ideal $\mathfrak{A}$ of R_S, $\mathfrak{A}^{ce} = \mathfrak{A}$;*
(ii) *for any ideal $\mathfrak{a}$ of R, $\mathfrak{a} \subseteq \mathfrak{a}^{ec}$, with equality if and only if no element of S is a zerodivisor on $R/\mathfrak{a}$.*

When equality holds in condition (ii), we shall say that S is $\mathfrak{a}$-*regular*. Thus S is $\mathfrak{a}$-regular iff for any $s \in S$, $x \in R$, $sx \in \mathfrak{a}$ implies $x \in \mathfrak{a}$.

Proof. (i) $\mathfrak{A}^{ce}$ is the ideal of R_S generated by $(\mathfrak{A}\lambda^{-1})\lambda$; clearly $(\mathfrak{A}\lambda^{-1})\lambda \subseteq \mathfrak{A}$, hence $\mathfrak{A}^{ce} \subseteq \mathfrak{A}$ and we must prove the reverse inclusion. Let $a/s \in \mathfrak{A}$; then $a \in \mathfrak{A}^c$ and $a/s \in \mathfrak{A}^{ce}$, hence $\mathfrak{A} = \mathfrak{A}^{ce}$ as claimed.

(ii) For any $x \in \mathfrak{a}$, $x/1 \in \mathfrak{a}^e$ and so $x \in \mathfrak{a}^{ec}$; this shows that $\mathfrak{a} \subseteq \mathfrak{a}^{ec}$. Now assume that $\mathfrak{a} = \mathfrak{a}^{ec}$ or more generally, $\mathfrak{a} = \mathfrak{A}^c$ for some ideal $\mathfrak{A}$ of R_S and let $sx \in \mathfrak{a}$, where $x \in R$, $s \in S$. Then $sx/1 \in \mathfrak{A}$, hence $x/1 = sx/s \in \mathfrak{A}$, so $x \in \mathfrak{a}$; this shows that S must be $\mathfrak{a}$-regular. Conversely, if S is $\mathfrak{a}$-regular, then $x \in \mathfrak{a}^{ec} \Leftrightarrow x/1 \in \mathfrak{a}^e \Leftrightarrow x/1 = a/s$ for

some $x \in \mathfrak{a}$, $s \in S$. This means that $(sx - a)t = 0$ for some $t \in S$, hence $xst \in \mathfrak{a}$ and by regularity we may cancel st and find that $x \in \mathfrak{a}$. Thus $\mathfrak{a}^{ec} \subseteq \mathfrak{a}$ and equality follows. ∎

This result shows that there is a bijection between the set of ideals of R_S and that of ideals $\mathfrak{a}$ of R for which S is $\mathfrak{a}$-regular. In particular, if $\mathfrak{a}$ is a prime ideal, S is $\mathfrak{a}$-regular iff $S \cap \mathfrak{a} = \varnothing$ and we obtain

Corollary 10.3.3. *With the notations of Proposition 10.3.2 the correspondence $\mathfrak{A} \leftrightarrow \mathfrak{A}^c$ is a bijection between the set of all prime ideals of R_S and the set of all prime ideals of R that are disjoint from S. In particular, for any prime ideal $\mathfrak{p}$ of R the correspondence $\mathfrak{A} \leftrightarrow \mathfrak{A}^c$ is a bijection between the set of all prime ideals of $R_\mathfrak{p}$ and that of all prime ideals of R contained in $\mathfrak{p}$.* ∎

This result shows that the localization of a local ring need not be local, e.g. invert $x + y$ in the power series ring $k[[x, y]]$, but it does hold for local Bezout domains, i.e. valuation rings.

The formation of fractions can also be extended to modules. Let R be a commutative ring and M be an R-module. Given any multiplicative subset S of R, we can define M_S as the set of equivalence classes on $M \times S$, where $(m, s) \sim (m', s')$ iff $(ms' - sm')t = 0$ for some $t \in S$. As before we can verify that this is indeed an equivalence and M_S becomes an R_S-module relative to the operations

$$m/s + m'/s' = (ms' + sm')/ss', \quad m/s.a/t = ma/st. \tag{10.3.5}$$

Of course we can equally well regard M_S as an R-module, using the equation

$$x.a = x(a\lambda), \quad x \in M_S, \quad a \in R,$$

to define the R-action in terms of the R_S-action. This is described as 'pulling the action of R_S back along λ', or more briefly defining the action of R by *pullback*. In detail, the action of R_S on M_S is defined by a homomorphism $R_S \to \mathrm{End}(M_S)$ (given by the second equation (10.3.5)); we define the action of R by pullback along λ, using the composite map $R \to R_S \to \mathrm{End}(M_S)$.

There is a canonical mapping $\mu : M \to M_S$ given by $m \mapsto m/1$, whose kernel is given by

$$\ker \mu = \{m \in M | ms = 0 \text{ for some } s \in S\}. \tag{10.3.6}$$

The universal property of M_S is described in

Proposition 10.3.4. *Let R be a commutative ring, S be a multiplicative subset and M be an R-module. Then there is an R-module M_S with a homomorphism $\mu : M \to M_S$ which is universal for homomorphisms of M into R-modules on which S acts by automorphisms.*

Proof. This is straightforward and may be left to the reader to prove. ∎

We remark that the correspondence $M \mapsto M_S$ is a functor; this is not hard to verify, and also follows from the fact (easily checked) that

$$M_S \cong M \otimes_R R_S. \tag{10.3.7}$$

In general the functor $M \mapsto M_S$ need not be faithful, since e.g. S may contain an element annihilating M. But we can obtain a faithful functor by localizing at all prime ideals, or even at all maximal ideals.

Proposition 10.3.5. *Let R be a commutative ring and X be the set of all maximal ideals of R. Then the functor*

$$M \mapsto \prod_{\mathfrak{m} \in X} M_{\mathfrak{m}} \tag{10.3.8}$$

is faithful.

Proof. Let $\alpha : M \to N$ be a non-zero mapping, say $x\alpha \neq 0$, and define $\mathrm{Ann}(x\alpha) = \{a \in R | (x\alpha)a = 0\}$. This is an ideal of R, proper because $x\alpha \neq 0$, and so is contained in some $\mathfrak{m} \in X$. Hence $x\alpha \notin \ker(N \to N_{\mathfrak{m}})$ and so the induced mapping $\alpha_{\mathfrak{m}} : M_{\mathfrak{m}} \to N_{\mathfrak{m}}$ is non-zero; this shows (10.3.8) to be faithful. ■

For any R-module M, the set of prime ideals $\mathfrak{p}$ in R such that $M_{\mathfrak{p}} \neq 0$ is called the *support* of M, written $\mathrm{Supp}(M)$. By (10.3.6) any $x \in M$ maps to 0 in $M_{\mathfrak{p}}$ iff $\mathrm{Ann}(x) \not\subset \mathfrak{p}$. Hence we see that for a finitely generated R-module M, $\mathrm{Supp}(M)$ consists of those $\mathfrak{p}$ for which $\mathfrak{p} \supseteq \mathrm{Ann}(M)$.

A further useful property of our functor is given by

Proposition 10.3.6. *Let R be a commutative ring and S be a multiplicative subset. Then the functor $M \mapsto M_S$ is exact.*

Proof. Given a sequence

$$A \xrightarrow{\alpha} B \xrightarrow{\beta} C, \tag{10.3.9}$$

exact at B, consider the corresponding sequence

$$A_S \xrightarrow{\alpha_S} B_S \xrightarrow{\beta_S} C_S. \tag{10.3.10}$$

We have to show that $\mathrm{im}\,\alpha_S = \ker \beta_S$. Clearly $\alpha_S\beta_S = 0$; conversely, let $b/s \in B_S$ and suppose that $(b/s)\beta_S = 0$, i.e. $b\beta/s = 0$. Then $b\beta.t = 0$ for some $t \in S$, hence $(bt)\beta = 0$. By exactness of (10.3.9), $bt = a\alpha$ for some $a \in A$ and so $b/s = (a/st)\alpha_S$, and this shows (10.3.10) to be exact. ■

If we recall (10.3.7), we see that tensoring with R_S is an exact functor, by Proposition 10.3.6. This is expressed by saying that R_S as left R-module is *flat*, a fact which can also be verified directly (see also FA).

We end with a useful relation between factorization in R and in R_S. The first part is well known, the converse is essentially due to Nagata [1957].

Theorem 10.3.7. *Let R be a commutative integral domain and S be a multiplicative subset of $R^\times$.*

(i) *If R is a UFD, then so is R_S.*
(ii) *If R is an atomic domain, S consists of products of primes and R_S is a UFD, then R is a UFD.*

Proof. (i) We may regard R as a subring of R_S, because the natural mapping $\lambda : R \to R_S$ is injective, and since the elements of S are units in R_S, every element of R_S is associated in R_S to an element of R. In R each element has a factorization into primes, so (i) will follow if we show that every prime p in R either becomes a unit in R_S or stays prime, depending on whether or not p divides an element of S.

If $p|s \in S$, then clearly p becomes a unit in R_S. Otherwise let $p|ab$ in R_S; here a, b may be taken in R, by passing to associates. Then $ab = pd$, where $d = c/s$ ($c \in R, s \in S$), hence $abs = pc$. By hypothesis, p does not divide s, but p is prime in R, hence $p|a$ or $p|b$, and this shows that p is prime in R_S. Thus every element of R_S has a factorization into primes, and this shows R_S to be a UFD.

(ii) We must show that every atom in R is prime. Thus let p be an atom in R; we claim that either p remains an atom in R_S or it becomes a unit. For let $p = ab$ in R_S; we shall show that a or b is a unit in R_S. Write $a = a'/s$, $b = b'/t$, where $a', b' \in R$, $s, t \in S$, so that $pst = a'b'$. By hypothesis st can be written as a product of primes: $st = q_1 \dots q_r$ and $pq_1 \dots q_r = a'b'$. Each q_i divides either a' or b' in R, and cancelling them one by one we obtain

$$p = a''b'', \tag{10.3.11}$$

where a'', b'' are the quotients of a', b' respectively, after division by the q_i. In R_S all the q_i are units, hence a'' is associated to a' and so also to a within R_S; similarly b'' is associated to b in R_S. Now p is an atom in R; by (10.3.11) either a'' or b'' is a unit, and accordingly, a or b is a unit in R_S. This shows p to be an atom or a unit in R_S.

We now treat these two cases separately.

(α) p remains an atom in R_S, hence p is prime in R_S because the latter is a UFD. If $p|ab$ in R, then $p|ab$ in R_S, hence p divides a or b in R_S, say the former: $a = pd$, $d = c/s$ ($c \in R, s \in S$), so

$$as = pc. \tag{10.3.12}$$

Now s is a product of primes in R : $s = s_1 \dots s_r$ say, and no s_i divides p in R, for if $s_i|p$, then p would be associated to s_i in R and so would become a unit in R, which is not the case. Hence by (10.3.12), $s_i|c$ in R for $i = 1, \dots, r$ and cancelling $s_1, \dots, s_r$ in turn from (10.3.12), we are left with the equation $a = pc'$, i.e. $p|a$ in R. This shows p to be prime in R.

(β) p becomes a unit in R. Then p divides some $s \in S$, say $s = pc$, and on cancelling the prime factors of s one by one we find that p is divisible by (and hence associated to) some prime factor s' of s, therefore p itself is prime in R. ∎

Our aim is to show that a polynomial ring over a UFD is again a UFD. We recall

Gauss's lemma (Lemma 7.7.1): if A is an integral domain, then every prime in A stays prime in $A[x]$.

Theorem 10.3.8. *If R is a UFD, then so is the polynomial ring $R[x]$.*

Proof. Let K be the field of fractions of R; by the Euclidean algorithm $K[x]$ is a PID and hence a UFD. Now $K[x] = R[x]_{R^\times}$ and $R^\times$ consists of products of primes, because R is a UFD. Moreover, by Gauss's lemma, these primes stay prime in $R[x]$. It only remains to show that $R[x]$ is atomic. Given $f = a_0x^n + \ldots + a_n$ ($a_0 \neq 0$), if a_0 can be written as a product of k prime factors, then f can be factorized into at most $k + n$ non-unit factors, for the leading term of each factor must either contain x or a non-unit factor of a_0. Thus $R[x]$ satisfies all the conditions of Theorem 10.3.7 relative to the multiplicative set $R^\times$, and so $R[x]$ is a UFD, as claimed. ■

Since a field is trivially a UFD, we obtain by induction,

Corollary 10.3.9. *For any field k, the polynomial ring $k[x_1, \ldots, x_n]$ in a number of indeterminates is a UFD.* ■

Exercises

1. Show that the set inverted in a homomorphism $R \to R'$ is multiplicative and saturated.
2. Show that every ring of fractions of $\mathbf{Z}/n$ has the form $\mathbf{Z}/m$. When n is given, what integers m can occur?
3. Let R be a ring and S be a multiplicative subset of R. Show that if R is Noetherian, then so is R_S.
4. Let $\mathfrak{a}$ be an ideal in R and S be a multiplicative subset. Show that in R_S, $(\sqrt{\mathfrak{a}})_S = \sqrt{\mathfrak{a}_S}$.
5. Prove the isomorphism (10.3.7) by defining a bilinear mapping from M and R_S to M_S.
6. Show that a local ring has characteristic 0 or a prime power.
7. Verify that for a finitely generated R-module M, $\mathrm{Supp}(M)$ consists of the prime ideals of R containing $\mathrm{Ann}(M)$. Give an example to show that this fails for general R-modules.
8. Show that for a short exact sequence $0 \to M' \to M \to M'' \to 0$, $\mathrm{Supp}(M) = \mathrm{Supp}(M') \cup \mathrm{Supp}(M'')$.
9. Let M be an R-module and A, B be submodules of M. Show that $A \subseteq B$ iff $A_\mathfrak{m} \subseteq B_\mathfrak{m}$ for all maximal ideals $\mathfrak{m}$ of R.
10. Let R be an integral domain with field of fractions K, so that every localization of R is embedded in K. Show that $\cap R_\mathfrak{m} = R$, where $\mathfrak{m}$ runs over all maximal ideals of R. (Hint. Apply Proposition 10.3.5 to the R-linear maps from R to K/R.)
11. In Proposition 10.3.2 show that $\mathfrak{a}^{ec} = \{a \in R | sa \in \mathfrak{a} \text{ for some } s \in S\}$.
12. Show that the following are equivalent, for any commutative ring R: (a) $R_\mathfrak{p}$ is a domain for each prime ideal $\mathfrak{p}$; (b)$R_\mathfrak{m}$ is a domain for each maximal ideal $\mathfrak{m}$; (c) if $ab = 0$ in R, then $\mathrm{Ann}(a)$ and $\mathrm{Ann}(b)$ are comaximal in R.

10.4 Noetherian Rings

We recall from Section 4.2 that a ring is *Noetherian* if every ascending chain of ideals breaks off, or equivalently, if every ideal is finitely generated. For example, any principal ideal domain is Noetherian, and so are many of the rings encountered in number theory and algebraic geometry.

In a Noetherian ring R the nilradical $\mathfrak{N}$ is nilpotent, i.e. $\mathfrak{N}^n = 0$ for some natural number n. For let $\mathfrak{N} = (a_1, \ldots, a_r)$ and suppose that $a_i^{\nu_i+1} = 0$; write $\nu = \sum \nu_i$ and consider $\mathfrak{N}^{\nu+1}$. It is spanned by the products of $\nu + 1$ factors a_i, hence some a_i must occur to a power $\geq \nu_i + 1$ and so each product is zero; thus $\mathfrak{N}^{\nu+1} = 0$. More generally, this argument shows that for any ideal $\mathfrak{a}$ of R, some power of $\sqrt{\mathfrak{a}}$ is contained in $\mathfrak{a}$.

As we saw in Section 10.2, every Noetherian domain is atomic. This also follows from the decomposition lemma (Lemma 3.2.7). The most important source of Noetherian rings is provided by the following result:

Theorem 10.4.1 (Hilbert basis theorem). *Let R be a Noetherian ring. Then the polynomial ring $R[x]$ is again Noetherian.*

Proof. (H. Sarges) We assume that $A = R[x]$ is not Noetherian and show that R cannot be Noetherian. Let $\mathfrak{a}$ be an ideal of A which is not finitely generated, and take a non-zero polynomial f_1 of least degree in $\mathfrak{a}$. By induction, if we have found $f_1, \ldots, f_k \in \mathfrak{a}$, we can take $f_{k+1} \in \mathfrak{a} \setminus \sum_1^k f_i A$ of least possible degree. Since $\mathfrak{a}$ is not finitely generated, we thus obtain an infinite sequence of polynomials $f_1, f_2, \ldots$ in $\mathfrak{a}$. Let f_i have degree n_i and leading coefficient a_i. We have $n_1 \leq n_2 \leq \ldots$ and we claim that

$$a_1 R \subset a_1 R + a_2 R \subset \ldots \tag{10.4.1}$$

is an infinite ascending chain, which will contradict the fact that R is Noetherian. If the chain (10.4.1) breaks off, we have $a_{k+1} = \sum_1^k a_i b_i$ for some $b_i \in R$, but then $f_{k+1} - \sum_1^k f_i x^{n_{k+1}-n_i} b_i$ would be an element in $\mathfrak{a} \setminus \sum_1^k f_i A$ of degree less than n_{k+1}, which contradicts the definition of f_{k+1}. This establishes the result. ∎

By induction on n we obtain

Corollary 10.4.2. *If R is Noetherian, then so is $R[x_1, \ldots, x_n]$. In particular, $k[x_1, \ldots, x_n]$ is Noetherian for any field k and any $n \geq 1$.* ∎

Of course this result does not extend to infinitely many indeterminates; thus for any non-zero ring R we have

$$R \subset R[x_1] \subset R[x_1, x_2] \subset \ldots,$$

and we can form the union $R[x_1, x_2, \ldots]$ as a polynomial ring in countably many indeterminates x_i ($i \in \mathbf{N}$). This is an integral domain if R is, but it is not Noetherian, for the ideal $(x_1, x_2, \ldots)$ cannot be finitely generated.

Proposition 10.4.3. *Let K be a commutative Noetherian ring. Then any commutative ring R which is finitely generated as a K-algebra is Noetherian.*

Proof. Let R be generated by $c_1, \ldots, c_n$ over K. Then R is a homomorphic image of the polynomial ring $K[x_1, \ldots, x_n]$ obtained by mapping $x_i \mapsto c_i$, say $R \cong K[x_1, \ldots, x_n]/\mathfrak{a}$. Thus the ideals of R correspond to the ideals of $K[x_1, \ldots, x_n]$ which contain $\mathfrak{a}$, and hence satisfy the maximum condition, because the polynomial ring does. ■

Exercises

1. Show that every Noetherian domain is atomic by considering the monoid of classes of associated elements and applying Lemma 3.2.7.
2. Show that if R is a Noetherian ring, then so is $R[[x]]$, the ring of formal power series in x.
3. Show that the polynomial ring $k[x_1, x_2, \ldots]$ in countably many indeterminates over a field k is a UFD.
4. If k is an infinite field, show that the subring $k + xR$ of $R = k[x, y]$ is not Noetherian.
5. Let R be a ring and $\mathfrak{a}_1, \ldots, \mathfrak{a}_n$ be ideals of R such that $\cap \mathfrak{a}_i = 0$. Show that if $R/\mathfrak{a}_i$ is Noetherian for $i = 1, \ldots, n$, then R is Noetherian.
6. Let $\mathfrak{a}$ be a finitely generated ideal in a ring R. Show that if $\mathfrak{a}^2 = \mathfrak{a}$, then $\mathfrak{a}$ is generated by a single idempotent element.
7. Show that the relation $\mathfrak{a} = \mathrm{Ann}(\mathfrak{b})$ between ideals $\mathfrak{a}$ and $\mathfrak{b}$ in a ring R defines a Galois connexion which is a lattice anti-isomorphism between annihilator ideals. Show that if every ideal of R is the annihilator of a finite set, then R is Artinian.

10.5 Dedekind Domains

The phenomenon of non-unique factorization was first encountered by Kummer in his work on Fermat's last theorem, a famous conjecture which states that the equation

$$x^n + y^n = z^n, \quad \text{where } n > 2, \tag{10.5.1}$$

has no solution in non-zero integers x, y, z. It was asserted by Pierre de Fermat without proof and after intensive efforts by many mathematicians was proved in 1994 by Andrew Wiles, using methods from algebraic geometry. Previously Gerd Faltings had shown in 1983 that for any $n \geq 3$ there are at most a finite number of solutions with x, y, z coprime.

An early method of attack was to factorize the left-hand side of (10.5.1) in $\mathbf{Z}[\zeta_n]$, where ζ_n is a primitive n-th root of 1. If $\mathbf{Z}[\zeta_n]$ is a UFD, this leads to a proof of Fermat's last theorem for this value of n. But in general $\mathbf{Z}[\zeta_n]$ is not a UFD, and Kummer's investigations of such rings led him to the creation of his theory of 'ideal numbers' and hence led Richard Dedekind to his ideal theory. One of

Kummer's discoveries was that every non-zero ideal in $\mathbf{Z}[\zeta_n]$ can be written uniquely as a finite product of prime ideals. Later, Dedekind showed that the same is true in the ring of integers of any algebraic number field, i.e. the integral closure of $\mathbf{Z}$ in a finite extension field of $\mathbf{Q}$. This property is also of interest in algebraic geometry, where it describes the rings of non-singular curves.

Before examining these rings abstractly let us look at some concrete examples. In Section 10.2 we saw that every Euclidean domain is a UFD; this enables us to show that the ring $\mathbf{Z}[i]$ of Gaussian integers is a UFD. We shall do this by verifying that the ring is Euclidean with respect to the usual norm

$$N(x+iy) = x^2 + y^2.$$

We have to show that for any $a, b \in \mathbf{Z}[i], b \neq 0$, there exists c such that

$$N(a - bc) < N(b). \tag{10.5.2}$$

We note that $\mathbf{Z}[i]$ is the precise ring of integers in $\mathbf{Q}(i)$. On dividing by b and recalling that N is multiplicative, we obtain $N(a/b - c) < 1$. In other words, we have to show that for each $\gamma \in \mathbf{Q}(i)$ there exists $c \in \mathbf{Z}[i]$ such that

$$N(\gamma - c) < 1. \tag{10.5.3}$$

Put $\gamma = x + iy$, where x, y are rational numbers. Every rational number is within 1/2 of an integer, so we can find rational integers u, v such that $|x - u| \leq 1/2$, $|y - v| \leq 1/2$, and on writing $c = u + iv$, we have

$$N(\gamma - c) = (x-u)^2 + (y-v)^2 \leq 1/4 + 1/4 = 1/2 < 1.$$

Thus (10.5.3) is established and it shows $\mathbf{Z}[i]$ to be Euclidean. The same method can be used on the ring of integers in $\mathbf{Q}(\sqrt{-d})$, for $d = 2, 3, 7, 11$ (see Exercise 6).

Next consider $\mathbf{Q}(\sqrt{-5})$; its ring of integers is $\mathbf{Z}[\sqrt{-5}]$ and we have

$$6 = 2.3 = (1 + \sqrt{-5})(1 - \sqrt{-5}). \tag{10.5.4}$$

By considering norms we see that $2, 3, 1 \pm \sqrt{-5}$ are atoms. For example, $N(2) = 4$, so if 2 is composite, then a proper factor must have norm 2, but the equation $u^2 + 5v^2 = 2$ has no solution in integers. Similarly for $3, 1 \pm \sqrt{-5}$; hence (10.5.4) represents two distinct (i.e. non-isomorphic) factorizations of 6, and this shows that $\mathbf{Z}[\sqrt{-5}]$ is not a UFD. However, the ideal (6) can be expressed uniquely as a product of maximal ideals:

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

We begin with the most general case. Let $\mathfrak{o}$ be an integral domain and K be its field of fractions. By a *fractional ideal* of $\mathfrak{o}$ we understand an $\mathfrak{o}$-submodule $\mathfrak{A}$ of K such that

$$z\mathfrak{o} \subseteq \mathfrak{A} \subseteq u\mathfrak{o}, \quad \text{for some } z, u \in K^\times. \tag{10.5.5}$$

We note that (10.5.5) certainly holds when $0 \neq \mathfrak{A} \subseteq \mathfrak{o}$. Thus an ordinary ideal $\mathfrak{a}$ of

$\mathfrak{o}$ is a fractional ideal iff it is non-zero; the non-zero ideals of $\mathfrak{o}$ will be called the *integral ideals*. The usual multiplication can be defined for fractional ideals:

$$\mathfrak{AB} = \left\{\sum x_i y_i | x_i \in \mathfrak{A}, y_i \in \mathfrak{B}\right\}, \tag{10.5.6}$$

and it is clear from (10.5.5) that this product is again a fractional ideal. Hence in any integral domain $\mathfrak{o}$ the fractional ideals form a monoid (with $\mathfrak{o}$ itself as neutral), which will be denoted by $I = I(\mathfrak{o})$.

For each fractional ideal $\mathfrak{A}$ we can define an 'inverse'

$$(\mathfrak{o} : \mathfrak{A}) = \{x \in K | x\mathfrak{A} \subseteq \mathfrak{o}\}.$$

If $z\mathfrak{o} \subseteq \mathfrak{A} \subseteq u\mathfrak{o}$, then $u^{-1}\mathfrak{o} \subseteq (\mathfrak{o} : \mathfrak{A}) \subseteq z^{-1}\mathfrak{o}$, and if $c \in \mathfrak{o}$, then $x\mathfrak{A} \subseteq \mathfrak{o}$ implies $cx\mathfrak{A} \subseteq x\mathfrak{A} \subseteq \mathfrak{o}$. This shows that $(\mathfrak{o} : \mathfrak{A})$ is again a fractional ideal. Any fractional ideal $\mathfrak{A}$ satisfies

$$\mathfrak{A}(\mathfrak{o} : \mathfrak{A}) \subseteq \mathfrak{o},$$

but here equality need not hold. If it does hold, then $\mathfrak{A}$ is said to be *invertible* and we also write $\mathfrak{A}^{-1}$ in place of $(\mathfrak{o} : \mathfrak{A})$. For example, any non-zero principal ideal $a\mathfrak{o}$ is invertible, with inverse $a^{-1}\mathfrak{o}$. We remark that if $\mathfrak{AB} = \mathfrak{o}$ for some fractional ideal $\mathfrak{B}$, then $\mathfrak{B} \subseteq (\mathfrak{o} : \mathfrak{A})$, hence $\mathfrak{o} = \mathfrak{AB} \subseteq \mathfrak{A}(\mathfrak{o} : \mathfrak{A})$, and it follows that $\mathfrak{A}(\mathfrak{o} : \mathfrak{A}) = \mathfrak{o}$, so that $\mathfrak{A}$ is then invertible. Thus the invertible fractional ideals are just the units of the monoid $I(\mathfrak{o})$.

We note that in any integral domain $\mathfrak{o}$ an ideal is isomorphic to $\mathfrak{o}$ iff it is non-zero principal; further, any fractional ideal is isomorphic to an integral ideal, for if $\mathfrak{A} \subseteq u\mathfrak{o}$, then $u^{-1}\mathfrak{A} \subseteq \mathfrak{o}$ and the ideals $\mathfrak{A}, u^{-1}\mathfrak{A}$ are isomorphic via the mapping $x \mapsto u^{-1}x$.

We first give some characterizations of Dedekind domains, first defined in Section 4.7, including one in homological terms.

Proposition 10.5.1. *For any integral domain $\mathfrak{o}$ the following conditions are equivalent:*

(a) *the set $I(\mathfrak{o})$ of fractional ideals is a group under the multiplication* (10.5.6),
(b) *every fractional ideal of $\mathfrak{o}$ is invertible,*
(c) *every integral ideal of $\mathfrak{o}$ is invertible,*
(d) *every ideal of $\mathfrak{o}$ is projective.*

Moreover, any ring satisfying (a)–(d) *is Noetherian.*

Proof. (a) $\Leftrightarrow$ (b) $\Rightarrow$ (c) are clear.

(c) $\Rightarrow$ (d). Let $\mathfrak{a}$ be an ideal in $\mathfrak{o}$; we may assume that $\mathfrak{a} \neq 0$. By hypothesis $\mathfrak{a}$ is invertible, say $\mathfrak{ab} = \mathfrak{o}$. Hence there exist $a_i \in \mathfrak{a}$, $b_i \in \mathfrak{b}$ $(i = 1, \ldots, n)$ such that $\sum a_i b_i = 1$ and $\mathfrak{a}b_i \subseteq \mathfrak{o}$. It follows that multiplication by b_i defines a homomorphism $\mathfrak{a} \to \mathfrak{o}$. Now for any $x \in \mathfrak{a}$, $b_i x \in \mathfrak{o}$ and

$$x = \sum a_i(b_i x); \tag{10.5.8}$$

therefore by the dual basis lemma (Proposition 4.7.5), $\mathfrak{a}$ is finitely generated projective.

(d) $\Rightarrow$ (b). Let (a_i), (f_i) $(i \in I)$ be dual bases for the non-zero projective ideal $\mathfrak{a}$; thus

$$x = \sum a_i(f_i, x). \tag{10.5.9}$$

Given $x, y \in \mathfrak{a}$, we have, for any $i \in I$, $(f_i, x)y = (f_i, xy) = (f_i, y)x$. If $x \neq 0$, we can write $b_i = (f_i, x)x^{-1}$; the b_i lie in K, almost all are 0 and $b_i y = (f_i, y)$ for all $y \in \mathfrak{a}$; hence $b_i\mathfrak{a} \subseteq \mathfrak{o}$ and (10.5.9) reduces to (10.5.8). Dividing by x, we have $\sum a_i b_i = 1$, so on putting $\mathfrak{b} = \sum b_i\mathfrak{o}$, we have $\mathfrak{ab} = \mathfrak{o}$, and this shows $\mathfrak{a}$ to be invertible.

This proves (a)–(d) to be equivalent; moreover, we saw that every invertible ideal is finitely generated, so $\mathfrak{o}$ must be Noetherian. ■

We remark that the proof shows an ideal in an integral domain to be invertible iff it is non-zero projective.

An integral domain satisfying any of these equivalent conditions is called a *Dedekind domain*. It is clear from (c) or (d) that any principal ideal domain is a Dedekind domain. A ring in which every ideal is projective is also called *hereditary*; thus Dedekind domains may be described as hereditary commutative integral domains.

If S is a multiplicative set in a Dedekind domain $\mathfrak{o}$, then the ideals of the ring of fractions $\mathfrak{o}_S$ correspond to the contracted ideals of $\mathfrak{o}$, by Proposition 10.3.2. Using Proposition 10.5.1(c), we obtain

Corollary 10.5.2. *Let $\mathfrak{o}$ be a Dedekind domain and S be a multiplicative subset. Then $\mathfrak{o}_S$ is again a Dedekind domain.* ■

Dedekind domains have been characterized by E. Noether as Noetherian integrally closed domains in which all non-zero prime ideals are maximal. To prove this result we need a variant of the decomposition lemma (Lemma 3.2.7), which has a similar proof.

Lemma 10.5.3. *Any ideal $\mathfrak{a}$ of a non-trivial Noetherian ring R contains a finite product of non-zero prime ideals:*

$$\mathfrak{a} \supseteq \mathfrak{p}_1 \ldots \mathfrak{p}_r, \quad \text{where } \mathfrak{p}_i \text{ is a prime ideal } \neq 0,$$

unless $\mathfrak{a} = 0$ and R is an integral domain, or $\mathfrak{a} = R$ and R is a field.

Proof. Assume the contrary and let $\mathfrak{a}$ be a 'maximal offender', i.e. $\mathfrak{a}$ is maximal among the ideals that do not contain a finite product of non-zero prime ideals. If $\mathfrak{a} = R$, there is nothing to prove, since every ring not zero or a field has non-zero prime ideals. If $\mathfrak{a} \neq R$, then either $\mathfrak{a}$ is prime, but then $\mathfrak{a} = 0$ and this leads to the excluded case, or there exist ideals $\mathfrak{b}_1, \mathfrak{b}_2$ such that $\mathfrak{b}_1\mathfrak{b}_2 \subseteq \mathfrak{a} \subset \mathfrak{b}_1, \mathfrak{b}_2$. By maximality, $\mathfrak{b}_1 \supseteq \mathfrak{p}_1 \ldots \mathfrak{p}_r$, $\mathfrak{b}_2 \supseteq \mathfrak{p}_{r+1} \ldots \mathfrak{p}_s$, where the $\mathfrak{p}_i$ are non-zero prime ideals. Hence $\mathfrak{a} \supseteq \mathfrak{b}_1\mathfrak{b}_2 \supseteq \mathfrak{p}_1 \ldots \mathfrak{p}_s$; this is a contradiction, and the conclusion follows. ■

We can now state E. Noether's result; it is convenient to include several characterizations.

Theorem 10.5.4. *Let $\mathfrak{o}$ be an integral domain. Then the following conditions are equivalent:*

(a) $\mathfrak{o}$ *is a Dedekind domain;*
(b) $\mathfrak{o}$ *is Noetherian and* $\mathfrak{o}_\mathfrak{m}$ *is a principal valuation ring (or a field) for all maximal ideals* $\mathfrak{m}$ *of* $\mathfrak{o}$*;*
(c) $\mathfrak{o}$ *is Noetherian, integrally closed and every non-zero prime ideal is maximal;*
(d) *every non-zero prime ideal of* $\mathfrak{o}$ *is invertible.*

Proof. (a) $\Rightarrow$ (b). Let $\mathfrak{o}$ be a Dedekind domain; by Proposition 10.5.1 it is Noetherian and by Corollary 10.5.2, $\mathfrak{o}_\mathfrak{m}$ is a Dedekind domain, for any maximal ideal $\mathfrak{m}$ of $\mathfrak{o}$. By Proposition 10.5.1(c), if $a, b \in \mathfrak{o}_\mathfrak{m}$ are non-zero and $\mathfrak{A} = a\mathfrak{o}_\mathfrak{m} + b\mathfrak{o}_\mathfrak{m}$, then $a(\mathfrak{o}_\mathfrak{m} : \mathfrak{A}) + b(\mathfrak{o}_\mathfrak{m} : \mathfrak{A}) = \mathfrak{o}_\mathfrak{m}$. But this ring is local, so one of the two ideals must be the whole ring. If $a(\mathfrak{o}_\mathfrak{m} : \mathfrak{A}) = \mathfrak{o}_\mathfrak{m}$, then $a^{-1} \in (\mathfrak{o}_\mathfrak{m} : \mathfrak{A})$ and so $a^{-1}b \in \mathfrak{o}_\mathfrak{m}$. Thus either $a^{-1}b$ or $b^{-1}a$ is in $\mathfrak{o}_\mathfrak{m}$, so $\mathfrak{o}_\mathfrak{m}$ is a valuation ring; being Noetherian, it must be either a field or a principal valuation ring.

(b) $\Rightarrow$ (c). Assume (b), thus $\mathfrak{o}$ is Noetherian and for every maximal ideal $\mathfrak{m}$, $\mathfrak{o}_\mathfrak{m}$ is a valuation ring. If $x \in \mathfrak{o}_\mathfrak{m}$, say $x = a/s$, where $a, s \in \mathfrak{o}, s \notin \mathfrak{m}$, then $s \in (\mathfrak{o} : x\mathfrak{o})$ and so $\mathfrak{o} \cap (\mathfrak{o} : x\mathfrak{o})$ meets $\mathfrak{o}\backslash\mathfrak{m}$. It follows that if $x \in \cap\mathfrak{o}_\mathfrak{m}$, then $\mathfrak{o} \cap (\mathfrak{o} : x\mathfrak{o})$ meets the complement of each maximal ideal and so must be the whole of $\mathfrak{o}$, hence $(\mathfrak{o} : x\mathfrak{o}) \supseteq \mathfrak{o}$, and so $x \in \mathfrak{o}$. This shows that $\mathfrak{o} = \cap\mathfrak{o}_\mathfrak{m}$, and so $\mathfrak{o}$ is integrally closed, by Theorem 9.4.4. Finally, if $\mathfrak{p}$ is a non-zero prime ideal of $\mathfrak{o}$, then $\mathfrak{p}$ is contained in a maximal ideal $\mathfrak{m}$ and so $\mathfrak{p}\mathfrak{o}_\mathfrak{m}$ is a non-zero prime ideal in $\mathfrak{o}_\mathfrak{m}$, which can only be $\mathfrak{m}\mathfrak{o}_\mathfrak{m}$, by (b); therefore $\mathfrak{p} = \mathfrak{m}$ by Corollary 10.3.3 and (c) follows.

(c) $\Rightarrow$ (d). Let $\mathfrak{p}$ be a non-zero prime ideal of $\mathfrak{o}$ and let $0 \neq a \in \mathfrak{p}$. By Lemma 10.5.3 we have $(a) \supseteq \mathfrak{p}_1 \ldots \mathfrak{p}_r$, where the $\mathfrak{p}_i$ are non-zero prime ideals. Let us choose r minimal; since $\mathfrak{p} \supseteq (a) \supseteq \mathfrak{p}_1 \ldots \mathfrak{p}_r$ and $\mathfrak{p}$ is prime, we have $\mathfrak{p} \supseteq \mathfrak{p}_i$ for some i, say $i = 1$. By (c), $\mathfrak{p}_1$ is maximal, so $\mathfrak{p} = \mathfrak{p}_1$ and by the minimality of r we can find $b \in \mathfrak{p}_2 \ldots \mathfrak{p}_r$ such that $b \notin (a)$. However, $b\mathfrak{p} \subseteq (a)$, i.e. $a^{-1}b\mathfrak{p} \subseteq \mathfrak{o}$, so $a^{-1}b \in (\mathfrak{o} : \mathfrak{p})$. Now $\mathfrak{p} \subseteq (\mathfrak{o} : \mathfrak{p})\mathfrak{p} \subseteq \mathfrak{o}$, hence $(\mathfrak{o} : \mathfrak{p})\mathfrak{p}$ must be $\mathfrak{p}$ or $\mathfrak{o}$. If $(\mathfrak{o} : \mathfrak{p})\mathfrak{p} = \mathfrak{p}$, then $a^{-1}b\mathfrak{p} = \mathfrak{p}$, and by Proposition 9.4.1(c), $a^{-1}b$ is then integral over $\mathfrak{o}$. Since $\mathfrak{o}$ is integrally closed, this means that $a^{-1}b \in \mathfrak{o}$, but this contradicts the fact that $b \notin (a)$. Therefore $(\mathfrak{o} : \mathfrak{p})\mathfrak{p} = \mathfrak{o}$ and $\mathfrak{p}$ is invertible, as claimed in (d).

(d) $\Rightarrow$ (a). Suppose that (d) holds and (a) fails. Then the set $\mathscr{P}$ of non-invertible integral ideals of $\mathfrak{o}$ is non-empty. Under the ordering by inclusion $\mathscr{P}$ is inductive, for if $\{\mathfrak{a}_\lambda\}$ is a chain in $\mathscr{P}$ and $\mathfrak{a} = \cup\mathfrak{a}_\lambda$ is invertible, then it is finitely generated and so $\mathfrak{a} = \mathfrak{a}_\lambda$ for some λ, a contradiction. Thus $\mathscr{P}$ is inductive and by Zorn's lemma it has a maximal member $\mathfrak{m}$, say. By (d), $\mathfrak{m}$ is not prime or equal to $\mathfrak{o}$, so there exist $a, b \notin \mathfrak{m}$ such that $ab \in \mathfrak{m}$, thus $b \in \mathfrak{m} : (a) \supset \mathfrak{m}$. By the maximality of $\mathfrak{m}$, $\mathfrak{m} + (a)$ and $\mathfrak{m} : (a)$ are both invertible, hence so is $a(\mathfrak{m} : (a)) = \mathfrak{m} \cap (a)$. Now consider the short exact sequence

$$0 \to \mathfrak{m} \cap (a) \xrightarrow{\lambda} \mathfrak{m} \oplus (a) \xrightarrow{\mu} \mathfrak{m} + (a) \to 0, \tag{10.5.10}$$

where $(x, ya)\mu = x - ya$. The third term $\mathfrak{m} + (a)$ is invertible and hence projective; therefore (10.5.10) splits and we have $\mathfrak{m} \oplus (a) \cong [\mathfrak{m} \cap (a)] \oplus [\mathfrak{m} + (a)]$. Here the

right-hand side is projective, hence $\mathfrak{m}$ is projective, which contradicts the fact that $\mathfrak{m} \in \mathscr{P}$. It follows that $\mathscr{P}$ is empty, as claimed. ∎

The advantage of Dedekind domains over UFDs is that the former survive integral extensions, whereas the latter may not, as we saw at the beginning of this section. We now apply Noether's characterization to prove that a finite separable integrally closed integral extension of a Dedekind domain is again Dedekind; this still holds in the inseparable case, but with a different proof (see Exercise 10).

Theorem 10.5.5. *Let $\mathfrak{o}$ be a Dedekind domain with field of fractions K. Given a finite separable extension L of K, the integral closure $\mathfrak{O}$ of $\mathfrak{o}$ in L is again a Dedekind domain.*

Proof. The ring $\mathfrak{O}$ is integrally closed by construction. To show that non-zero prime ideals are maximal, let $\mathfrak{P} \subseteq \mathfrak{P}'$ be prime ideals in $\mathfrak{O}$ such that $\mathfrak{P} \cap \mathfrak{o} = \mathfrak{P}' \cap \mathfrak{o}$. Given $x \in \mathfrak{P}'$, we take a monic equation for x over $\mathfrak{o}$ and consider it mod $\mathfrak{P}$:

$$x^n + a_1 x^{n-1} + \ldots + a_n \equiv 0 \pmod{\mathfrak{P}}, \quad \text{where } a_i \in \mathfrak{o}.$$

Choose such a congruence for which n has its least value. We have $a_n \in \mathfrak{P}' \cap \mathfrak{o} = \mathfrak{P} \cap \mathfrak{o}$, hence $x(x^{n-1} + a_1 x^{n-2} + \ldots + a_{n-1}) \in \mathfrak{P}$, and since $\mathfrak{P}$ is prime, we have $x \in \mathfrak{P}$, by the minimality of n. This shows that $\mathfrak{P}' = \mathfrak{P}$. Since every non-zero prime ideal of $\mathfrak{o}$ is maximal, the same holds for $\mathfrak{O}$.

It remains to show that $\mathfrak{O}$ is Noetherian. Let $u_1, \ldots, u_n$ be a K-basis for L; on multiplying by suitable elements of K we may assume that $u_i \in \mathfrak{O}$. Since the trace $T(xy)$ is a non-singular pairing on L (see Proposition 7.9.5), we can find a dual basis $v_1, \ldots, v_n$ for the u_i relative to T; thus we have $T(u_i v_j) = \delta_{ij}$. Now if $x \in \mathfrak{O}$, we have $x = \sum \alpha_i v_i$ for some $\alpha_i \in K$, and in fact $\alpha_i = T(u_i x) \in \mathfrak{o}$, hence $\mathfrak{O} \subseteq \sum \mathfrak{o} v_i$. Therefore $\mathfrak{O}$ is a submodule of a finitely generated $\mathfrak{o}$-module, and since $\mathfrak{o}$ is Noetherian, it follows that every ideal in $\mathfrak{O}$ is finitely generated, hence $\mathfrak{O}$ is Noetherian. ∎

As a consequence, the ring $\mathfrak{O}$ of integers of an algebraic number field L is a Dedekind domain, since $\mathbf{Z}$ is clearly a Dedekind domain with field of fractions $\mathbf{Q}$. Similarly the ring of functions integral over $k[x]$ in a finite separable extension of $k(x)$ is a Dedekind domain.

We now come to the unique factorization property of Dedekind domains mentioned at the beginning of this section; in fact it yields another characterization.

Theorem 10.5.6. *For any integral domain $\mathfrak{o}$, the following properties are equivalent:*

(a) *$\mathfrak{o}$ is a Dedekind domain,*
(b) *every integral ideal of $\mathfrak{o}$ can be expressed uniquely as a finite product of maximal ideals,*
(c) *every integral ideal of $\mathfrak{o}$ can be expressed as a finite product of prime ideals.*

Proof. (a) ⇒ (b). Let $\mathfrak{o}$ be a Dedekind domain and I_0 be the set of integral ideals of $\mathfrak{o}$ which can be expressed as a finite product of maximal ideals. Clearly I_0 contains $\mathfrak{o}$, as

the empty product. If some integral ideal does not belong to I_0, then since $\mathfrak{o}$ is Noetherian, we can find a maximal such ideal $\mathfrak{a}$, say. Since $\mathfrak{a} \neq \mathfrak{o}$, $\mathfrak{a} \subseteq \mathfrak{m} \subset \mathfrak{o}$ for some maximal ideal $\mathfrak{m}$. By Proposition 10.5.1(c), $\mathfrak{a} = \mathfrak{m}\mathfrak{b}$ for some integral ideal $\mathfrak{b}$, and since $I(\mathfrak{o})$ is a group and $\mathfrak{m} \neq \mathfrak{o}$, we have $\mathfrak{a} \neq \mathfrak{b}$. Thus $\mathfrak{a} \subset \mathfrak{b}$, but $\mathfrak{b} \in I_0$ by the maximality of $\mathfrak{a}$, hence $\mathfrak{a} = \mathfrak{m}\mathfrak{b} \in I_0$, a contradiction. Thus every integral ideal can be expressed as a finite product of maximal ideals. Such an expression is unique, for if $\mathfrak{p}_1 \dots \mathfrak{p}_r = \mathfrak{q}_1 \dots \mathfrak{q}_s$, then since $\mathfrak{p}_1 \supseteq \mathfrak{q}_1 \dots \mathfrak{q}_s$ and $\mathfrak{p}_1$ is prime, we have $\mathfrak{p}_1 \supseteq \mathfrak{q}_i$ for some i. But $\mathfrak{q}_i$ is maximal, so $\mathfrak{p}_1 = \mathfrak{q}_i$ and we may cancel these terms, because $I(\mathfrak{o})$ is a group. By induction it follows that $r = s$ and the expression is unique up to the order of the factors and this proves (b).

(b) $\Rightarrow$ (c) is clear. To prove (c) $\Rightarrow$ (a), we first show that any invertible prime ideal $\mathfrak{p}$ of $\mathfrak{o}$ is maximal. If $\mathfrak{p}$ is not maximal, then there exists $a \in \mathfrak{o}$ such that $\mathfrak{p} \subset \mathfrak{p} + (a) \subset \mathfrak{o}$, and by (c) we can write $\mathfrak{p} + (a) = \mathfrak{p}_1 \dots \mathfrak{p}_r$, $\mathfrak{p} + (a^2) = \mathfrak{q}_1 \dots \mathfrak{q}_s$, where $\mathfrak{p}_i, \mathfrak{q}_j$ are prime ideals containing $\mathfrak{p}$. In $\bar{\mathfrak{o}} = \mathfrak{o}/\mathfrak{p}$ we then have $(\bar{a}) = \bar{\mathfrak{p}}_1 \dots \bar{\mathfrak{p}}_r$, $(\bar{a}^2) = \bar{\mathfrak{q}}_1 \dots \bar{\mathfrak{q}}_s = \bar{\mathfrak{p}}_1^2 \dots \bar{\mathfrak{p}}_r^2$. Let $\bar{\mathfrak{n}}$ be minimal among the $\bar{\mathfrak{p}}_i, \bar{\mathfrak{q}}_j$, say $\bar{\mathfrak{n}} = \bar{\mathfrak{p}}_1$. Since $\bar{\mathfrak{p}}_1 \supseteq \bar{\mathfrak{q}}_j$ for some j, we find that $\bar{\mathfrak{p}}_1 = \bar{\mathfrak{q}}_j$ and so $\mathfrak{p}_1 = \mathfrak{q}_j$. Further, $\bar{\mathfrak{p}}_1$ is invertible as factor of the invertible ideal $(\bar{a})$ and so may be cancelled. By induction we find that $s = 2r$ and each $\bar{\mathfrak{p}}_i$ is equal to two of the $\bar{\mathfrak{q}}_j$, thus we have $(\mathfrak{p}_1 \dots \mathfrak{p}_r)^2 = \mathfrak{q}_1 \dots \mathfrak{q}_s$. It follows that

$$\mathfrak{p} \subseteq \mathfrak{p} + (a)^2 = [\mathfrak{p} + (a)]^2 \subseteq \mathfrak{p}^2 + (a);$$

because $a \notin \mathfrak{p}$, we have $\mathfrak{p} \cap (a) = \mathfrak{p}.(a)$, and so

$$\mathfrak{p} = \mathfrak{p} \cap (\mathfrak{p}^2 + (a)) = \mathfrak{p}^2 + (\mathfrak{p} \cap (a)) = \mathfrak{p}^2 + \mathfrak{p}.(a).$$

Since $\mathfrak{p}$ is invertible, we can cancel it and find that $\mathfrak{p} + (a) = \mathfrak{o}$, which is a contradiction. This shows $\mathfrak{p}$ to be maximal.

Now let $\mathfrak{m}$ be any non-zero prime ideal of $\mathfrak{o}$. If $0 \neq a \in \mathfrak{m}$, then $\mathfrak{m} \supseteq (a) = \mathfrak{p}_1 \dots \mathfrak{p}_r$, hence $\mathfrak{m} \supseteq \mathfrak{p}_i$ for some i. But $\mathfrak{p}_i$ as factor of (a) is invertible, therefore it is maximal and so $\mathfrak{m} = \mathfrak{p}_i$, which shows $\mathfrak{m}$ to be invertible. Thus every non-zero prime ideal of $\mathfrak{o}$ is invertible, hence $\mathfrak{o}$ is Dedekind, by Theorem 10.5.4. ■

In spite of this unique factorization property Dedekind domains by no means include all UFDs, as will be clear from Corollary 10.5.8 below.

Corollary 10.5.7. *A ring is a principal ideal domain if and only if it is a unique factorization domain in which all prime ideals are maximal.*

Proof. The condition is clearly necessary. When it holds for a ring R, we see from the remarks following Theorem 10.2.10 that every minimal non-zero prime ideal in a UFD is principal. Hence every non-zero prime ideal of R is principal, so invertible, and by Theorem 10.5.4, R is a Dedekind domain. By Theorem 10.5.6 every non-zero ideal can be written as a product of (principal) prime ideals, and so is principal; thus R is a principal ideal domain. ■

This result clearly has the following further consequence:

Corollary 10.5.8. *A Dedekind domain is a unique factorization domain if and only if it is a principal ideal domain.* ∎

We can also use Theorem 10.5.6 to describe the group of fractional ideals more closely.

Proposition 10.5.9. *In a Dedekind domain $\mathfrak{o}$ the set $I(\mathfrak{o})$ of fractional ideals is a free abelian group on the non-zero prime ideals.*

Proof. Let $\{\mathfrak{m}\}$ be the set of all non-zero prime ideals of $\mathfrak{o}$ and take a fractional ideal $\mathfrak{A} \subseteq u\mathfrak{o}$. For some $v \neq 0$ we have $vu \in \mathfrak{o}$ and $v \in \mathfrak{o}$, hence $v\mathfrak{A} \subseteq \mathfrak{o}$ and we can write $v\mathfrak{A} = \prod \mathfrak{m}^{e_\mathfrak{m}}$, $(v) = \prod \mathfrak{m}^{f_\mathfrak{m}}$, where $e_\mathfrak{m}, f_\mathfrak{m} \geq 0$ and so $\mathfrak{A} = \prod \mathfrak{m}^{e_\mathfrak{m} - f_\mathfrak{m}}$, hence the $\mathfrak{m}$'s generate $I(\mathfrak{o})$. If there is a relation between them, we may write this as $\prod \mathfrak{m}^{a_\mathfrak{m}} = \prod \mathfrak{m}^{b_\mathfrak{m}}$, where $a_\mathfrak{m}, b_\mathfrak{m} \geq 0$; by Theorem 10.5.6, $a_\mathfrak{m} = b_\mathfrak{m}$, hence the $\mathfrak{m}$'s are indeed free (abelian) generators of $I(\mathfrak{o})$. ∎

We can now give a precise description of the valuations on a Dedekind domain. Let $\mathfrak{o}$ be a Dedekind domain, not a field, and K be its field of fractions. For each maximal ideal $\mathfrak{p}$ of $\mathfrak{o}$ there is a mapping $v_\mathfrak{p} : K^\times \to \mathbf{Z}$ given by $v_\mathfrak{p}(u) = e_\mathfrak{p}$ if $(u) = \prod \mathfrak{m}^{e_\mathfrak{m}}$. It is easily verified that $v_\mathfrak{p}$ is a valuation on $\mathfrak{o}$, called the $\mathfrak{p}$-*adic* valuation. Let us show that every non-trivial valuation v whose valuation ring contains $\mathfrak{o}$ is of this form. Let V be the valuation ring of v and $\mathfrak{m}$ be its maximal ideal; then $\mathfrak{p} = \mathfrak{m} \cap \mathfrak{o}$ is a prime ideal of $\mathfrak{o}$, and $\mathfrak{p} \neq 0$, because v is non-trivial. By Theorem 10.5.4(b), $\mathfrak{o}_\mathfrak{p}$ is a principal valuation ring and $\mathfrak{o}_\mathfrak{p} \subseteq V \subset K$. By Theorem 9.4.6, $\mathfrak{o}_\mathfrak{p}$ is a maximal valuation ring of K, so $\mathfrak{o}_\mathfrak{p} = V$ and it follows that v is equivalent to $v_\mathfrak{p}$. With this information we can derive an approximation theorem which strengthens Theorem 9.2.5.

Theorem 10.5.10 (Strong approximation theorem). *Let $\mathfrak{o}$ be a Dedekind domain with field of fractions K. Given any non-trivial inequivalent valuations $v_1, \ldots, v_r$ on $\mathfrak{o}$, $n_1, \ldots, n_r \in \mathbf{Z}$ and $x_1, \ldots, x_r \in K$, there exists $x \in K$ such that $v_i(x - x_i) \geq n_i$ for $i = 1, \ldots, r$ and $v(x) \geq 0$ for any valuation v on $\mathfrak{o}$ not equivalent to any v_i.*

Proof. Without loss of generality we may assume that $n_i \geq 0$ $(i = 1, \ldots, r)$. By the remark preceding the theorem we know that $v_i = v_{\mathfrak{p}_i}$ for non-zero prime ideals $\mathfrak{p}_1, \ldots, \mathfrak{p}_r$ of $\mathfrak{o}$. Let us write $x_i = y_i s^{-1}$, where $y_i, s \in \mathfrak{o}$. We have $(s) = \prod \mathfrak{p}^{v_\mathfrak{p}(s)}$, so $v_\mathfrak{p}(s) = 0$ for almost all $\mathfrak{p}$, and by increasing r we may assume that $v_\mathfrak{p}(s) = 0$ for $v_\mathfrak{p} \neq v_1, \ldots, v_r$. Now put $m_i = n_i + v_i(s)$; the $\mathfrak{p}_i^{m_i}$ are clearly pairwise comaximal, so by Theorem 4.5.2 there exists $y \in \mathfrak{o}$ such that $y \equiv y_i \pmod{\mathfrak{p}_i^{m_i}}$. Now it follows that $x = ys^{-1}$ has all the desired properties. ∎

Corollary 10.5.11. *If in Theorem 10.5.10 the v_i are normalized valuations, then there exists $x \in K$ such that $v_i(x) = n_i$ and $v(x) \geq 0$ for all other valuations v.*

Proof. We can find $x_i \in K$ such that $v_i(x_i) = n_i$; by the theorem there exists $x \in K$ such that $v_i(x - x_i) > n_i$, hence $v_i(x) = v_i(x_i) = n_i$. ∎

Let $\mathfrak{o}$ be a Dedekind domain with field of fractions K. The group $I(\mathfrak{o})$ has a subgroup consisting of all principal ideals, and the corresponding quotient group is called the *ideal class group* of $\mathfrak{o}$, written $C(\mathfrak{o})$. The mapping which associates with each $a \in K^\times$ the fractional ideal (a) gives rise to an exact sequence

$$1 \to U \to K^\times \to I(\mathfrak{o}) \to C(\mathfrak{o}) \to 1,$$

where U denotes the group of units of $\mathfrak{o}$. It is clear that $\mathfrak{o}$ is a principal ideal domain precisely when $C(\mathfrak{o}) = 1$, so this group measures the departure from being principal. It is a remarkable fact, proved by Kummer in 1847, that the ring of integers in an algebraic number field has a finite ideal class group (see e.g. Cohn (1991) Section 3.4). By contrast, the class groups of general Dedekind domains include all abelian groups (see Fossum (1973) §14). The structure of U was determined by Dirichlet, who showed that for an algebraic number field with r_1 real and $2r_2$ complex conjugates, U is the direct product of a finite cyclic group and a free abelian group of rank $r_1 + r_2 - 1$ (see Cohn (1991) Section 3.3).

If L is a finite extension of K and $\mathfrak{O}$ is the integral closure of $\mathfrak{o}$ in L, then by Theorem 10.5.5 and Exercise 10 below, $\mathfrak{O}$ is again a Dedekind domain. This shows in particular that any (integrally closed) ring of algebraic integers is a Dedekind domain. For any non-zero prime ideal $\mathfrak{p}$ in $\mathfrak{o}$ we can write $\mathfrak{p}\mathfrak{O} = \mathfrak{P}_1^{e_1} \ldots \mathfrak{P}_r^{e_r}$, $e_i > 0$. Thus $\mathfrak{p}$ splits into a finite number of prime ideals in $\mathfrak{O}$; we indicate that $\mathfrak{P}_i$ is a factor of $\mathfrak{p}$ by writing $\mathfrak{P}_i | \mathfrak{p}$. If $e_i > 1$, $\mathfrak{p}$ is said to be *ramified* at $\mathfrak{P}_i$; in fact e_i is the ramification index of the extension of the $\mathfrak{p}$-adic valuation on K to the $\mathfrak{P}_i$-adic valuation on L. The homomorphism from $I(\mathfrak{o})$ to $I(\mathfrak{O})$ defined by $a \mapsto a\mathfrak{O}$ is called the *conorm mapping*; in terms of the free generators it sends $\mathfrak{p}$ to $\prod \mathfrak{P}_i^{e_i}$. Further it maps principal ideals to principal ideals, and so induces a homomorphism $C(\mathfrak{o}) \to C(\mathfrak{O})$, which may be summed up in the commutative diagram

$$\begin{array}{ccccccccc} 1 & \to & U_\mathfrak{o} & \to & K^\times & \to & I(\mathfrak{o}) & \to & C(\mathfrak{o}) \to 1 \\ & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ 1 & \to & U_\mathfrak{o} & \to & L^\times & \to & I(\mathfrak{O}) & \to & C(\mathfrak{O}) \to 1. \end{array}$$

Exercises

1. In an integral domain show that if $\mathfrak{a}\mathfrak{b} = (c)$ is principal, then $\mathfrak{a}$ is invertible, and express its inverse in terms of $\mathfrak{b}$ and c.
2. Show that in the ring $\mathfrak{o} = k[x, y]$ the ideal (x, y) is not invertible. (Hint. Find $(\mathfrak{o} : (x, y))$.)
3. Show that for a square-free integer $d \neq 1$, the number 2 is not prime in $\mathbf{Z}[\sqrt{d}]$. Show also that 2 is an atom if $d \leq -3$.
4. Find the kernel of the homomorphism $\mathbf{Z}[\sqrt{-1}] \to \mathbf{Z}/(10)$ given by $\sqrt{-1} \mapsto 3$, and deduce that $\mathbf{Z}[\sqrt{-1}]/(3 - \sqrt{-1}) \cong \mathbf{Z}/(10)$.
5. Let d be a square-free integer and let $\mathfrak{O}$ be the ring of integers in $\mathbf{Q}(\sqrt{d})$, i.e. the integral closure of $\mathbf{Z}$ in $\mathbf{Q}(\sqrt{d})$. Show that if $d \equiv 2$ or $3 \pmod 4$, $\mathfrak{O}$ has a $\mathbf{Z}$-basis $1, \sqrt{d}$, and if $d \equiv 1 \pmod 4$, $\mathfrak{O}$ has a $\mathbf{Z}$-basis $1, (1 + \sqrt{d})/2$. Determine the ramified primes in each case.

6. Find the norm of a general element of the ring of integers in $\mathbf{Q}(\sqrt{-d})$, when $d \equiv 3 \pmod 4$, and show that this ring is Euclidean with respect to the norm for $d = 2, 3, 7, 11$.
7. Let d be a square-free integer > 11 or $d = 5$. Show that the ring $\mathfrak{o}$ of integers in $\mathbf{Q}(\sqrt{-d})$ is not Euclidean for any function φ such that $\varphi(ab) \geq \varphi(a)$ for $a, b \neq 0$. (Hint. Take $\gamma \neq 0$ with minimal $\varphi(\gamma)$ and show that $|\mathfrak{o}/\gamma\mathfrak{o}| = N(\gamma)$; deduce that γ must be a unit.)
8. Show that for an odd prime p, the p-adic valuation on $\mathbf{Q}$ has two extensions to $\mathbf{Q}(\sqrt{-1})$ if $p \equiv 1 \pmod 4$ and one if $p \equiv 3 \pmod 4$. Show that all such extensions are unramified. By decomposing p into prime ideals over $\mathbf{Z}[\sqrt{-1}]$ show that p can be written as a sum of two squares iff $p \equiv 1 \pmod 4$ or $p = 2$ (Fermat). (Hint. Recall that $\mathbf{U}(p)$ is a cyclic group.)
9. Show that the group of units in the ring of integers in $\mathbf{Q}(\sqrt{-d})$, for a square-free positive integer d, is $\mathbf{C}_2 = \{\pm 1\}$ except when $d = 1$ or 3, and determine the group in these cases.
10. Let $\mathfrak{o}$ be a Dedekind domain with field of fractions K, let L be a finite extension of K and let $\mathfrak{O}$ be the integral closure of $\mathfrak{o}$ in L. Prove that $\mathfrak{O}$ is a Dedekind domain. (Hint. By Theorem 10.5.5 reduce to the purely inseparable case, say $L \subseteq K^{1/q}$. Since $K \cong K^q$, the image $\mathfrak{o}^{1/q}$ of $\mathfrak{o}$ is a Dedekind domain such that $\mathfrak{o}^{1/q} \supseteq \mathfrak{O}$. If $\mathfrak{A}$ is a fractional ideal of $\mathfrak{O}$, then we have $\sum a_i b_i = 1$, where $a_i \in \mathfrak{A}$, $b_i \in (\mathfrak{o}^{1/q} : \mathfrak{A})$, because $\mathfrak{o}^{1/q}$ is Dedekind. Now raise this equation to the q-th power and note that $\sum a_i . a_i^{q-1} b_i^q = 1$.)
11. For any field k of characteristic not 2 show that the k-algebra generated by x, y with the defining relations $x^2 + y^2 = 1$ is a Dedekind domain, but the k-algebra generated by x, y, z with the defining relation $x^2 + y^2 + z^2 = 1$ is not. Find the class group of $k[x, y]/(x^2 + y^2 - 1)$.
12. Let $\mathfrak{o}$ be a Dedekind domain with field of fractions K. Show that any ring between $\mathfrak{o}$ and K is a Dedekind domain.
13. (B. Iversen) Show that a fractional ideal $\mathfrak{A}$ in an order $\mathfrak{o}$ of a quadratic algebraic number field is invertible iff $\mathrm{End}(\mathfrak{A}) \cong \mathfrak{o}$.
14. Let K be a field and $\mathscr{F}$ be a family of principal valuations on K such that (i) for any $x \in K$, $v(x) \geq 0$ for almost all $v \in \mathscr{F}$ and (ii) given $v, v' \in \mathscr{F}$ and $n \geq 0$, there exists $\alpha \in K$ such that $v(\alpha - 1) > n, v'(\alpha) > n$, and $w(\alpha) \geq 0$ for all $w \neq v, v'$. If R is the intersection of the corresponding valuation rings, show how to describe a fractional ideal in terms of the valuations in $\mathscr{F}$. Deduce that the fractional ideals form a group, and hence R is a Dedekind domain.

10.6 Modules over Dedekind Domains

We have seen that an integral domain is Dedekind iff it is hereditary. This also leads to a convenient description of projective modules.

Proposition 10.6.1. *Every finitely generated projective module over a Dedekind domain can be written as a direct sum of modules isomorphic to ideals.*

Proof. We remark that the ideals can be taken to be integral since every fractional ideal is isomorphic to an integral ideal. Now let $\mathfrak{o}$ be a Dedekind domain and P be a finitely generated projective $\mathfrak{o}$-module; P is submodule of a free module of finite rank, say $P \subseteq \mathfrak{o}^r$. We project on the last factor $\mathfrak{o}$ and denote the image, an ideal of $\mathfrak{o}$, by $\mathfrak{a}$. Thus we have a surjective map $\lambda : P \to \mathfrak{a}$, which leads to an exact sequence

$$0 \to P' \to P \to \mathfrak{a} \to 0, \tag{10.6.1}$$

where $P' = \ker \lambda = P \cap \mathfrak{o}^{r-1}$. By induction on r, P' is isomorphic to a direct sum of ideals, and since $\mathfrak{a}$ is projective, the sequence (10.6.1) splits and we obtain the desired decomposition for P. ■

We note that the proof shows every submodule of a free module of finite rank to be projective; this actually holds for all submodules of free modules, and it accounts for the name 'hereditary'.

To study modules over Dedekind domains, we need a reduction theorem, which allows us in many cases to replace the ring by a principal ideal domain (PID).

Proposition 10.6.2. *A Dedekind domain with only finitely many prime ideals is a principal ideal domain.*

Proof. Let $\mathfrak{o}$ be a Dedekind domain whose non-zero prime ideals are $\mathfrak{p}_1, \ldots, \mathfrak{p}_r$ and denote by v_i the valuation associated with $\mathfrak{p}_i$. Given any integers $n_1, \ldots, n_r \geq 0$, there exists $a \in \mathfrak{o}$ such that $v_i(a) = n_i$, by Corollary 10.5.11, and clearly

$$(a) = \mathfrak{p}_1^{n_1} \ldots \mathfrak{p}_r^{n_r};$$

hence every ideal of $\mathfrak{o}$ is principal. ■

Corollary 10.6.3. *Let $\mathfrak{o}$ be a Dedekind domain and $\mathfrak{a}$ be an integral ideal of $\mathfrak{o}$. Then $\mathfrak{o}/\mathfrak{a}$ is a principal ideal ring.*

Proof. Write $\mathfrak{a} = \mathfrak{p}_1^{n_1} \ldots \mathfrak{p}_r^{n_r}$, where the $\mathfrak{p}_i$ are distinct prime ideals containing $\mathfrak{a}$. Let S be the set of elements of $\mathfrak{o}$ which become units mod $\mathfrak{a}$, i.e. elements x such that $\mathfrak{a} + (x) = \mathfrak{o}$. Then S is multiplicative and $\mathfrak{o}_S$ is a domain whose group of fractional ideals is freely generated (as abelian group) by $\mathfrak{p}_1, \ldots, \mathfrak{p}_r$. Hence it is a Dedekind domain with finitely many prime ideals, and so is a PID, by Proposition 10.6.2. Now the natural homomorphism $\mathfrak{o} \to \mathfrak{o}/\mathfrak{a}$ is S-inverting, and hence can be taken via $\mathfrak{o}_S$, and as homomorphic image of $\mathfrak{o}_S$, $\mathfrak{o}/\mathfrak{a}$ is principal. ■

Of course $\mathfrak{o}/\mathfrak{a}$ will not in general be a domain; in fact it has zerodivisors unless it is a field (the case $r = 1$, $n_1 = 1$).

Corollary 10.6.4. *Every integral ideal $\mathfrak{a}$ of a Dedekind domain $\mathfrak{o}$ can be generated by two elements, one of which may be chosen arbitrarily as non-zero element of $\mathfrak{a}$. Moreover, this element may be chose prime to a given element, itself prime to $\mathfrak{a}$.*

Proof. Take $0 \neq a \in \mathfrak{a}$; then $\mathfrak{a}/(a)$ is an ideal of $\mathfrak{o}/(a)$ and therefore principal. If $\mathfrak{a} \equiv (b) \pmod{(a)}$, then $\mathfrak{a} = (a, b)$. Moreover, given $c \in \mathfrak{o}$, prime to $\mathfrak{a}$, we have $cu + a = 1$ for some $u \in \mathfrak{o}$, $a \in \mathfrak{a}$ and we can start with this a. ■

Corollary 10.6.5. *Let* $\mathfrak{a}$, $\mathfrak{b}$ *be fractional ideals in a Dedekind domain* $\mathfrak{o}$ *with field of fractions* K, *and suppose that* $\mathfrak{b}$ *is integral. Then there exists* $u \in K^\times$ *such that*

$$u\mathfrak{a} + \mathfrak{b} = \mathfrak{o}. \tag{10.6.2}$$

Proof. Choose $c \in K^\times$ such that $c\mathfrak{a}$ is integral. Then $c\mathfrak{a}/c\mathfrak{a}\mathfrak{b}$ is an ideal in $\mathfrak{o}/c\mathfrak{a}\mathfrak{b}$, and hence is principal: $c\mathfrak{a} = (v) + c\mathfrak{a}\mathfrak{b}$. On multiplying by $(c\mathfrak{a})^{-1}$ we obtain $(vc^{-1})\mathfrak{a}^{-1} + \mathfrak{b} = \mathfrak{o}$, and now (10.6.2) follows if we put $u = vc^{-1}$ and replace $\mathfrak{a}$ by $\mathfrak{a}^{-1}$. ∎

We recall that for abelian groups we have the basis theorem (Theorem 2.4.1) which tells us that every finitely generated abelian group is a direct sum of cyclic groups. The generalization to modules over a PID is well known; such a module, if finitely generated, is again a direct sum of cyclic modules, in particular, every finitely generated torsion-free module over a PID is free. We shall find that corresponding results hold for finitely generated modules over Dedekind domains, with projective modules taking the place of free modules.

Proposition 10.6.6. *A finitely generated module over a Dedekind domain is projective if and only if it is torsion-free.*

Proof. Let $\mathfrak{o}$ be the ring and K be its field of fractions. If M is projective, then it is a submodule of a free module and hence torsion-free. Conversely, let M be finitely generated torsion-free. Then the natural mapping

$$M \to M \otimes_{\mathfrak{o}} K \tag{10.6.3}$$

is an embedding, and $M \otimes K$ is a finitely generated K-module, hence isomorphic to K^r, for some $r \geq 0$. Let M be generated by $u_1, \ldots, u_n$ and let c be a common denominator for the expressions of the images of the u_i in the embedding (10.6.3). Then $cM \subseteq \mathfrak{o}^r$, hence M is isomorphic to a submodule of a free module; since $\mathfrak{o}$ is hereditary, it follows that M is projective (see the remark after Proposition 10.6.1). ∎

Corollary 10.6.7. *Let* M *be a finitely generated module over a Dedekind domain. Then* $M = tM \oplus P$ *where* P *is a torsion-free submodule of* M *and* tM *is the torsion submodule.*

Proof. We have the exact sequence

$$0 \to tM \to M \to M/tM \to 0;$$

here M/tM is torsion-free and finitely generated, because M is. Hence it is projective and so the sequence splits: $M = tM \oplus P$, where $P \cong M/tM$. ∎

This result shows that we may consider the torsion part and the torsion-free part separately; we begin with the torsion part. Thus let M be a finitely generated torsion module over a Dedekind domain $\mathfrak{o}$. If $u_1, \ldots, u_n$ is a generating set and $\mathfrak{n}_i$ is the annihilator of u_i $(i = 1, \ldots, n)$, then $\mathfrak{n} = \mathfrak{n}_1 \ldots \mathfrak{n}_n$ annihilates every element of M.

Let S be the set of all elements of $\mathfrak{o}$ prime to $\mathfrak{n}$; then for any $s \in S$ there exists $t \in \mathfrak{o}$ and $a \in \mathfrak{n}$ such that $st + a = 1$, hence $xst = x$ for all $x \in M$. Thus the action of s on M is an automorphism with inverse t. It follows that the natural mapping

$$M \to M \otimes \mathfrak{o}_S \tag{10.6.4}$$

is an isomorphism of abelian groups. More precisely, (10.6.4) provides a natural transformation from $\mathfrak{o}$-modules to $\mathfrak{o}_S$-modules, which for modules annihilated by $\mathfrak{n}$ is an isomorphism. Now $\mathfrak{o}_S$ has only finitely many prime ideals, hence by Proposition 10.6.2, $\mathfrak{o}_S$ is a PID, so we can apply the basis theorem for modules over PIDs (see Cohn (2000) Chapter 3), which states that any finitely generated module over a PID is a direct sum of cyclic modules, each a homomorphic image of the next, and a free module. So we have

$$M \otimes \mathfrak{o}_S \cong \mathfrak{o}_S/(a_1) \oplus \ldots \oplus \mathfrak{o}_S/(a_r), \quad a_i | a_{i+1},$$

and this decomposition is unique up to isomorphism. It follows that M admits a corresponding decomposition into cyclic $\mathfrak{o}$-modules, with ideals of $\mathfrak{o}$ corresponding to the a_i. Thus we obtain

Theorem 10.6.8. *Any finitely generated torsion module M over a Dedekind domain $\mathfrak{o}$ is a direct sum of cyclic $\mathfrak{o}$-modules:*

$$M \cong \mathfrak{o}/\mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{o}/\mathfrak{a}_r, \quad \mathfrak{a}_i \supseteq \mathfrak{a}_{i+1}, \tag{10.6.5}$$

and this decomposition is unique up to isomorphism. ■

As in the case of abelian groups we can, for any torsion module, indeed for any module, define for each non-zero prime ideal $\mathfrak{p}$, the $\mathfrak{p}$-*primary part*

$$M = \{x \in M | x\mathfrak{p}^n = 0 \text{ for some } n \geq 1\},$$

and in the same way as for abelian groups prove

Proposition 10.6.9. *Any torsion module M over a Dedekind domain is a direct sum of its primary parts, in a unique way:*

$$M = \oplus M_{\mathfrak{p}},$$

and when M is finitely generated, only finitely many terms on the right are different from zero. ■

It remains to classify the torsion-free modules. This is done in the next theorem, which is essentially due to Ernst Steinitz [1912]. Before stating it we note a lemma which is needed in the proof.

Lemma 10.6.10. *Let R be an integral domain with field of fractions K. If $\mathfrak{a}$, $\mathfrak{b}$ are fractional ideals of R and $f : \mathfrak{a} \to \mathfrak{b}$ is a homomorphism of R-modules, then there exists $c \in K$ such that $xf = cx$ for all $x \in \mathfrak{a}$. In particular, f is either zero or injective.*

Proof. For $a, x \in \mathfrak{a}$ we have

$$af.x = (ax)f = xf.a.$$

Fix $a \neq 0$ and write $c = af.a^{-1}$; then $c \in K$ and $xf = cx$. ∎

In particular we see that $\mathfrak{a} \cong \mathfrak{b}$ iff $\mathfrak{a} = c\mathfrak{b}$ for some $c \in K^\times$.

Theorem 10.6.11. *Let $\mathfrak{o}$ be a Dedekind domain. Any finitely generated torsion-free $\mathfrak{o}$-module has the form $\mathfrak{o}^r \oplus \mathfrak{a}$, where $\mathfrak{a}$ is an ideal of $\mathfrak{o}$, and*

$$\mathfrak{o}^r \oplus \mathfrak{a} \cong \mathfrak{o}^s \oplus \mathfrak{b} \Leftrightarrow r = s \text{ and } \mathfrak{a} \cong \mathfrak{b}.$$

More explicitly, if $\mathfrak{a}_1, \ldots, \mathfrak{a}_r, \mathfrak{b}_1, \ldots, \mathfrak{b}_s$ are any fractional ideals of $\mathfrak{o}$, then

$$\mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_r \cong \mathfrak{b}_1 \oplus \ldots \oplus \mathfrak{b}_s \Leftrightarrow r = s \text{ and } \mathfrak{a}_1 \ldots \mathfrak{a}_r \cong \mathfrak{b}_1 \ldots \mathfrak{b}_s. \tag{10.6.6}$$

Proof. By Propositions 10.6.6 and 10.6.1, any finitely generated torsion-free module is a direct sum of ideals. Let us first prove the necessity of the conditions in (10.6.6). In Lemma 10.6.10 we have seen that any homomorphism $\mathfrak{a} \to \mathfrak{b}$ is obtained by multiplying by an element of K, hence any homomorphism

$$\gamma : \mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_r \to \mathfrak{b}_1 \oplus \ldots \oplus \mathfrak{b}_s$$

is obtained by multiplying by an $s \times r$ matrix over K : $\mathfrak{b}_i = \sum c_{ij}\mathfrak{a}_j$. Clearly γ is an isomorphism iff the matrix $C = (c_{ij})$ has an inverse; in particular we must then have $r = s$. We claim further that in this case

$$\mathfrak{b}_1 \ldots \mathfrak{b}_r = (\det C)\mathfrak{a}_1 \ldots \mathfrak{a}_r. \tag{10.6.7}$$

For, given $a_j \in \mathfrak{a}_j$, we have $c_{ij}a_j \in \mathfrak{b}_i$, hence $(\det C)\mathfrak{a}_1 \ldots \mathfrak{a}_r \subseteq \mathfrak{b}_1 \ldots \mathfrak{b}_r$. By symmetry we have the reverse inclusion, and hence equality in (10.6.7). This proves the necessity of the conditions (in either formulation); we observe that it holds in any integral domain, not necessarily Dedekind.

To establish the converse we shall show that $\mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_r \cong \mathfrak{o}^{r-1} \oplus \mathfrak{a}_1 \ldots \mathfrak{a}_r$; clearly the desired conclusion follows from this. Consider first the case $r = 2$:

$$\mathfrak{a} \oplus \mathfrak{b} \cong \mathfrak{o} \oplus \mathfrak{ab}. \tag{10.6.8}$$

On multiplying $\mathfrak{a}$, $\mathfrak{b}$ by elements of K we may assume them to be integral, and then by Corollary 10.6.5 we may further assume that $\mathfrak{a} + \mathfrak{b} = \mathfrak{o}$. Then we have an exact sequence

$$0 \to \ker \lambda \longrightarrow \mathfrak{a} \oplus \mathfrak{b} \xrightarrow{\lambda} \mathfrak{o} \to 0, \tag{10.6.9}$$

where $\lambda : (x, y) \mapsto x - y$ and $\ker \lambda = \mathfrak{a} \cap \mathfrak{b} = \mathfrak{ab}$, because $\mathfrak{a}$, $\mathfrak{b}$ are comaximal. Now (10.6.9) splits and so we obtain (10.6.8). It follows that

$$\mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_r \cong \mathfrak{o} \oplus \mathfrak{a}_1\mathfrak{a}_2 \oplus \mathfrak{a}_3 \oplus \ldots \oplus \mathfrak{a}_r \cong \mathfrak{o}^{r-1} \oplus \mathfrak{a}_1 \ldots \mathfrak{a}_r,$$

by induction on r, which is what we had to show. ∎

The result may be summed up by saying that a complete set of invariants for the projective module $\mathfrak{a}_1 \oplus \ldots \oplus \mathfrak{a}_r$ consists of the integer r and the isomorphism class of the fractional ideal $\mathfrak{a}_1 \ldots \mathfrak{a}_r$.

Exercises

1. Show that every projective module over a Dedekind domain, which is countably but not finitely generated, is free. (This actually holds without the countability restriction, by a theorem of Kaplansky [1958].)
2. Show that if the definition of torsion-free module in the text is applied to general rings, then a ring which is not an integral domain has no non-zero torsion-free modules. (A module over a general ring R is sometimes called *torsion-free* if multiplication by any non-zerodivisor of R is injective, but we shall have no need to use this definition.)
3. Show that any module over a Dedekind domain has the form $M = I \oplus E$, where I is injective and E has no non-zero injective submodules.
4. Let $\mathfrak{o}$ be a Dedekind domain and K be its field of fractions. Show that for any prime ideal $\mathfrak{p} \neq 0$ the $\mathfrak{p}$-primary part of $K/\mathfrak{o}$ is indecomposable injective. It is called a module of *type* $\mathfrak{p}^\infty$. Show that every divisible $\mathfrak{o}$-module is a direct sum of a K-space and of modules of type $\mathfrak{p}^\infty$, for different $\mathfrak{p}$. Deduce that every divisible $\mathfrak{o}$-module is injective.
5. Show that a finitely generated indecomposable module over a Dedekind domain $\mathfrak{o}$ is either torsion-free or of type $\mathfrak{o}/\mathfrak{p}^n$.

10.7 Algebraic Equations

Algebraic geometry may be defined as the study of the solutions of polynomial equations over fields. An equation $f(x) = 0$ in one variable has a finite number of roots, but the solutions of an equation in several variables form an algebraic variety, e.g. the equation $x^2 + y^2 = 1$ defines a circle in the (x, y)-plane. Our task will be to explain how rings arise in this context.

Let K be a commutative ring and consider a system of polynomial equations over K, in the indeterminates $x_1, \ldots, x_n$. This means that we have a family of equations

$$E : f_\lambda(x_1, \ldots, x_n) = 0 \quad (\lambda \in I), \tag{10.7.1}$$

where the left-hand sides are members of the polynomial ring $K[x_1, \ldots, x_n]$. Let L be a K-algebra; by a *solution* of the system E in L we understand an n-tuple $(a_1, \ldots, a_n)$ of elements of L such that $f_\lambda(a_1, \ldots, a_n) = 0$ for all $\lambda \in I$. The set of all such solutions is denoted by $V_L(E)$.

Two systems E and F in the same variables over K are said to be *equivalent* if $V_L(E) = V_L(F)$ for all K-algebras L. Among all systems equivalent to E there is a maximal one, namely the ideal generated by E in the polynomial ring:

Theorem 10.7.1. *Let E be a system of equations in $x_1, \ldots, x_n$ over a ring K, denote by $\mathfrak{a}$ the ideal of $K[x_1, \ldots, x_n]$ generated by the left-hand sides of E and put $A = K[x_1, \ldots, x_n]/\mathfrak{a}$. Then each solution of E in a K-algebra L may be described by a K-algebra homomorphism from A to L and the mapping so obtained is a natural bijection:*

$$V_L(E) \cong \mathrm{Hom}_K(A, L). \tag{10.7.2}$$

The algebra A is called the *function ring* or *coordinate ring* of the system E.

Proof. We shall verify the universal property of A. Any homomorphism from A to L gives rise to a homomorphism $\varphi : K[x_1, \ldots, x_n] \to L$ mapping E to 0. If φ maps

$$x_i \mapsto \xi_i, \tag{10.7.3}$$

then $(\xi_1, \ldots, \xi_n)$ is the corresponding solution. Conversely, let $(\xi_1, \ldots, \xi_n)$ be a solution of the system E in L and let φ be the unique K-algebra homomorphism from $K[x_1, \ldots, x_n]$ to L defined by (10.7.3). Then $E \subseteq \ker \varphi$, because $(\xi_1, \ldots, \xi_n)$ was a solution of E. Hence φ can be factored by the natural homomorphism $K[x_1, \ldots, x_n] \to A$ to give a homomorphism $A \to L$, and these two constructions are evidently mutually inverse. ∎

A system E is said to be *consistent* if it has a solution in some non-trivial K-algebra, i.e. $V_L(E) \neq \varnothing$ for some $L \neq 0$; if $V_L(E) = \varnothing$, E is *inconsistent*. By Theorem 10.7.1. we have

Corollary 10.7.2. *A system E of equations over K is consistent if and only if the ideal generated by the left-hand sides is proper in $K[x_1, \ldots, x_n]$.*

Proof. Clearly the ideal in question is proper iff the algebra A in Theorem 10.7.1 is non-trivial. Now assume $V_L(E) \neq \varnothing$; then by Theorem 10.7.1, $\mathrm{Hom}(A, L) \neq \varnothing$ for some non-trivial L, hence $A \neq 0$. Conversely, if $A \neq 0$, then $V_A(E) = \mathrm{Hom}(A, A) \neq \varnothing$, since there is always the identity mapping. ∎

We can apply the Hilbert basis theorem and obtain

Proposition 10.7.3. *Let K be a Noetherian ring. Then every system of equations in $x_1, \ldots, x_n$ over K is equivalent to a finite system.*

Proof. By the Hilbert basis theorem (Theorem 10.4.1), $R = K[x_1, \ldots, x_n]$ is Noetherian. Let E be any subset of R and $\mathfrak{a}$ be the ideal generated by E. As a system of equations, E is equivalent to $\mathfrak{a}$, and since R is Noetherian, we can write $\mathfrak{a} = (f_1, \ldots, f_r)$, so $\mathfrak{a}$ is equivalent to the finite system $f_1, \ldots, f_r$. ∎

Theorem 10.7.1 leads to a correspondence between systems of equations and function rings which have their counterparts in the geometrical objects formed by the solution sets. Thus let k be a field and E be a system of equations in $x_1, \ldots, x_n$ over k. If L is a field containing k, then $V_L(E)$, the set of solutions of E in L, is a subset of L. When k is a subfield of the real numbers and $n = 2$ or 3, we thus obtain a representation of E in the plane or in space, and the geometric language

suggested by this example is used even when $n > 3$ and the coordinates lie in a general field.

To give an illustration, the equation

$$x^2 + y^2 = 1 \tag{10.7.4}$$

defines a circle in the plane and its function ring is generated by x, y over k, subject to the relation (10.7.4). This function ring is Noetherian, by Proposition 10.4.3, but not a UFD, for x^2 has the two factorizations $x.x = (1-y)(1+y)$, and it is not hard to show (e.g. using the norm over $k(x)$) that $x, 1+y, 1-y$ are atoms in the ring, but not primes.

The circle in space corresponding to (10.7.4) is given by the system

$$x^2 + y^2 = 1, \quad z = 0.$$

Clearly it has the same function ring as the circle in the plane. The geometric object may consist of several parts, e.g. the system

$$x(x^2 + y^2 + z^2 - 1) = y(x^2 + y^2 + z^2 - 1) = 0 \tag{10.7.5}$$

consists of a sphere and a line (the z-axis).

Let us consider more closely the correspondence between subsets of n-dimensional space and ideals in $R = k[x_1, \ldots, x_n]$. For simplicity we shall work over k itself (rather than an extension), so that our space is k^n. To each subset E of R there corresponds the set $V_k(E)$, or simply $V(E)$, of solutions in k. A subset of k^n is said to be *closed* or an *algebraic set* if it has the form $V(E)$ for some $E \subseteq R$. Likewise with each subset S of k^n we associate the set $I(S)$ of all polynomials vanishing on S, and thus obtain a Galois connexion:

$$V(E) = \{p \in k^n | f(p) = 0 \text{ for all } f \in E\},$$

$$I(S) = \{f \in R | f(p) = 0 \text{ for all } p \in S\}.$$

It is clear that $I(S)$ is an ideal in R; moreover, it coincides with its own radical, for if $f^r \in I(S)$, then $f(p)^r = 0$ for all $p \in S$ and so $f(p) = 0$, i.e. $f \in I(S)$. By Proposition 10.7.3, every algebraic set can be defined by a finite set of equations.

As in every Galois connexion we have the following rules:

(i) $VI(S) \supseteq S, \qquad IV(\mathfrak{a}) \supseteq \mathfrak{a},$

(ii) $S_1 \subseteq S_2 \Rightarrow I(S_1) \supseteq I(S_2), \qquad \mathfrak{a}_1 \subseteq \mathfrak{a}_2 \Rightarrow V(\mathfrak{a}_1) \supseteq V(\mathfrak{a}_2),$

(iii) $IVI(S) = I(S), \qquad VIV(\mathfrak{a}) = V(\mathfrak{a}).$

Proof. (i) $f \in I(S)$ means that $f(p) = 0$ for all $p \in S$, hence $p \in S$ implies $p \in VI(S)$; similarly for the other inclusion.

(ii) If $f(p) = 0$ for all $p \in S_2$, then this holds for all $p \in S_1$, hence $I(S_1) \supseteq I(S_2)$; again the second relation follows similarly.

Now (iii) follows from (i) and (ii), as we have seen in Section 7.6. ∎

What has been proved shows that the relation $f(p) = 0$ defines an order-inverting bijection between the subsets of k^n of the form $V(a)$, the algebraic sets, and the

subsets of R of the form $I(S)$. We have seen that each $I(S)$ is an ideal equal to its own radical; later we shall find, as a consequence of the Hilbert Nullstellensatz, in Section 10.10, that over an algebraically closed field every ideal equal to its own radical is of this form.

We observe that the descending chains of closed sets in k^n correspond to ascending chains of ideals in $k[x_1, \ldots, x_n]$, and since the latter ring is Noetherian, the ideal chains break off. It follows that k satisfies the descending chain condition on closed subsets. Of course the ascending chain condition will not generally hold, since e.g. every finite set is closed.

A closed set S in k is said to be *reducible* if $S = \varnothing$ or $S = S_1 \cup S_2$, where S_1, S_2 are proper closed subsets of S; otherwise it is called *irreducible* or a *variety*. Thus the circle (10.7.4) is a variety, but the sphere-and-line (10.7.5) is reducible. However, even a variety may consist of several pieces in real space (see Exercise 5). There is a simple algebraic criterion for irreducibility:

Proposition 10.7.4. *A closed set S in k^n is irreducible if and only if $I(S)$ is a prime ideal in $k[x_1, \ldots, x_n]$.*

Proof. Write $\mathfrak{a} = I(S)$ and suppose that $\mathfrak{a}$ is not prime; if $S = \varnothing$, there is nothing to prove; otherwise $\mathfrak{a}$ is proper and there exist $f_1, f_2 \notin \mathfrak{a}$ but $f_1 f_2 \in \mathfrak{a}$. Hence there exist $p_1, p_2 \in S$ such that $f_i(p_i) \neq 0$. Put $S_i = V(\mathfrak{a} + (f_i))$ $(i = 1, 2)$; then S_i is closed and a proper subset of S. Moreover, $S_1 \cup S_2 = V((\mathfrak{a} + (f_1))(\mathfrak{a} + (f_2))) = V(\mathfrak{a}^2 + (f_1 f_2)) = S$, so S is reducible. Conversely, suppose that $S = S_1 \cup S_2$, where S_i is closed and $S_i \subset S$, and write $\mathfrak{a}_i = I(S_i)$. Then $\mathfrak{a}_i \supset \mathfrak{a}$, hence there exists $f_i \in \mathfrak{a}_i \backslash \mathfrak{a}$. Now $f_1 f_2 \in \mathfrak{a}_1 \cap \mathfrak{a}_2 = \mathfrak{a}$, and this shows that $\mathfrak{a}$ is not prime. ∎

If V is a variety in k^n, then $I(V)$ is a prime ideal and hence there is an extension field F of k containing the function ring of V, namely the field of fractions of $k[x_1, \ldots, x_n]/I(V)$. If ξ_i is the image in F of x_i under the natural homomorphism, then every point of V can be obtained by 'specializing' the point $(\xi_1, \ldots, \xi_n) \in F$ (in a sense which needs to be made precise); this point is therefore called a *generic point* of V. Conversely, any algebraic set with a generic point is a variety, e.g. a generic point for the circle (10.7.4) is $((1 - t^2)/(1 + t^2), 2t/(1 + t^2))$.

Exercises

1. Show that a system of equations is consistent iff it has a common solution over some extension field.
2. Show that every algebraic set in k^n can be written as a union of a finite number of irreducible sets. (Hint. Use the Hilbert basis theorem.)
3. Show that over an algebraically closed field k, the irreducible subsets of k^2 are points, curves and k^2.
4. Let φ be the homomorphism from $k[x, y]$ to $k[x] \times k[y]$ given by $f(x, y) \mapsto (f(x, 0), f(0, y))$. Show that $\ker \varphi = (xy)$, $\operatorname{im} \varphi = (f(x), g(y))$ with $f(0) = g(0)$.
5. Verify that the curve $y^2 = x^3 - x$ is irreducible, but consists of two pieces (in the real plane).

6. Find the function ring of the hyperbola $xy = 1$ (over a field of characteristic not 2) and compare it with that of a straight line.
7. Find the function ring of the 'semicubical parabola' $y^2 = x^3$, and show that it is not integrally closed. Find the integral closure, and a curve of which it is the function ring.
8. Verify that $V(\mathfrak{a}\mathfrak{b}) = V(\mathfrak{a}) \cup V(\mathfrak{b})$, $V(\cup \mathfrak{a}_\lambda) = \cap V(\mathfrak{a}_\lambda)$, $V(\{1\}) = \varnothing$, $V(\{0\}) = k^n$. (This means that the algebraic sets form the closed sets of a topology on k, known as the *Zariski topology*, see Section 10.10 below.)
9. Show that every covering by open sets of k^n has a finite subcovering (since k^n is not Hausdorff, this is known as *quasicompactness* of k^n).

10.8 The Primary Decomposition

In **Z** or, more generally, in any UFD, each non-zero element has a factorization into primes:

$$a = up_1^{\alpha_1} \dots p_r^{\alpha_r} \quad (u \text{ a unit}, p_i \text{ distinct primes}, \alpha_i > 0, r \geq 0). \tag{10.8.1}$$

In terms of ideals this can be written as

$$(a) = (p_1^{\alpha_1}) \cap \dots \cap (p_r^{\alpha_r}). \tag{10.8.2}$$

Of course (10.8.2) is less precise than (10.8.1), but it has the advantage of holding for a much wider class of rings. We shall find that a decomposition of this form is true for any Noetherian ring. For $k[x_1, \dots, x_n]$ this was proved in 1905 by Emanuel Lasker (also world chess champion from 1894 to 1921) and extended to general Noetherian rings by Emmy Noether in 1921. Following Bourbaki we shall derive a decomposition for modules, which of course includes ideals as a special case.

Let R be a commutative ring and M be an R-module. A prime ideal $\mathfrak{p}$ of R is said to be *associated* to M if it occurs as the annihilator of an element: $\mathfrak{p} = \mathrm{Ann}(u)$ for some $u \in M$. The set of all associated primes of M is denoted by $\mathrm{Ass}(M)$ (the 'assassinator' of M). If $\mathfrak{p} = \mathrm{Ann}(u)$, the mapping $x \mapsto ux$ of R into M shows that $R/\mathfrak{p}$ is embedded in M and conversely, if $R/\mathfrak{p}$ is embedded in M, then $\mathfrak{p} = \mathrm{Ann}(u)$ for some $u \in M$; hence $\mathfrak{p}$ is associated to M iff $R/\mathfrak{p}$ is embedded in M. We also note that $\mathrm{Ass}(R/\mathfrak{p}) = \{\mathfrak{p}\}$, since every non-zero element of $R/\mathfrak{p}$ has annihilator $\mathfrak{p}$.

Proposition 10.8.1. *Let M be an R-module; any maximal member of the set of annihilators of non-zero elements of M is a prime ideal of R.*

Proof. The annihilator of $u \neq 0$ is a proper ideal. Suppose that $\mathfrak{p} = \mathrm{Ann}(u)$ is maximal and let $ab \in \mathfrak{p}$, $a \notin \mathfrak{p}$. Then $ua \neq 0$, $uab = 0$, hence $b \in \mathrm{Ann}(ua) \supseteq \mathfrak{p}$. By the maximality of $\mathfrak{p}$ we have equality, hence $b \in \mathfrak{p}$ and this shows $\mathfrak{p}$ to be prime. ■

The converse is false: a member of $\mathrm{Ass}(M)$ need not be maximal, but if $\mathrm{Ann}(u) \in \mathrm{Ass}(M)$, then $\mathrm{Ann}(u)$ *is* maximal among the annihilators of submodules of uR. Of course there may be no such maximal elements (see Exercise 3), but if R

is Noetherian and $M \neq 0$, then we can be sure of such maximal annihilators, and they will be prime, by Proposition 10.8.1. Together with the obvious fact that $\mathrm{Ass}(0) = \varnothing$, this proves

Corollary 10.8.2. *Let R be a Noetherian ring. Then for any R-module M, $\mathrm{Ass}(M) = \varnothing$ if and only if $M = 0$.* ∎

Let us call $a \in R$ a *zerodivisor on M* if it annihilates some non-zero element of M. Since every annihilator is contained in some maximal annihilator, we obtain

Corollary 10.8.3. *Let R be a Noetherian ring and M be an R-module. Then the set of zerodivisors on M is $\cup\{\mathfrak{p} | \mathfrak{p} \in \mathrm{Ass}(M)\}$.* ∎

Let M be an R-module and M' be a submodule; we have the formula

$$\mathrm{Ass}(M') \subseteq \mathrm{Ass}(M) \subseteq \mathrm{Ass}(M') \cup \mathrm{Ass}(M/M'). \tag{10.8.3}$$

The first inclusion is clear. To prove the second, let $\mathfrak{p} \in \mathrm{Ass}(M)$, say $\mathfrak{p} = \mathrm{Ann}(u)$. Either $uR \cap M' \neq 0$, and since any non-zero element of $uR \cong R/\mathfrak{p}$ has annihilator $\mathfrak{p}$, we see that $\mathfrak{p} \in \mathrm{Ass}(M')$; or $uR \cap M' = 0$, in which case the natural mapping $M \to M/M'$, restricted to uR, is injective, so uR is embedded in M/M' and $\mathfrak{p} \in \mathrm{Ass}(M/M')$. ∎

We remark that $\mathrm{Ass}(M)$ is the set of prime ideals which *equal* some $\mathrm{Ann}(x)$ $(x \in M)$, while $\mathrm{Supp}(M)$ is the set of prime ideals which *contain* some $\mathrm{Ann}(x)$ $(x \in M)$. A Noetherian module M has finite length iff $\mathrm{Ass}(M)$, or equivalently, $\mathrm{Supp}(M)$, consists entirely of maximal ideals.

We can now show that for a finitely generated module M over a Noetherian ring, $\mathrm{Ass}(M)$ is finite. More precisely, we have

Proposition 10.8.4. *Let R be a Noetherian ring and M be a finitely generated R-module. Then there is a finite chain of submodules*

$$0 = M_0 \subset M_1 \subset \ldots \subset M_r = M, \tag{10.8.4}$$

such that $M_i/M_{i-1} \cong R/\mathfrak{p}_i$ for some prime ideal $\mathfrak{p}_i$ of R, and

$$\mathrm{Ass}(M) \subseteq \{\mathfrak{p}_1, \ldots, \mathfrak{p}_r\}.$$

Proof. For $M = 0$ there is nothing to prove. Otherwise M has a submodule M_1 of the form $R/\mathfrak{p}$, by Corollary 10.8.2, and we can take $\mathfrak{p}_1 = \mathfrak{p}$. If we have found a chain $M_1 \subset \ldots \subset M_r$ with $M_i/M_{i-1} \cong R/\mathfrak{p}_i$ and $M_r \neq M$, then M/M_r has a submodule of the form $R/\mathfrak{p}$, where $\mathfrak{p} \in \mathrm{Ass}(M/M_r)$, say $M_{r+1}/M_r \cong R/\mathfrak{p}$. We can then put $\mathfrak{p}_{r+1} = \mathfrak{p}$ and continue the chain. Since R is Noetherian, the chain must break off and we obtain (10.8.4). Now by induction we have, from (10.8.3),

$$\mathrm{Ass}(M) \subseteq \mathrm{Ass}(M_1/M_0) \cup \ldots \cup \mathrm{Ass}(M_r/M_{r-1}) = \{\mathfrak{p}_1, \ldots, \mathfrak{p}_r\}.$$ ∎

Let M be an R-module; an element $a \in R$ is said to be *locally nilpotent* on M if for each $x \in M$ there exists $n = n(x)$ such that $xa^n = 0$. If M is finitely generated, by $u_1, \ldots, u_r$ say, and $u_i a^{n_i} = 0$, then by taking $n = \max\{n_1, \ldots, n_r\}$ we find that a^n annihilates M, so in this case the action of a on M is nilpotent.

When R is Noetherian, we observe that a is locally nilpotent on M iff $a \in \cap\{\mathfrak{p} | \mathfrak{p} \in \mathrm{Ass}(M)\}$. For if a is locally nilpotent, then for any $\mathfrak{p} \in \mathrm{Ass}(M)$, $a^n \in \mathfrak{p}$ for some n, hence $a \in \mathfrak{p}$. Conversely, if a is not locally nilpotent, then there exists $x \in M$ such that $xa^n \neq 0$ for all n. Among such x choose one, say x_1, with maximal annihilator $\mathfrak{p} = \mathrm{Ann}(x_1)$. We claim that $\mathfrak{p}$ is prime. Clearly $\mathfrak{p} \neq R$; if $bc \in \mathfrak{p}$, then $x_1 bc = 0$, hence $c \in \mathrm{Ann}(x_1 b) \supseteq \mathfrak{p}$. Either $x_1 ba^n \neq 0$ for all n; then by maximality, $c \in \mathfrak{p}$. Or $x_1 ba^n = 0$ for some n, in which case $b \in \mathrm{Ann}(x_1 a^n) \supseteq \mathfrak{p}$, and $x_1 a^n . a^m \neq 0$ for all m, so $b \in \mathfrak{p}$, again by maximality. This shows that $\mathfrak{p}$ is prime. Clearly $\mathfrak{p} \in \mathrm{Ass}(M)$ and $a \notin \mathfrak{p}$. This proves

Proposition 10.8.5. *For a Noetherian ring R and any R-module M, an element a of R is locally nilpotent on M if and only if $a \in \cap\{\mathfrak{p} | \mathfrak{p} \in \mathrm{Ass}(M)\}$.* ■

Let R be a Noetherian ring and M be an R-module. A submodule Q of M is said to be *primary* (in M) if $\mathrm{Ass}(M/Q)$ consists of a single element. If $\mathrm{Ass}(M/Q) = \{\mathfrak{p}\}$, we also say that Q is $\mathfrak{p}$-*primary*. In particular, this defines primary ideals in R. Thus an ideal $\mathfrak{q}$ in R is primary precisely if $\mathfrak{q} \neq R$ and every zerodivisor in $R/\mathfrak{q}$ is nilpotent; then $\sqrt{\mathfrak{q}} = \mathfrak{p}$ is a prime ideal and $\mathfrak{q}$ is $\mathfrak{p}$-primary. For example, in $\mathbf{Z}$, the p-primary ideals are (p^r), $r = 1, 2, \ldots$, but in general the $\mathfrak{p}$-primary ideals need not be powers of $\mathfrak{p}$, e.g. in $k[x, y]$, (x, y^2) is (x, y)-primary, but $(x, y^2) \neq (x, y)^r$ for all r. Neither is it true that each power of a prime ideal is necessarily primary (see Exercise 9).

Lemma 10.8.6. *Let R be a Noetherian ring and M be an R-module. If $Q_1, \ldots, Q_r$ are submodules of M which are $\mathfrak{p}$-primary for the same prime ideal $\mathfrak{p}$, then $Q_1 \cap \ldots \cap Q_r$ is also $\mathfrak{p}$-primary.*

Proof. We have a natural homomorphism

$$M/(Q_1 \cap \ldots \cap Q_r) \to M/Q_1 \oplus \ldots \oplus M/Q_r, \tag{10.8.5}$$

obtained by composing the mappings from the left-hand side to M/Q_i. If the module on the right is written N, then $\mathrm{Ass}(N) = \{\mathfrak{p}\}$ by (10.8.3), and since (10.8.5) is clearly injective, the same holds for the module on the left, hence $Q_1 \cap \ldots \cap Q_r$ is $\mathfrak{p}$-primary, as claimed. ■

Our objective will be to obtain an embedding of our module into a direct sum of quotients by primary submodules:

$$M \to M/Q_1 \oplus \ldots \oplus M/Q_r. \tag{10.8.6}$$

Such a mapping clearly exists for any primary submodules $Q_1, \ldots, Q_r$ and (10.8.6) will be injective precisely when

$$Q_1 \cap \ldots \cap Q_r = 0. \tag{10.8.7}$$

Such a representation of 0 as intersection of primary submodules is called a *primary decomposition* in M. To show that such decompositions exist let us define a submodule N of M to be *meet-reducible* if $N = M$ or $N = N_1 \cap N_2$, where $N_i \supset N$; otherwise N is *meet-irreducible*. By the decomposition lemma (Lemma 3.2.7), any submodule of a Noetherian module can be written as a finite intersection of meet-irreducible submodules. It remains to observe that meet-irreducible submodules are primary:

Lemma 10.8.7. *Let R be a Noetherian ring and M be any R-module. Then any meet-irreducible submodule of M is primary.*

Proof. Let N be a submodule of M; we assume that N is not primary and show it to be meet-reducible. If $N \neq M$, then $\mathrm{Ass}(M/N)$ contains at least two primes $\mathfrak{p}_1, \mathfrak{p}_2$ say; hence M has submodules $N_i \cong R/\mathfrak{p}_i$. Every element of $N_i \backslash N$ has annihilator $\mathfrak{p}_i$, hence $N_1 \cap N_2 = N$, and this shows N to be meet-reducible, as claimed. ∎

Thus we find that every finitely generated module over a Noetherian ring has a primary decomposition (10.8.7). Of course the decomposition will not usually be unique, e.g. some terms in (10.8.6) might be redundant. To obtain some uniqueness we shall modify (10.8.6) as follows. Firstly, using Lemma 10.8.6, we collect together primary submodules for the same prime ideal, so that if Q_i is $\mathfrak{p}_i$-primary, all the $\mathfrak{p}_i$ are distinct. Secondly we may suppose that

$$\text{no } Q_i \text{ contains } \bigcap_{j \neq i} Q_j, \tag{10.8.8}$$

for otherwise we could omit Q_i from (10.8.7) without affecting the result. If the family $\{Q_1, \ldots, Q_r\}$ satisfies these two conditions, the primary decomposition is called *irredundant*. For such decompositions we have the following first uniqueness theorem:

Theorem 10.8.8. *Let R be a Noetherian ring and M be a finitely generated R-module. Then there is a primary decomposition in M:*

$$0 = Q_1 \cap \ldots \cap Q_r, \quad \textit{where } Q_i \textit{ is } \mathfrak{p}_i\textit{-primary}, \tag{10.8.9}$$

and (10.8.9) *may be chosen irredundant, i.e. so that the $\mathfrak{p}_i$ are distinct and* (10.8.8) *holds. When this is so, the $\mathfrak{p}_i$ are uniquely determined as the members of* $\mathrm{Ass}(M)$.

Proof. We have seen that decompositions (10.8.9) always exist, and that we can modify such a decomposition so as to become irredundant. As a result we have the embedding (10.8.6) and it follows that $\mathrm{Ass}(M) \subseteq \{\mathfrak{p}_1, \ldots, \mathfrak{p}_r\}$. For $r \leq 1$ there is nothing to prove. If $r > 1$, put $N = Q_2 \cap \ldots \cap Q_r$; then $N \neq 0$ by irredundancy. We have $N = N/(Q_1 \cap N) \cong (Q_1 + N)/Q_1 \subseteq M/Q_1$, hence $\mathrm{Ass}(N) \subseteq \mathrm{Ass}(M/Q_1) = \{\mathfrak{p}_1\}$, so $\mathfrak{p}_1$ is associated with N and hence with M; similarly for the other $\mathfrak{p}_i$. ∎

For an ideal $\mathfrak{a}$ in R we can, by taking $M = R/\mathfrak{a}$ in Theorem 10.8.8, obtain the usual primary decomposition for $\mathfrak{a}$, generalizing (10.8.2):

$$\mathfrak{a} = \mathfrak{q}_1 \cap \ldots \cap \mathfrak{q}_r, \quad \text{where } \mathfrak{q}_i \text{ is } \mathfrak{p}_i\text{-primary.} \tag{10.8.10}$$

From (10.8.10) it follows that

$$\sqrt{\mathfrak{a}} = \mathfrak{p}_1 \cap \ldots \cap \mathfrak{p}_r. \tag{10.8.11}$$

We note that if in (10.8.10) $\mathfrak{p}_1 \subseteq \mathfrak{p}_i$ for all $i > 1$, then $\sqrt{\mathfrak{a}}$ is prime, but $\mathfrak{a}$ need not be primary.

From (10.8.11) we find in particular, taking $\mathfrak{a} = 0$ and remembering the definition of the $\mathfrak{p}_i$ as annihilators:

Corollary 10.8.9. *In a Noetherian ring R the nilradical $\mathfrak{N}$ can be written as intersection of finitely many prime ideals: $\mathfrak{N} = \bigcap_1^r \mathfrak{p}_i$, and then $\bigcup_1^r \mathfrak{p}_i$ is the set of all zerodivisors in R.* ∎

Theorem 10.8.8 does not tell us whether the primary components of $\mathfrak{a}$ are unique, and in general the answer is 'no'. Thus in the polynomial ring $k[x, y]$ we have

$$(x^2, xy) = (x) \cap (x, y)^2 = (x) \cap (x^2, y).$$

The primary component for (x) is the same in each decomposition, but not that for (x, y). Geometrically this ideal defines the y-axis and the origin with an 'infinitesimal arrow', represented by a nilpotent in the local ring. Without trying to make this idea more precise, we note that it was the non-maximal geometrical component which failed to be unique. To obtain a more precise uniqueness result, we shall need a lemma which is also useful elsewhere, the prime avoidance lemma:

Lemma 10.8.10. *Let R be any (commutative) ring.*

(i) *If $\mathfrak{p}$ is a prime ideal in R and*

$$\mathfrak{p} \supseteq \bigcap_{i=1}^{n} \mathfrak{a}_i, \tag{10.8.12}$$

then $\mathfrak{p} \supseteq \mathfrak{a}_i$ for some i. Moreover, if $\mathfrak{p} = \cap \mathfrak{a}_i$, then $\mathfrak{p} = \mathfrak{a}_i$ for some i. Thus every prime ideal in R is meet-irreducible.

(ii) *If $\mathfrak{a}$ is an ideal in R and*

$$\mathfrak{a} \subseteq \bigcup_{i=1}^{n} \mathfrak{p}_i, \tag{10.8.13}$$

where each $\mathfrak{p}_i$ is a prime ideal, then $\mathfrak{a} \subseteq \mathfrak{p}_i$ for some i.

Note that in (ii) $\cup \mathfrak{p}_i$ is not generally an ideal, and the result is false with $\sum \mathfrak{p}_i$ in place of $\cup \mathfrak{p}_i$.

Proof. (i) Suppose that $\mathfrak{p}$ contains no $\mathfrak{a}_i$ and choose $x_i \in \mathfrak{a}_i \backslash \mathfrak{p}$. Then $x = x_1 \ldots x_n \in \cap \mathfrak{a}_i$ but $x \notin \mathfrak{p}$ because $\mathfrak{p}$ is prime; this contradicts (10.8.12), so $\mathfrak{p} \supseteq \mathfrak{a}_i$ for some i. If

now $\mathfrak{p} = \cap \mathfrak{a}_i$, then $\mathfrak{p} \subseteq \mathfrak{a}_i$ for all i, and by what has already been proved, equality must hold for some i.

(ii) We use induction on n, the case $n = 1$ being trivial. Suppose that no $\mathfrak{p}_i$ contains $\mathfrak{a}$; then by induction, no $\cup_{j \neq i} \mathfrak{p}_j$ contains $\mathfrak{a}$ (for $i = 1, \ldots, n$); hence there exists $x_i \in \mathfrak{a}$, $x_i \notin \mathfrak{p}_j$ for all $j \neq i$, and so $x_i \in \mathfrak{p}_i$ by (10.8.13). Consider $y = x_1 + x_2 \ldots x_n$; by construction $x_i \in \mathfrak{a}$, so $y \in \mathfrak{a}$. But $x_1 \in \mathfrak{p}_1, x_2 \ldots x_n \notin \mathfrak{p}_1$, so $y \notin \mathfrak{p}_1$ and for $i > 1$, $x_1 \notin \mathfrak{p}_i$ while $x_2 \ldots x_n \in \mathfrak{p}_i$, so again $y \notin \mathfrak{p}_i$. Thus y lies in $\mathfrak{a}$ but in no $\mathfrak{p}_i$, a contradiction, and the result follows. ■

For any ideal $\mathfrak{a}$ of R, the minimal terms in $\mathrm{Ass}(R/\mathfrak{a})$ are called the *minimal* or *isolated components* of $\mathfrak{a}$, while the other terms are said to be *embedded.* Thus in the above example, (x, y) is an embedded prime ideal and (x) is isolated. More generally, a subset Σ of $\mathrm{Ass}(R/\mathfrak{a})$ is called *isolated* if any $\mathfrak{p} \in \mathrm{Ass}(R/\mathfrak{a})$ such that $\mathfrak{p} \subseteq \mathfrak{p}' \in \Sigma$ satisfies $\mathfrak{p} \in \Sigma$; in other words, Σ is a lower segment in $\mathrm{Ass}(R/\mathfrak{a})$.

We remark that the isolated primes of $\mathfrak{a}$ are precisely the primes that are minimal among all the primes containing $\mathfrak{a}$. For if $\mathfrak{p}$ is any prime containing $\mathfrak{a}$, then $\mathfrak{p} \supseteq \mathfrak{a} = \cap \mathfrak{q}_i$, $\mathfrak{p} = \sqrt{\mathfrak{p}} \supseteq \sqrt{\cap\, \mathfrak{q}_i} = \cap \mathfrak{p}_i$, hence $\mathfrak{p} \supseteq \mathfrak{p}_i$ for some i, by Lemma 10.8.10.

We can now state the second uniqueness theorem.

Theorem 10.8.11. *Let $\mathfrak{a} = \mathfrak{q}_1 \cap \ldots \cap \mathfrak{q}_r$ be an irredundant primary decomposition of an ideal $\mathfrak{a}$ in a ring R, and let $\sqrt{\mathfrak{q}_i} = \mathfrak{p}_i$ be the associated prime ideal. If $\{\mathfrak{p}_{i_1}, \ldots \mathfrak{p}_{i_t}\}$ is an isolated subset, then $\mathfrak{q}_{i_1} \cap \ldots \cap \mathfrak{q}_{i_t}$ is independent of the choice of the decomposition for $\mathfrak{a}$. In particular, for any minimal $\mathfrak{p}_i$, $\mathfrak{q}_i$ is uniquely determined.*

Proof. To simplify the notation, let us number the $\mathfrak{p}_i$ so that the isolated subset is $\{\mathfrak{p}_1, \ldots, \mathfrak{p}_t\}$. The set $S = R \backslash \mathfrak{p}_1 \cup \ldots \cup \mathfrak{p}_t$ is multiplicative and by hypothesis, if $\mathfrak{p}_i \cap S = \varnothing$, then $\mathfrak{p}_i \subseteq \mathfrak{p}_1 \cup \ldots \cup \mathfrak{p}_t$, hence $i \leq t$. Thus S meets $\mathfrak{p}_{t+1}, \ldots, \mathfrak{p}_r$ and no others. Now form R_S; since $\mathfrak{p}_i \cap S = \varnothing$ for $i \leq t$, we have $\mathfrak{q}_i^{ec} = \mathfrak{q}_i$ (recall Proposition 10.3.2), while for $i > t$, $\mathfrak{p}_i \cap S \neq \varnothing$. Take $a \in \mathfrak{p}_i \cap S$; then $a^n \in \mathfrak{q}_i$ for some n, but a^n is a unit in R_S, hence $(\mathfrak{q}_i)_S = (1)$ and it follows that

$$\mathfrak{a}_S = (\mathfrak{q}_1)_S \cap \ldots \cap (\mathfrak{q}_t)_S, \quad \mathfrak{a}^{ec} = \mathfrak{q}_1 \cap \ldots \cap \mathfrak{q}_t.$$

This provides a description which is independent of the decomposition. ■

If $\mathfrak{p}$ is a prime ideal in R and $S = R \backslash \mathfrak{p}$, then $(\mathfrak{p}^r)_S \cap R = \mathfrak{p}^{(r)}$ is sometimes called the *symbolic r-th power* of $\mathfrak{p}$; we note that it is the $\mathfrak{p}$-primary component of $\mathfrak{p}^r$.

Exercises

1. Find all possible primary decompositions of (x^2, xy) in $k[x, y]$. Do the same in $\mathbf{Z}[x, y]$.
2. Show that the members of $\mathrm{Ass}(M)$ can also be defined as the ideals $\mathfrak{a}$ for which there is a submodule L of M such that $\mathfrak{a} = \mathrm{Ann}(L')$ for every non-zero submodule L' of L.

3. Let $R = k[x_1, x_2, \dots]$ and take M to be the R-module generated by elements $u_1, u_2, \dots$ with defining relations $u_r x_s = 0$ $(s \leq r)$. Show that $\mathrm{Ass}(M)$ has no maximal element.
4. Let A be a local ring with maximal ideal $\mathfrak{m}$. For a given A-module M show that if $\mathfrak{m}$ consists of zerodivisors on M, then $\mathfrak{m} \in \mathrm{Ass}(M)$.
5. Let R be a ring and S be a multiplicative subset. Show that for any non-zero R-module M, any maximal ideal of the form $\mathrm{Ann}(u)$ $(0 \neq u \in M)$ and disjoint from S is prime. By taking a Noetherian localization of R show that such maximal ideals exist, and deduce that $\mathrm{Ass}(M) \neq \varnothing$.
6. In a ring R let $\mathfrak{q}_i$ be $\mathfrak{p}_i$-primary $(i = 1, 2)$, where $\mathfrak{p}_1 \supset \mathfrak{p}_2$ but neither of $\mathfrak{q}_1, \mathfrak{q}_2$ contains the other. Show that $\mathfrak{a} = \mathfrak{q}_1 \cap \mathfrak{q}_2$ is not primary although $ab \equiv 0 \pmod{\mathfrak{a}}$ implies $a^r \equiv 0$ or $b^s \equiv 0 \pmod{\mathfrak{a}}$ for some r, s. (An ideal with this property is sometimes called *quasiprimary*.)
7. Show that if $\sqrt{\mathfrak{a}}$ is maximal, then $\mathfrak{a}$ is primary.
8. Let $\mathfrak{a}$ be an ideal and $\mathfrak{p}_1, \dots, \mathfrak{p}_r$ be prime ideals in R. Given $c \in R$, if $cR + \mathfrak{a} \not\subset \cup \mathfrak{p}_i$, show that $c + a \notin \cup \mathfrak{p}_i$ for some $a \in \mathfrak{a}$.
9. Consider the mapping $f : k[x, y, z] \to k[t]$ given by $x \mapsto t^3$, $y \mapsto t^4$, $z \mapsto t^5$. Show that its kernel is $\mathfrak{p} = (y^2 - xz, yz - x^3, z^2 - x^2 y)$. Verify that $\mathfrak{p}$ is prime and that $\mathfrak{p}^2$ is quasiprimary but not primary.
10. Show that if in a module M, Q_1, Q_2 are primary submodules of which neither contains the other and $Q_1 \cap Q_2$ is $\mathfrak{p}$-primary, then Q_1, Q_2 are both $\mathfrak{p}$-primary.
11. Show that if $N = A \cap B$ is an irredundant primary decomposition in a module and $B \subset B_1$, then B is meet-reducible.
12. Show that in Lemma 10.8.10(ii) the conclusion still holds if all but at most two of the $\mathfrak{p}_i$ are prime.
13. Show that if an ideal $\mathfrak{a}$ of a ring R has no embedded components, then $\mathfrak{p} \in \mathrm{Ass}(R/\mathfrak{a})$ iff $\mathfrak{a} \subseteq \mathfrak{p}$, $\mathfrak{a} \subset \mathfrak{a} : \mathfrak{p}$.
14. Show that a ring has a unique prime ideal iff it is *completely primary*, i.e. every non-unit is nilpotent.
15. Let R be a Noetherian ring and M be a finitely generated R-module. Show that if an ideal $\mathfrak{a}$ consists entirely of zerodivisors on M, then there exists $u \in M$, $u \neq 0$, such that $u\mathfrak{a} = 0$.
16. Let R be the function ring of the quadric $xy = z^2$. Show that $\mathfrak{p} = (x, z)$ is a prime ideal, but $\mathfrak{p}^2$ is not primary.
17. In any ring R, if $\mathfrak{q}$ is a $\mathfrak{p}$-primary ideal, show that for any ideal $\mathfrak{a}$, (i) if $\mathfrak{a} \subset \mathfrak{q}$, then $\mathfrak{q} : \mathfrak{a} = (1)$; (ii) if $\mathfrak{a} \not\subset \mathfrak{p}$, then $\mathfrak{q} : \mathfrak{a} = \mathfrak{q}$; (iii) if $\mathfrak{a} \subseteq \mathfrak{p}$ but $\mathfrak{a} \not\subset \mathfrak{q}$, then $\mathfrak{q} \subset \mathfrak{q} : \mathfrak{a} \subset (1)$. Deduce that if $\mathfrak{a}$ has an irredundant primary decomposition (10.8.10), then for any ideal $\mathfrak{b}$, $\mathfrak{a} : \mathfrak{b} = \mathfrak{a}$ iff $\mathfrak{b} \not\subset \mathfrak{p}_i$ for all i.
18. Let R be a ring and $\mathfrak{a}$ be an ideal of R. Show that if $\mathfrak{m}^r \subseteq \mathfrak{a} \subseteq \mathfrak{m}$ for some maximal ideal $\mathfrak{m}$ and some $r \geq 1$, then $\mathfrak{a}$ is $\mathfrak{m}$-primary.

10.9 Dimension

The dimension of an algebraic variety may be defined as the maximum length of a chain of subvarieties. Since varieties correspond to prime ideals in the coordinate

ring (Proposition 10.7.4), this suggests defining the dimension of a commutative ring in terms of chains of prime ideals.

In any commutative ring R, consider a chain of prime ideals

$$\mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \ldots \subset \mathfrak{p}_r. \tag{10.9.1}$$

This chain is said to have *length* r; note that r is the number of links in the chain. Now the *Krull dimension*, or simply the *dimension*, of R, written dim R, is defined as the supremum of the lengths of chains of prime ideals in R (possibly infinite). For example, any field has the dimension 0, dim $\mathbf{Z} = 1$, more generally, any PID has dimension 1, and as we shall see in Section 10.10, dim $k[x_1, \ldots, x_n] = n$. The dimension of an R-module M may be defined as the supremum of the lengths of chains of prime ideals in Supp(M).

Our first result is the observation that an integral extension does not raise the dimension; this will follow from the going-up theorem.

Let R be a ring, R' be a ring containing R as a subring and such that R' is integral over R. We shall say more briefly: R' is an *integral extension* of R; here neither R nor R' need be an integral domain. We first examine the behaviour of integral extensions under formation of quotient rings and rings of fractions.

Proposition 10.9.1. *Let R' be an integral extension of R. Then* (i) *if $\mathfrak{A}$ is an ideal in R' and $\mathfrak{a} = \mathfrak{A} \cap R$, then $R'/\mathfrak{A}$ is an integral extension of $R/\mathfrak{a}$;* (ii) *if S is a multiplicative subset of R, then R'_S is an integral extension of R_S.*

Proof. (i) The inclusion $R \to R'$ induces a homomorphism $R/\mathfrak{a} \to R'/\mathfrak{A}$, given by $a \mapsto \bar{a}$, with kernel $(\mathfrak{A} \cap R)/\mathfrak{a} = 0$, hence an injection. Let $\xi \in R'/\mathfrak{A}$, say $\xi = \bar{x}$, where $x \in R'$ satisfies the monic equation $x^n + a_1 x^{n-1} + \ldots + a_n = 0$ $(a_i \in R)$; when reduced mod $\mathfrak{A}$ this becomes $\xi^n + \bar{a}_1 \xi^{n-1} + \ldots + \bar{a}_n = 0$, a monic equation for ξ, which is therefore integral over $R/\mathfrak{a}$.

(ii) It is clear that the induced mapping $R_S \to R'_S$ is injective, for any element of R_S has the form a/s, $a \in R$, $s \in S$, and $a/s = 0$ iff $at = 0$ for some $t \in S$. Now consider an element of R'_S, say x/s, where $x \in R'$, $s \in S$. If x satisfies the equation $x^n + a_1 x^{n-1} + \ldots + a_n = 0$, then $x/s = y$ satisfies $y^n + a_1/s . y^{n-1} + \ldots + a_n/s^n = 0$, and $a_i/s^i \in R_S$, so x/s is integral over R_S. ∎

The key lemma for what follows is the next result:

Lemma 10.9.2. *Let R' be an integral extension of R and assume that R' is an integral domain. Then R' is a field if and only if R is a field.*

Proof. Assume that R is a field and let $y \in R'$, $y \neq 0$. Then y satisfies an equation $y^n + a_1 y^{n-1} + \ldots + a_n = 0$, where $a_i \in R$, and since R' is an integral domain, we may take $a_n \neq 0$. It follows that $a_n^{-1} \in R$, and so $y^{-1} = -a_n^{-1}(y^{n-1} + a_1 y^{n-2} + \ldots + a_{n-1}) \in R'$. Conversely, if R' is a field, take $x \in R$, $x \neq 0$. Then $x^{-1} \in R'$ and so we have an equation $x^{-m} + c_1 x^{1-m} + \ldots + c_m = 0$, where $c_i \in R$. Therefore $x^{-1} = -(c_1 + c_2 x + \ldots + c_m x^{m-1}) \in R$. ∎

For arbitrary integral extensions this yields

Corollary 10.9.3. *Let R' be an integral extension of R, $\mathfrak{P}$ be a prime ideal in R' and $\mathfrak{p} = \mathfrak{P} \cap R$. Then $\mathfrak{P}$ is maximal in R' if and only if $\mathfrak{p}$ is maximal in R.*

Proof. Since $R'/\mathfrak{P}$ is an integral domain, integral over $R/\mathfrak{p}$ by Proposition 10.9.1, we can apply the lemma to reach the conclusion. ∎

Given $R \subset R'$ and ideals $\mathfrak{p}$ in R, $\mathfrak{P}$ in R', we shall say that $\mathfrak{P}$ *lies over* $\mathfrak{p}$ if $\mathfrak{P} \cap R = \mathfrak{p}$.

Corollary 10.9.4. *Let R' be an integral extension of R and let $\mathfrak{p}$ be a prime ideal in R. If $\mathfrak{P} \subseteq \mathfrak{P}'$ are two prime ideals in R' both lying over $\mathfrak{p}$, then $\mathfrak{P}' = \mathfrak{P}$; thus the prime ideals lying over a given prime ideal of R are pairwise incomparable.*

Proof. Put $S = R\backslash\mathfrak{p}$; then R'_S is integral over R_S $(= R_\mathfrak{p})$ and the latter is a local ring with maximal ideal $\mathfrak{p}_S$. If $\mathfrak{m}$, $\mathfrak{m}'$ are the extensions of $\mathfrak{P}$, $\mathfrak{P}'$ to R'_S, then both contract to $\mathfrak{p}_S$ in R_S, so by Corollary 10.9.3 both are maximal. But $\mathfrak{m} \subseteq \mathfrak{m}'$, hence they are equal and $\mathfrak{P} = \mathfrak{m} \cap R = \mathfrak{m}' \cap R = \mathfrak{P}'$. ∎

The essential idea of this proof already occurred in the proof of Theorem 10.5.5 (cf. the proof of Lemma 10.9.2).

The existence of prime ideals lying over a given prime ideal is assured by

Lemma 10.9.5. *Let R' be an integral extension of R and let $\mathfrak{p}$ be a prime ideal in R. Then there exists a prime ideal $\mathfrak{P}$ in R' lying over $\mathfrak{p}$.*

Proof. Put $S = R\backslash\mathfrak{p}$ and form R_S, R'_S with canonical mappings $\lambda : R \to R_S$, $\lambda' : R' \to R'_S$. If $\mathfrak{m}$ is any maximal ideal in R'_S, then $\mathfrak{m} \cap R_S$ is maximal in R_S, by Corollary 10.9.3, so $\mathfrak{m} \cap R_S = \mathfrak{p}_S$. Now the inverse image of $\mathfrak{m}$ in R' is a prime ideal: $\mathfrak{P} = \mathfrak{m}\lambda'^{-1}$, and by the commutativity of the diagram below, we have $\mathfrak{P} \cap R = \mathfrak{p}$, so $\mathfrak{P}$ is the desired prime ideal.

$$\begin{array}{ccc} R' & \xrightarrow{\lambda'} & R'_S \\ \uparrow & & \uparrow \\ R & \xrightarrow{\lambda} & R_S \end{array}$$

∎

Theorem 10.9.6 (Going-up theorem). *Let R' be an integral extension of R. Given a chain of prime ideals in R:*

$$\mathfrak{p}_1 \subset \mathfrak{p}_2 \subset \ldots \subset \mathfrak{p}_n, \tag{10.9.2}$$

and a chain

$$\mathfrak{P}_1 \subset \mathfrak{P}_2 \subset \ldots \subset \mathfrak{P}_m \ (m \leq n) \tag{10.9.3}$$

in R', where $\mathfrak{P}_i \cap R = \mathfrak{p}_i (i = 1, \ldots, m)$, we can find prime ideals $\mathfrak{P}_{m+1}, \ldots, \mathfrak{P}_n$ lying over $\mathfrak{p}_{m+1}, \ldots, \mathfrak{p}_n$ respectively, to extend the chain (10.9.3).

Proof. When $m = 0$, the second chain is absent and we can find $\mathfrak{P}_1$ lying over $\mathfrak{p}_1$ by Lemma 10.9.5. Now let $m \geq 1$, put $\overline{R} = R/\mathfrak{p}_m$, $\overline{R}' = R'/\mathfrak{P}_m$, so that $\overline{R}'$ is an integral domain integral over $\overline{R}$, and $\overline{\mathfrak{p}}_{m+1} = \mathfrak{p}_{m+1}/\mathfrak{p}_m$ is a prime ideal of $\overline{R}$. Then by Lemma 10.9.5 there is a prime ideal $\overline{\mathfrak{P}}_{m+1}$ in $\overline{R}'$ lying over $\overline{\mathfrak{p}}_{m+1}$. The inverse image of $\overline{\mathfrak{P}}_{m+1}$ in R' is a prime ideal $\mathfrak{P}_{m+1}$ containing $\mathfrak{P}_m$ such that $\mathfrak{P}_{m+1} \cap R = \mathfrak{p}_{m+1}$. Now the result follows by induction on m. ∎

With the help of this result it is easy to show that integral extensions preserve the dimension.

Corollary 10.9.7. *If R' is an integral extension of R, then* $\dim R' = \dim R$.

Proof. Let $\mathfrak{P}_0 \subset \mathfrak{P}_1 \subset \ldots \subset \mathfrak{P}_r$ be a prime ideal chain of length r in R' and put $\mathfrak{p}_i = \mathfrak{P}_i \cap R$; then by Corollary 10.9.4, the $\mathfrak{p}_i$ are distinct and we get a chain

$$\mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \ldots \subset \mathfrak{p}_r \tag{10.9.4}$$

in R; this shows that $\dim R' \leq \dim R$. Conversely, over any chain (10.9.4) in R we can by Theorem 10.9.6 find a chain in R' and so conclude that $\dim R' = \dim R$. ∎

For a prime ideal $\mathfrak{p}$ in a ring the *height* of $\mathfrak{p}$ is defined as the supremum of the lengths of chains of prime ideals below $\mathfrak{p}$.

A natural question at this point concerns the relation between $\dim R$ and $\dim R[x]$, where R is a ring and x is an indeterminate. Given a prime ideal chain of length r in R, say (10.9.4), let $\mathfrak{P}_i$ be the ideal of $R[x]$ generated by $\mathfrak{p}_i$; then $R[x]/\mathfrak{P}_i \cong (R/\mathfrak{p}_i)[x]$; since the latter is an integral domain, $\mathfrak{P}_i$ is a prime ideal. Likewise $\mathfrak{P}_i + (x)$ is prime, because $R[x]/(\mathfrak{P}_i + (x)) \cong R/\mathfrak{p}_i$. Thus

$$\mathfrak{P}_0 \subset \mathfrak{P}_1 \subset \ldots \subset \mathfrak{P}_r \subset \mathfrak{P}_r + (x) \tag{10.9.5}$$

is a prime ideal chain of length $r + 1$ in $R[x]$, for the inclusions in (10.9.5) are clearly proper. It follows that

$$\dim R[x] \geq \dim R + 1. \tag{10.9.6}$$

For Noetherian rings equality holds in (10.9.6) (see Matsumura (1985); Nagata (1962)), but not in general (Seidenberg).

It is easy to determine the zero-dimensional rings completely. For this task we shall need

Lemma 10.9.8. *Let R be a ring in which 0 can be written as a product of maximal ideals. Then any R-module is Artinian if and only if it is Noetherian.*

Proof. Assume that $\mathfrak{m}_1 \ldots \mathfrak{m}_r = 0$, where the $\mathfrak{m}_i$ are maximal ideals of R, not necessarily distinct, and for any R-module M consider the chain

$$M \supseteq M\mathfrak{m}_1 \supseteq M\mathfrak{m}_1\mathfrak{m}_2 \supseteq \ldots \supseteq M\mathfrak{m}_1 \ldots \mathfrak{m}_r = 0. \tag{10.9.7}$$

Write $k_i = R/\mathfrak{m}_i$; since $\mathfrak{m}_i$ is maximal in R, k_i is a field, and the quotient $M/M\mathfrak{m}_1$ can be considered as a k_1-module, i.e. a vector space over k_1, for two elements of R have

the same effect on $M/M\mathfrak{m}_1$ if they are congruent mod $\mathfrak{m}_1$. Such a vector space is finite-dimensional iff it satisfies the maximum condition on subspaces, or equivalently the minimum condition, so we have a composition series from M to $M\mathfrak{m}_1$ iff the modules between M and $M\mathfrak{m}_1$ satisfy the maximum or equivalently, the minimum condition. The other links in the chain (10.9.7) can be treated similarly, hence M has a composition series (of R-modules) iff it satisfies the maximum or the minimum condition on submodules. ■

Theorem 10.9.9. *Let R be any (commutative) ring. Then R is Artinian if and only if R is Noetherian and zero-dimensional.*

Proof. Suppose that R is Artinian. We first note that an Artinian domain is a field, for if $a \in R$, $a \neq 0$, then $(a) \supseteq (a^2) \supseteq \ldots$ is a descending chain, which must terminate, say $(a^r) = (a^{r+1})$, hence $a^r = a^{r+1}b$, and so $ab = 1$. It follows that in an Artinian ring every prime ideal is maximal, hence dim $R = 0$. Further, R is Noetherian, by Hopkins' theorem (Corollary 5.3.10).

Conversely, when R satisfies these conditions, take a primary decomposition of 0: $0 = \mathfrak{q}_1 \cap \ldots \cap \mathfrak{q}_r$. This shows that R has finitely many minimal prime ideals $\mathfrak{p}_i = \sqrt{\mathfrak{q}_i}$; but each $\mathfrak{p}_i$ is also maximal because dim $R = 0$. Hence $\mathfrak{N} = \mathfrak{p}_1 \cap \ldots \cap \mathfrak{p}_r$ is the nilradical, and since R is Noetherian, $\mathfrak{N}^k = 0$ for some k (see Section 10.4), so again $(\mathfrak{p}_1 \ldots \mathfrak{p}_r)^k = 0$ and by Lemma 10.9.8 we find that R is Artinian. ■

To study 1-dimensional rings we first look at the form taken by the primary decomposition. We recall from Section 4.5 that for a family of pairwise comaximal ideals, their intersection equals their product. Moreover, if $\sqrt{\mathfrak{a}}$, $\sqrt{\mathfrak{b}}$ are comaximal, then so are $\mathfrak{a}$, $\mathfrak{b}$. For if $\mathfrak{a} + \mathfrak{b}$ is proper, it is contained in some maximal ideal $\mathfrak{m}$, say, thus $\mathfrak{a} \subseteq \mathfrak{m}$ and so $\sqrt{\mathfrak{a}} \subseteq \sqrt{\mathfrak{m}} = \mathfrak{m}$, and similarly $\sqrt{\mathfrak{b}} \subseteq \mathfrak{m}$, but this contradicts the comaximality of $\sqrt{\mathfrak{a}}$ and $\sqrt{\mathfrak{b}}$.

Proposition 10.9.10. *In a Noetherian domain of dimension 1, every non-zero ideal can be written uniquely as a product of primary ideals with distinct radicals.*

Proof. Any non-zero ideal $\mathfrak{a}$ has a primary decomposition $\mathfrak{a} = \mathfrak{q}_1 \cap \ldots \cap \mathfrak{q}_r$ which may be taken irredundant, so the $\mathfrak{p}_i = \sqrt{\mathfrak{q}_i}$ are all distinct. The $\mathfrak{p}_i$ are non-zero, hence maximal, and so coprime. By the remark preceding Proposition 10.9.10 the $\mathfrak{q}_i$ are pairwise comaximal, hence $\mathfrak{a} = \mathfrak{q}_1 \ldots \mathfrak{q}_r$. If we also have $\mathfrak{a} = \mathfrak{q}'_1 \ldots \mathfrak{q}'_s$, where the $\sqrt{\mathfrak{q}'_i}$ are distinct, then this is $\mathfrak{q}'_1 \cap \ldots \cap \mathfrak{q}'_s$ and here the $\mathfrak{q}'_i$ are uniquely determined as the isolated components of $\mathfrak{a}$, hence $s = r$ and the $\mathfrak{q}'_i$ and $\mathfrak{q}_i$ coincide except for order. ■

To get a more precise description we want to be able to say that the $\mathfrak{q}_i$ are actually powers of prime ideals; for this to hold the ring must be integrally closed and we thus again reach the class of Dedekind domains (Theorem 10.5.4).

Theorem 10.9.11. *A Noetherian domain of dimension 1 is Dedekind if and only if every primary ideal is a prime power.*

Proof. Let R be a 1-dimensional Noetherian domain in which every primary ideal is a power of a prime ideal. By Proposition 10.9.10 every non-zero ideal is then uniquely expressible as a product of powers of distinct prime ideals, hence R is then a Dedekind domain, by Theorem 10.5.6. Conversely, let R be a Dedekind domain and $\mathfrak{a}$ be a non-zero ideal in R. By Theorem 10.5.6 we can write $\mathfrak{a} = \mathfrak{p}_1^{n_1} \dots \mathfrak{p}_r^{n_r}$, where the $\mathfrak{p}_i$ are maximal and $n_i > 0$. Clearly $\sqrt{\mathfrak{a}} = \mathfrak{p}_1 \dots \mathfrak{p}_r$, so $\mathfrak{a}$ is primary iff $r = 1$, and then $\mathfrak{a}$ is a power of a prime ideal. ∎

Exercises

1. Let S be an integral domain and R be a subring. If $\mathfrak{a} \neq 0$ is a finitely generated ideal in R and $c \in S$ is such that $c\mathfrak{a} \subseteq \mathfrak{a}$, show that c is integral over R.
2. Let R' be an integral domain and R be a subring and define the *conductor* of R in R' as $\mathfrak{f} = \{x \in R | xR' \subseteq R\}$. Show that $\mathfrak{f}$ is the largest ideal in R which is also an ideal in R'. Let $\mathfrak{f}$ be the conductor of a ring in its integral closure; show that for any multiplicative set S in R, if $\mathfrak{f} \cap S \neq \varnothing$, then R_S is integrally closed. Is the converse true?
3. Give an example of a local ring of infinite dimension.
4. Show that the ring $k[x_1, x_2, \dots | x_i x_j = 0 \text{ for } i \neq j]$ is zero-dimensional but not Artinian (so not Noetherian).
5. Show that an Artinian ring has only finitely many prime ideals. (Hint. Use Lemma 10.8.10.) Deduce that any Artinian (commutative) ring is a finite direct product of local rings.
6. Let K be a field and k be a subfield. Show that if K is finitely generated as k-algebra, then K is algebraic over k.
7. Let R be a Noetherian ring and $\mathfrak{a}$ be an ideal of R. Show that $R/\mathfrak{a}$ has finite length as R-module iff the associated prime ideals are all maximal. Show that the same holds for a finitely generated R-module M iff every ideal in $\mathrm{Supp}(M)$ is maximal.
8. Give a direct proof that $\dim k[x_1, \dots, x_n] \geq n$.

10.10 The Hilbert Nullstellensatz

We now return to the subject of Section 10.7; our object will be to establish a bijection between algebraic sets and radical ideals. We begin with a lemma on finitely generated extensions of fields.

Lemma 10.10.1 (Noether normalization lemma). *Let R be an integral domain, finitely generated as a ring over a field k. Then there exist $x_1, \dots, x_n \in R$, algebraically independent over k, such that R is integral over $k[x_1, \dots, x_n]$.*

Proof. Let R be generated by $y_1, \dots, y_n$ over k; if the y_i are algebraically independent, there is nothing to prove. Otherwise there will be a relation $f(y_1, \dots, y_n) = 0$. We write

$$z_i = y_i - y_1^{r^{i-1}} \quad (i = 2, \dots, m),$$

where the integer r is to be determined. Then our relation becomes

$$f(y_1, z_2 + y_1^r, \ldots, z_m + y_1^{r^{m-1}}) = 0. \tag{10.10.1}$$

Each monomial $\prod y_i^{b_i}$ in f gives rise to a term y_1^c, where

$$c = b_1 + rb_2 + \ldots + r^{m-1}b_m, \tag{10.10.2}$$

plus terms in z_i but of lower degree in y_1. We now choose r so that all sums (10.10.2) corresponding to the different terms in f are distinct; this can be done by avoiding the positive integers t satisfying the equations $\sum t^{i-1}(b_i - b_i') = 0$, corresponding to terms $\prod y_i^{b_i}, \prod y_i^{b_i'}$ in f. Then (10.10.1) is an equation for y_1 over $k[z_2, \ldots, z_m]$ which on multiplying by an element of k becomes monic, because by construction only a single term in f contributes to the leading term in y_1; hence y_1 is integral over $k[z_2, \ldots, z_m]$. Now the result follows by induction on m and the transitivity of integral dependence. ∎

Corollary 10.10.2. *For any field k and indeterminates $x_1, \ldots, x_n$,*

$$\dim k[x_1, \ldots, x_n] = n.$$

Proof. Write $R = k[x_1, \ldots, x_n]$; we know from (10.9.6) that $\dim R \geq n$. Now let

$$0 = \mathfrak{p}_0 \subset \mathfrak{p}_1 \subset \ldots \subset \mathfrak{p}_r \tag{10.10.3}$$

be a prime ideal chain in R and consider $\overline{R} = R/\mathfrak{p}_1$. If $a \mapsto \bar{a}$ is the residue class map, then $\overline{R}$ is generated by $\bar{x}_1, \ldots, \bar{x}_n$ over k and by Lemma 10.10.1, there exist $y_1, \ldots, y_m \in \overline{R}$, algebraically independent over k, such that $\overline{R}$ is integral over $S = k[y_1, \ldots, y_m]$. Moreover, the $\bar{x}_i$ are algebraically dependent, because $\mathfrak{p}_1 \neq 0$, hence $m < n$. Using induction on n, Corollary 10.9.7 and the chain (10.10.3), we obtain

$$n > m = \dim S = \dim \overline{R} \geq r - 1,$$

hence $r \leq n$ and it follows that $\dim R = n$. ∎

The Hilbert Nullstellensatz (literally: zero-point theorem or solution-set theorem) shows that a family of polynomials over an algebraically closed field has a common solution whenever the ideal they generate in the polynomial ring is proper.

Theorem 10.10.3 (Nullstellensatz, weak form). *Let k be an algebraically closed field. Then each maximal ideal $\mathfrak{m}$ in $R = k[x_1, \ldots, x_n]$ has the form*

$$\mathfrak{m} = (x_1 - \alpha_1, \ldots, x_n - \alpha_n), \quad \text{where } \alpha_i \in k. \tag{10.10.4}$$

In particular, $V(\mathfrak{a}) \neq \varnothing$ for any proper ideal $\mathfrak{a}$ of R.

Proof. The ideal $\mathfrak{m}$ given by (10.10.4) is maximal, because it is the kernel of the surjective mapping $x_i \mapsto \alpha_i$ from R to k.

Conversely, let $\mathfrak{m}$ be a maximal ideal in R. Then $K = R/\mathfrak{m}$ is a field, finitely generated as k-algebra; hence by Lemma 10.10.1, K is integral over $k[y_1, \ldots, y_r]$,

for suitable $y_1, \ldots, y_r$ algebraically independent over k. By Lemma 10.9.2, $k[y_1, \ldots, y_r]$ is a field, which means that $r = 0$, so K is integral over k, i.e. algebraic. But k is algebraically closed, hence $K = k$. If $x_i \mapsto \alpha_i \in k$ in the natural homomorphism from R to $R/\mathfrak{m} = k$, then $\mathfrak{m} \supseteq (x_1 - \alpha_1, \ldots, x_n - \alpha_n)$. Here the right-hand side is maximal, so equality holds and $\mathfrak{m}$ has the form (10.10.4).

Now let $\mathfrak{a}$ be any proper ideal in R. Then $\mathfrak{a} \subseteq \mathfrak{m}$ for some maximal ideal $\mathfrak{m}$, hence $V(\mathfrak{a}) \supseteq V(\mathfrak{m}) = (\alpha) \neq \varnothing$. ∎

It is now an easy matter to deduce that every radical ideal is of the form $I(S)$ for some subset S of k^n.

Theorem 10.10.4 (Nullstellensatz, full form). *Let $R = k[x_1, \ldots, x_n]$ be a polynomial ring over an algebraically closed field k. Then for any ideal $\mathfrak{a}$ in R,*

$$IV(\mathfrak{a}) = \sqrt{\mathfrak{a}}.$$

Proof. ('Rabinowitsch trick') Clearly $\sqrt{\mathfrak{a}} \subseteq IV(\mathfrak{a})$, for if $V(\mathfrak{a}) = S$ and $f \in \sqrt{\mathfrak{a}}$, then $f^r \in \mathfrak{a}$ say, and so $f(\alpha)^r = 0$ for all $\alpha \in S$, hence $f(\alpha) = 0$, i.e. $f \in I(S)$.

Conversely, assume that $f(\alpha) = 0$ for all $\alpha \in V(\mathfrak{a})$; we have to show that $f \in \sqrt{\mathfrak{a}}$. We form the ring $R' = R[x_0]$ in a further indeterminate x_0 and in it consider the ideal $\mathfrak{b} = \mathfrak{a}R' + (1 - x_0 f)$. If $\mathfrak{b}$ were proper, it would have a zero α, by Theorem 10.10.3. This means in particular that all elements of $\mathfrak{a}$ vanish at α, so $f(\alpha) = 0$; but also $1 - \alpha_0 f(\alpha) = 0$, a contradiction. Hence $\mathfrak{b}$ is improper, i.e. if $\mathfrak{a} = (g_1, \ldots, g_t)$, then

$$1 = \sum g_i h_i + (1 - x_0 f)h, \quad \text{where } h, h_i \in R'.$$

We replace x_0 by $1/f$; then the second term on the right reduces to 0, while each h_i is now a polynomial in $x_1, \ldots, x_n$ and $1/f$. If we clear the denominators, we obtain an equation

$$f^r = \sum g_i h_i', \quad \text{where } h_i' \in R.$$

Thus $f^r \in \mathfrak{a}$ and so $f \in \sqrt{\mathfrak{a}}$, as required. ∎

We thus have an order-reversing bijection between the closed sets on k^n and the radical ideals in $k[x_1, \ldots, x_n]$. Moreover, the closed sets satisfy the following relations:

(i) $\cap V(\mathfrak{a}_i) = V\left(\sum \mathfrak{a}_i\right)$,
(ii) $V(\mathfrak{a}_1) \cup V(\mathfrak{a}_2) = V(\mathfrak{a}_1 \cap \mathfrak{a}_2) = V(\mathfrak{a}_1\mathfrak{a}_2)$,
(iii) $V(0) = k^n$, $V(1) = \varnothing$.

These relations are easily verified and they show that the sets of the form $V(\mathfrak{a})$ may be regarded as the closed sets of a topology on k^n. This is called the *Zariski topology*. Although very different from the classical topologies (it is non-Hausdorff), it is useful in discussing relations in k^n. For example, a subset S of k^n is said to be *dense* in k^n if its closure (in the Zariski topology) is the whole of k^n; this means

that any polynomial vanishing on S vanishes on all of k^n. If f is a non-zero polynomial over an infinite field k, then the points where f does not vanish form a dense subset of k^n. This is just a restatement of the density property for algebraic inequalities (see Section 7.9).

Exercises

1. Show that the Nullstellensatz holds for any function ring of an algebraic set in the following form: Let $k[X]$ be the function ring of a subset X of k^n, where k is an algebraically closed field. If $f \in k[X]$ vanishes at all the points of X where $g_1, \ldots, g_r$ vanish, then $f \in \sqrt{(g_1, \ldots, g_r)}$.
2. Given an algebraic set X over an algebraically closed field k with function ring $k[X]$, let f, $g \in k[X]$ be such that f/g is defined at each point of X. Show that there exists $h \in k[X]$ such that $f = gh$. (Hint. Write $f/g = f_x/g_x$, where f_x, g_x are defined at $x \in X$ and use the fact that the g_x are elements of $k[X]$ which have no common zero.)
3. Show that the Nullstellensatz is false over any field that is not algebraically closed.
4. Let R be an integral domain generated by $x_1, \ldots, x_n$ over a ring K. Show that there exist polynomials $u_1, \ldots, u_r$ in the x_i with integer coefficients, which are algebraically independent, and a non-zero element $a \in K$ such that $R[a^{-1}]$ is integral over $K[a^{-1}, u_1, \ldots, u_r]$.
5. Show that if a ring R is finitely generated (over $\mathbf{Z}$ or a field), then its Jacobson radical is nilpotent.
6. Give a direct proof of the Nullstellensatz for primary ideals and deduce the general form by applying Theorem 10.9.9 and the fact that for any $\mathfrak{p}$-primary ideal $\mathfrak{q}$ over a Noetherian ring, $\mathfrak{p}^r \subseteq \mathfrak{q} \subseteq \mathfrak{p}$ for some r.

Further Exercises for Chapter 10

1. Let m, n be positive integers. When is it true that $(m) : (n) = (m/n)$? When is $(m) : (n) = (m)$? What other cases can arise?
2. Show that the set of all zerodivisors and 0 in any (commutative) ring is a union of prime ideals.
3. Let R be a reduced ring. In $R[x]$, if $fg = 0$, where $f = \sum a_i x^i$, $g = \sum b_j x^j$, show that $a_i b_j = 0$ for all i, j.
4. (M. Zafrullah) Show that the set of all polynomials in an indeterminate x over the real numbers, with rational constant term, is an atomic integral domain, but not a UFD.
5. Let us call a polynomial *full* if the ideal generated by its coefficients is the whole ring. Show that for any ring R and any $f, g \in R[x]$, fg is full iff f and g are both full.
6. Let S be the set of all full polynomials in $R[x]$. Show that S is a multiplicative set which contains no zerodivisors. (Hint. Observe that if $f \in R[x]$ is a zerodivisor, then it annihilates an element of R.) Deduce that if R is a local ring, then so is $R[x]_S$.

7. (M. Nagata) Let $q(x_1, \ldots, x_n)$ be a non-singular quadratic form in $n \geq 5$ variables over $\mathbf{C}$. Show that the ring $R = \mathbf{C}[x_1, \ldots, x_n]/(q)$ is a UFD. (Hint. Write q as $x_1x_2 + q_1(x_3, \ldots, x_n)$, where q_1 is non-singular and use Theorem 10.3.7.)
8. Show that an infinite UFD with a finite group of units has infinitely many primes.
9. Let R be a ring and S be a multiplicative set. Given R-modules M, N, where M is finitely presented, and any homomorphism $\varphi : M_S \to N_S$, find a homomorphism $f : M \to N$ and $s \in S$ such that $(x/1)\varphi = xf/s$ for all $x \in M$. Deduce that a finitely presented R-module M is projective if $M_{\mathfrak{m}}$ is free for all maximal ideals $\mathfrak{m}$.
10. Let R be a ring and M be a finitely generated R-module. Show that if $\mathfrak{a}$ is an ideal such that $M\mathfrak{a} = M$, then there exists $a \in \mathfrak{a}$ such that $1 + a$ annihilates M. (Hint. If $u = (u_1, \ldots, u_n)$ is a row of elements generating M, find a matrix C over $\mathfrak{a}$ such that $uC = u$ and hence obtain $M.\,det\,(I - C) = 0$.) Deduce a proof of Nakayama's lemma (Lemma 5.3.6) for commutative rings.
11. Let R be a Noetherian ring, $\mathfrak{a}$ be an ideal in R and M be a finitely generated R-module. Defining $M_\omega = \cap M\mathfrak{a}^n$, show that $M_\omega = \{x \in M | x(1 - a) = 0$ for some $a \in \mathfrak{a}\}$. Deduce that if $M_\omega \subseteq N \subseteq M$, then $M_\omega = N_\omega$. (Hint. Use Exercise 10.)
12. (Krull's intersection theorem) Let R be a Noetherian ring and $\mathfrak{a}$ be an ideal in R. Show that $\cap\mathfrak{a}^n = 0$ iff there is no zerodivisor $\equiv 1 \pmod{\mathfrak{a}}$. Deduce that (i) in a Noetherian domain every proper ideal $\mathfrak{a}$ satisfies $\cap\mathfrak{a}^n = 0$, and (ii) in a Noetherian local ring the maximal ideal $\mathfrak{m}$ satisfies $\cap\mathfrak{m}^n = 0$.
13. Let $\mathfrak{o}$ be a Dedekind domain with field of fractions K, let L be a finite separable extension of K and $\mathfrak{O}$ be the integral closure of $\mathfrak{o}$ in L. For any subset X of L define the *complementary set* X' as $X' = \{y \in L | T(xy) \in \mathfrak{o}$ for all $x \in X\}$, where T is the trace $T_{L/K}$. Show (i) X' is an $\mathfrak{o}$-submodule of L, (ii) $X \subseteq Y \Rightarrow X' \supseteq Y'$, (iii) $zX \subseteq X \Rightarrow zX' \subseteq X'$, (iv) $(zX)' = z^{-1}X'$. Deduce that the complementary set of a fractional ideal is again a fractional ideal, and if $\mathfrak{O} = \mathfrak{o}[\alpha]$ and f is the minimal polynomial of α over $\mathfrak{o}$, then $(\mathfrak{O}')^{-1} = (f'(\alpha))$, where f' is the usual derivative (here $(\mathfrak{O}')^{-1}$ is called the *different*). (Hint. If $[L : K] = n$, then $1, \alpha, \ldots, \alpha^{n-1}$ is an $\mathfrak{o}$-basis of $\mathfrak{O}$; if $f(x)/(x - \alpha) = b_0 + b_1x + \ldots + b_{n-1}x^{n-1}$, use the Lagrange interpolation formula to show that the $b_i/f'(\alpha)$ form a dual basis.)
14. Show that an ideal maximal among the non-finitely generated ideals in a ring is prime. Deduce I. S. Cohen's theorem, that if all prime ideals of a ring are finitely generated, then the ring is Noetherian. Similarly, prove that if all the prime ideals of a ring are principal, the ring is a principal ideal ring. (Hint. Recall the proof of Theorem 10.5.4 (d) $\Rightarrow$ (a).)
15. Prove the identity $(\mathfrak{a} + \mathfrak{b} + \mathfrak{c})(\mathfrak{bc} + \mathfrak{ca} + \mathfrak{ab}) = (\mathfrak{b} + \mathfrak{c})(\mathfrak{c} + \mathfrak{a})(\mathfrak{a} + \mathfrak{b})$ between ideals in any (commutative) ring. Deduce that R is a Dedekind domain if it is a Noetherian domain in which every 2-generator ideal $\neq 0$ is invertible (Dedekind).
16. Show that in a UFD an ideal is projective iff it is principal.

17. Let $\mathfrak{o}$ be a Dedekind domain distinct from its field of fractions K. Show that a non-zero $\mathfrak{o}$-submodule of K is a fractional ideal iff it is finitely generated. Give an example of a non-zero $\mathfrak{o}$-submodule of K which is not a fractional ideal.
18. Show that if $\mathfrak{o}$ is a Dedekind domain and $\mathfrak{a}$ is an integral ideal of $\mathfrak{o}$, then $\mathfrak{o}/\mathfrak{a}$ is Artinian.
19. Let $\mathfrak{o}$ be a ring and $f : \mathfrak{a} \to M$ be a homomorphism from an invertible ideal $\mathfrak{a}$ into an $\mathfrak{o}$-module M. Show that if $\sum a_i b_i = 1$, where $a_i \in \mathfrak{a}$, $b_i \in \mathfrak{a}^{-1}$, then the map $x \mapsto \sum a_i f . b_i x$ is a homomorphism from $\mathfrak{o}$ to M. Deduce that every divisible module over a Dedekind domain is injective.
20. Show that for a Dedekind domain $\mathfrak{o}$, the monoid of projectives under direct sums has cancellation and so has a group of fractions $K_0(\mathfrak{o})$. Show that $K_0(\mathfrak{o})$ becomes a ring under the tensor product and that $K_0(\mathfrak{o}) \cong \mathbf{Z} \oplus C(\mathfrak{o})$, where $C(\mathfrak{o})$ is the projective class group, i.e. the quotient of $K_0(\mathfrak{o})$ by the submonoid of free modules, with zero multiplication. (Hint. Use (10.6.8) to verify the isomorphism.)
21. Let R be the function ring over k of an algebraic set X. Show that the functions which vanish at a given point p of X form a prime ideal $\mathfrak{p}$ in R, and that $R_{\mathfrak{p}}$ (the 'local ring at p') may be interpreted as the ring of functions defined at p.
22. A mapping $f : X \to Y$ between algebraic sets X, Y is called *regular* if for any polynomial function p on Y, the map $x \mapsto p(xf)$ is a polynomial function on X. Verify that the correspondence which assigns to each algebraic set its function ring is a contravariant functor from algebraic sets and regular mappings to k-algebras and homomorphisms. Show that injective homomorphisms correspond to regular mappings with dense image (in the sense of the Zariski topology). Verify that the projection of the hyperbola $xy = 1$ on the x-axis has a dense image but is not surjective.
23. Let R be a ring and S be a multiplicative subset. For any R-module M show that $\mathfrak{p} \mapsto \mathfrak{p}_S$ defines an injection from the subset of $\mathrm{Ass}(M)$ consisting of primes disjoint from S to $\mathrm{Ass}(M_S)$. Show that moreover, this is a bijection when R is Noetherian.
24. Show that $\mathrm{Ass}(M) \subseteq \mathrm{Supp}(M)$ for any module M. Further show that if R is Noetherian, then $\mathrm{Ass}(M)$ and $\mathrm{Supp}(M)$ have the same minimal members.
25. Let $\mathrm{Ass}(M) = S' \cup S''$, where $S' \cap S'' = \varnothing$. Show that the family $\mathscr{F}$ of submodules M' such that $\mathrm{Ass}(M') \subseteq S'$ is inductive. Verify that for a maximal M' in $\mathscr{F}$, $\mathrm{Ass}(M/M') \subseteq S''$. Deduce that $\mathrm{Ass}(M') = S'$, $\mathrm{Ass}(M/M') = S''$.
26. Let $f_1, \ldots, f_m$ be polynomials in $x_1, \ldots x_n$ over an algebraically closed field K. Show that the equations $f_1 = \ldots = f_m = 0$ may have no solution in any extension field of K, but when they do, they already have a solution in K.
27. Show that if $\mathfrak{a}$ is any ideal in $k[x_1, \ldots, x_n]$, where k is an algebraically closed field, then $\mathfrak{a} = \cap\{\mathfrak{m} | \mathfrak{m}$ maximal ideal and $\mathfrak{m} \supseteq \mathfrak{a}\}$. (Geometrically this states that an algebraic variety is the union of its points.)

11 Infinite Field Extensions

Chapter 7 was almost entirely devoted to field extensions of finite degree and concentrated on Galois theory. However, even an introductory account should make some mention of infinite field extensions, and we shall discuss them in the present chapter, including transcendental extensions (Section 11.3) and infinite Galois theory (Section 11.8). The notion of algebraic dependence has similarities to linear dependence, which are described in abstract form in Section 11.1 and applied in Section 11.2. In addition there are concise accounts of topics that are useful in commutative ring theory and algebraic geometry, besides being of independent interest: separability (Sections 11.4 and 11.5), the interactions of two or more subfields (Sections 11.6 and 11.7), applications of Galois theory (Section 11.9) and abelian extensions of finite exponent (Section 11.10).

11.1 Abstract Dependence Relations

The notion of algebraic dependence is similar to that of linear dependence in a vector space, but there is not the same intuitive picture. For this reason it is advisable to begin by presenting an account of abstract dependence relations; this concept also occurs elsewhere, e.g. in some geometrical constructions, and in this section we shall derive their general properties.

Let S be a set; by a *dependence relation* on S we understand a rule which associates with each finite subset X of S certain elements of S, said to be *dependent* on X, and subject to the following conditions:

D.0 *If* $X = \{x_1, \ldots, x_n\}$, *then each* x_i *is dependent on* X.

D.1 *(Transitivity) If* z *is dependent on* Y *and each member of* Y *is dependent on* X, *then* z *is dependent on* X.

D.2 *(Exchange property) If* y *is dependent on* $\{x_1, \ldots, x_n\}$, *but not on* $\{x_2, \ldots, x_n\}$, *then* x_1 *is dependent on* $\{y, x_2, \ldots, x_n\}$.

We note that by **D.0** and **D.1**, if y is dependent on X' and $X' \subseteq X$, then y is dependent on X.

The notion of dependence can be extended to infinite sets by saying that y is dependent on an infinite subset X of S if there is a finite subset X' of X such that y is dependent on X'. Then it is easily seen that if the exchange property **D.2**

holds for finite subsets, it also holds for infinite subsets. Similarly **D.0** and **D.1** hold for arbitrary subsets.

As an example let us consider linear dependence in a vector space over a field or, more generally, in a module M over a ring R. Given a left R-module M and a subset X of M, an element y of M is said to be linearly dependent on X if

$$y = \sum c_i x_i \text{ for some } c_i \in R, \text{ where } x_i \in X. \tag{11.1.1}$$

The element y is said to satisfy a *non-trivial relation* with the elements of X if

$$by + \sum a_i x_i = 0 \text{ for some } a_i, b \in R, b \neq 0, \text{ where } x_i \in X. \tag{11.1.2}$$

It is clear that (11.1.1) implies (11.1.2) (with $b = 1$, $a_i = -c_i$). Conversely, if (11.1.2) holds with an invertible element b, then (11.1.1) follows by taking $c_i = -b^{-1}a_i$. This shows in particular that (11.1.1) and (11.1.2) are equivalent for vector spaces over a field.

Now **D.0–D.2** are easily verified for linear dependence as defined by (11.1.1) or (11.1.2) over a field. We prove **D.2** as an example. If y is dependent on $x_1, \ldots, x_n$ then we have a relation (11.1.2), and if y is not dependent on $x_2, \ldots, x_n$ then a_1 cannot be zero, so (11.1.2) shows x_1 to be dependent on $y, x_2, \ldots, x_n$.

For modules over a ring, (11.1.1) and (11.1.2) are no longer equivalent. The relation (11.1.1) always satisfies **D.0** and **D.1** but not always **D.2**. For example, in $\mathbf{Z}^2$ (as $\mathbf{Z}$-module) $(2,2)$ depends on $(1,0)$ and $(0,1)$ but not on $(0,1)$ alone, yet $(1,0)$ does not depend on $(2,2)$ and $(0,1)$, i.e. there is no relation $(1,0) = m(2,2) + n(0,1)$, $m, n \in \mathbf{Z}$. Similarly the relation (11.1.2) satisfies **D.0** and **D.2** but not necessarily **D.1** (see Exercise 2).

Given a dependence relation on a set S, we define for each $X \subseteq S$ the *span* of X as the set $\langle X \rangle$ of all elements of S dependent on X; we also say that $\langle X \rangle$ is *spanned* by X. From **D.0** we see that $\langle X \rangle \supseteq X$ and **D.1** shows that every element dependent on $\langle X \rangle$ itself belongs to $\langle X \rangle$.

A family $X = \{x_i | i \in I\}$ of elements of S is said to be *independent* if no x_i is dependent on $\{x_j | j \neq i\}$; otherwise X is called *dependent*. Clearly an independent family consists of distinct elements. Moreover, being independent is a property of finite character, for a family is dependent whenever this is true of some finite subfamily. An independent family which spans S is called a *basis* of S. This of course generalizes the usual definition of a basis for a vector space; as in that case, we can recognize bases in the following way:

Proposition 11.1.1. *Let S be a set with a dependence relation. Then for any subset X of S the following conditions are equivalent:*

(a) *X is a maximal independent subset of S,*
(b) *X is a minimal spanning set,*
(c) *X is a basis of S.*

Proof. (a) $\Leftrightarrow$ (c). Let X be a maximal independent subset of S; we have to show that every element of S is dependent on X. Given $y \in S$, if $y \in X$, then y is dependent on X

by **D.0**, and if $y \notin X$, then $X \cup \{y\}$ is dependent, by maximality, so some element is dependent on the rest, say $x \in X$ is dependent on $X' \cup \{y\}$, where $X' = X \backslash \{x\}$. Since X is independent, x is not dependent on X', hence by **D.2**, y is dependent on $X' \cup \{x\} = X$, and this shows that X spans S. Conversely, if X is a basis, it is independent, but every element not in X is dependent on X, so X is maximal independent.

(b) $\Leftrightarrow$ (c). Let X be a minimal spanning set; we have to show that X is independent. If X were dependent, say $x \in X$ is dependent on the rest of X, then we could omit x from X and still have a spanning set, by **D.1**. This contradicts the minimality, hence X is a basis. Conversely, if X is a basis, it is independent, so no element is dependent on the rest, and hence X is a minimal spanning set. ■

It is instructive to go through this proof in a particular case, and the reader is advised to do this, say for vector spaces over a skew field, taking first the case where the basis is finite and then the general case.

Our main aim will be to show that every set with a dependence relation has a basis, and that any two bases of a given set have the same number of elements. This will again show that a vector space, even infinite-dimensional, has a basis, and a uniquely defined dimension. We first obtain an important consequence of **D.2.**

Lemma 11.1.2 (Exchange lemma). *Let S be a set with a dependence relation. If X is an independent subset and Y is a spanning set of S, then we can complete X to a basis of S by elements from Y, i.e. there is a subset Y' of Y such that $X \cap Y' = \varnothing$ and $X \cup Y'$ is a basis of S. Moreover, if Y is finite, then X is finite and*

$$|X \cup Y'| \leq |Y|. \tag{11.1.3}$$

Proof. Consider the collection $\mathscr{I}$ of all independent subsets Z of S such that

$$X \subseteq Z \subseteq X \cup Y. \tag{11.1.4}$$

Since independence is a property of finite character, $\mathscr{I}$ is inductive, and so by Zorn's lemma $\mathscr{I}$ has a maximal member B, i.e. a maximal independent set B such that $X \subseteq B \subseteq X \cup Y$. This shows that B consists of X together with some elements of Y, thus $B = X \cup Y'$; by omitting from Y' elements that are also in X we ensure that $X \cap Y' = \varnothing$. Now every element of Y depends on B, by maximality, hence $\langle B \rangle \supseteq \langle Y \rangle = S$, therefore B is a basis, as claimed.

Assume now that Y is finite. We shall show that starting from any finite subset X' of X we can obtain a basis by adjoining at most $|Y| - |X'|$ elements of Y. This will show that $|X'| \leq |Y|$ and for $X' = X$ we reach the desired conclusion (11.1.3). We shall use induction on $|X'|$; when $|X'| = 0$, we can by the first part choose a subset of Y which is a basis. Now suppose that $|X'| = r > 0$ and that we have a basis $B_{r-1} = \{x_1, \ldots, x_{r-1}, y_r, \ldots, y_n\}$, where $x_i \in X', y_j \in Y$. With another element x_r of X' we form the spanning set $B_{r-1} \cup \{x_r\}$ which contains the independent subset $\{x_1, \ldots, x_r\}$. By the first part it contains a basis B_r including $x_1, \ldots, x_r$, but B_r cannot include all of $y_r, \ldots, y_n$ because the family $\{x_1, \ldots, x_r, y_r, \ldots, y_n\}$ is dependent (x_r is dependent on the rest); hence B_r has the required form and it contains at most n elements. By induction we obtain a basis of at most $|Y|$ elements,

consisting of X' together with some elements of Y. In particular, $|X'| \leq |Y|$ and since X' was any finite subset of X, it follows that X is finite and if the basis is $X \cup Y'$, then $|X \cup Y'| \leq |Y|$. ∎

By the last part, the number of elements in an independent family is bounded by the number of elements in a finite spanning set. In particular, given two bases B, C, if B is finite, then $|C| \leq |B|$ and so by symmetry, $|B| = |C|$. Thus we have

Corollary 11.1.3. *Let S be a set with a dependence relation. If S has a finite spanning set C, then any independent family has at most $|C|$ members and any two bases have the same number of elements.* ∎

Our next task is to show that any two bases, finite or not, have the same cardinal. For the case of infinite bases this is true in a rather wider context, which we shall now explain, as it is sometimes useful.

Suppose that instead of a dependence relation we have a relation satisfying only **D.0** and **D.1** (not necessarily **D.2**); let us call this a *spanning relation*. We can as before define the *span* of a subset X as the set $\langle X \rangle$ of all elements dependent on X. If we examine the proof of Proposition 11.1.1 we find that for a spanning relation (c) $\Leftrightarrow$ (b) $\Rightarrow$ (a), but a maximal independent set need no longer be a basis. In fact the term 'basis' in this context tends to mislead and we shall only speak of minimal spanning sets. The example of the rational numbers $\mathbf{Q}$ (as $\mathbf{Z}$-module) shows that minimal spanning sets need not exist. But when they do, we have the following relation, generalizing the situation for modules (see Proposition 4.6.4).

Proposition 11.1.4. *Let S be a set with a spanning relation and suppose that S has a minimal spanning set X. If X is infinite, of cardinal α, then any spanning set of S has at least α elements. In particular, S has no finite spanning set and any two minimal spanning sets of S have the same cardinal.*

Proof. This is similar to the module case and again uses the relation $\aleph_0\beta = \beta$ for infinite cardinals (Proposition 1.2.7). Let Y be any spanning set of S, of cardinal β. Each $y \in Y$ is dependent on X and hence is dependent on a finite subset X_y of X. We claim that

$$X = \bigcup_{y \in Y} X_y. \tag{11.1.5}$$

For the span of the right-hand side contains each $y \in Y$ and so must be S. Thus $\cup X_y$ is a subset of X spanning S, and (11.1.5) follows by the minimality of X. If Y were finite, then (11.1.5) would express X as a finite union of finite sets, but this contradicts the fact that X is infinite. We thus have

$$\alpha = |X| \leq \sum |X_y| \leq \aleph_0\beta = \beta.$$

Hence $\alpha \leq \beta$; if Y is also minimal, we can interchange the roles of X and Y and so conclude that $\alpha = \beta$. ∎

This result applies in particular to modules, since as we observed earlier, the relation 'y lies in the submodule generated by X' is a spanning relation. We thus obtain Proposition 4.6.4 as a special case.

Applying Proposition 11.1.4 to the case of dependence relations, we obtain

Theorem 11.1.5. *Let S be any set with a dependence relation. Then S has a basis; more precisely, if X is an independent subset and Y is a spanning set of S such that $X \subseteq Y$, then there is a basis B such that*

$$X \subseteq B \subseteq Y, \tag{11.1.6}$$

and any two bases have the same cardinal.

Proof. Let X, Y be as stated; then $X \cup Y = Y$ and by Lemma 11.1.2 there exists a basis B satisfying (11.1.6). In particular, taking $X = \varnothing$, $Y = S$, we see that there always is a basis. The last part follows by Corollary 11.1.3 and Proposition 11.1.4. ∎

As a consequence we see that every vector space has a basis and that any two bases have the same cardinal. For example, **R** considered as a **Q**-module has a basis, but the proof (involving Zorn's lemma) is non-constructive, and no explicit construction of such a basis is known. Any **Q**-basis of **R** is called a *Hamel basis* of the real numbers.

Occasionally we shall have to deal with a relation satisfying **D.0** and **D.2** but only a weakened form of **D.1** obtained by assuming Y independent. Thus our relation satisfies

D.1′ *If z is dependent on the independent set Y and each element of Y is dependent on X, the z is dependent on X.*

We still find that bases are just maximal independent sets, for the proof (a) $\Leftrightarrow$ (c) of Proposition 11.1.1 did not use **D.1**, and maximal independent sets always exist, by Zorn's lemma, so every independent subset is contained in a basis. Now the proof of Lemma 11.1.2, in the case where X, Y are finite bases, uses **D.1** only in the weaker form **D.1′**, so we still have the following conclusion:

Corollary 11.1.6. *Let S be a set with a relation satisfying* **D.0**, **D.1′** and **D.2**. *Then any independent subset is contained in a basis, and if S has a finite basis, then any two bases have the same cardinal.* ∎

The axiomatic study of dependence relations turns out to have connexions with many other fields such as graph theory, combinatorial theory, block designs etc. See Welsh (1976), White (1986).

Exercises

1. In a set with a dependence relation prove the exchange property for infinite subsets: if y is dependent on $X \cup \{z\}$ but not on X, then z is dependent on $X \cup \{y\}$.
2. Verify that the relation defined by (11.1.2) over a commutative integral domain satisfies **D.0**–**D.2**. Give an example of a commutative ring with zerodivisors for which **D.1** fails.

3. Check that in the proof of Proposition 11.1.1, (c) $\Rightarrow$ (a), (b) only uses **D.0**; (a) $\Rightarrow$ (c) only uses **D.0** and **D.2**; and (b) $\Rightarrow$ (c) only uses **D.0** and **D.1**. Give examples to show that the axioms listed here cannot be omitted.
4. Let S be a set with a dependence relation and for each subset X write $r(X)$ for the cardinal of a basis of $\langle X\rangle$. Show that (i) $r(\varnothing) = 0$; (ii) $r(X) \leq r(X \cup \{y\}) \leq r(X) + 1$; (iii) if $r(X) = r(X \cup \{y\}) = r(X \cup \{z\})$, then $r(X \cup \{y, z\}) = r(X)$.

   Show that any function $r(X)$ on the subsets of a set S with integer values, satisfying (i)–(iii) leads to a dependence relation on S.
5. Show that for any $n \geq 1$ there is a minimal spanning set for $\mathbf{Z}$ (as $\mathbf{Z}$-module) of n elements.
6. How many dependence relations are there on a set of 3 elements?
7. Two dependence relations on a set S are called *equivalent* if we can pass from one to the other by a permutation of S. Show that there are 3 equivalence classes of dependence relations on a set of 3 elements. How many classes are there on a 4-element set?
8. Show that any dependence relation on a set S is determined by the collection of bases of S.
9. Show that any dependence relation on a set S is determined by the collection of all minimal dependent subsets.
10. Show that every additive mapping of $\mathbf{R}$ into itself is $\mathbf{Q}$-linear, but not necessarily $\mathbf{R}$-linear. (Hint. Use a Hamel basis.)

11.2 Algebraic Dependence

There is a notion of algebraic dependence that is entirely analogous to the notion of linear dependence. Like the latter, it satisfies axioms **D.0–D.2** for an abstract dependence relation, so we shall be able to use the consequences proved in Section 11.1.

Given any field extension E/k, let $u_1, \ldots, u_m, v \in E$; we shall say that v is *algebraically dependent* on $u_1, \ldots, u_m$ over k if v is algebraic over the field $k(u_1, \ldots, u_m)$. This just means that v satisfies an equation with coefficients in $k(u_1, \ldots, u_m)$; on multiplying by a common denominator, we obtain an equation for v:

$$a_0 v^d + a_1 v^{d-1} + \ldots + a_d = 0, \tag{11.2.1}$$

where $a_i \in k[u_1, \ldots, u_m]$ and $a_0 \neq 0$. It is easily seen that this dependence relation satisfies rules **D.0–D.2** of Section 11.1:

D.0 *Each u_i is algebraically dependent on $u_1, \ldots, u_m$.*

D.1 *If w is algebraically dependent on $v_1, \ldots, v_n$ and each v_i is algebraically dependent on $u_1, \ldots, u_m$ then w is algebraically dependent on $u_1, \ldots, u_m$.*

D.2 *If v is algebraically dependent on $u_1, \ldots, u_m$ but not on $u_2, \ldots, u_m$ then u_1 is algebraically dependent on $v, u_2, \ldots, u_m$.*

D.0 holds trivially. To prove **D.1**, we are given that v_i is algebraic over $k(u_1, \ldots, u_m)$; hence in the tower

$$k(u_1, \ldots, u_m) \subseteq k(u_1, \ldots, u_m, v_1) \subseteq \ldots \subseteq k(u_1, \ldots, u_m, v_1, \ldots, v_n)$$

each extension is of finite degree; moreover, w is algebraic over $k(v_1, \ldots, v_n)$ and a fortiori also over $k(u_1, \ldots, u_m, v_1, \ldots, v_n)$. By the product formula for the degrees, $k(u_1, \ldots, u_m, v_1, \ldots, v_n, w)$ is of finite degree over $k(u_1, \ldots, u_m)$, and this shows w to be algebraically dependent on $u_1, \ldots, u_m$.

Finally, to prove **D.2**, we have by hypothesis a relation (11.2.1), i.e. a linear dependence over k between the power products in $u_1, \ldots, u_m, v$. Let us arrange the left-hand side as a polynomial in u_1:

$$b_0 u_1^r + b_1 u_1^{r-1} + \ldots + b_r = 0, \tag{11.2.2}$$

where $b_i \in k[u_2, \ldots, u_m, v]$. By hypothesis, $a_0 \neq 0$ in (11.2.1); if $b_0, b_1, \ldots, b_{r-1}$ all vanish, then (11.2.2) reduces to $b_r = 0$, a polynomial relation between $u_2, \ldots, u_m, v$ and since this is obtained by rearranging (11.2.1), it would show that v is algebraically dependent on $u_2, \ldots, u_m$, against the hypothesis. Hence $b_0, \ldots, b_{r-1}$ do not all vanish and now (11.2.2) shows that u_1 is algebraically dependent on $u_2, \ldots, u_m, v$.

Thus algebraic dependence is a dependence relation in the sense of Section 11.1 and we can use the concepts introduced there, and the facts proved about them. From the definitions given there we see that a subset X of E is *algebraically independent over* k if no element of X is algebraically dependent on the rest. If $X = \{x_1, \ldots, x_n\}$, this means that the power products $x_1^{a_1} \ldots x_n^{a_n}$ are linearly independent over k, but the definition applies even if X is infinite. A subset B of E which is algebraically independent and such that every element of E is algebraically dependent on B over k is called a *transcendence basis* of E over k. From Lemma 11.1.2, Corollary 11.1.3 and Proposition 11.1.4 we now obtain

Theorem 11.2.1. *Given any field extension E/k, any algebraically independent subset of E can be completed to a transcendence basis of E/k and any two such bases have the same number of elements.* ■

The number of elements in a transcendence basis of E/k is called the *transcendence degree* or *dimension* of E/k and is written $\mathrm{tr.deg}(E/k)$. An extension of transcendence degree zero is called *algebraic*. Thus an extension E/k is algebraic precisely if every element of E is algebraic over k. As we saw in Corollary 7.1.5, it is enough to assume that E has a generating set of algebraic elements, for E/k to be algebraic. We remark that a field extension E/k which is finitely generated, by $x_1, \ldots, x_n$ say, is of finite degree over k iff each x_i is algebraic over k, but in general an extension may well be algebraic and yet of infinite degree over k. For example, the algebraic closure of a field is always algebraic but not generally of finite degree.

The transcendence degree of a repeated extension is given by the following formula:

Proposition 11.2.2. *For any fields $k \subseteq E \subseteq F$ we have*

$$\mathrm{tr.deg}(F/k) = \mathrm{tr.deg}(F/E) + \mathrm{tr.deg}(E/k). \tag{11.2.3}$$

Proof. Let X be a transcendence basis for E/k and Y be one for F/E. Then no element of Y is in E, but every element of X is, hence $X \cap Y = \varnothing$ and to establish (11.2.3) it will be enough to show that $X \cup Y$ is a transcendence basis for F/k. Any element z of F is algebraic over $E(Y)$; by writing down an equation for z we see that z is algebraically dependent on a finite set over k, say $u_1, \ldots, u_r, y_1, \ldots, y_s$, where $u_i \in E, y_j \in Y$. Each u_i is algebraically dependent on X over k, hence (by transitivity) z itself is algebraically dependent on $X \cup Y$ over k. Secondly we must show that $X \cup Y$ is algebraically independent. Suppose not; then there is a non-trivial polynomial relation

$$f(x_1, \ldots, x_r, y_1, \ldots, y_s) = 0, \quad x_i \in X, y_j \in Y, \tag{11.2.4}$$

with coefficients in k, i.e. the power products in the x_i, y_j are linearly dependent over k. Here some y_i must occur because X is algebraically independent over k. Since $x_i \in E$, (11.2.4) is an algebraic relation between the y_i over E; but Y was algebraically independent over E, so if we rewrite (11.2.4) as a polynomial in the y_i, all the coefficients vanish. But these coefficients are polynomials in the x_i, so we obtain an algebraic relation between the x_i, which is a contradiction. Thus $X \cup Y$ is algebraically independent over k, hence it is a transcendence basis for F/k and (11.2.3) follows. ∎

In particular, for algebraic extensions this shows the truth of

Corollary 11.2.3. *An algebraic extension of an algebraic extension is again algebraic.* ∎

A field extension is called *transcendental* if it is not algebraic; thus a transcendental field extension is one of positive transcendence degree. An extension E/k is said to be *purely transcendental* if there is an algebraically independent subset X of E such that $E = k(X)$. Such a purely transcendental extension of k is determined up to isomorphism by its transcendence degree: given any cardinal α, take a set X of cardinal α, construct the polynomial ring $k[X]$ in the indeterminates X over k and form its field of fractions $k(X)$. This field, consisting of all rational functions in the indeterminates X over k, is a purely transcendental extension of tr.deg.α over k, and it is clearly the only one, up to isomorphism.

If $E = k(x_1, \ldots, x_r)$ is a purely transcendental extension of k, then any permutation of $x_1, \ldots, x_r$ defines a k-automorphism of E; thus any permutation group of the x_i acts in a natural way as a group of k-automorphisms of E. It is an old problem (known as E. Noether's problem) whether the fixed field under a given permutation group is a purely transcendental extension of k. This was answered negatively by Richard G. Swan [1969] for $n = 47$ and a cyclic transitive group. More general results have been obtained by V. E. Voskresenskii [1973] and H. W. Lenstra [1974]. The latter gives a complete solution of Noether's problem for abelian transitive groups in a most readable account which also includes a survey of earlier results.

Exercises

1. Show that the fixed field in $k(x_1, \ldots, x_n)$ under the group of all permutations of the x_i is a purely transcendental extension of k, of transcendence degree n.
2. Show that the extension of $\mathbf{C}$ defined by x, y subject to $x^2 + y^2 = 1$ is purely transcendental, but for the equation $x^3 + y^3 = 1$ it is not. (Hint. Write $x = f/h$, $y = g/h$ and examine the degrees of f, g, h. Try differentiating.)
3. Let E be the extension of $\mathbf{R}$ defined by x, y subject to $x^2 + y^2 + 1 = 0$. Show that any element of $E \backslash \mathbf{R}$ is transcendental over $\mathbf{R}$ and deduce that $E/\mathbf{R}$ is not purely transcendental.
4. Show that $\mathrm{tr.deg}(\mathbf{C}/\mathbf{Q}) = \mathfrak{c}$, where $\mathfrak{c} = 2^{\aleph_0}$. Deduce that the automorphism group of $\mathbf{C}$ has order $2^{\mathfrak{c}}$.
5. Show that a field extension E/k is algebraic iff every subalgebra of E over k is a subfield.
6. Let F/k be any finitely generated field extension and $k \subseteq E \subseteq F$. Show that E is finitely generated over k.
7. Let k be any field and $E = k((x))$ be the field of formal Laurent series $\sum_{-N}^{\infty} a_\nu x^\nu$ in an indeterminate x. Show that E has $k(x)$ as a subfield and that E is not algebraic over $k(x)$, even though it contains elements that are algebraic over $k(x)$.

11.3 Simple Transcendental Extensions

We have seen that a purely transcendental extension of a field k, of finite transcendence degree d, is just the field $k(x_1, \ldots, x_d)$ of all rational functions in d independent indeterminates. The general study of such fields raises some difficult questions, but in the case $d = 1$ the situation is rather better. We then have a field of rational functions in one variable $k(x)$, also called a *simple transcendental extension*, and we shall now examine this case.

Let $E = k(x)$ be a simple transcendental extension; its elements are rational functions in x over k, and so may be written

$$u = \frac{f(x)}{g(x)}, \quad \text{where } f, g \in k[x], (f, g) = 1 \text{ and } g \text{ is monic.} \tag{11.3.1}$$

The *degree* of u is defined as $\deg u = \max\{\deg f, \deg g\}$. It is positive unless u is constant, i.e. $u \in k$, and although the representation (11.3.1) depends on the choice of the generator x, the degree of u is independent of this choice, by the next result. Given any field extension E/k, we shall say that k is *relatively algebraically closed* in E if every element of $E \backslash k$ is transcendental over k; in other words, any element of E algebraic over k already lies in k.

Proposition 11.3.1. *Let $k(x)$ be a simple transcendental extension of k. Then k is relatively algebraically closed in $k(x)$, and for any $u \in k(x) \backslash k$ the degree of u is given by*

$$\deg u = [k(x) : k(u)]. \tag{11.3.2}$$

Proof. If u is given by (11.3.1), x satisfies the equation in the indeterminate X:

$$f(X) - ug(X) = 0. \tag{11.3.3}$$

Let b_i be a non-zero coefficient in $g(X)$ and a_i be the corresponding coefficient in $f(X)$; then the corresponding coefficient in (11.3.3) is $a_i - ub_i$. This is non-zero, for otherwise $u = a_i/b_i \in k$ against the hypothesis. Thus not all the coefficients of (11.3.3) vanish, and it follows that x is algebraic over $k(u)$. If u were algebraic over k, then x would also be algebraic over k, which is clearly not the case; hence u is transcendental over k.

The left-hand side of (11.3.3), as a polynomial in X, is irreducible over $k(u)$, for if it could be factorized over $k(u)[X]$, then by Theorem 7.7.2 it can be factorized over $k[u, X]$. Since it is linear in u, one factor must be independent of u and the other linear in u, but this is impossible, because f, g are coprime. So (11.3.3) is an irreducible polynomial for x over $k(u)$. It has degree deg u and so (11.3.2) follows. ■

The case $n = 1$ is of particular importance; it applies iff u generates $k(x)$, so we obtain

Corollary 11.3.2. *Let u be an element of a simple transcendental extension $k(x)$ of k. Then $k(u) = k(x)$ if and only if*

$$u = \frac{ax + b}{cx + d}, \quad \text{where } a, b, c, d \in k \text{ and } ad - bc \neq 0.$$

Proof. By Theorem 11.3.2, $k(u) = k(x)$ iff deg $u = 1$, i.e. f and g are linear; now the condition $ad \neq bc$ just ensures that $u \notin k$. ■

This result makes it easy to determine the automorphisms of $k(x)$ over k. Any such automorphism must map x to a generator, and conversely, any mapping from x to a generator determines a unique automorphism. Thus we have

Proposition 11.3.3. *The group of automorphisms of a simple transcendental extension $k(x)/k$ is the group $\mathbf{PGL}_2(k)$ of projective linear transformations*

$$x \mapsto \frac{ax + b}{cx + d}, \quad \text{where } a, b, c, d \in k \text{ and } ad - bc \neq 0. \quad ■ \tag{11.3.4}$$

We observe that $\mathbf{PGL}_2(k)$ is the quotient of the general linear group $\mathbf{GL}_2(k)$ by its centre; each matrix A in $\mathbf{GL}_2(k)$ defines a transformation (11.3.4), which is the identity precisely when A lies in the centre, i.e. $A = \lambda I$ for some $\lambda \in k^\times$. Thus the constants in (11.3.4) are determined up to a constant multiple; this means that we can normalize our transformations, at least in the case $k = \mathbf{C}$, by requiring that $ad - bc = 1$.

Proposition 11.3.1 has an important generalization:

Theorem 11.3.4 (Lüroth's theorem). *Let $k(x)$ be a simple transcendental extension of a field k. Then every field E such that $k \subset E \subseteq k(x)$ is of the form $E = k(u)$, where u is transcendental over k.*

Proof. Since $E \supset k$, there exists $u \in E \backslash k$; in Proposition 11.3.1 we saw that x is algebraic over $k(u)$, hence x is algebraic over E. Let the minimal polynomial for x over E be

$$\varphi(X) = X^n + u_1 X^{n-1} + \ldots + u_n, \quad \text{where } u_i \in E.$$

Here the u_i are rational functions of x and on multiplying by a lowest common denominator we obtain a polynomial

$$\Phi(X, x) = v_0(x)X^n + v_1(x)X^{n-1} + \ldots + v_n(x), \tag{11.3.5}$$

in which the coefficients v_i are polynomials in x without a common (non-constant) factor, i.e. Φ is primitive, as a polynomial in X. Since x is transcendental over k, the $u_i = v_i/v_0$ are not all in k; we pick $u_j \notin k$, so that $m = \deg u_j > 0$. By Proposition 11.3.1, x has degree m over $k(u_j)$, while its degree over E is n, thus

$$m = [k(x) : k(u_j)] = [k(x) : E][E : k(u_j)] = n[E : k(u_j)],$$

and it follows that

$$[E : k(u_j)] = m/n. \tag{11.3.6}$$

Now the proof will be complete if we can show that $m = n$, for then $E = k(u_j)$, by (11.3.6). Let us write u_j in lowest terms as

$$u_j = \frac{f(x)}{g(x)}, \tag{11.3.7}$$

where f, g are coprime and g is monic, while $m = \max\{\deg f, \deg g\}$, by the definition of m. The polynomial $f(X) - u_j g(X)$ does not vanish identically, but has x as a zero, and so is divisible by $\varphi(X)$. If we replace u_j by its value from (11.3.7) and multiply up by g, we obtain by (11.3.5),

$$f(X)g(x) - g(X)f(x) = Q(X, x)\Phi(X, x), \tag{11.3.8}$$

where Q is a polynomial in X and by the primitivity of Φ, also one in x. The left-hand side of (11.3.8) has degree m in x, and on the right Φ has degree at least m in x. Hence Q must be independent of x; if Q has positive degree in X, it has a zero α in an algebraic extension of k. Thus $f(\alpha)g(x) - g(\alpha)f(x) = 0$, and $f(\alpha), g(\alpha) \neq 0$, because f, g are coprime, so

$$\frac{f(\alpha)}{g(\alpha)} = \frac{f(x)}{g(x)} = u_j,$$

hence u_j is algebraic over k, which is a contradiction. This shows Q to be independent of X as well. Now the left-hand side of (11.3.8) has degree m in X, while the right-hand side has degree n in X, by (11.3.5). Hence $m = n$ and so $E = k(u_j)$, as claimed. ∎

For fields in two variables $k(x, y)$ the analogue need not hold, though it does hold when k is algebraically closed (Theorem of Castelnuovo–Zariski). However, for more than two variables the result fails even over algebraically closed fields (see Deligne [1973]).

Let us return to the subject of Proposition 11.3.3. By Galois theory, any finite subgroup of $\mathbf{PGL}_2(k)$ corresponds to a subfield F of $k(x)$ such that $[k(x) : F]$ is finite. In the special case $k = \mathbf{C}$ any such subfields may be determined as follows.*

By stereographic projection the unit-sphere in real 3-space $\mathbf{R}^3$ corresponds to the complex line, i.e. the real plane, with the North pole corresponding to the point at infinity. Any rotation in $\mathbf{R}^3$ corresponds to a fractional linear transformation (of the type (11.3.4)) of the complex line with two fixed points, corresponding to the axis of rotation. Taking the fixed points to be 0, ∞, we obtain the form

$$x' = \kappa x, \quad \text{where } |\kappa| = 1, \tag{11.3.9}$$

for our transformation; this is known as an *elliptic* transformation of the line. In the form (11.3.4) (with $ad - bc = 1$) an elliptic transformation is characterized by the condition that $a + d$ is real and $|a + d| < 2$; for this is an invariant condition which just ensures that $|\kappa| = 1$ in (11.3.9). Conversely, an elliptic transformation corresponds to a rotation, as we see by transforming the fixed points to 0, ∞, respectively. Now any transformation (11.3.4) of finite order is elliptic. To see this we observe that the fixed points of (11.3.4) are given by

$$\frac{x}{y} = \frac{ax + by}{cx + dy}, \quad \text{i.e. } A\begin{pmatrix} x \\ y \end{pmatrix} = \lambda\begin{pmatrix} x \\ y \end{pmatrix}, \quad \text{where } A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

If $A^n = I$, then the minimal equation of A has distinct roots, so A is similar to a diagonal matrix, which gives the fixed points 0, ∞ on the line. Moreover, the roots are roots of 1, reciprocal and so conjugate over $\mathbf{R}$, hence $a + d$ is real and $|a + d| < 2$.

Thus it follows that any finite subgroup of $\mathbf{PGL}_2(\mathbf{C})$ corresponds to a finite group of rotations in $\mathbf{R}^3$. These groups are well known; they are the cyclic and dihedral groups, and the tetrahedral, octahedral and icosahedral groups. The latter three groups are isomorphic to Alt_4, Sym_4 and Alt_5 respectively; thus the only non-abelian simple group that occurs is the alternating group of degree 5 (for more details, see Klein (1884); Weber (1894)).

By contrast, there are many different finite extensions of $k(x)$; they form the field of study of the theory of algebraic functions of one variable, of importance in function theory, as well as algebraic geometry (see Chevalley (1951); Cohn (1991)).

* The remarks that follow assume some acquaintance with the rudiments of complex function theory; they will not be used later in this work.

Exercises

1. Show that if φ, ψ are any rational functions of degrees m, n (as defined after (11.3.1)), then $\varphi(\psi(x))$ is of degree mn.
2. Show that any simple subextension of a purely transcendental extension is purely transcendental.
3. Let $k(x)/k$ be a simple transcendental extension. Show that for any $n \geq 1$ there exists an extension E of $k(x)$ of degree n, such that E is again simple transcendental over k. Use this result to prove the existence of an infinite algebraic extension F of $k(x)$ such that any finitely generated subextension of F is simple transcendental over k.
4. Show that $\alpha : x \mapsto x^{-1}$ and $\beta : x \mapsto 1 - x$ generate an automorphism group $\Gamma \cong \mathrm{Sym}_3$ of $k(x)$ and that the fixed field is $k((x^2 - x + 1)^3/x^2(x-1)^2)$.
5. Let k be a finite field of q elements. Show that $\mathrm{Gal}(k(x)/k)$ has order $q^3 - q$ and find the exact fixed field.
6. Let k be a field and x be an indeterminate. Show that a k-subalgebra R of $k[x]$ is of the form $k[y]$ or k iff R is integrally closed. (Hint. Use Lüroth's theorem to show that R has a field of fractions of the form $k(y)$ (or k if $R = k$) and prove that R is contained in all valuation subrings of $k(y)$ except at most one. Now use the characterization of integrally closed rings in Theorem 9.4.4.)
7. (E. Noether) Show that a subfield E of $k(x)$ has a generator in $k[x]$ iff $E \cap k[x] \supset k$. (Hint. Use Exercise 6.)

11.4 Separable and p-radical Extensions

Separable extensions were defined in Section 7.4; we shall now take a closer look at the inseparable case, bearing in mind that it only arises in prime characteristic. Over a field of characteristic 0 every extension is separable.

We have seen in Theorem 7.5.4 that finite separable extensions may be characterized as finite extensions E/k possessing exactly $[E : k]$ k-homomorphisms into a normal closure of E/k. Here we may instead of a normal closure take an algebraic closure of E, for the latter contains a normal closure.

Let us define, for any finite extension E/k, the *separable degree* $[E : k]_s$ as the number of k-homomorphisms of E into an algebraic closure of E. Then we have, by Theorem 7.5.4,

$$[E : k]_s \leq [E : k],$$

with equality holding precisely when E/k is separable. Our first observation is that the separable degree is again multiplicative:

Proposition 11.4.1. *Let k be a field and $F \supseteq E \supseteq k$ be algebraic extensions. Then*

$$[F : k]_s = [F : E]_s [E : k]_s, \tag{11.4.1}$$

whenever either side is finite.

Proof. Let F_a be an algebraic closure of F; since F/E is algebraic, F_a is also an algebraic closure of E. If $s = [F : E]_s$, we have s E-homomorphisms of F into F_a; it follows that any isomorphism $\varphi : E \to E'$ can be extended in just s ways to a homomorphism of F into an algebraic closure of E', for if we use φ to identify E with E', the problem is reduced to extending the identity mapping on E, i.e. to finding E-homomorphisms of F, and we saw that there are just s such extensions. We now apply this argument to a k-homomorphism ψ of E into F_a. There are $[E : k]_s = r$ such mappings, and each can be extended in s ways, giving rs k-homomorphisms of F into F_a. But every k-homomorphism of F into F_a can be obtained in this way; thus there are just rs of them and so (11.4.1) follows whenever either side is finite. ■

In contrast to the separable extensions we now consider the *p-radical* or *purely inseparable* extensions, defined as algebraic extensions E/k of prime characteristic p, such that there is only one k-homomorphism of E into an algebraic closure of E. In particular this means that each element α of E has only one conjugate over k, for if α had a conjugate $\alpha' \neq \alpha$, then the k-isomorphism $k(\alpha) \cong k(\alpha')$ in which α maps to α' could be extended to a k-homomorphism different from 1 on E, which contradicts the hypothesis.

Let E/k be any p-radical extension, where $E \supset k$. This means that $p = \text{char } k$ is a prime and the minimal polynomial f of any $\alpha \in E$ has only one zero. If its degree is $n = mq$, where $q = p^r$ and m is prime to p, let us write $\alpha^q = a$. The minimal polynomial for α is $f = (x - \alpha)^n = (x^q - a)^m = x^{mq} - max^{(m-1)q} + \ldots$; it follows that $a \in k$ and so $m = 1$ and the minimal polynomial for α is of the form

$$f = (x - \alpha)^q = x^q - a, \quad \text{where } \alpha^q = a \in k \text{ and } q = p^r. \tag{11.4.2}$$

An element α of E is said to be *p-radical* or *purely inseparable* over k if $\alpha^q \in k$, where $q = p^r$ ($p = \text{char } k$). Here the least value of r is sometimes called the *height* of α. By (11.4.2) every element of a p-radical extension is p-radical; the converse also holds, in a slightly sharper form:

Proposition 11.4.2. *An extension E/k is p-radical if and only if it is generated by p-radical elements.*

Proof. We have seen that a p-radical extension consists entirely of p-radical elements. Conversely, assume that E is generated over k by a set X of elements that are p-radical over k. In any k-homomorphism of E, any $x \in X$ must map to a root of the minimal equation of x and hence remains fixed. Therefore $E = k(X)$ has only one k-homomorphism into an algebraic closure of E, so E/k is p-radical, as claimed. ■

Our next task is to see how a general algebraic extension is built up from separable and p-radical extensions. We shall need:

Lemma 11.4.3. *Let α be algebraic over a field k, where* char $k = p$. *Then for some $q = p^r$, α^q is separable over k.*

Proof. We shall use induction on the degree of α over k. Let f be the minimal polynomial of α over k. If f has distinct zeros, it is separable and the result follows. Otherwise f has a repeated zero and so is not prime to its derivative f'. Since f is irreducible over k, this means that $f|f'$, and by comparing degrees we see that this can happen only when $f' = 0$. If $f = \sum a_i x^i$, this means that $ia_i = 0$; hence $a_i = 0$ unless $p|i$ and so $f(x) = g(x^p)$ for some polynomial g of lower degree than f. Now α^p is a zero of g, hence by induction on the degree of f, for some power q of p, $(\alpha^p)^q = \alpha^{pq}$ is separable over k, as claimed. ∎

Let E/k be an algebraic extension. Then the elements of E that are separable over k form a subfield, by Theorem 7.5.4, which will be denoted by k_s and called the *separable closure* of k in E. We note that E is p-radical over k_s, for if $\alpha \in E$, then for some $q = p^r$, α^q is separable over k, by Lemma 11.4.3, and so α^q lies in k_s. Thus we have $[E : k_s]_s = 1$; now if E/k is finite, then $[E : k]_s = [E : k_s]_s[k_s : k]_s$ by (11.4.1), and of course $[k_s : k]_s = [k_s : k]$, hence

$$[k_s : k] = [E : k]_s. \tag{11.4.3}$$

We shall define the *degree of inseparability* or simply the *inseparable degree* of E/k as

$$[E : k]_i = [E : k_s]. \tag{11.4.4}$$

Like the total and the separable degree it is again multiplicative. This is clear from the definition, which may be written

$$[E : k] = [E : k]_s[E : k]_i. \tag{11.4.5}$$

We remark that $[E : k]_i$ must be a power of p, if finite, since it is the degree of the p-radical extension E/k. The situation may be summed up as follows:

Theorem 11.4.4. *Every algebraic extension E/k, where* char $k = p$ *is prime, may be obtained by taking a separable algebraic extension followed by a p-radical extension: $k \subseteq k_s \subseteq E$. Moreover, when E/k is finite, then*

$$[E : k]_s = [k_s : k], [E : k]_i = [E : k_s]. \quad ∎ \tag{11.4.6}$$

The set of all p-radical elements of an algebraic extension E/k also forms a subfield k_p say, called the *perfect closure* of k in E. For if $\alpha, \beta \in k$, say $\alpha^{p^r}, \beta^{p^s} \in k$ and $r \leq s$ say, then $(\alpha - \beta)^{p^s} = \alpha^{p^s} - \beta^{p^s} \in k$ and of course $(\alpha\beta)^{p^s} = \alpha^{p^s}\beta^{p^s} \in k$. Strictly speaking, k_p should be described as the *relative* perfect closure of k in E, to distinguish it from the following absolute notion.

Let k be any field of prime characteristic p. A *perfect closure* of k is a field $\hat{k}$ containing k as a subfield such that $\hat{k}$ is perfect and $\hat{k}/k$ is p-radical.

Proposition 11.4.5. *Every field k of prime characteristic p has a perfect closure, unique up to isomorphism.*

Proof. We define $\hat{k}$ as the relative perfect closure of k in its algebraic closure. Every element of $\hat{k}$ is p-radical over k, hence $\hat{k}/k$ is p-radical, and $\hat{k}$ is perfect because every

element of $\hat{k}$ has a p-th root in the algebraic closure of k, and so this root also lies in $\hat{k}$. The uniqueness is clear: if E, F are two perfect closures of k and $\alpha \in E$, then $\alpha^q \in k$ for some $q = p^r$ and there is a unique element $\alpha' \in F$ such that $\alpha'^q = \alpha^q$. The correspondence $\alpha \mapsto \alpha'$ is a homomorphism, for if $\beta \mapsto \beta'$, then for a high enough power $q = p^r$ we have $\alpha^q, \beta^q \in k$ and so $(\alpha + \beta)^q = \alpha^q + \beta^q = \alpha'^q + \beta'^q = (\alpha' + \beta')^q$, therefore $\alpha + \beta \mapsto \alpha' + \beta'$ and similarly $\alpha\beta \mapsto \alpha'\beta'$. Clearly it is bijective and preserves 1, so it is an isomorphism and $E \cong F$. ■

This result may also be proved directly, without recourse to the algebraic closure of k (see Exercise 3). The relation of the perfect closure to its relative version should be clear: if $\hat{k}$ denotes the perfect closure of a field k and E is an extension field of k, then the relative perfect closure of k in E is the largest subfield of E which is k-isomorphic to a subfield of $\hat{k}$.

Let E/k be an algebraic extension. We have in E the separable closure k_s and the perfect closure k_p of k. As we saw in Theorem 11.4.4, E/k_s is then p-radical, so E can be obtained by a separable extension followed by a p-radical extension. However, E/k_p need not be separable, so the order of the extensions cannot in general be reversed (see Exercise 4). In the special case where it can, we have a more precise description of the extension. First we prove an auxiliary result. We write $E^n = \{x^n | x \in E\}$ for any $n \in \mathbf{N}$.

Lemma 11.4.6. *Let E/k be a finite separable extension of characteristic p and let $u_1, \ldots, u_m$ be a basis. Then $u_1^q, \ldots, u_m^q$ is again a basis, for any $q = p^r, r \geq 1$.*

Proof. Since $E \supseteq E^q k \supseteq k$, it follows that E is p-radical as well as separable over $E^q k$, hence $E = E^q k$. Hence $u_1^q, \ldots, u_m^q$ again spans E over k and so is a basis. ■

We can now achieve our aim, to prove that a p-radical extension followed by a separable extension always decomposes as a tensor product.

Theorem 11.4.7. *Given extensions $k \subseteq E \subseteq F$, where E/k is p-radical and F/E is separable, we have*

$$F \cong E \otimes_k k_s, \tag{11.4.7}$$

where k_s is the separable closure of k in F.

Proof. In proving (11.4.7) we may assume F/k to be finite; for when that case has been proved, the general case follows because the algebraic extensions F/k, E/k can be written as ascending unions of their finite subextensions, and tensor products preserve ascending unions.

Further, if we can find a k-basis for k_s which is also an E-basis for F, then the natural homomorphism $E \otimes_k k_s \to F$ is injective, by the independence property of the tensor product. Now the image Ek_s in F is a k-algebra, finite-dimensional over k and without zerodivisors, and therefore is a field. Moreover, F/E is separable and F/k_s is p-radical, so F is both separable and p-radical over Ek_s and therefore $F = Ek_s$. So it only remains to find a k-basis of k_s which is an E-basis for F.

Let $u_1, \ldots, u_m$ be any k-basis of k_s; then $u_1, \ldots, u_m$ span F over E, by what we have seen, and we claim that they are linearly independent over E. Suppose that $\sum a_i u_i = 0$ $(a_i \in E)$. Since E/k is p-radical, $a_i^q \in k$ for some $q = p^r$, so $\sum a_i^q u_i^q = 0$, but the u_i^q are linearly independent over k, by Lemma 11.4.6, hence $a_i^q = 0$, so $a_i = 0$ and this proves that the u_i are linearly independent over E. ∎

It is clear that in Theorem 11.4.7, E is just the perfect closure of k in F. The result can be applied to describe the structure of normal extensions. We recall from Section 7.2 that an extension E/k is called *normal* (or sometimes *quasi-Galois*) if it is algebraic and every irreducible polynomial over k which has a zero in E splits completely over E. We shall need a generalization of a property of normal extensions, proved in Corollary 7.2.5, to the infinite case.

Proposition 11.4.8. *Let Ω/k be a normal extension. Then any two isomorphic subextensions are conjugate, i.e. any isomorphism $\varphi : E \to F$ of subextensions can be extended to a k-automorphism of Ω.*

Proof. Consider the family of all k-homomorphisms $\varphi_\lambda : E_\lambda \to \Omega$, where $E_\lambda \supseteq E$ and $\varphi_\lambda | E = \varphi$. These homomorphisms can be partially ordered by writing $(\varphi_\lambda, E_\lambda) \leq (\varphi_\mu, E_\mu)$ if $E_\lambda \subseteq E_\mu$ and $\varphi_\mu | E_\lambda = \varphi_\lambda$. This family is clearly inductive; by Zorn's lemma there is a maximal element $\varphi_0 : E_0 \to \Omega$, and we claim that $E_0 = \Omega$. If $E_0 \neq \Omega$, take $\alpha \in \Omega \backslash E_0$, let f be the minimal polynomial of α over E_0 and g its minimal polynomial over k. Then $g\varphi_0 = g$ and so both $f, f\varphi_0$ divide g. Since f has the zero α in Ω, hence so does $f\varphi_0$. We can therefore extend φ_0 to $E_0(\alpha)$ by mapping α to a zero of $f\varphi_0$. This contradicts the maximality of φ_0, hence $E_0 = \Omega$ and φ_0 is the desired automorphism of Ω over k. ∎

This result can be used to characterize normal extensions:

Corollary 11.4.9. *An algebraic extension E/k is normal if and only if every k-automorphism of any field Ω containing E maps E into itself.*

Proof. The condition is necessary because when E/k is normal, then every element of E satisfies an irreducible equation over k whose roots are permuted by any k-automorphism of Ω. Conversely, when it holds, let f be any irreducible polynomial over k which has a zero α in E. Take a normal extension Ω/k containing E/k such that f splits completely in Ω (e.g. a normal closure of the splitting field of f over E). If β is any zero of f in Ω, then $\alpha \mapsto \beta$ under an isomorphism $k(\alpha) \cong k(\beta)$, hence α and β are conjugate in Ω, by Proposition 11.4.8, and so $\beta \in E$. Thus f splits completely over E and this shows E/k to be normal. ∎

We now come to the structure theorem for normal extensions. For simplicity we shall limit ourselves to the finite case; the general case is quite analogous but requires some ideas from Section 11.8 below.

Theorem 11.4.10. *Let E/k be a finite normal extension and denote by k_s the separable closure and by k_p the perfect closure of k in E. Then k_s is a Galois extension of k and we have*

$$E \cong k_s \otimes_k k_p.$$

Proof. Any k-automorphism of an extension of E maps E into itself, hence it maps k_s into itself. Therefore k_s/k is normal, and being also separable it is Galois (Proposition 7.6.1). Now the fixed field under all automorphisms is k_p, so E/k_p is Galois and the result follows from Theorem 11.4.7. ∎

Exercises

1. Show that a finite extension E/k of characteristic p is separable iff $E = E^p k$.
2. Show that for a given extension E/k, the perfect closure of k in E need not be perfect. Under what conditions on E is the perfect closure of every subfield perfect?
3. Let k be a field of prime characteristic p. Show that there is a tower of fields $k = k_0 \subseteq k_1 \subseteq \ldots$, such that the p-th power mapping $x \mapsto x^p$ induces an isomorphism $k_r \to k_{r-1}$, and that the union $\cup k_r$ is the perfect closure of k.
4. Let p be an odd prime and $k = \mathbf{F}_p(s, t)$, where s, t are indeterminates. Show that if α is a root of the equation

$$x^{2p} + 2sx^p + t = 0,$$

   then the perfect closure of k in $k(\alpha)$ is k itself, but that $k(\alpha)/k$ is not separable (such extensions are sometimes called *exceptional*).
5. Let Ω/k be a normal extension and E, F be any subfields of Ω that are conjugate. Show that any normal extension of k containing E also contains F.
6. Show that if F/k is generated by finitely many algebraic elements, all but at most one separable over k, then F/k is simple. (Hint. Use Steinitz' criterion, Theorem 7.9.3.)
7. For a finite p-radical extension F/k define the *height* as the least exponent r such that $F^{p^r} \subseteq k$. Show that for any finite extension F/k there exists a p-radical extension E of k, of the same height as F over k_s, such that F/k is contained in $k_s \otimes_k E$.

11.5 Derivations

It is well known and easily checked that a polynomial f has no repeated factor iff it is prime to f', its formal derivative. From this fact it is not hard to derive a criterion for the separability of simple extensions in terms of derivations, and it is reasonable to expect that derivations can also be used to study more general extensions. In a ring R generated by a set X, any derivation is uniquely determined by its values on X, but usually not every map of X into R can be extended to a derivation. We shall return to this topic in FA Chapter 2, but in this section one of our main tasks will be to find conditions for this to be possible, for certain field extensions.

The derivations used here are always (1,1)-derivations, in the terminology of Section 6.2, i.e. they are mappings $a \mapsto da$ satisfying

$$d(ab) = da.b + a.db,$$

besides being additive. We recall that the kernel of a derivation d is a subring, the *ring of constants* for d, and a derivation d of a k-algebra is said to be *over* k if $k \subseteq \ker d$. When the derivation is defined on a field, the set of constants actually forms a field.

We begin by noting how a derivation on an integral domain extends to the field of fractions.

Proposition 11.5.1. *Let R be an integral domain with field of fractions K. Then any derivation d on R can be extended in just one way to a derivation of K, again written d, and given by the formula*

$$d\left(\frac{a}{b}\right) = \frac{da.b - a.db}{b^2}. \tag{11.5.1}$$

Proof. The mapping $a \mapsto \begin{pmatrix} a & da \\ 0 & a \end{pmatrix}$ of R into $\begin{pmatrix} K & K \\ 0 & K \end{pmatrix}$ maps each non-zero element of R to an invertible element and so extends to a unique homomorphism $u \mapsto \begin{pmatrix} u & du \\ 0 & u \end{pmatrix}$. Hence $u \mapsto du$ is a derivation of K, whose precise form (11.5.1) follows by evaluating $\begin{pmatrix} a & da \\ 0 & a \end{pmatrix}.\begin{pmatrix} b & db \\ 0 & b \end{pmatrix}^{-1}$. ∎

The chain rule, familiar from calculus, can be extended to arbitrary derivations, where it takes the following form. Let $R = K[x_1, \ldots, x_n]$ be a polynomial ring in n variables over a ring K. We shall write D_1 or $\partial/\partial x_1$ for the derivation with respect to x_1; this is the derivation over K which maps x_1 to 1 and the other variables to 0. Since R may be regarded as a polynomial ring in x_1 over $K[x_2, \ldots, x_n]$, this derivation certainly exists. Now $D_i = \partial/\partial x_i$ $(i = 2, \ldots, n)$ is defined similarly. Next consider an arbitrary commutative ring R, generated by n elements $a_1, \ldots, a_n$ over K. Any element c of R may be written (possibly in more than one way) as $c = f(a_1, \ldots, a_n)$, where f is a polynomial with coefficients in K. If d is a derivation of R and f^d denotes the polynomial obtained from f by applying d to the coefficients, then the chain rule asserts

$$df(a_1, \ldots, a_n) = f^d(a_1, \ldots, a_n) + \sum_1^n D_i f(a_1, \ldots, a_n).da_i. \tag{11.5.2}$$

Here $D_i f$ is understood to mean formal differentiation with respect to the i-th argument in f. For the case when f is a monomial, (11.5.2) follows by induction on the degree, and the general case follows by linearity. In fact the result holds even for rational functions, as we see by writing $f = g/h$ and using the formula (11.5.1), but we shall not need this fact. We now establish conditions under which a derivation can be extended; the reader should compare this with the parallel discussion in Section 6.2.

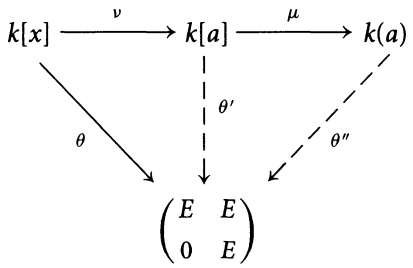
Theorem 11.5.2. *Let $E = k(a_1, \ldots, a_n)$ be a finitely generated field extension of k, denote by $\mathfrak{a}$ the ideal in $k[x_1, \ldots, x_n]$ of all polynomials $f(x_1, \ldots, x_n)$ vanishing for $x_i = a_i$ and let $\{f_\lambda(x)\}$ be a generating set for $\mathfrak{a}$. Given any derivation d on k and any $u_1, \ldots, u_n \in E$, the derivation d can be extended to a derivation d^* of E such that $d^*a_i = u_i (i = 1, \ldots, n)$ if and only if*

$$f_\lambda^d(a) + \sum_i D_i f_\lambda(a) u_i = 0. \tag{11.5.3}$$

Moreover, d^* is then uniquely determined and is given by

$$d^* f(a) = f^d(a) + \sum D_i f(a) u_i. \tag{11.5.4}$$

Proof. Consider the diagram



where ν is the reduction mod $\mathfrak{a}$ and μ is the inclusion, while θ maps x to $\begin{pmatrix} a_i & u_i \\ 0 & a_i \end{pmatrix}$. The subring of $\begin{pmatrix} E & E \\ 0 & E \end{pmatrix}$ generated by these elements over k is commutative and hence this mapping extends to a homomorphism $\theta : f(x) \mapsto \begin{pmatrix} f(a) & d'f(a) \\ 0 & f(a) \end{pmatrix}$, where d' is the unique (ν, ν)-derivation extending d such that $a_i \mapsto u_i$. Now the chain rule shows that

$$d'f(a) = f^d(a) + \sum_i D_i f(a) u_i.$$

Hence if an extension d^* is to exist, we must have $d'f = 0$ for all $f \in \mathfrak{a}$ and (11.5.3) follows. Conversely, when (11.5.3) holds for a generating set $\{f_\lambda\}$ of $\mathfrak{a}$, then $\ker \theta \supseteq \mathfrak{a}$, hence by the factor theorem we obtain a vertical dotted arrow θ', and now by Proposition 11.5.1 we obtain the arrow θ''. This is a homomorphism of the form $\theta'' : c \mapsto \begin{pmatrix} c & d^*c \\ 0 & c \end{pmatrix}$, and d^* is the required derivation. ∎

Consider in particular a simple extension $k(\alpha)$. If α is transcendental over k, there is no condition, so any derivation of k has a unique extension mapping α to a prescribed element of $k(\alpha)$. If α is algebraic over k, let f be its minimal polynomial; the condition that there should exist an extension of d mapping α to $u \in k(\alpha)$ is that

$$f^d(\alpha) + f'(\alpha) u = 0. \tag{11.5.5}$$

Now two cases can arise. If α is separable over k, then $f'(\alpha) \neq 0$ and we see that d has an extension iff $u = -f'(\alpha)^{-1} f^d(\alpha)$, and with this value the extension is unique. If α is inseparable, then $f'(\alpha) = 0$ and (11.5.5) holds precisely when $f^d(\alpha) = 0$, and u can then be chosen arbitrarily. But f^d is a polynomial of lower degree than f, for the highest coefficient of f is 1 and so maps to 0. Hence $f^d(\alpha) = 0$ only when f^d vanishes identically, i.e. when all the coefficients of f are constant. We can sum up the result as

Theorem 11.5.3. *Let k be a field with a derivation d and let α be an element of some extension field E of k. Then*

(i) *if α is transcendental over k and u is any element of E, then d extends to a unique derivation of $k(\alpha)$ into E such that $d\alpha = u$;*

(ii) *if α is separably algebraic over k with minimal polynomial f, then d extends in just one way to a derivation of $k(\alpha)$ into E and the value of $d\alpha$ then satisfies*

$$f^d(\alpha) + f'(\alpha).d\alpha = 0;$$

(iii) *if α is inseparable over k, then d can be extended to $k(\alpha)$ if and only if the minimal polynomial for α has constant coefficients, and when this is so, there is just one extension for each value of $d\alpha$, where $d\alpha$ can be chosen arbitrarily.* ■

By induction we obtain in the special case where $k \subseteq \ker d$:

Corollary 11.5.4. *Let E/k be a finitely generated field extension. Then there is no non-zero derivation of E over k if and only if E is separably algebraic over k.* ■

This leads to another useful criterion for separability:

Proposition 11.5.5. *Given a finitely generated extension $E = k(\alpha_1, \ldots, \alpha_n)$ over a field k, if there are polynomials f_i $(i = 1, \ldots, n)$ in n variables over k such that $f_i(\alpha_1, \ldots, \alpha_n) = 0$ $(i = 1, \ldots, n)$ and $\det(\partial f_i/\partial \alpha_j) \neq 0$, then E/k is separably algebraic.*

Proof. By Corollary 11.5.4 we need only show that every derivation of E over k is zero. Let d be a derivation of E/k; then $df_i(\alpha) = 0$, hence by the chain rule, $\sum_j D_j f_i(\alpha).d\alpha_j = 0$. Since $\det(D_j f_i) \neq 0$ by hypothesis, these equations have only the zero solution, hence $d\alpha_j = 0$ and d must be zero on E. ■

Exercises

1. Let k be a field of odd prime characteristic p. Show that any extension generated by α, β such that $\alpha^p + a\beta^2 = 1$, $b\alpha^2 + \beta^p = 1$ $(a, b \in k)$ is separably algebraic iff $ab \neq 0$.
2. Complete the details in the proof of Proposition 11.5.1.
3. Use Proposition 11.5.5 to establish the transitivity of separably algebraic extensions.

4. Prove Corollary 11.5.4.
5. Let E/k be a separably algebraic extension. Given an E-space V, show that any derivation of k into V has a unique extension to a derivation of E into V. (Hint. Adapt the proof of Theorem 11.5.3 and use induction on the degree.)
6. Let E/k be a finitely generated extension of characteristic p. Show that an element of E lies in $E^p k$ iff it is constant for every derivation of E/k.

11.6 Linearly Disjoint Extensions

Let k be a field and Ω be any field extension; in this section we examine the way in which different extensions of k in Ω interact. All our extensions will be algebraic and we may take Ω to be an algebraic closure of k, but this is not essential.

We recall from Section 5.4 that two subalgebras A, B of Ω are called *linearly disjoint* over k if the mapping

$$A \otimes_k B \to \Omega : \sum a_i \otimes b_i \mapsto \sum a_i b_i \tag{11.6.1}$$

is injective. We shall be concerned with conditions for linear disjointness of subalgebras, and in particular, of subextension fields of Ω/k.

It is clear that if A, B are linearly disjoint and $A' \subseteq A$, $B' \subseteq B$, then A', B' are again linearly disjoint. In the opposite direction, A, B are linearly disjoint whenever all finitely generated subalgebras of A and B are linearly disjoint; this follows because the condition involves only finite sets of elements. We also note that any subalgebra of Ω is an integral domain; if it is algebraic, it must be a field, for then any element α satisfies an equation

$$\alpha^n + c_1\alpha^{n-1} + \ldots + c_n = 0,$$

where $c_n \neq 0$ if $\alpha \neq 0$, and then $\alpha^{-1} = -c_n^{-1}(\alpha^{n-1} + c_1\alpha^{n-2} + \ldots + c_{n-1})$.

From the structure of tensor products we obtain the following equivalent conditions for linear disjointness:

Proposition 11.6.1. *For any k-algebras A and B in Ω the following conditions are equivalent:*

(a) *A, B are linearly disjoint over k;*
(b) *any k-basis of A remains linearly independent over B;*
(c) *any k-basis of B remains linearly independent over A;*
(d) *if (u_i) is a k-basis for A and (v_j) is a k-basis for B, then the products $u_i v_j$ are linearly independent over k;*
(e) *if K, L are the subfields of Ω generated by A, B respectively, then K, L are linearly disjoint over k.*

Proof. The equivalence of (a)–(d) is an immediate consequence of the independence property of the tensor product (Proposition 4.8.6). (e) $\Rightarrow$ (a) is clear, and to prove (a) $\Rightarrow$ (e), let $a_i \in K$, $b_i \in B$ be such that $\sum a_i b_i = 0$. Write $a_i = a_i'/s$, where a_i', $s \in A$; then $\sum a_i' b_i = \sum (a_i b_i)s = 0$, so if A, B are linearly disjoint, it follows that

$\sum a_i' \otimes b_i = 0$, hence $\sum a_i \otimes b_i = 0$. Thus if A, B are linearly disjoint, then so are K, B, and by another application, so are K and L. ■

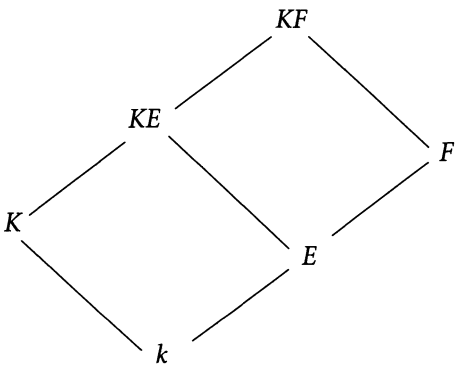
As an example, suppose that K/k is p-radical and L/k is separable, then $KL = \left\{\sum a_i b_i | a_i \in K, b_i \in L\right\}$ is separable over K, and as we shall see in Proposition 11.6.3, K and L are linearly disjoint over k (see also Further Exercise 12).

We record the following transitivity property:

Proposition 11.6.2. *Let $k \subseteq K, k \subseteq E \subseteq F$ be extension fields of k. Then K and F are linearly disjoint over k if and only if K and E are linearly disjoint over k and KE and F are linearly disjoint over E.*

Proof. We have the evident relation

$$K \otimes_k F \cong (K \otimes_k E) \otimes_E F.$$



Thus the mapping

$$K \otimes_k F \to \Omega \tag{11.6.2}$$

can be carried out in two stages, first mapping $K \otimes_k E$ to Ω, and then $KE \otimes_E F$ to Ω, and here it is immaterial, by Proposition 11.6.1, whether we take KE to be the subfield or the subalgebra generated by K and E. Thus the mapping (11.6.2) is injective iff the latter two mappings are, and this proves the assertion. ■

Proposition 11.6.3. *Given subfields K, L of Ω, if K/k is p-radical and L/k is separable, then K, L are linearly disjoint over k and L is the separable closure of k in KL.*

Proof. Put $E = KL$; since L/k is separable, so is E/K, and the result will follow from Theorem 11.4.7 if we can show that L is the separable closure of k in E. If this separable closure is denoted by k_s, then clearly $L \subseteq k_s$ and we have to prove equality. Thus, given $a \in k_s$, we must show that $a \in L$, and here we may assume K, L to be finitely generated, and hence finite over k. Any k-basis of L spans E over K, and its elements are separable over K, hence $[E : K]_s \leq [L : k]_s$. Now $[E : k]_s = [E : K]_s[K : k]_s = [E : K]_s \leq [L : k]_s \leq [E : L]_s[L : k]_s = [E : k]_s$, hence equality holds throughout, and $[E : L]_s = 1$, so L has no separable extension in E. It follows that any $a \in k_s$ lies in L, as claimed. ■

Let K, L be two algebraic subextensions of Ω/k; to find out whether K, L are linearly disjoint, we consider the natural mapping

$$K \otimes_k L \to \Omega. \tag{11.6.3}$$

If the tensor product on the left is a field, then the kernel must be trivial and so K, L are then linearly disjoint. Conversely, when K, L are linearly disjoint, then $K \otimes_k L$ can be embedded in Ω by (11.6.3). As a subalgebra of Ω generated by algebraic elements it is algebraic and hence a field. Thus we have

Proposition 11.6.4. *Two algebraic subextensions K, L of Ω/k are linearly disjoint if and only if $K \otimes_k L$ is a field.* ∎

It is clear that for linearly disjoint extensions K, L of k we must have $K \cap L = k$. In an important case this necessary condition is also sufficient.

Theorem 11.6.5. *Let K, L be two subextensions of Ω/k of which one is normal and one (possibly the same) is separable. Then K, L are linearly disjoint over k if and only if $K \cap L = k$.*

In order to prove this result we restate it in absolute form:

Theorem 11.6.5′. *Let K/k, L/k be two extensions of which one is normal and one (possibly the same) is separable. Then $K \otimes L$ is a field if and only if K, L have no isomorphic subfields properly containing k.*

Proof. Suppose that K, L each have a subfield isomorphic to a proper extension of k. We may as well take this to be simple, say $k(\alpha)$, where α is algebraic of degree $n > 1$. Let

$$c_0 x^n + c_1 x^{n-1} + \ldots + c_n = 0 \quad (c_0 = 1)$$

be the minimal equation of α over k, so $\sum c_i \alpha^{n-i} = 0$. Consider the product

$$\left(\sum_{i=0}^{n-1} \sum_{j=1}^{n-i} c_i \alpha^{n-i-j} \otimes \alpha^{j-1} \right) (\alpha \otimes 1 - 1 \otimes \alpha). \tag{11.6.4}$$

The coefficient of c_i is

$$\sum_i (\alpha^{n-i-j+1} \otimes \alpha^{j-1} - \alpha^{n-i-j} \otimes \alpha^j) = \alpha^{n-i} \otimes 1 - 1 \otimes \alpha^{n-i}.$$

Hence (11.6.4) becomes

$$\sum_i c_i \alpha^{n-i} \otimes 1 - 1 \otimes \sum_i c_i \alpha^{n-i} = 0.$$

This shows that $K \otimes L$ has zerodivisors, so it cannot be a field.

To prove the converse it will be enough to show that under the given conditions $K \otimes L$ has no zerodivisors, and for this we may take K, L to be finitely generated and

hence finite over k. Suppose first that K/k is both separable and normal, hence Galois, and so simple, say $K = k(\alpha)$, where α has the minimal polynomial f over k. Over L we have a factorization

$$f = p_1 \dots p_r, \quad p_i \text{ irreducible over } L, \tag{11.6.5}$$

and $K \otimes L$ is a field iff $r = 1$. Assume that $K \otimes L$ is not a field; then $r > 1$ and if L_1 is the subfield of L generated by the coefficients of all the p_i, then $L_1 \neq k$. Let us enlarge L to a minimal splitting field of f; this will contain a splitting field L_0 of f over k and clearly $L_0 \supseteq L_1$. But K as a Galois extension of k generated by α is also a minimal splitting field of f over k. By uniqueness, $K \cong L_0$; now L_1 is a subfield of L and is contained in L_0, hence it is isomorphic to a subfield of K. This contradicts the hypothesis and it shows the condition to be necessary.

There remains the case when one of K, L, say K is normal and L is separable. Taking K again finite, we have by Theorem 11.4.10,

$$K = k_s \otimes_k k_p, \tag{11.6.6}$$

where k_s is the separable closure and k_p is the perfect closure of k in K. Here k_s is Galois and we have

$$K \otimes_k L = k_s \otimes k_p \otimes L.$$

Under the given hypothesis k_s and L have no isomorphic subfield $\supset k$, hence $k_s \otimes L$ is a field, by what has been proved. It is generated by separable elements over k and hence is a separable extension. Therefore $k_s \otimes L \otimes k_p$ is a field, by Proposition 11.6.3, and thus $K \otimes L$ is a field, as required. ∎

It is now easy to deduce Theorem 11.6.5. If K, L are subfields of Ω having isomorphic subextensions K_1, L_1 and K is normal, then L_1 is isomorphic to K_1, hence conjugate to K_1 in Ω, by Proposition 11.4.8, and so $L_1 \subseteq K$ by the normality of K. It follows that $K \cap L \neq k$. Conversely, when $K \cap L \neq k$, then it is clear that K, L have isomorphic subextensions $\supset k$; now we can apply Theorem 11.6.5′ to complete the proof. ∎

Here is an example to show that normality cannot be omitted in Theorem 11.6.5. Let E/k be an extension of degree n without proper subextensions (e.g. E corresponds to a maximal subgroup of index n within a Galois extension). Let E' be conjugate to E within Ω but distinct from it. Then $E \cap E' = k$ but $[EE' : k] \leq n(n-1)$, hence E and E' are not linearly disjoint.

Linear disjointness may be used to study field extensions in prime characteristic p, and here it is useful to define separability in a wider setting. The remainder of this section will not be needed in what follows and so may be omitted without loss of continuity.

We recall that a commutative ring A is said to be *reduced* if it has no nilpotent elements $\neq 0$, i.e. if given $x \in A$, $x^n = 0$ for some $n > 0$ implies $x = 0$. Now a commutative k-algebra A is called *separable* if $A \otimes_k F$ is reduced for all field extensions F of k. For example, any purely transcendental extension is separable, because $k(x) \otimes F = F(x)$ for an indeterminate x, and similarly for several indeterminates.

It is clear that a k-algebra is separable iff all its finitely generated subalgebras are separable. Our first concern is to show that for fields the terminology agrees with that introduced earlier.

Proposition 11.6.6. *An algebraic field extension E/k is separable if and only if E is separable as a k-algebra.*

Proof. By the remark just made we can take E/k to be finitely generated and hence finite. If E/k is separable as field extension, it is simple (Theorem 7.9.2) and so has the form $E = k(\alpha)$, where α has a minimal polynomial f with distinct zeros in any field. Let F be an extension field of k and suppose that $f = p_1 \dots p_r$ is the complete factorization of f over F. Then $E \otimes F \cong A_1 \times \dots \times A_r$, where $A_i \cong F[x]/(p_i)$ is a field; hence $E \otimes F$ is reduced and so E is separable as k-algebra. On the other hand, if E/k is not separable as field extension, then k_s, the separable closure of k in E is distinct from E. For any $\alpha \in E \backslash k_s$ the minimal polynomial f of α over k_s has the form $f = (x - \alpha)^q$, where $q = p^t > 1$. Moreover, we have

$$E \otimes_k E \cong E \otimes_k k_s \otimes_{k_s} E,$$

and $\alpha^q \in k_s$, hence $(\alpha \otimes 1 - 1 \otimes \alpha)^q = \alpha^q \otimes 1 - 1 \otimes \alpha^q = 0$ in $(E \otimes k_s) \otimes E$, while $\alpha \otimes 1 - 1 \otimes \alpha \neq 0$. So in this case E is not separable as k-algebra. ∎

We observe that in a commutative ring the set $\mathfrak{N}$ of all nilpotent elements is an ideal, the nilradical, which may also be expressed as the intersection of all prime ideals (see Proposition 10.2.9). As a result we have

Theorem 11.6.7. *A commutative ring A is reduced if and only if it is a subring of a direct product of fields. If A is a finite-dimensional k-algebra and is reduced, then it is a direct product of fields.*

Proof. Let A be reduced; by the result quoted, we have $\cap \mathfrak{p}_\lambda = 0$, where $\mathfrak{p}_\lambda$ ranges over all prime ideals of A. The quotient algebra $A/\mathfrak{p}_\lambda$ is an integral domain and so has a field of fractions E_λ. The natural mappings $f_\lambda : A \to E_\lambda$ can be combined to a homomorphism $f : A \to \prod E_\lambda$ whose kernel is $\cap \mathfrak{p}_\lambda = 0$, so A has been embedded in a direct product of fields. The converse is clear.

If A is a finite-dimensional k-algebra which is reduced, then it is Artinian with zero radical and so can be expressed as a direct product of fields, by Wedderburn's theorem (Theorem 5.2.4). ∎

Corollary 11.6.8. *If A, B are separable k-algebras, then so is $A \otimes_k B$.*

Proof. We first show that if A is separable and B is reduced, then $A \otimes B$ is again reduced. By Theorem 11.6.7 we have an embedding $f : B \to \prod E_\lambda$, where the E_λ are fields; let $f_\lambda : B \to E_\lambda$ be the projection on E_λ. Then $1 \otimes f_\lambda : A \otimes B \to A \otimes E_\lambda$ is a homomorphism into a reduced k-algebra $A \otimes E_\lambda$. Combining these mappings we obtain a mapping $\varphi : A \otimes B \to \prod (A \otimes E_\lambda)$, which is an embedding in a reduced k-algebra; its kernel is $\cap (A \otimes \ker f_\lambda) = 0$, and so $A \otimes B$ is reduced.

Now if A and B are separable, then $B \otimes E$ is reduced for any field E, hence $A \otimes B \otimes E$ is reduced, by what we have shown, and this means that $A \otimes B$ is separable, as claimed. ∎

In characteristic 0 every field extension is separable, as algebra; this is clear by writing it as an algebraic extension of a purely transcendental extension. Now let A be a reduced k-algebra, where k is of characteristic 0. Then $A \otimes E$ is reduced for any field E, by Corollary 11.6.8, hence we obtain

Corollary 11.6.9. *Let k be a field of characteristic 0. Then every reduced k-algebra is separable; in particular every field extension is separable.* ∎

In testing for separability it is enough to take p-radical field extensions. This is made precise in

Theorem 11.6.10. *Let A be any commutative k-algebra, where k is a field of prime characteristic p, denote by k_p the perfect closure of k and put $k^{1/p} = \{x \in k_p | x^p \in k\}$. Then the following conditions are equivalent:*

(a) *A is separable;*
(b) *$A \otimes k_p$ is reduced;*
(c) *$A \otimes k^{1/p}$ is reduced;*
(d) *if a family (a_i) of elements of A is linearly independent over k, then so is the family (a_i^p).*

Proof. Since $k_p \supseteq k^{1/p}$, it follows that (a) $\Rightarrow$ (b) $\Rightarrow$ (c). To prove that (c) $\Rightarrow$ (d), take any linearly independent family $a_i \in A$ and assume that $\sum \lambda_i a_i^p = 0$. In $k^{1/p}$ there exists μ_i such that $\mu_i^p = \lambda_i$ and we have $(\sum \mu_i \otimes a_i)^p = \sum \lambda_i a_i^p = 0$, hence $\sum \mu_i \otimes a_i = 0$ by (c), and so $\mu_i = 0$. It follows that $\lambda_i = \mu_i^p = 0$, i.e. (d).

(d) $\Rightarrow$ (a). Let (u_i) be a k-basis of A and suppose that E/k is a field extension. To prove that A is separable we must show that $A \otimes E$ is reduced, i.e. $x^p = 0$ implies $x = 0$, for any $x \in A \otimes E$. We can write $x = \sum \lambda_i \otimes u_i$ $(\lambda_i \in E)$, and we have $0 = x^p = \sum \lambda_i^p \otimes u_i^p$; by (d) the u_i^p are linearly independent, so $\lambda_i^p = 0$, hence $\lambda_i = 0$ and so $x = 0$. This shows A to be separable. ∎

We can also write down an intrinsic condition for separability. Let $E = k(x_1, \ldots, x_n)$ be a finitely generated field extension; we shall say that E is *separably generated* over k if there is a transcendence basis $u_1, \ldots, u_r$ such that E is separably algebraic over $k(u_1, \ldots, u_r)$. The u_i are then called a *separating transcendence basis* for E/k. To find conditions for such a basis to exist we can of course restrict ourselves to prime characteristic p.

Theorem 11.6.11 (Mac Lane's criterion). *Let $k \subseteq E \subseteq \Omega$, where Ω is algebraically closed, of prime characteristic p and write k_p for the perfect closure of k in Ω. Then the following conditions are equivalent:*

(a) *E/k is separable;*
(b) *E is linearly disjoint from k_p over k;*
(c) *E is linearly disjoint from $k^{1/p}$ over k;*
(d) *every subfield of E containing k and finitely generated over k is separably generated.*

Here (d) is often used as general definition of separable. Condition (c) is sometimes called 'Mac Lane separable'.

Proof. (a) $\Rightarrow$ (b) $\Rightarrow$ (c) is again clear. To prove (c) $\Rightarrow$ (d), let $F = k(x_1, \ldots, x_n) \subseteq E$; then F is linearly disjoint from $k^{1/p}$ because E is. We use induction on n, the case $n = 0$ being trivial. Likewise, if the x_i are algebraically independent over k, there is nothing to prove. Otherwise choose a polynomial of least degree:

$$f(x_1, \ldots, x_n) = 0. \tag{11.6.7}$$

If f involves x_1 and not merely as p-th power, i.e. $\partial f/\partial x_1 \neq 0$, this is a separable equation for x_1 over $k(x_2, \ldots, x_n)$, thus F is separable over $k(x_2, \ldots, x_n)$, the latter is separably generated over k by induction on n, and now the result follows because separability is transitive. A similar argument holds if $\partial f/\partial x_i \neq 0$ for some $i = 2, \ldots, n$. This only leaves the case where f is a polynomial in $x_1^p, \ldots, x_n^p$. Then we can write f as $f = \sum a_\lambda M_\lambda^p$, where each M_λ is a monomial in the x_i and $a_\lambda \in k$. By the minimality of deg f the monomials M_λ are linearly independent over k, so they are also linearly independent over $k^{1/p}$, by (c), but we have the relation

$$\sum b_\lambda \otimes M_\lambda = 0,$$

where $b_\lambda^p = a_\lambda$. Hence this must be trivial, i.e. $b_\lambda = 0$, so $a_\lambda = 0$ for all λ, which is a contradiction. So this case cannot occur.

(d) $\Rightarrow$ (a). To prove (a) we can limit ourselves to finitely generated subfields, so let E be separably generated over k, say E is separably algebraic over $k(u_1, \ldots, u_r) = k(u)$, where the u_i are algebraically independent over k. We claim that $k(u)$ and $k^{1/p}$ are linearly disjoint over k. For the distinct monomials M_λ in the u_i form a basis for $k[u_1, \ldots, u_r]$ and they are linearly independent over $k^{1/p}$ because if

$$\sum c_\lambda M_\lambda = 0, \quad c_\lambda^p \in k, \tag{11.6.8}$$

then

$$\sum c_\lambda^p M_\lambda^p = 0.$$

Here the M_λ^p are distinct and hence linearly independent over k, so (11.6.8) must be trivial. Next E and $k^{1/p}(u)$ are linearly disjoint over $k(u)$, for the first is separable and the second is p-radical over $k(u)$ (Proposition 11.6.3). Now we can apply Proposition 11.6.2 with $k^{1/p}$, $k(u)$, E for K, E, F respectively, and conclude that E and $k^{1/p}$ are linearly disjoint over k; then $E \otimes k^{1/p}$ is a field and hence by Theorem 11.6.10, E is separable, so (a) holds. ■

The following immediate consequences are often useful.

Corollary 11.6.12. *Any finitely generated subextension of a separably generated extension is again separably generated.* ■

Corollary 11.6.13 (F. K. Schmidt). *If k is perfect, then any finitely generated extension is separably generated.* ■

In conclusion we examine briefly a class of separable extensions which is of importance in algebraic geometry. A field extension E/k is said to be *regular* if $E \otimes F$ is an integral domain for any field extension F/k. It is again easy to check that E/k is regular iff this is so for any finitely generated subextension. Our object is to give some alternative characterizations, but to do so we shall need a lemma. We recall Gauss's lemma (Lemma 7.7.1) which states that for any integral domain A, a prime element of A remains prime in $A[x]$. Hence for any UFD A with field of fractions K, a polynomial $f \in A[x]$ is irreducible over A iff it is irreducible over K and primitive in $A[x]$; this follows because every factorization over $K[x]$ can be pulled down to $A[x]$, i.e. $A[x]$ is inert in $K[x]$ (Theorem 7.7.2).

Lemma 11.6.14. *Let E/k be a field extension, where k is relatively algebraically closed in E. If $y_1, \ldots, y_n$ are indeterminates over E and f is a polynomial in x with coefficients in $k(y_1, \ldots, y_n)$ which is irreducible over this field, then it is irreducible over $E(y_1, \ldots, y_n)$.*

Proof (H. Flanders) We have to show that $f \in k(y_1, \ldots, y_n)[x]$ is irreducible over $E(y_1, \ldots, y_n)$. By the remark just made, we may multiply f by a unit of $k(y_1, \ldots, y_n)$ so that f becomes an irreducible element of $k[y_1, \ldots, y_n, x]$. We shall show that f is still irreducible in $E[y_1, \ldots, y_n, x]$. For convenience we denote x by y_0 and write $Y = \{y_0, \ldots, y_n\}$. Suppose that $f = gh$ in $E[Y]$; choose an integer $N > \deg f$ and consider the homomorphism $\varphi : E[Y] \to E[t]$ defined by $y_i \mapsto t^{N^i}$. We have $f^\varphi = g^\varphi h^\varphi$ and on multiplying g by a unit in E and h by its inverse, we may assume that the leading coefficients of g^φ and h^φ lie in k. By taking a complete factorization of f^φ over an algebraic closure of E we see that the coefficients of g^φ and h^φ are algebraic over k and lie in E, hence g^φ, $h^\varphi \in k[t]$, by the hypothesis on k. But φ is injective on polynomials of degree less than N, because $y_0^{c_0} \ldots y_n^{c_n} \mapsto t^C$, where $C = c_0 + c_1 N + \ldots + c_n N^n$. It follows that g and h have coefficients in k, so one of g, h must be a unit and we have proved that f is irreducible in $E[Y]$. By Gauss's lemma it now follows that f is irreducible in $E(y_1, \ldots, y_n)[x]$. ■

For any field E let us denote its algebraic closure by E_a. If $k \subseteq E$, we may clearly take k_a to be a subfield of E_a.

Theorem 11.6.15. *For any field extension E/k the following conditions are equivalent:*

(a) *E/k is regular,*
(b) *E and k_a are linearly disjoint in E_a,*
(c) *k is relatively algebraically closed in E and E is separable over k.*

Proof. (a) $\Rightarrow$ (b). Assume that E/k is regular. Then $E \otimes_k k_a$ is an integral domain,

and hence a field F, say. If F_a is an algebraic closure, then it is clear that E and k_a are linearly disjoint in F_a, and hence also in an algebraic closure of E, so (b) holds.

(b) $\Rightarrow$ (c). When E and k_a are linearly disjoint, it is clear that k is relatively algebraically closed in E, and since k_a contains a perfect closure k_p of k, Mac Lane's criterion is satisfied, hence E/k is separable.

(c) $\Rightarrow$ (a). To prove that E/k is regular we may assume that it is finitely generated; thus we may assume E to be separably generated over k. Our aim is to prove that $E \otimes F$ is a domain for any $F \supseteq k$, and here we may also assume F/k to be finitely generated, with transcendence basis Y say. If we put $K = k(Y)$, then F/K is a finitely generated algebraic extension, hence finite. Denote by K_s the separable closure of K in F; then $K_s = K[\alpha]$ for some $\alpha \in K_s$ and so $K_s = K[x]/(f)$, where f is the minimal polynomial for α over K. By Lemma 11.6.14, f is still irreducible over $E(Y)$, so $K_s \otimes_{k(Y)} E(Y) \cong E(Y)[x]/(f)$ and this is a field, which may be denoted by $E(Y)[\alpha]$. Since E/k is separably generated, it follows that $E(Y)[\alpha]/k(Y)[\alpha]$ is again separably generated. Now $F \subseteq (K_s)_p$, therefore by Mac Lane's criterion, $F \otimes_{K_s} E(Y)[\alpha]$ is a domain. But

$$F \otimes_k E = F \otimes_{K_s} (K_s \otimes_K (K \otimes_k E)) \subseteq F \otimes_{K_s} (K_s \otimes_K E(Y)) = F \otimes_{K_s} E(Y)[\alpha],$$

and it follows that $F \otimes_k E$ is a domain. Therefore E/k is regular, as claimed. ∎

The reader is advised to draw a diagram.

For an algebraically closed ground field the conclusion can clearly be simplified:

Corollary 11.6.16. *If k is algebraically closed, then any extension of k is regular.* ∎

Comparing Theorem 11.6.11 and Theorem 11.6.15 we see that whereas *separability* of E/k means that E is linearly disjoint from the p-radical elements over k, *regularity* means that E is linearly disjoint from all the algebraic elements over k. In particular, the notion of regularity is not vacuous even in characteristic zero.

Exercises

1. Show that in a finitely generated extension which is separably generated, every generating set contains a separating transcendence basis.
2. Let $K = k(x, y)$, where x, y are transcendental over k and satisfy $x^2 = y^3 - 1$. Show that if char $k = 2$, then $K/k(y)$ is not separable, but that $K/k(x)$ is and hence K/k is separably generated. If char $k = 2$ and α, β are non-squares in k, show that $k(x, y)$, where $\alpha x^2 + \beta y^2 = 1$, is not separably generated over k.
3. Let K/k be a function field of one variable, i.e. a finitely generated extension of transcendence degree 1. Show that if k is perfect, then for any $x \in K$ transcendental over k there exists $y \in K$ such that $K = k(x, y)$. If k is imperfect, say $\alpha \notin k^p$, where char $k = p$ and $K = k(x, y)$, where $x^p + y^p = \alpha$, show there is no such pair including x^p which generates K.
4. Show that a root α of an irreducible equation with insoluble group over $\mathbf{Q}$ is such that α^n is irrational for all $n \neq 0$. With such α and a transcendental element

u of $\mathbf{C}$ put $E = \mathbf{Q}(u)$, $F = \mathbf{Q}(\alpha u)$. Verify that $E \cap F = \mathbf{Q}$ but that E, F are not linearly disjoint over $\mathbf{Q}$. (Hint. To show that $E \cap F = \mathbf{Q}$ use formal Laurent series in u.)

5. Two extensions E, F of k, both contained in some field Ω, are said to be *free* over k if any set of elements of E which is algebraically independent over k is also algebraically independent over F. Show that if E, F are free over k, then so are F, E.
6. Show that if E, F are linearly disjoint over k, then they are free over k. Does the converse hold? (Hint. Any algebraic extensions of k are free over k.)
7. Show that if E, F are free over k and E is separably generated over k, then EF is separably generated over F.
8. Show that if E/F and F/k are regular, then so is E/k.
9. Let A be an integral domain with field of fractions K and suppose that A has an irreducible element (i.e. an atom) c such that $A/(c)$ is not reduced. Show that $A[x]$ has irreducible polynomials which become reducible over K.
10. Let $F = \mathbf{F}_p$, $E = F(c_0, \ldots, c_{p-1}, a)$, where the c_i and a are indeterminates, and define $k = F(c_0^p, \ldots, c_{p-1}^p, a)$, $K = k(\alpha, \beta)$, where α is a root of $x^p = a$ and β is a root of $x^p = \sum_0^{p-1} c_i^p a^i$. Show that K/k, E/k are both normal, but $K \otimes_k E$ has zerodivisors.
11. Show that an extension E/k is separable iff $E \otimes_k E$ is reduced.
12. An extension E/k is called *self-regular* if $E \otimes E$ is an integral domain. Show that if E/k is self-regular, then k is relatively algebraically closed in E. (Hint. Look at the proof of Theorem 11.6.5.)
13. Find an example of a self-regular extension which is not regular.
14. Show that if E, F are free over k, then $E \cap F$ is algebraic over k.
15. Let Ω/k be a field extension, E, F be subextensions and X be a transcendence basis of E/k. Show that E, F are free over k iff $k(X)$ and F are linearly disjoint over k.

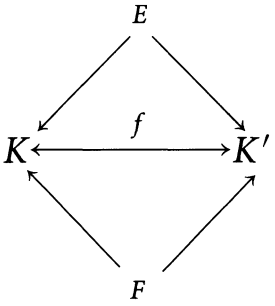
11.7 Composites of Fields

In the last section we discussed how two extensions of k which are both subfields of a given field can interact. We now consider the case where there is no common field containing them and examine what fields can be formed to contain them both.

Given two fields E, F with a common subfield k, we can always find a common extension field: the tensor product $R = E \otimes_k F$ is a commutative algebra, and if $\mathfrak{m}$ is a maximal ideal, then $R/\mathfrak{m} = K$ is a field with homomorphisms of E and F into K, given by $x \mapsto [x \otimes 1]$, $y \mapsto [1 \otimes y]$, where $x \in E$, $y \in F$ and $[\ ]$ denotes the residue class mod $\mathfrak{m}$. Like every homomorphism between fields, these mappings are embeddings, so E and F have been embedded in K. At this point three questions arise. Firstly we shall want to know how far K is determined by E and F; secondly we may ask under what conditions $E \otimes F$ is itself a field or at least an integral domain; and thirdly there is the question of the role played by the common subfield k.

The third point is easily dealt with. If two fields E and F are to be subfields of a third, they must have the same characteristic, hence their prime subfields are isomorphic and so may be identified. This shows that it represents no loss of generality to assume that E and F have a common subfield. The second question will be answered in Proposition 11.7.3 and we now turn to the first question, concerning the different possible composites; we have to begin by defining this term.

Let E, F be two extension fields of k. By a *composite* of E and F over k we understand a field extension K/k with k-homomorphisms $E \to K$, $F \to K$ such that K is generated (as a field) by the images of E and F. Two composites K and K' are said to be *equivalent* over k if there is a k-isomorphism $f : K \to K'$ such that the diagram below commutes:



Our first result gives a survey of the different composites. Given an integral domain R, we shall write $\mathscr{F}(R)$ for its field of fractions.

Theorem 11.7.1. *Let k be any field and E, F be two extension fields of k. Then there exists a composite of E and F over k. Moreover, the equivalence classes of composites of E and F over k correspond to the different prime ideals in $R = E \otimes_k F$.*

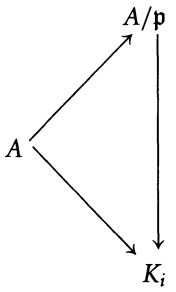
Proof. We have already seen that composites arise as homomorphic images of R by maximal ideals, which exist by Krull's theorem (Theorem 4.2.6). Now let a composite K of E and F be given, with embeddings $\alpha : E \to K$, $\beta : F \to K$. Then there is a homomorphism $\gamma : R \to K$, given by $\gamma = \alpha \otimes \beta$. If $\mathfrak{p} = \ker \gamma$, then $R/\mathfrak{p}$ is embedded in K, hence an integral domain, so $\mathfrak{p}$ is a prime ideal in R, and from the definition of K we see that $K = \mathscr{F}(R/\mathfrak{p})$. Conversely, every prime ideal $\mathfrak{p}$ of R gives rise to a composite $\mathscr{F}(R/\mathfrak{p})$ of E and F in this way. For $\mathscr{F}(R/\mathfrak{p})$ is a field and we have k-homomorphisms from E and F to $\mathscr{F}(R/\mathfrak{p})$ by combining the inclusion in R with the residue class mapping. Moreover, equivalent composites clearly have the same kernel and conversely, the kernel determines the composite up to equivalence. ■

If one of E, F is finite-dimensional over k, the result can still be simplified. We shall need a lemma on general commutative algebras, which uses the results of Section 5.3:

Lemma 11.7.2. *Any finite-dimensional non-trivial commutative k-algebra A has only finitely many prime ideals, all maximal,* $\mathfrak{m}_1, \ldots, \mathfrak{m}_r$ *say. The quotients* $A/\mathfrak{m}_i = K_i$ *are extension fields of k and if* $\mathfrak{N} = \mathfrak{m}_1 \cap \ldots \cap \mathfrak{m}_r$ *is the radical of A, then*

$$A/\mathfrak{N} \cong K_1 \times \ldots \times K_r \tag{11.7.1}$$

Proof. Any finite-dimensional k-algebra which is an integral domain must be a field (by Proposition 5.5.1), therefore any prime ideal in A is maximal. The radical $\mathfrak{N}$ of A is the intersection of all its maximal ideals and is nilpotent; further, $A/\mathfrak{N}$ is semi-simple and so is a direct product of a finite number of fields, by Wedderburn's theorem (Theorem 5.2.4), so we obtain a representation (11.7.1). The homomorphism from A to the right-hand side of (11.7.1) combined with the projection on K_i is a homomorphism onto K_i with kernel $\mathfrak{m}_i$, a maximal ideal, and $\mathfrak{N} = \mathfrak{m}_1 \cap \ldots \cap \mathfrak{m}_r$ is the kernel of the homomorphism to the product on the right of (11.7.1). It remains to show that there are no other prime ideals. Let $\mathfrak{p}$ be any prime ideal of A; then $A/\mathfrak{p}$ is an integral domain, hence a field, and the mapping $A \to A/\mathfrak{p}$ induces on each K_i either an isomorphism or the zero mapping.



It must be an isomorphism for at least one i; thus we have the commutative triangle shown and it follows that the kernels of the residue class mappings agree; hence $\mathfrak{p} = \mathfrak{m}_i$. Thus $\mathfrak{m}_1, \ldots, \mathfrak{m}_r$ are the only prime ideals of A. ∎

If $[F : k] < \infty$, then $E \otimes F$ is a finite-dimensional E-algebra and we can apply Lemma 11.7.2 with E for k:

Proposition 11.7.3. *Let E, F be fields over k, one of which, say F, is finite-dimensional over k. Then the number of composites of E and F is finite, say* $K_1, \ldots, K_r$ $(r \geq 1)$, *and*

$$(E \otimes_k F)/\mathfrak{N} \cong K_1 \times \ldots \times K_r, \tag{11.7.2}$$

where $\mathfrak{N}$ *is the radical of* $E \otimes_k F$. *In particular, E, F have a single composite precisely if* $E \otimes F$ *has a prime ideal consisting of nilpotent elements.* ∎

Specializing further, let us take F/k to be a simple extension, say $F = k(a)$, where a has minimal polynomial f over k. Then $F = k[x]/(f)$, so we have an exact sequence

$$0 \to (f) \to k[x] \to F \to 0.$$

Tensoring with E we obtain the exact sequence

$$0 \to (f) \to E[x] \to E \otimes F \to 0,$$

and so $E \otimes F \cong E[x]/(f)$. Let the decomposition of f into irreducible factors over E be $f = p_1^{\alpha_1} \dots p_r^{\alpha_r}$; then the maximal ideals in $E \otimes F$ are (p_i), $i = 1, \dots, r$, and their intersection is $\mathfrak{N} = (p_1 \dots p_r)/(f)$. In particular, if F/k is separable, then all the α's are 1 and $\mathfrak{N} = 0$. In that case (11.7.2) shows that

$$[F : k] = [E \otimes F : E] = \sum_i [K_i : E].$$

Hence we obtain

Corollary 11.7.4. *If F/k is a finite separable field extension, E/k is an arbitrary field extension and $K_1, \dots, K_r$ are the inequivalent composites of E and F over k, then the K_i correspond to the irreducible factors over E of the minimal polynomial for a generator of F over k, and*

$$E \otimes_k \mathrm{F} \cong \mathrm{K}_1 \times \dots \times \mathrm{K}_\mathrm{r}; \tag{11.7.3}$$

moreover, $[F : k] = \sum [K_i : E]$.

Proof. This follows from what has been said, if we remember that any finite separable field extension is simple (see Section 7.9). ■

In particular, if $K_i = E$ for all i, then (11.7.3) expresses $E \otimes F$ as a direct product of copies of E; we then say that *E/k splits* F.

When F/k is a Galois extension and $E = F$, then the minimal polynomial for a generator of F splits into linear factors over F (because F/k is normal); hence we obtain

Corollary 11.7.5. *If F/k is a finite Galois extension of degree n, then*

$$F \otimes_k F \cong F_1 \times \dots \times F_n, \quad \text{where } F_i \cong F.$$ ■

Exercises

1. Show that if $E \otimes_k F$ is a field, then either E or F must be algebraic over k.
2. Show that two fields can be embedded in the same field iff they have the same characteristic.
3. Let E, F be finite separable field extensions of k, with generators having minimal polynomials f, g respectively. Show that f splits into as many irreducible factors over F as g does over E. Deduce that for any irreducible separable polynomials f, g over k, f remains irreducible over $k[x]/(g)$ iff g remains irreducible over $k[x]/(f)$. What happens when f, g are no longer separable?
4. Let $E = k(\alpha), F = k(\beta)$, where $\alpha^n = \beta^n = a \in k$ and n is prime to char k. Show that an idempotent in $E \otimes F$ is $\frac{1}{n}(1 + \lambda + \dots + \lambda^{n-1})$, where $\lambda = \alpha/\beta$.
5. Let F/k be a finite separable field extension of degree > 1, let f be the minimal polynomial of a generator of F and let E be a minimal splitting field of f over

k. Suppose further that the Galois group of f over k is 2-fold transitive. Show that E and F have composites that are k-isomorphic as fields, but inequivalent.

6. Let F/k be a finite separable field extension of degree > 1 and suppose that the minimal polynomial for a generator of F has a 2-fold transitive Galois group. Show that $F \otimes F$ is the direct product of F and another field. Illustrate this by considering $x^3 - 2$ over $\mathbf{Q}$.
7. Show that if E/k is purely inseparable and F/k arbitrary, then $E \otimes F$ is primary. Deduce that there is just one composite of E and F over k. If moreover, F/k is separable, then E, F are linearly disjoint in this composite.

11.8 Infinite Algebraic Extensions

We recall from Section 7.3 that a field K is said to be *algebraically closed* if every polynomial equation $f(x) = 0$ over K has a root in K. By the remainder theorem this amounts to requiring every polynomial (in one variable) of positive degree over K to have a linear factor, and an induction on the degree shows that then every polynomial in x over K can be written as a product of linear factors. Given a field k, an algebraic closure k_a is a field containing k which is algebraic over k and algebraically closed. In Section 7.3 we saw that every field has an algebraic closure. Below we shall give another existence proof. We begin by establishing the uniqueness.

Proposition 11.8.1. *Let k be any field. Then any two algebraic closures of k are isomorphic.*

Proof. Let K, K' be algebraic closures of k and consider the algebra $A = K \otimes_k K'$. This is a k-algebra containing K, via the embedding $c \mapsto c \otimes 1$, and likewise K', and A is generated by them; hence every element is algebraic over k. Let $\mathfrak{m}$ be a maximal ideal of A and consider the field $L = A/\mathfrak{m}$. The natural homomorphism $K \to A \to L$ is an embedding, because the kernel is a proper ideal of K. Thus L again contains K and K' and is generated by them and is algebraic over k. Let $c \in L$ and consider its minimal polynomial f over k. By hypothesis f splits into linear factors over K : $f(x) = \prod (x - \alpha_i)$; hence $\prod (c - \alpha_i) = 0$ and since L is a field, we have $c = \alpha_i$ for some i. Thus $c \in K$ and this shows that $K = L$. By symmetry we also have $K' = L$, so the identity mapping of L is a k-isomorphism between K and K'. ∎

Next we note a property which will facilitate the existence proof of the algebraic closure.

Proposition 11.8.2. *Let E/k be an algebraic extension. If every polynomial equation over k has a root in E, then E is an algebraic closure of k.*

Proof. We have to show that every polynomial over k splits into linear factors over E, and here we may clearly limit ourselves to irreducible polynomials. Let f be an irreducible polynomial over k and denote its minimal splitting field by F. Since F/k is normal, we can write $F = L \otimes_k M$, where L is Galois and M is p-radical over k,

by Theorem 11.4.10. By the theorem of the primitive element (Theorem 7.9.2) L/k is generated by a single element α, say. The minimal polynomial g of α over k is irreducible and so has a zero β in E; moreover, $k(\beta) \cong k(\alpha) = L$, so E contains a subfield isomorphic to L. Any element γ of M is p-radical over k, and so is a root of an equation $x^q = c$ over k, where $q = p^r$. It follows that there is a unique element γ' of E with the same minimal equation as γ, and the correspondence $\gamma \mapsto \gamma'$ is clearly a homomorphism $M \to E$. Thus E contains a subfield isomorphic to M. By Proposition 11.6.3, E contains a subalgebra isomorphic to $L \otimes M \cong F$, and it follows that f splits over E. Now any element algebraic over E is algebraic over k, and so by what has been proved, belongs to E; hence E is algebraically closed. It is also algebraic over k and hence is an algebraic closure of k. ∎

Theorem 11.8.3 (Steinitz). *Let k be any field. Then k has an algebraic closure and any two algebraic closures are isomorphic over k.*

Proof (E. Artin) Let $\mathscr{F}$ be the set of all irreducible polynomials over k of degree at least two, and let $X = \{x_f\}(f \in \mathscr{F})$ be a family of indeterminates indexed by $\mathscr{F}$. We form the polynomial ring $R = k[X]$ in these indeterminates and consider the ideal $\mathfrak{a}$ of R generated by the elements $f(x_f), f \in \mathscr{F}$. This ideal is proper, for if not, we would have an equation

$$1 = u_1 f_1(x_{f_1}) + \ldots + u_n f_n(x_{f_n}), \quad \text{where } u_i \in R. \tag{11.8.1}$$

If we adjoin to k zeros $\alpha_1, \ldots, \alpha_n$ of $f_1, \ldots, f_n$ respectively, we obtain an algebraic extension k' of k and replacing x_{f_i} in (11.8.1) by α_i and x_f by 0 for $f \neq f_i$, we obtain the equation $1 = 0$ in k', a contradiction. Hence $\mathfrak{a}$ is proper; it is therefore contained in a maximal ideal $\mathfrak{m}$ of R (Theorem 4.2.6). The residue class ring $K = R/\mathfrak{m}$ is a field because $\mathfrak{m}$ was maximal, and the homomorphism $k \to k[X] \to K$ obtained by composing the inclusion with the natural mapping is an embedding of fields. Now in K every irreducible polynomial over k has a zero, the zero of f being the image of x_f in K. Hence every polynomial over k has a zero in K; moreover, K is generated by the images of the x_f, algebraic over k, and so K is itself algebraic over k. By Proposition 11.8.2, K is an algebraic closure of k, and by Proposition 11.8.1 it is unique up to isomorphism. ∎

We shall usually denote an algebraic closure of k by k_a. Although any two algebraic closures of k are isomorphic, the isomorphism between them is not generally unique, for k_a usually has many k-automorphisms. To study these automorphisms let us define an *algebraic Galois extension* or simply *Galois extension* as an extension K/k, where k is the fixed field of a group G of automorphisms of K, acting with finite orbits. It follows that K/k is algebraic, for if $c \in K$ lies in the orbit $\{c_1, \ldots, c_r\}$, then $f = \prod (x - c_i)$ is a polynomial over k with c as zero. As in the finite case, an algebraic extension K/k is said to be *separable* if every element of K is separable over k.

Our first task is to extend the characterization of Galois extensions to the infinite case. We shall again write $\mathrm{Gal}(K/k)$ for the group of k-automorphisms of a Galois extension K/k:

Theorem 11.8.4. *For any algebraic extension K/k the following conditions are equivalent:*

(a) *K/k is Galois,*
(b) *K/k is normal and separable,*
(c) *K is the minimal splitting field of a family of separable polynomials over k.*

Proof. (a) $\Rightarrow$ (b). Let f be an irreducible monic polynomial with a zero $\alpha \in K$. If the orbit of α under the group of K/k is $\{\alpha_1, \ldots \alpha_r\}$, we form $g = \prod (x - \alpha_i)$; now g is fixed by all k-automorphisms of K and so has coefficients in k, hence $f \mid g$, but g was constructed as a factor of f, therefore $f = g$. This shows that f splits over K and has distinct zeros, so K/k is normal and separable.

(b) $\Rightarrow$ (c). Let K be generated as k-algebra by a family $\{\alpha_i\}$ of separable algebraic elements, and denote the minimal polynomial of α_i over k by f_i. If K/k is normal, f_i splits over K, hence K is the minimal splitting field of the family $\{f_i\}$ of separable polynomials.

(c) $\Rightarrow$ (a). Let G be the group of all k-automorphisms of K and suppose that K is the minimal splitting field of a family $\{f_i | i \in I\}$ of separable polynomials. Any element α of K clearly lies in the minimal splitting field E of some finite subfamily $\{f_i | i \in I'\}$. Now E/k is normal and by hypothesis also separable, hence Galois. So if $\alpha \notin k$, then there exists $\sigma \in \mathrm{Gal}(E/k)$ such that $\alpha^\sigma \neq \alpha$. By Proposition 11.4.8, σ can be extended to a k-automorphism σ' of K. It follows that the fixed field of G is k; moreover, the conjugates of any element satisfy the same equation over k and so are finite in number. Hence G acts with finite orbits and K/k is Galois. ■

Corollary 11.8.5. *Any separable extension E/k is contained in a Galois extension K/k.*

Proof. Let E be generated by $X = \{x_i\}$, take the minimal polynomial f_i of x_i over k and let K be a minimal splitting field of the f_i over k. Then K is a field containing E and K/k is normal and separable, hence Galois, by Theorem 11.8.4. ■

Let us now investigate how finite Galois theory carries over to the algebraic case. The results are best stated in topological terms, but only a basic acquaintance with topological notions will be needed. Given an algebraic Galois extension K/k with group G, we define for any intermediate field L,

$$L^* = \{\sigma \in G | x^\sigma = x \text{ for all } x \in L\}, \tag{11.8.2}$$

and for any subgroup H of G we put

$$H^* = \{x \in K | x^\sigma = x \text{ for all } \sigma \in H\}. \tag{11.8.3}$$

As in the finite case we have a Galois connexion

$$L_1 \subseteq L_2 \Rightarrow L_1^* \supseteq L_2^*, \tag{11.8.4}$$

$$H_1 \subseteq H_2 \Rightarrow H_1^* \supseteq H_2^*, \tag{11.8.5}$$

$$L \subseteq L^{**}, H \subseteq H^{**} \tag{11.8.6}$$

and hence

$$L^{***} = L^*, H^{***} = H^*. \tag{11.8.7}$$

Moreover, K/L is Galois, by Theorem 11.8.4(b), so $L^* \cong \mathrm{Gal(K/L)}$ and thus $L^{**} = L$. However, in contrast to the finite case, H^{**} need not equal H, as we shall soon see. Let us therefore denote H^{**} by $H^\blacklozenge$ and call it the *closure* of H in G. If $H^\blacklozenge = H$, H is said to be *closed* in G. We can now state the fundamental theorem for algebraic Galois extensions.

Theorem 11.8.6. *Let K/k be an algebraic Galois extension with group G. Then the mappings $L \mapsto L^*$, $H \mapsto H^*$ define a Galois connexion between the set of all fields between k and K and the set of all closed subgroups of G. An extension L/k is normal if and only if L^* is a normal subgroup of G, and when this is so, then $\mathrm{Gal}(L/k) \cong G/L^*$.*

Every closed subgroup H of finite index in G corresponds to a finite extension L such that $[L : k] = (G : H)$. Moreover, if $\mathscr{F}$ denotes the set of all closed normal subgroups of finite index in G, then for any subgroup H of G, its closure is given by

$$H^\blacklozenge = \bigcap_{N \in \mathscr{F}} HN. \tag{11.8.8}$$

Proof. By the definition of closure we see that $H^{**} = H$ for any closed subgroup H of G. If L is an intermediate field, then L^* is a closed subgroup by (11.8.7), and we have seen that $L^{**} = L$, so we have a Galois connexion.

For any $\sigma \in G$, $(L^\sigma)^* = \sigma^{-1}L^*\sigma$, so L/k is normal (and hence Galois) iff $L^* \lhd G$. When this holds, there is a map $G \to \mathrm{Gal}(L/k)$ defined by $\sigma \mapsto \sigma|L$; clearly this is a homomorphism with kernel L^*. As in the proof of Theorem 11.8.4 we see that this homomorphism is surjective, so we conclude that $G/L^* \cong \mathrm{Gal}(L/k)$.

Suppose now that L/k is a finite extension, where $L \subseteq K$. We can find a finite Galois extension E/k such that $L \subseteq E \subseteq K$. By the previous result, $\mathrm{Gal}(E/k) \cong G/E^*$ and $\mathrm{Gal}(E/L) \cong L^*/E^*$, so by finite Galois theory, $[L : k] = (G/E^* : L^*/E^*) = (G : L^*)$. Conversely, if H is a closed subgroup of finite index in G, then the cosets $H\sigma$ give rise to distinct k-homomorphisms $H^* \to K$, hence $(G : H) \leq [H^* : k]_s$. Similarly distinct k-homomorphisms $H^* \to K$ give rise to distinct cosets of $H^{**} = H$ in G, therefore $[H^* : k]_s \leq (G : H^{**}) = (G : H)$.

Finally, let H be an arbitrary subgroup of G. If $N \in \mathscr{F}$, then N is closed of finite index in G, hence it is open (as the complement of a finite collection of closed subsets) and so HN is also open and hence closed of finite index. Therefore $H^\blacklozenge \subseteq HN$, and it follows that $H^\blacklozenge \subseteq \cap HN$, where the intersection is taken over all $N \in \mathscr{F}$. To prove equality, suppose that $\sigma \notin H^\blacklozenge$, i.e. for some $\alpha \in H^*$, $\alpha^\sigma \neq \alpha$. Let E/k be a finite Galois extension such that $k(\alpha) \subseteq E \subseteq K$; then $E^* \in \mathscr{F}$, but H and E^* fix α, hence $\sigma \notin HE^* \supseteq \cap HN$, and this establishes equality in (11.8.8). ∎

We remark that the 'closed' subgroups arise indeed as the closed subgroups of a certain topology on G in which the group operations are continuous, i.e. we have a topological group. This topology, called the *profinite* or *Krull topology*, may be defined by taking as the base for the open sets the collection of all cosets $H\sigma$,

where $H \in \mathscr{F}$, $\sigma \in G$. Since any algebraic extension is the direct limit of its finite extensions, its Galois group is the inverse limit of a system of finite groups, i.e. a *profinite* group. As an example let us show that the topology of a Galois group is always Hausdorff. This means that distinct points have disjoint neighbourhoods, and it will follow if we show that $\cap\mathscr{F} = 1$. For if $\sigma \neq \tau$, then $\mathscr{F}$ contains H such that $\sigma\tau^{-1} \notin H$, hence $H\sigma \cap H\tau = \varnothing$ and we have found disjoint neighbourhoods for σ, τ. Now $\cap\mathscr{F} = 1$ just expresses the fact that $1 = K^*$ is closed (Theorem 11.8.6). We also remark that G is compact, as an inverse limit of finite groups. On the other hand, not every subgroup of finite index need be closed (see Exercise 8).

Any profinite group of automorphisms of a field gives rise to a Galois group, under suitable conditions, by the following result.

Proposition 11.8.7. *Let K be a field with a profinite group G of automorphisms and fixed field k. If the stabilizer of any $x \in K$, $S(x) = \{\sigma \in G | x^\sigma = x\}$ is an open subgroup of G and $\cap S(x) = 1$, then $G = \mathrm{Gal}(K/k)$.*

Proof. For any finite set of elements $a_1, \ldots, a_n \in K$, the subgroup $H = S(a_1) \cap \ldots \cap S(a_n)$ is again open, hence so is the intersection N of all its conjugates. The finite group G/N acts faithfully on $E = k(a_1^G, \ldots, a_n^G)$ and has fixed field k. Hence E/k is a finite extension with group G/N. Now K is the union of all such fields E, and the intersection of all such subgroups N is 1; hence $\mathrm{Gal}(K/k) \cong \lim_{\leftarrow} \mathrm{Gal}(E/k) \cong \lim_{\leftarrow} G/N \cong G$. ∎

As an illustration of algebraic Galois extensions let us take the field $\mathbf{F}_p$ and form its algebraic closure $(\mathbf{F}_p)_a$. We know from Section 7.8 that any finite extension of $\mathbf{F}_p$ has the form $\mathbf{F}_q$, where $q = p^n$ for some $n \geq 1$, and this field may be described as the minimal splitting field of $x^q - x$. Since any algebraic extension is determined by its finite subextensions, every subextension of $(\mathbf{F}_p)_a$ is the minimal splitting field of a family $\{x^{p^n} - x | n \in I\}$for some family I of positive integers. From the results of Section 7.8 it is clear that we do not change the field by assuming I to be closed under taking factors and LCMs, and we then have a bijection between such sets of integers and algebraic extensions of $\mathbf{F}_p$. To describe this situation concisely we introduce the notion of a supernatural number. Let $p_1, p_2, \ldots$ be all the prime numbers in some order. By a *supernatural number* (also called *Steinitz number*) we understand a formal product

$$N = \prod p_i^{\alpha_i} \quad (\alpha_i \in \mathbf{N}_0 \text{ or } \alpha_i = \infty). \tag{11.8.9}$$

When all the α_i are finite and almost all are 0, this reduces to an ordinary positive integer. Given two supernatural numbers, N given by (11.8.9) and $P = \prod p_i^{\beta_i}$, we write $N|P$ to mean $\alpha_i \leq \beta_i$ for all i. It is easy to see that any supernatural number is completely determined by its finite divisors. Hence an algebraic extension E of $\mathbf{F}_p$ corresponds to a unique supernatural number N such that E has a finite subfield of degree n over $\mathbf{F}_p$ iff $n|N$, and in this way the algebraic extensions of $\mathbf{F}_p$ are described by supernatural numbers; conversely, to each supernatural number there corresponds such an extension.

Let $G = \mathrm{Gal}((\mathbf{F}_p)_a/\mathbf{F}_p)$; to give an example of a non-closed subgroup of G, consider the subgroup H generated by the Frobenius endomorphism $x \mapsto x^p$. Clearly $H^* = \mathbf{F}_p$, hence $H^{**} = G$; but for any infinite algebraic extension $E/\mathbf{F}_p$ such that $E \neq (\mathbf{F}_p)_a$, Theorem 11.8.4 shows that there exists $\sigma \in G$ such that $\sigma \neq 1$ and σ fixes E. It follows that $\sigma \notin H$ and so $H \neq G$. We note that by Theorem 11.8.6, $H^\blacklozenge = G$; this may be expressed by saying that H is *dense* in G.

Exercises

1. A group is called *locally cyclic* if every finitely generated subgroup is cyclic. Show that a locally cyclic torsion group is of the form $A = \oplus A_p$, where A_p, the p-primary component, is either cyclic of order p^m (for some $m \geq 0$) or of type $\mathbf{Z}(p^\infty)$. Deduce that each locally cyclic torsion group can be described up to isomorphism by the supernatural number whose factors are the orders of its cyclic subgroups.
2. Let K be any field. Show that the torsion subgroup of $K^\times$ is locally cyclic of type N, where N is divisible by 2 if char $K = 0$, and N is of the form

$$N = \lim_{m|M} (p^m - 1),$$

   (where M is supernatural) if char $K = p \neq 0$. Verify that all supernatural numbers can occur.
3. Show that any union of a tower of abelian extensions is abelian.
4. Let p be a prime and $k = k_0 \subset k_1 \subset \ldots$ be a tower of fields such that k_n/k is cyclic of degree p^n. Show that $K = \cup k_n$ is an abelian extension of k and that all finite quotients of $\mathrm{Gal}(K/k)$ have p-power order. Verify that $\mathrm{Gal}(K/k) = \mathbf{Z}_p$, the additive group of p-adic integers.
5. Show that $\mathrm{Gal}((\mathbf{F}_p)_a/\mathbf{F}_p) = \hat{\mathbf{Z}} = \lim_{\leftarrow} \mathbf{Z}/n$, and deduce that the dual of this group is $\mathrm{Hom}(\hat{\mathbf{Z}}, \mathbf{Q}/\mathbf{Z}) \cong \mathbf{Q}/\mathbf{Z}$.
6. Show that $\hat{\mathbf{Z}} = \lim_{\leftarrow} \mathbf{Z}/n$ is a topological ring and $\hat{\mathbf{Z}} \cong \prod \mathbf{Z}_p$ is a ring homomorphism. Verify that every closed subgroup of $\hat{\mathbf{Z}}$ is an ideal.
7. Let E/k be a Galois extension with group G. Given $u_1, \ldots, u_r \in E$, if

$$\sum b_i(u_i^\alpha - u_i^\beta) = 0,$$

   for fixed $b_i \in E$ not all 0, a fixed $\alpha \in G$ and all $\beta \in G$, then $1, u_1, \ldots, u_r$ are linearly dependent over k.
8. Let E be the extension of $\mathbf{Q}$ obtained by adjoining all square roots. Show that $G = \mathrm{Gal}(E/\mathbf{Q})$ is a 2-group, uncountable, hence with an uncountable basis, as vector space over $\mathbf{F}_2$. Thus G has uncountably many subgroups of index 2, but $\mathbf{Q}$ has only countably many quadratic extensions. Deduce that G has non-closed subgroups of finite index.

11.9 Galois Descent

Let E be a field with an automorphism σ and let V be a vector space over E. By a *σ-semilinear transformation* of V we understand a mapping $\alpha_\sigma : V \to V$ such that

$$(x+y)\alpha_\sigma = x\alpha_\sigma + y\alpha_\sigma, \quad x, y \in V, \tag{11.9.1}$$

$$(\lambda x)\alpha_\sigma = \lambda^\sigma x\alpha_\sigma, \quad \lambda \in E. \tag{11.9.2}$$

To give an example, taking k to be a subfield fixed by σ, let V_0 be a vector space over k and put $V = E \otimes_k V_0$. In terms of a basis (u_i) of V_0 we can write every element of V uniquely as $x = \sum \xi_i u_i$, where $\xi_i \in E$. Now the mapping α_σ defined by

$$x\alpha_\sigma = \sum \xi_i^\sigma u_i \tag{11.9.3}$$

is σ-semilinear, as is easily verified. Suppose that E/k is a Galois extension with group G, and for some k-space V_0 the mappings α_σ for all $\sigma \in G$ are defined on $V = E \otimes_k V_0$ by (11.9.3). Then V_0 can be recovered as the set of elements of V fixed by α_σ for all $\sigma \in G$; for clearly, by (11.9.3), $x\alpha_\sigma = x$ holds iff $\xi_i^\sigma = \xi_i$ and this holds for all $\sigma \in G$ iff $\xi_i \in k$. This process of passing from V to V_0 by means of the α_σ is called *Galois descent.* A vector space V over E which is of the form $V = E \otimes_k V_0$ for some k-space V_0 is called a space with a *k-form V_0 and any k-subspace fixed by G* is said to be *k-rational.* In such a vector space we always have a family of semilinear mappings α_σ defined as in (11.9.3), and moreover, as is easily seen,

$$\alpha_{\sigma\tau} = \alpha_\sigma \alpha_\tau, \quad \alpha_1 = 1. \tag{11.9.4}$$

Conversely, any such family of mappings on an E-space V defines a k-form on V. For the proof we recall that if $u_1, \ldots, u_n$ is a basis of a finite Galois extension E/k, then the matrix (u_i^σ) $(\sigma \in G)$ is invertible. Given a vector space V over E with a family of mappings α_σ $(\sigma \in G)$, satisfying (11.9.1), (11.9.2), (11.9.4), this defines an action of G on V by semilinear transformations, and V will be called a *space with a G-action.* The subset fixed by all the α_σ is called the set *fixed* by the G-action; it turns out that this fixed subset is actually a k-form:

Theorem 11.9.1. *Let E/k be a finite Galois extension with group G. Given a vector space V over E with a G-action, denote by V_0 the fixed subspace of V. Then V_0 is a k-form of V. Moreover, any subspace W of V admitting the G-action is a direct summand:*

$$V = W \oplus W', \tag{11.9.5}$$

with a complement W' also admitting the G-action.

Proof. It is clear that V_0 is a k-subspace of V and the mapping

$$\lambda \otimes x \mapsto \lambda x \tag{11.9.6}$$

of $E \otimes V_0$ into V is k-linear; we claim that it is an isomorphism. To find its kernel, take a basis (γ_i) of E over k. Any element of $E \otimes V_0$ can be written as $\sum \gamma_i \otimes x_i$ $(x_i \in V_0)$ and if $\sum \gamma_i x_i = 0$, then $\sum \gamma_i^\sigma x_i = 0$ for all $\sigma \in G$. But these equations

only have the trivial solution because the matrix is non-singular, by Proposition 7.6.6, so (11.9.6) is injective. Now let f be a linear form on V which vanishes on V_0 and take $x \in V$. For any $\lambda \in E$, $u_\lambda = \sum_\sigma (\lambda x)\alpha_\sigma \in V_0$, hence $f(u_\lambda) = 0$, i.e. $\sum \lambda^\sigma f(x\alpha_\sigma) = 0$. By Dedekind's lemma (Lemma 7.5.1), $f(x\alpha_\sigma) = 0$ for all $\sigma \in G$. Taking $\sigma = 1$, we find that $f(x) = 0$, and so $f = 0$. This shows the mapping (11.9.6) to be an isomorphism.

Now let W be a subspace of V admitting the G-action. By applying the first part of the proof to W, we see that it has the form $W = W_0 \otimes E$, where W_0, the subset of W fixed by G, is a k-form. Now W_0 is a k-subspace of V_0, so it has a complement: $V_0 = W_0 \oplus W_0'$, and so (11.9.5) follows by tensoring with E. ∎

As a consequence there is a useful reduction theorem for families of matrices indexed by the Galois group, which was proved by Andreas Speiser in 1919.

Theorem 11.9.2. *Let E/k be a finite Galois extension with group G. Given a matrix family $U_\sigma \in \mathbf{GL}_r(E)$, $\sigma \in G$, there exists $P \in \mathbf{GL}_r(E)$ such that*

$$U_\sigma = P^\sigma P^{-1} \tag{11.9.7}$$

if and only if

$$U_{\sigma\tau} = U_\sigma^\tau U_\tau, \quad \sigma, \tau \in G. \tag{11.9.8}$$

Equations (11.9.8) are often called *Noether's equations*, at least in the scalar case.

Proof. If (11.9.7) holds, then $U_\sigma^\tau U_\tau = P^{\sigma\tau}(P^\tau)^{-1}P^\tau P^{-1} = U_{\sigma\tau}$, so (11.9.8) is necessary. Conversely, assume (11.9.8) and consider the elements of E^r as rows, with the k-form defined by $x^\sigma = (x_1^\sigma, \ldots, x_r^\sigma)$. The mappings

$$x \mapsto x^\sigma U_\sigma \tag{11.9.9}$$

form a family of linear mappings satisfying (11.9.1), (11.9.2), (11.9.4), as is easily verified. Hence the subspace V_0 fixed by these mappings is a k-form, by Theorem 11.9.1. Take a basis $u_1, \ldots, u_r$ of V_0; thus each $u_i \in E^r$. Let A be the matrix formed by these rows. Then $A = (u_i^\sigma)$ is invertible and $A^\sigma U_\sigma = A$, because the u_i are fixed under (11.9.9); now put $P = A^{-1}$. Then $U_\sigma = P^\sigma P^{-1}$, i.e. (11.9.7). ∎

Consider the special case of Theorem 11.9.2 where G is cyclic of order n, with generator σ. If $U_\sigma = C$, then by induction on i we have

$$U_{\sigma^i} = C^{\sigma^{i-1}} \ldots C^\sigma C,$$

and so in particular,

$$I = U_{\sigma^n} = C^{\sigma^{n-1}} \ldots C^\sigma C. \tag{11.9.10}$$

Now Theorem 11.9.2 shows that any matrix C over E satisfying (11.9.10) has the form $C = P^\sigma P^{-1}$. In (11.9.10) let us take $r = 1$; then we have an element of norm 1 and Theorem 11.9.2 takes the following form. It was first stated as Theorem 90 of Hilbert's *Zahlbericht* (1897) and is generally known as 'Hilbert's Theorem 90':

Corollary 11.9.3. *Let E/k be a cyclic Galois extension and σ be a generator of the Galois group. Then any $c \in E$ has the form $c = a^\sigma a^{-1}$ for some $a \in E$ if and only if $N(c) = 1$. Moreover, a is unique up to a factor in k.*

Proof. Only the last part still needs proof, and this is clear, because if $a^\sigma a^{-1} = b^\sigma b^{-1}$, then $(b^{-1}a)^\sigma = b^{-1}a$, hence $b^{-1}a \in k$. ∎

Corresponding results exist for the action of G on the additive group of E; in order to prove them we need another result of independent interest, the normal basis theorem. Given a finite Galois extension E/k, any basis of the form $c^\sigma (\sigma \in G)$, consisting of all the conjugates of a single element c of E, is called a *normal basis*. We know that the extension E/k is simple, so there exists $c \in E$ such that the c^σ are all distinct, but there is nothing so far to tell us when we have a normal basis. In fact every finite Galois extension has a normal basis. To prove this we begin with a preparatory result on the algebraic independence of automorphisms.

Let E/k be a finite Galois extension and take any k-basis $e_1, \ldots, e_n$ of E. Any element $a \in E$ has the form $\sum \alpha_i e_i$, where $\alpha_i \in k$, and so

$$a^\sigma = \sum \alpha_i e_i^\sigma,$$

where σ ranges over $G = \mathrm{Gal}(E/k)$. We know from Proposition 7.6.6 that the matrix (e_i^σ) is invertible. Conversely, given n elements $e_1, \ldots, e_n \in E$ such that (e_i^σ) is invertible, it follows that $e_1, \ldots, e_n$ are linearly independent over k, and hence form a basis.

Proposition 11.9.4. *Let k be an infinite field and E/k be a Galois extension with group $G = \{\sigma_1, \ldots, \sigma_n\}$. If $f \in E[x_1, \ldots, x_n]$ satisfies $f(u^{\sigma_1}, \ldots, u^{\sigma_n}) = 0$ for all $u \in E$, then $f = 0$.*

Proof. Let us take a basis $e_1, \ldots, e_n$ of E/k and define a polynomial g by the equation

$$g(y_1, \ldots, y_n) = f\left(\sum_i y_i e_i^{\sigma_1}, \ldots, \sum_i y_i e_i^{\sigma_n}\right).$$

Then $g(a_1, \ldots, a_n) = 0$ for all $a_i \in k$, by hypothesis. Since k is infinite, the polynomial g is identically zero. But the matrix $(e_i^{\sigma_j})$ is invertible; if its inverse is (p_{ij}) and $x_i = \sum y_i e_i^{\sigma_j}$, then $y_j = \sum x_i p_{ij}$ and

$$f(x_1, \ldots, x_n) = g(y_1, \ldots, y_n) = 0,$$

so f vanishes identically. ∎

We can now prove

Theorem 11.9.5 (Normal basis theorem). *Any finite Galois extension E/k has a normal basis.*

Proof. (i) Assume that k is infinite and let G be the Galois group. We know that a family of n elements $e_1, \ldots, e_n$ of E is a basis iff the matrix (e_i^σ) $(\sigma \in G)$ is invertible,

hence the sequence (a^σ) is a normal basis iff $(a^{\sigma\tau})$ is an invertible matrix. Let us put

$$\sigma\tau = \varphi(\sigma, \tau), \tag{11.9.11}$$

and consider the following polynomial over k:

$$f(x^\sigma, x^\tau, \dots) = \det(x^{\varphi(\sigma,\tau)}).$$

We claim that f does not vanish identically. For by (11.9.11) we see that for any τ, $\varphi(\sigma, \tau) = \varphi(\sigma', \tau) \Rightarrow \sigma = \sigma'$; thus for fixed τ the mapping $\sigma \mapsto \varphi(\sigma, \tau)$ is a permutation of G. Hence each row of the matrix of the determinant $f(1, 0, \dots, 0)$ has one non-zero entry, equal to 1, and likewise each column, so $f(1, 0, \dots, 0)$ is the determinant of a permutation matrix, hence equal to ± 1. This establishes the claim that $f \neq 0$. Now

$$f(x^\sigma, x^\tau, \dots) = \det(x^{\varphi(\sigma,\tau)}),$$

and by Proposition 11.9.4 there exists $c \in E$ such that $\det(c^{\varphi(\sigma,\tau)}) \neq 0$, hence (c^σ) is then a normal basis.

(ii) k is finite. In this case E/k is cyclic, with generator σ say. We may regard σ as a linear transformation of E as k-space. Let $\mu(x)$ be its minimal polynomial; then $\deg \mu \leq n = [E : k]$. But by Dedekind's lemma, $1, \sigma, \dots, \sigma^{n-1}$ are linearly independent over E, hence also over k, so there exists $a \in E$ such that $a, a^\sigma, \dots, a^{\sigma^{n-1}}$ are linearly independent over k, and by counting we see that they form a basis. ∎

We also note the additive analogue. Here the place of the norm is taken by the trace: $T(a) = \sum a^\sigma$.

Theorem 11.9.6. *Let E/k be a finite Galois extension with group G. Given a family (b_σ) of elements of E indexed by G, there exists $a \in E$ such that $b_\sigma = a^\sigma - a$ if and only if*

$$b_{\sigma\tau} = b_\sigma^\tau + b_\tau, \quad \sigma, \tau \in G. \tag{11.9.12}$$

Proof. Suppose that $b_\sigma = a^\sigma - a$; then $b_\sigma^\tau + b_\tau = (a^\sigma - a)^\tau + a^\tau - a = a^{\sigma\tau} - a = b_{\sigma\tau}$ and so (11.9.12) holds. Conversely, assume (11.9.12); by Dedekind's lemma there exists $c \in E$ such that $T(c) \neq 0$. Put $b_\sigma = c^\sigma - c$, $d = -T(c)^{-1}$, and consider $a = d\sum b_\sigma c^\sigma$. We have

$$\begin{aligned} a^\tau - a &= d\sum_\sigma (b_\sigma^\tau c^{\sigma\tau} - b_\sigma c^\sigma) \\ &= d\sum_\sigma (b_\sigma^\tau c^{\sigma\tau} - b_{\sigma\tau} c^{\sigma\tau}) \\ &= -d\sum_\sigma b_\tau c^{\sigma\tau} \\ &= -dT(c)b_\tau = b_\tau. \end{aligned}$$

∎

We note again the case of a cyclic Galois group. If c is such that $T(c) = 0$, then (11.9.12) holds for $b_1 = 0$, $b_{\sigma^i} = c + c^\sigma + \ldots c^{\sigma^{i-1}}$. Hence we obtain

Corollary 11.9.7. *For any cyclic extension E/k with generating automorphism σ, an element $b \in E$ has the form $b = a^\sigma - a$ for some $a \in E$ if and only if $T(b) = 0$.* ■

For example, if E/k is a Galois extension of degree p, where char $k = p$, this shows that there exists $a \in E$ such that $a^\sigma = a + 1$.

For infinite algebraic Galois extensions similar results may be obtained, using the topological Galois groups introduced in Section 11.8 (see Serre (1979)).

Exercises

1. Show that Proposition 11.9.4 fails when k is finite.
2. State and prove an analogue of Theorem 11.9.6 for matrices.

11.10 Kummer Extensions

Abelian field extensions are of importance because they include all extensions of finite fields, and in the case of an algebraic number field K it is possible to classify the abelian extensions of K by subgroups of $K^\times$ itself. This is the subject of class field theory, an extensive theory which is beyond the framework of this book (see Neukirch (1986)), but the special case of Kummer theory can be described in elementary terms.

We recall that a group G is said to be of *exponent* m if $x^m = 1$ for all $x \in G$ and m is the least positive integer with this property. A Galois extension E/k is said to be of *exponent* m if its group has exponent dividing m; this then means that $\sigma^m = 1$ for all $\sigma \in \mathrm{Gal}(E/k)$. By a *Kummer extension* one understands an abelian extension E/k of finite exponent m, where k contains exactly m distinct m-th roots of 1. It follows that such an extension has characteristic prime to m.

Throughout the discussion the ground field k will be fixed, and all extensions lie in a given algebraic closure k_a of k. Moreover, we shall assume that k contains all m-th roots of $1 : \omega_1 = 1, \omega_2, \ldots, \omega_m$ say. Then for any $a \in k$, the roots of $x^m = a$ are $\alpha = \omega_1\alpha, \omega_2\alpha, \ldots, \omega_m\alpha$. When $\alpha \neq 0$, these roots are distinct, and each generates the same field over k, which we may unambiguously write as $k(a^{1/m})$. We shall put

$$k^{\times m} = \{c^m | c \in k^\times\},$$

and for any subgroup P of $k^\times$ containing $k^{\times m}$ we shall write $k(P^{1/m})$ or E_P for the subfield of k_a generated by all subfields $k(c^{1/m})$ for $c \in P$. By what has been said, E_P is determined uniquely as a subfield of k, and it is normal over k, for if $c \in P$,

then $x^m - c$ splits in E_P. Since its zeros are distinct, E_P/k is also separable, and is therefore a Galois extension. It is in fact a Kummer extension, for it is generated over k by the elements α such that $\alpha^m \in P$. If $\sigma, \tau \in G = \mathrm{Gal}(E_P/k)$, then $\alpha^\sigma = \omega\alpha$, $\alpha^\tau = \omega'\alpha$, hence $\alpha^{\sigma\tau} = \alpha^{\tau\sigma} = \omega\omega'\alpha$ and $\alpha^{\sigma^m} = \omega^m\alpha = \alpha$, so that $\sigma^m = 1$. This shows that G is abelian of exponent dividing m.

Given any Kummer extension E/k of exponent m, let us write $N = \{a \in k^\times | a = \alpha^m$ for some $\alpha \in E\}$. In the particular case where $E = E_P$, it is easily verified that $P \subseteq N$; below, in Theorem 11.10.1, we shall see that equality holds here. Let us put $G = \mathrm{Gal}(E/k)$; then there is a natural homomorphism $N \to \hat{G} = \mathrm{Hom}(G, k^\times)$ defined as follows. Given $a \in N$, choose $\alpha \in E$ such that $\alpha^m = a$; then for any $\sigma \in G$ we have $\alpha^\sigma = \omega_\sigma\alpha$, where $\omega_\sigma^m = 1$. Thus the mapping

$$f_a : \sigma \mapsto \omega_\sigma = \alpha^\sigma\alpha^{-1} \tag{11.10.1}$$

is a homomorphism $G \to k^\times$, because $\alpha^{\sigma\tau} = (\omega_\sigma\alpha)^\tau = \omega_\sigma\omega_\tau\alpha$. Moreover, ω_σ depends only on a and σ, not on α; if α' is another root of $x^m = a$, then $\alpha' = \lambda\alpha$, where $\lambda^m = 1$, and so $\lambda \in k$, hence $\alpha'^\sigma = \lambda\alpha^\sigma = \omega_\sigma\lambda\alpha = \omega_\sigma\alpha'$. Thus $f_a \in \hat{G}$; moreover, the mapping $a \mapsto f_a$ is a homomorphism, for if $\alpha^m = a$, $\beta^m = b$, then $(\alpha\beta)^m = ab$ and $(\alpha\beta)^\sigma(\alpha\beta)^{-1} = \alpha^\sigma\alpha^{-1}\beta^\sigma\beta^{-1}$, so $f_{ab} = f_af_b$. This may be summed up by saying that the mapping $G \times P \to k^\times$ given by

$$\sigma, a \mapsto \langle\sigma, a\rangle = \alpha^\sigma\alpha^{-1}, \quad \text{where } \alpha^m = a, \tag{11.10.2}$$

is *bimultiplicative*, i.e. multiplicative in each argument. This mapping (11.10.2) can be used to give the following description of Kummer extensions.

Theorem 11.10.1. *Let k be a field containing m distinct m-th roots of 1 (and hence of characteristic prime to m), let P be a subgroup of $k^\times$ including $k^{\times m}$ and write $E_P = k(P^{1/m})$. Then E_P/k is a Kummer extension of exponent m and if $\mathrm{Gal}(E_P/k) = G$, we have an exact sequence*

$$1 \to k^{\times m} \longrightarrow P \xrightarrow{f} \hat{G} \to 1, \tag{11.10.3}$$

where f is the mapping $a \mapsto f_a$ defined by (11.10.1). *The extension E_P/k is finite if and only if the group $P/k^{\times m}$ is finite, and when this is so, then*

$$|G| = [E_P : k] = (P : k^{\times m}). \tag{11.10.4}$$

Moreover, every Kummer extension of exponent m is of this form, so that $P \mapsto E_P$ is a bijection.

Proof. We have already seen that E_P/k is a Kummer extension of exponent m, and it is clear that N defined before satisfies $N \supseteq P$. Consider the mapping (11.10.2); if we think of G and P as being written additively, this is a bilinear mapping and it therefore defines a duality. The kernel in G is 1, for if $\langle\sigma, a\rangle = 1$ for all $a \in P$, then σ fixes E_P and so $\sigma = 1$. The kernel in E is $k^{\times m}$, for if $\langle\sigma, a\rangle = 1$ for all σ, then $\alpha = a^{1/m}$ satisfies $\alpha^\sigma = \alpha$ for all σ and so $\alpha \in k$, i.e. $a \in k^{\times m}$. It follows that the groups

$P/k^{\times m}$ and G are dual to each other, in particular, $P/k^{\times m} \cong \hat{G}$ whenever either side is finite, and then

$$(P : k^{\times m}) = |\hat{G}| = |G| = [E_P : k], \tag{11.10.5}$$

where the last equality follows by Galois theory and the second by duality theory. We remark that $P/k^{\times m} \cong \hat{G}$ even in the infinite case, provided that G is correctly interpreted, see Exercise 5 below.

Now consider the homomorphism $f : N \to \hat{G}$ defined by (11.10.1). If $\chi \in \hat{G}$, then since $\chi_\sigma \in k$, we have

$$\chi_{\sigma\tau} = \chi_\sigma \chi_\tau = \chi_\sigma^\tau \chi_\tau,$$

hence χ satisfies Noether's equations and so by Corollary 11.9.3, $\chi_\sigma = \alpha^\sigma \alpha^{-1}$ for some $\alpha \in E_P$. Since $\chi_\sigma^m = 1$, it follows that $(\alpha^m)^\sigma = \alpha^m$, i.e. $\alpha^m \in k$, so f is surjective. The kernel consists of all $a \in N$ such that $\alpha^\sigma = \alpha$ for $\alpha = a^{1/m}$ and all $\sigma \in G$, i.e. $a \in k^{\times m}$. This establishes the exact sequence (11.10.3), with N in place of P. It follows that

$$(N : k^{\times m}) = |\hat{G}|,$$

and for finite extensions this together with (11.10.5) shows that $N = P$. In general E_P is the union of its subextensions $E_{P'}$ for P' such that $(P' : k^{\times m})$ is finite, and from this the equality $N = P$ follows in the general case. Finally, if E/k is a Kummer extension of exponent m and N is defined as before, then it is clear that $E_N = k(N^{1/m}) = E$, so the correspondence $P \mapsto E_P$ is indeed a bijection. ■

In a field of characteristic $p \neq 0$ there is an analogous theory for extensions of exponent p^n. For the case $n > 1$ this requires Witt vectors, and the description would take us too far afield, but we shall briefly outline the case $n = 1$.

Let E/k be a cyclic extension of degree p, where char $k = p$. The element $1 \in k$ satisfies $T(1) = p.1 = 0$, hence by Corollary 11.9.7 there exists $\gamma \in E$ such that $\gamma^\sigma - \gamma = 1$, where σ is a generator of $\mathrm{Gal}(E/k)$. Thus we have $\gamma^\sigma = \gamma + 1$, so the conjugates of γ are $\gamma, \gamma + 1, \ldots, \gamma + p - 1$ and the minimal equation for γ is

$$x^p - x - c = 0, \quad \text{where } c = \prod_{i=1}^{p} (\gamma + i). \tag{11.10.6}$$

This is to be regarded as the analogue of the binomial equation in characteristic p. We shall write $\wp x = x^p - x$; then (11.10.6) may be written $\wp x = c$ and a root of this equation will be denoted by $\wp^{-1}c$. Let P be the subgroup of the additive group k^+ containing $\wp k = \{\wp a | a \in k\}$ and put $E_P = k(\wp^{-1}P)$, the field obtained by adjoining all the elements $\wp^{-1}a$, $a \in P$, to k. Writing $G = \mathrm{Gal}(E_P/k)$, we have a mapping $G \times P \to \mathbf{F}_p$ such that

$$\sigma, a \mapsto [\sigma, a\rangle = \alpha^p - \alpha, \quad \text{where } \wp\alpha = a. \tag{11.10.7}$$

It is easily verified that this mapping is multiplicative in the first and additive in the second argument. Moreover, its kernel in G is 1 and in P it is $\wp k$. We thus obtain the following analogue of Theorem 11.10.1:

Theorem 11.10.2. *Given a field k of characteristic $p \neq 0$ and a subgroup P of k^+ containing $\wp k$, define $E_P = k(\wp^{-1}P)$. Then E_P/k is an abelian extension of exponent p, and if $G = \mathrm{Gal}(E_P/k)$, we have an exact sequence*

$$0 \to \wp k^+ \longrightarrow P \xrightarrow{g} \hat{G} \to 1, \tag{11.10.8}$$

where g is the mapping $a \mapsto g_a$ and $g_a : \sigma \mapsto [\sigma, a\rangle$. The extension E_P/k is finite if and only if the group $P/\wp k$ is finite, and when this is so, then

$$|G| = [E_P : k] = (P : \wp k).$$

Moreover, every abelian extension of k of exponent p is of this form.

Proof. This is exactly analogous to that of Theorem 11.10.1, using Corollary 11.9.7 instead of Corollary 11.9.3. ∎

Extensions E/k of the form described in Theorem 11.10.2 are sometimes called *Artin–Schreier extensions.*

Exercises

In Exercises 1–3 k is any field containing a primitive m-th root of 1.

1. Show that when $x^m - a$ and $x^m - b$ are irreducible, they have the same minimal splitting field over k iff $ba^r \in k^{\times m}$ for some r prime to m.
2. Show that the finite Kummer extensions of exponent m are just the minimal splitting fields of $(x^m - a_1)\ldots(x^m - a_r)$. What are the degrees of these extensions?
3. Show that any field k as above is contained in a unique maximal Kummer extension K of exponent m, of degree $(k^\times : k^{\times m})$. Deduce that $\mathbf{Q}$ has infinitely many non-isomorphic quadratic extensions.
4. Let char $k = p \neq 0$. Show that $x^p - x = a$ is a normal equation over k, for any $a \in k$.
5. Let E_P/k be an infinite Kummer extension of exponent m. Show that every homomorphism from $P/k^{\times m}$ into the group $\mathbf{U}_m(k)$ of m-th roots of 1 in k is realized by the form (11.10.2), for a suitable element σ of $G = \mathrm{Gal}(E_P/k)$, while a homomorphism σ of G into $\mathbf{U}_m(k)$ is realized by (11.10.2), for some $a \in k$ iff σ is continuous, as a homomorphism of G (with the Krull topology) into $\mathbf{U}_m(k)$ with the discrete topology.
6. Let E_P/k be an infinite Artin–Schreier extension of exponent p. Show that every homomorphism from P to $\mathbf{F}_p$ is realized by the form (11.10.7) for a suitable element σ of $G = \mathrm{Gal}(E_P/k)$; and a homomorphism σ of G into $\mathbf{F}_p$ is realized by the form (11.10.7) for some $a \in k$ iff σ is continuous.

Further Exercises for Chapter 11

1. Let S be a set with a dependence relation. A subset of S is said to be *closed* if it contains all the elements dependent on it. Show that the intersection of any family of closed sets is closed. Find conditions on a dependence relation under which the union of any two closed sets is closed.
2. Let S be a set with a spanning relation (**D.0** and **D.1**) and let A be the set of all $a \in S$ which can be omitted from any spanning set, i.e. if $X \cup \{a\}$ spans S, then so does X. Show that A is the intersection of all maximal closed proper subsets of S (see also Section 2.6).
3. Show that a non-zero vector space of dimension α over a countable field has cardinal $\max\{\alpha, \aleph_0\}$. Deduce that the Hamel basis of $\mathbf{R}$ has $2^{\aleph_0}$ elements.
4. (H. Whitney) Let S be a finite set with a dependence relation Φ and let Φ^* be the set of complements of bases of S for the relation Φ. Show that there is a dependence relation on S whose collection of bases is Φ^*. (Hint. Show that there is a dependence relation whose independent sets are the complements of the spanning sets of S.)
5. (Mac Lane–Ingleton) Let $S = \{1, 2, \ldots, 9\}$; show that there is a dependence relation on S with minimal dependent families $\{1, 2, 3\}$, $\{4, 5, 6\}$, $\{1, 7, 5\}$, $\{1, 8, 6\}$, $\{2, 7, 4\}$, $\{2, 9, 6\}$, $\{3, 8, 4\}$, $\{3, 9, 5\}$. Show that there is no set of vectors in a vector space with this dependence relation. (Hint. Use Pappus' theorem to show that $\{7, 8, 9\}$ would have to be dependent.)
6. A *matroid* may be defined as a finite set M with a family of subsets called *circuits* such that (i) no proper subset of a circuit is a circuit, and (ii) if X, Y are distinct circuits and $a \in X \cap Y$, then $X \cup Y \backslash \{a\}$ contains a circuit. Verify that the minimal cycles in a graph form a matroid. Show also that the minimal dependent subsets in a vector space form a matroid.

   Given a matroid M, let us call $a \in M$ *dependent* on a set $X \subseteq M$ if X contains a subset X' such that $X' \cup \{a\}$ is a circuit. Show that this defines a dependence relation on M. (Thus the theory of dependence relations is coextensive with the theory of matroids.)
7. Show that two uncountable algebraically closed fields are isomorphic iff they have the same characteristic and the same cardinal. Does this remain true in the countable case?
8. Let $\mathbf{F}_2(x, y)$ be the rational function field in two indeterminates x, y over the field $\mathbf{F}_2$ of two elements, and consider the subset $S = \{x, y, xy, x + y\}$. Show that under algebraic dependence the independent subsets are the subsets of S of at most two elements. Show that there is no vector space over $\mathbf{F}_2$ with a subset $\{a, b, c, d\}$ such that the linearly independent subsets are precisely the subsets of cardinal at most 2.
9. Find the subfields of $\mathbf{C}(x)$ corresponding to the cyclic and the dihedral groups.
10. (S. Abhyankar) Let f, g be polynomials over a field k of degrees m, n respectively. Given a simple transcendental extension $k(x)$, write $u = f(x)$, $v = g(x)$. Show that u, v satisfy an equation of degree n in u and m in v, monic in v if f was monic. (Hint. Use the resultant to express the fact that $f(x) - u$ and $g(x) - v$ have a common zero.)

11. Let k be a field of prime characteristic p with perfect closure k_p. Show that an irreducible polynomial f over k can over k_p be written as $f = g^q$, where $q = p^r$ and g is irreducible with distinct zeros.
12. Let E/k be separable and F/k be p-radical and form $L = E \otimes_k F$. Verify that L is p-radical over E; deduce that L has no zerodivisors and so is a field.
13. Let E/k be any extension in prime characteristic p, $F = k(\alpha)$ be a p-radical extension with minimal equation $\alpha^q = a \in k$ $(q = p^n)$ and let $\alpha^r = \alpha_1$ be the least power of α in E. Show that $E \otimes_k F$ is the E-algebra generated by x, y subject to $x^r = \alpha_1 - y, y^{q/r} = 0$. (Hint. Prove the case $r = 1$ first.)
14. Let k be any field. Show that the field of formal Laurent series $k((x))$ has non-trivial derivations over the rational function field $k(x)$.
15. Let E be a field of prime characteristic p. Given $x_1, \ldots, x_n, y \in E$, y is said to be *p-dependent* on $x_1, \ldots, x_n$ if $y \in E^p(x_1, \ldots, x_n)$, and otherwise *$p$-independent*. Verify that this is a dependence relation. Show that E/k is separable iff any p-independent family in k remains p-independent in E.
16. Let L/k be a finite field extension containing subextensions E/k, F/k of degrees r, s respectively. Show that if r, s are coprime, then the subfield generated by E and F has degree rs over k. Deduce that E and F are linearly disjoint over k.
17. Let L/k be a finite Galois extension and E/k, F/k be subextensions. Put $G = \mathrm{Gal}(L/k)$ and let H, K be the subgroups of G fixing E, F respectively. Show that EF has the group $H \cap K$ and that E, F are linearly disjoint over k iff $HK = G$. (Hint. Express the condition $[EF : k] = [E \otimes F : k]$ in terms of G.)
18. With the notation of Exercise 17, show that E, F have no non-trivial isomorphic subextensions iff H and $\sigma^{-1}K\sigma$ generate G, for any $\sigma \in G$. Find subgroups H, K of Sym_4 to satisfy this condition, but $HK \neq G$. Deduce the existence of a Galois extension with group Sym_4 and two subfields whose tensor product has zerodivisors.
19. Show that if E/k is regular, then for any commutative k-algebra A there is a natural bijection between the set of prime ideals of A and those of $A \otimes E$. Show that if k is separably closed in E, then for any F/k, F is separably closed in $E \otimes F$. Further show that if E/k, F/k are regular, then the field of fractions of $E \otimes F$ is regular over k.
20. Let k be a field of characteristic 0 and $K = k((x))$ be the field of formal Laurent series in an indeterminate x. Given $f = a_r x^r + a_{r+1}x^{r+1} + \ldots \in K$, where $r \neq 0$, find $y = x + c_2x^2 + \ldots \in K$ such that $f = a_r y^r$. Deduce that for any $f \in K \backslash k$ there is an automorphism of K fixing k but moving f.
21. (G. Cauchon) Let E be a field, α be an automorphism of E of infinite order and F be the fixed subfield of α. Show that for any $n > 1$ there is at most one subextension of E of degree n over F, namely the fixed field of α^n.
22. (W. Waterhouse) Let G be any profinite group and F/k be a Galois extension with a surjective homomorphism $f : G \to \mathrm{Gal}(F/k)$. With a set X indexed by the transversals of open normal subgroups of G form $L = F(X)$ (treating the elements of X as indeterminates) and define an action of G on L by $(N\sigma)^\tau = N\sigma\tau$ for $N\sigma \in X$ and $\tau \in G$, and $a^\sigma = a^{\sigma f}$ for $a \in F$. Verify that this action is faithful and the stabilizer of any element is an open subgroup. Use Proposition 11.8.7 to show that $G \cong \mathrm{Gal}(L/E)$, where E is the fixed subfield of G and $E \cap F = k$.

23. Given a finite Galois extension E/k, let f be a polynomial over E which is such that its zeros are permuted by any element of $\mathrm{Gal}(E/k)$. Show that there exists $c \in E^\times$ such that cf has coefficients in k.
24. Examine Kummer extensions of finite fields.
25. Let k be a field of characteristic $p \neq 0$. Show that if $x^p - x - a$ is irreducible over k and has a zero α in an extension field, then $x^p - x - \alpha^{p-1}a$ is irreducible over $k(\alpha)$. Deduce that if k has a separable extension of degree p, then it has separable extensions of arbitrarily high degree.

Bibliography

This is primarily a list of books where the topics are pursued further (and which were often used as sources); it is followed by a list of papers referred to in the text as well as a small selection of articles having a bearing on the text.

FA refers to the sequel of this book: *Further Algebra and Applications*. See below.

Anderson, F. W. and Fuller, K. R. (1973) *Rings and Categories of Modules*, Graduate Texts in Mathematics 13, Springer Verlag, Berlin.

Artin, E. (1948) *Galois Theory*, Notre Dame Math. Lectures No.2, Notre Dame, IN.

Artin, E. (1957) *Geometric Algebra*, Interscience, New York.

Barwise, J. (ed.) (1977) *Handbook of Logic*, North-Holland, Amsterdam.

Birkhoff, G. (1967) *Lattice Theory* (3rd edn), AMS, Providence, RI.

Bourbaki, N. (1961–80) *Algèbre*, Chs. 1–10, Hermann, Paris, later Masson, Paris.

Bourbaki, N. (1984) *Éléments d'Histoire de Mathématiques*, Masson, Paris.

Burnside, W. (1911) *Theory of Groups of Finite Order* (2nd edn), Cambridge University Press; reprinted 1955, Dover, New York.

Chase, S. U., Harrison, D. K. and Rosenberg, A. (1965) *Galois Theory and Cohomology Theory of Commutative Rings*, Mem. Amer. Math. Soc. 52, AMS, Providence, RI.

Chevalley, C. (1951) *Introduction to the Theory of Algebraic Functions of One Variable*, No.24, AMS Colloquium Publications, Providence, RI.

Cohen, P. J. (1966) *Set Theory and the Continuum Hypothesis*, Benjamin, New York.

Cohn, P. M. (1981) *Universal Algebra* (2nd edn), Reidel, Dordrecht.

Cohn, P. M. (1985) *Free Rings and Their Relations* (2nd edn), LMS Monographs No.19, Academic Press, New York.

Cohn, P. M. (1991) *Algebraic Numbers and Algebraic Functions*, Chapman & Hall/CRC Press.

Cohn, P. M. (1995) *Skew Fields, Theory of General Division Rings*, Encyclopedia of Mathematics and its Applications, Vol. 57, Cambridge University Press.

Cohn, P. M. (2000) *Introduction to Ring Theory*, SUMS, Springer Verlag, London.

Cohn, P. M. (2003) *Further Algebra and Applications*, Springer Verlag, London, referred to as FA.

Dedekind, R. (1894) *Über die Theorie der ganzen algebraischen Zahlen, XI.* Supplement zu Dirichlets Vorlesungen über Zahlentheorie, 2. Aufl.; reprinted 1964, Vieweg, Braunschweig.

Endler, O. (1972) *Valuation Theory*, Springer Verlag, Berlin.
Fossum, R. M. (1973) *The Divisor Class Group of a Krull Domain*, Springer Verlag, Berlin.
Fuchs, L. (1970, 1973) *Abelian Groups I, II*, Academic Press, New York.
Galois, E. (1951) *Oeuvres Mathématiques*, Gauthier-Villars, Paris.
Hall, M. Jr. (1959) *The Theory of Groups*, Macmillan, New York.
Hartshorne, R. (1977) *Algebraic Geometry*, Graduate Texts in Math. 52, Springer Verlag, Heidelberg.
Hilbert, D. (1897) *Bericht über die Theorie der algebraischen Zahlkörper, Jahrber.* DMV iv; reprinted in Vol. 1 of the Collected Works.
Huppert, B. (1967) *Endliche Gruppen I*, Grundl. d. math. Wiss. 134, Springer Verlag, Berlin.
Jacobson, N. (1985, 1989) *Basic Algebra* (2nd edn), *I, II*, W. H. Freeman, New York.
Kaplansky, I. (1972) *Set Theory and Metric Spaces*, Allyn & Bacon, Boston.
Klein, F. (1884) *Lectures on the Icosahedron*; reprinted 1956, Dover, New York.
Lam, T. Y. (1980) *The Algebraic Theory of Quadratic Forms*, Adv. Book Progr., Benjamin/Cummings, Reading, MA.
Lang. S. (1970) *Algebraic Number Theory*, Addison-Wesley, Reading, MA.
Lang, S. (1984) *Algebra* (2nd edn), Addison-Wesley, Reading, MA.
Lang, S. (2002) *Algebra* (revised 3rd edn) Springer Verlag, Berlin.
Lidl, R. and Pilz, G. (1984) *Applied Abstract Algebra*, Springer Verlag, Berlin.
Mac Lane, S. (1971) *Categories for the Working Mathematician*, Springer Verlag, Berlin.
Mahler, K. (1981) *p-adic Numbers and their Functions* (2nd edn), Cambridge University Press.
Matsumura, H. (1985) *Commutative Rings*, Cambridge University Press.
Nagata, M. (1962) *Local Rings*, Interscience, New York.
Neukirch, J. (1986) *Class Field Theory*, Grundl. d. math. Wiss. 280, Springer Verlag, Heidelberg.
Ore, O. (1953) *Cardano, the Gambling Scholar*, Princeton University Press, Princeton, NJ.
Rotman, J. J. (1965) *The Theory of Groups, An Introduction*, Allyn & Bacon, Boston.
Rowen, L. H. (1988) *Ring Theory I, II*, Academic Press, New York.
Rudin, W. (1966) *Real and Complex Analysis*, McGraw-Hill, New York.
Scharlau, W. (1985) *Quadratic and Hermitian Forms*, Grundl. d. math. Wiss. 270, Springer Verlag, Heidelberg.
Semple, J. G. and Roth, L. (1949) *Introduction to Algebraic Geometry*; reprinted 1987, Clarendon Press, Oxford.
Serre, J.-P. (1979) *Local Fields*, Graduate Texts in Math. 67, Springer Verlag, Heidelberg.
Sierpiński, W. (1956) *Cardinal and Ordinal Numbers*, Pan. Wyd. Nauk, Warsaw.
van der Waerden, B. L. (1971, 1976) *Algebra I, II*, Springer Verlag, Berlin.
Weber, H. (1894, 1896, 1908) *Lehrbuch der Algebra I–III*, Teubner, Leipzig; reprinted 1963, Chelsea, New York.
Welsh, D. J. A. (1976) *Matroid Theory*, LMS Monographs 8, Academic Press, London.

White, N. (ed.) (1986) *Theory of Matroids*, Encyclopedia of Mathematics and its Applications, Vol. 26, Cambridge University Press.

List of Papers

Bass, H. [1960] Finitistic dimension and a homological generalization of semi-primary rings, *Trans. Amer. Math. Soc.* 95, pp. 466–488.

Cohn, P. M. [1966] Some remarks on the invariant basis property, *Topology* 5, pp. 215–228.

Cohn, P. M. [1973] Unique factorization domains, *Amer. Math. Monthly* 80, pp. 1–17.

Cohn, P. M. [1997] Cyclic Artinian modules without a composition series, *J. London Math. Soc.* (2) 55, pp. 231–235.

Deligne, P. R. [1973] *Variétés unirationnelles non rationnelles*, Sém. Bourbaki 1971/2, Exp. 402, Lecture Notes in Math. 317, Springer Verlag, Heidelberg.

Eilenberg, S. and Mac Lane, S. [1945] General theory of natural equivalences, *Trans. Amer. Math. Soc.* 58, pp. 231–294.

Hartley, B. [1977] Uncountable Artinian modules and uncountable soluble groups satisfying Min-n, *Proc. London Math. Soc.* (3) 35, pp. 55–75.

Hodges, W. A. [1974] Six impossible rings, *J. Algebra* 31, pp. 218–244.

Kaplansky, I. [1958] Projective modules, *Ann. Math.* 68, pp. 372–377.

Lenstra Jr., H. W. [1974] Rational functions invariant under a finite abelian group, *Invent. Math.* 25, pp. 299–325.

Nagata, M. [1957] A remark on the unique factorization theorem, *J. Math. Soc. Japan* 9, pp. 143–145.

Pierce, R. S. [1967] *Modules over commutative regular rings*, Memoirs of the AMS No.70, AMS, Providence, RI.

Rota, G.-C. [1964] On the foundations of combinatorial theory I. Möbius functions, *Z. Wahrsch.* 2, pp. 340–368.

Schur, I. [1905] Neue Begründung der Theorie der Gruppencharaktere, *Sitzungsber. d. Preuss. Akad. d. Wiss.*, pp. 406–432.

Steinitz, E. [1910] Algebraische Theorie der Körper, *J. Reine Angew. Math.* 137, pp. 167–309; reprinted 1930, Teubner, Leipzig, 1950, Chelsea, New York.

Steinitz, E. [1911, 1912] Rechteckige Systeme und Moduln in algebraischen Zahlkörpern I, II, *Math. Ann.* 71, pp. 328–354, 72, pp. 297–345.

Swan, R. G. [1969] Invariant rational functions and a problem of Steenrod, *Invent. Math.* 7, pp. 148–158.

Voskresenskii, V. E. [1973] Fields of invariants of abelian groups, *Uspekhi Mat. Nauk SSSR* 28, pp. 77–102 (in Russian).

Witt, E. [1931] Über die Kommutativität endlicher Schiefkörper, *Hamb. Abh.* 8, p. 413.

List of Notations

In some cases a page number is given where the term is first used or defined.

Number Systems

$\mathbf{N}$	the natural numbers
$\mathbf{N}_0$	the natural numbers with 0
$\mathbf{Z}$	the integers
$\mathbf{Q}$	the rational numbers
$\mathbf{Q}_+$	the non-negative rational numbers
$\mathbf{R}$	the real numbers
$\mathbf{C}$	the complex numbers
$\mathbf{U}_m$	the group of m-th roots of unity
$\mathbf{Z}(p^\infty)$	the group of all p^n-th roots of 1, for $n = 1, 2, \ldots$
$\mathbf{Z}/(n)$ or $\mathbf{Z}/n$	the integers mod n 27
$\mathbf{U}(n)$	the group of units mod n 221
$\mathbf{F}_q$	the field of q elements 224
$\mathbf{Z}_p$	the p-adic integers 315
$\mathbf{Q}_p = \mathbf{Z}_p[p^{-1}]$	the p-adic numbers 315

Set Theory

$\varnothing$	the empty set xi, 1
$\lvert X\rvert$	cardinal of the set X 2
$\mathscr{P}(X)$	power set (set of all subsets) of X 6
$X \backslash Y$	complement of Y in X xi
Y^X	set of all mappings from X to Y 5
$\aleph_0$	aleph-null, the cardinal of $\mathbf{N}$ 2

Number Theory

$\max(a, b)$	the larger of a, b
$\min(a, b)$	the smaller of a, b
$a\|b$	a divides b
(a, b)	highest common factor (HCF) of a and b
$[a, b]$	least common multiple (LCM) of a and b
δ_{ij}	Kronecker delta xi
$\mu(n)$	Möbius function 158
$\varphi(m)$	Euler function 161
$\Phi_m(x)$	cyclotomic polynomial 219

Group Theory

Sym_n	symmetric group of degree n 32
Alt_n	alternating group of degree n 33
$\mathrm{sgn}\ \sigma$	sign of the permutation σ 33
$\mathbf{C}_n$	cyclic group of order n 27
$\mathbf{D}_m$	dihedral group of order $2m$ 26
G'	derived group of G 39
$N \lhd G$	N is a normal subgroup of G 28
$(G : H)$	index of H in G 28
$\mathbf{GL}_n(R)$	general linear group over a ring R
$\mathbf{SL}_n(R)$	special linear group over a ring R
$\mathbf{Aff}_n(k)$	affine group over a field k
$\mathrm{Sp}_{2m}(k)$	symplectic group over a field k 299

Rings and Modules

${}^mV^n$	space of all $m \times n$ matrices over V 97
${}^mV$	space of m-component columns over V $(= {}^mV^1)$ 97
V^n	space of n-component rows over V $(= {}^1V^n)$ 97
$\mathfrak{M}_n(R)$ or R_n	$n \times n$ matrix ring over R 97
$\mathrm{Lat}(M)$	lattice of all submodules of M 89
$\mathrm{Hom}(U, V)$	set of all homomorphisms from U to V 83
$\mathrm{End}(U)$	ring of all endomorphisms of U 83
$U \otimes V$	tensor product of U and V 117
tM	torsion submodule of M 90
R^O	opposite of the ring R 82
$R^\times$	set of non-zero elements in R 80
A^1	augmented algebra of A 132
$\mathrm{Ann}(X)$	annihilator of X 84
$\mathrm{Ass}(M)$	assassinator of M 380

$\mathrm{Supp}(M)$	support of M 358
R_S or $R_\mathfrak{p}$	localization of R at S (or at the complement of $\mathfrak{p}$) 354f
$\sqrt{\mathfrak{a}}$	radical of an ideal $\mathfrak{a}$ 353
$K[x]$	polynomial ring on x over K 166
$K[[x]]$	formal power series ring on x over K
$K\langle X\rangle$	free K-algebra on X 134
${}_S\mathrm{Mod}_R$	category of (S, R)-bimodules 86
$\mathscr{F}(R)$	field of fractions of commutative integral domain R 428
$\prod M_i$	direct product of modules 87
$\coprod M_i$	direct sum (coproduct) of modules 87
$\mathfrak{T}_n(R)$	upper triangular matrices over R 133

Field Theory

$[V : k]$	dimension of the k-space V 190
$k(\alpha)$	field generated by α over k 191
$k[\alpha]$	ring generated by α over k 191
$\mathrm{Gal}(E/F)$	group of the Galois extension E/F 211
$T(x)$	trace of x 153
$N(x)$	norm of x 153
$U \perp V$	orthogonal sum of U and V 252f
$U^\perp$	orthogonal complement of U 251
$\langle a_1, \ldots, a_n\rangle$	quadratic form (in diagonal form) 254

Categories (mappings resp. homomorphisms are understood)

Ens	sets 65
Gp	groups 65
Ab	abelian groups 66
Rg	rings 65
Top	topological spaces 65
Mod	modules 65
vec	vector spaces 68
col	column vectors 68
$\mathrm{Fun}(I, \mathscr{A})$	functors from I to $\mathscr{A}$ 69

Author Index

Abel, Niels Henrik (1802–29) 238
Abhyankar, Shreeram S. (1931–) 445
Adyan, Sergei I. 28
Akgül, M. 110
Amitsur, Shimshon A. (1921–94) 146
Arf, Cahit (1910–) 303
Artin, Emil (1898–1962) 89, 139, 209, 216, 279, 283, 285, 316, 319, 432

Baer, Reinhold (1902–79) 116, 130
Banach, Stefan (1892–1945) 321
Bass, Hyman (1932–) 144
Becker, Eberhard 285
Bernstein, Felix (1878–1956) 4,77
Binet, Jacques P. M. (1786–1856) 182
Bolzano, Bernard (1781–1848) 2
Boole, George (1815–64) 70ff.
Bourbaki, Nicolas (190▮–) 380
Brauer, Richard D. (1901–77) 152
Burali-Forti, Cesare (1861–1931) 6
Burnside, William (1852–1927) 28, 178

Cantor, Georg F.L.P. (1845–1918) 2, 4, 7f., 14, 275
Capelli, Alfredo (1858–1916) 247
Cardano, Girolamo (1501–76) 238
Cartan, Henri P. (1904–) 258
Castelnuovo, Guido (1865–1952) 408
Cauchon, Gérard 446
Cauchy, Augustin Louis (1789–1857) 182, 275f.
Cayley, Arthur (1821–95) 32, 135
Chevalley, Claude C (1909–84) 227, 333
Clifford, William K. (1845–79) 260, 265
Cohen, Irving S. (1917–55) 395
Cohen, Paul J. (1934–) 7, 10
Cohn, Paul M. 108, 146, 185

De Morgan, Augustus (1806–71) 70
Dedekind, J.W.Richard (1831–1916) 2, 55, 115, 206, 209, 216, 275,351, 362f., 365, 395
Deligne, Pierre René (1945–) 408
Descartes, René (1596–1650) 287
Dieudonné, Jean A. (1906–92) 137, 258
Dilworth, Robert P. (1914–93) 18
Diophantos (~250) 238
Dirichlet, Peter Gustav Lejeune (1805–59) 2, 216, 222, 370

Eilenberg, Samuel (1913–98) 69
Eisenstein, Ferdinand Gotthold M. (1823–52) 199f.
Erdős, Paul (1913–96) 21
Euler, Leonhard (1707–83) 17, 161, 242, 247f.
Faltings, Gerd (1954–) 362
Fermat Pierre de (1601–65) 242, 362, 371
Ferrari, Lodovico (1522–65) 238
Ferro, Scipio del (1465–1526) 238
Fitting, Hans (1906–38) 48
Flanders, Harley (1925–) 425
Franke, E. 186
Frattini, Giovanni (1852–1925) 44, 46f.
Frobenius, F. Georg (1849–1917) 203, 224

Galois, Evariste (1811–32) 206, 211ff., 216, 238f., 242, 244f.
Gauss, Carl F. (1777–1855) 134, 217, 242, 248, 338
Gelfand, Izrael M. (1913–) 321
Gödel, Kurt (1906–78) 7, 10
Golod, Evgenii S. (1935–) 176ff., 179
Goodearl, Kenneth R. (1945–) 129
Gottschalk, Walter H. 78
Grassmann, Hermann G. (1809–77) 184
Gregory, James (1638–75) 238
Grothendieck, Alexander (1928–) 291

Hahn, Hans (1879–1934) 321
Hall, Philip (1904–82) 18,44, 49
Halmos, Paul R.(1914–) 19
Hamel, Georg (1877–1954) 401
Hamilton, Sir William R.(1805–65) 148, 299
Harriot, Thomas (1560–1621) 287
Harrison, David K. (1931–90) 297
Hartley, Brian (1939–94) 146
Hasse, Helmut (1898–1980) 296
Hausdorff, Felix (1868–1942) 22, 435
Hensel, Kurt (1861–1941) 311, 322, 340f.
Hilbert, David (1863–1941) 174f., 221, 347, 361, 392
Hodges, Wilfrid A. (1941–) 60
Hölder, Otto (1859–1937) 36
Hopkins, Charles (1902–39) 139, 145f.
Huntingdon, Edward V. (1874–1952) 78

Ingleton, Aubrey W. (1921–2000) 445
Iversen, Birger 371

Jacobi, Carl Gustav Jacob (1804–51) 43
Jacobson, Nathan (1910–99) 142f.
Jordan, Camille (1838–1922) 36

Kaplansky, Irving (1917–) 376
Klein, Avraham A. 146
Klein, C. Felix (1849–1925) 26, 185

Knebusch, Martin 285
König, Dénes (1884–1944) 21
König, Gyula (Julius) (1849–1913) 5, 246
Krasner, Mark A. (1920–97) 345
Kronecker, Leopold (1823–91) 195, 221, 246
Krull, Wolfgang (1899–1971) 90, 96, 395, 434
Kummer, Ernst-Eduard (1810–93) 347, 351, 362f., 441
Kuratowski, Kazimierz (1896–1980) 10
Kurosh, Aleksandr G. (1908–71) 178

Lagrange, Joseph L. (1736–1813) 28, 104, 237
Laplace, Pierre S. Marquis de (1749–1827) 154, 184
Lasker, Emanuel (1868–1941) 380
Laurent, Pierre Alphonse (1803–54) 133, 166
Legendre, Adrien-Marie (1752–1833) 247f., 290
Leibniz, Gottfried Wilhelm, Freiherr von (1646–1716) 173
Leicht, Johann B. 297
Lenstra Jr., Hendrik W. 404
Levitzki, Jacob (1904–56) 139, 146
Lindemann, C. L. Ferdinand von (1852–1939) 194
Liouville, Joseph (1809–82) 322
Lorenz, Falko 297
Lubell, David 22
Lüroth, Jakob (1844–1910) 407

Mac Lane, Saunders (1909–) 69, 423f., 445
Mahler, Kurt (1903–88) 328
Maschke, Heinrich (1853–1908) 162
Mazur, Stanislaw (1905–) 321
Merkuryev, A. 305
Moebius, August Ferdinand (1790–1868) 158f.
Moore, Eliakim Hastings (1862–1932) 224
Morita, Kiiti (1915–95) 100
Mumford, David B. (1937–) 311

Nagata, Masayoshi (1927–) 185, 395
Nakayama, Tadasi (1912–64) 144, 395
Newton, Sir Isaac (1642–1727) 325
Noether, A. Emmy (1882–1935) 61, 89, 139, 265, 365f., 380, 391, 409, 438
Novikov, Petr S. 28

Ore, Oystein (1899–1968) 59
Ornstein, Donald S. 45
Ostrowski, Alexander (1893–1986) 319f.

Perlis, Sam (1913–) 142
Pfister, Albrecht (1934–) 305
Pierce, Richard S. (1927–92) 104
Plücker, Julius (1801–68) 185
Poincaré, J. Henri (1854–1912) 32, 174

Rabinowitsch, J. L. 393
Rados, G. 246
Ramsey, Frank Plumpton (1903–30) 19f.
Riemann, Bernhard (1826–66) 203, 309
Roos, Jan-Erik. 96
Rota, Gian-Carlo (1932–99) 157
Ruffini, Paolo (1765–1822) 238
Russell, (Lord) Bertrand A. W. (1872–1970) 1, 10

Sarges, Heidrun 361
Scharlau, Winfried 283
Schmidt, Otto Yu. (1891–1956) 96
Schmidt, Friedrich Karl (1901–77) 425
Schreier, Otto (1901–29) 37, 56, 279
Schröder, F.W.K.Ernst (1841–1902) 4, 77
Schur, Issai (1875–1941) 137, 139, 220
Seidenberg, Abraham(1916–) 389
Serre, Jean-Pierre (1926–) 174
Shafarevich, Igor R. (1923–) 176f.
Sheffer, H. M. (1883–1964) 78
Sierpiński, Wacław (1882–1969) 22
Skolem, A. Thoralf (1887–1963) 265
Speiser, Andreas (1885–1970) 438
Sperner, Emanuel (1905–80) 22
Spitzlay, K.-E. 285
Steinitz, Ernst (1871–1928) 228, 238, 374, 432, 435
Stone, Marshall H. (1903–89) 76
Sturm, Jacques-Charles-François (1803–55) 288ff.
Swan, Richard G. (1933–) 404
Sylow, P. Ludvig M. (1832–1918) 37
Sylvester, James Joseph (1814–97) 186, 285f.
Szekeres, George (1911–) 21

Tarski, Alfred (1902–83) 22
Tartaglia, Niccolo (1500–57) 238

Vahlen, Karl Theodor (1869–1945) 247
Vandermonde, Alexandre Théophile (1735–96) 232
Viète, François (1540–1603) 238
Vinberg, Ernest B. 177
Voskresenskii, Valentin E. 404

Wall, Charles Terence Clegg (1936–) 296
Wantzel, Pierre L. (1814–48) 194
Warning, Ewald 227
Waterhouse, William C. 446
Weber, Heinrich (1842–1913) 221
Wedderburn Joseph H. Maclagan (1882–1948) 137ff., 156, 226
Weisner, Louis (1899–1988) 162
Whaples, George (1914–81) 316
Whitney, Hassler (1907–89) 445
Wielandt, Helmut W. (1910–2001) 45, 47
Wiles, Andrew (1953–) 362
Witt, Ernst (1911–91) 43, 226, 268ff., 291ff.

Yoneda, Nobuo 78

Zafrullah, Muhammad (1942–) 394
Zariski, Oscar (1899–1986) 380, 393, 408
Zassenhaus, Hans J. (1912–91) 37, 56
Zech, Theodor 224
Zelmanov, Efim I. (1955–) 28
Zermelo, Ernst F.F. (1871–1953) 10
Zorn, Max A. (1906–93) 10

Subject Index

Generally, non-X, un-X, in-X is listed under X.

abelian group 25
abelianization 67
absolute value 273, 312
absorptive law 53
acquaintanceship graph 16
acyclic 17
addition (mod 2) 81
additive function 173
  – functor 110
adjacency matrix 23
adjoint associativity 119, 123
affine group 227, 244
aleph 2
algebra 131
algebraic element 192
  – equation 376
  – extension 193, 403
  – integer 134
  – set 378
algebraically closed 201, 285, 431
  – dependent 402
alternating form 256, 298, 301
  – matrix 298
  – group 33
anisotropic part 257, 270, 293
annihilator 84
anti-chain 8, 63f.
anticommutative 166f.
antiderivation 180
antihomomorphism 66, 83
approximation theorem 316, 369
archimedean absolute value 313
  – ordering 277
Arf invariant 303
arrow 17
Artin's theorem 209
Artin-Schreier extension 444
Artin-Schreier theory 279ff.
Artinian module, ring 89
assassinator 380
associated elements 349
  – prime ideal 380
associative law 25, 53
atom 75, 193, 349
atomic Boolean algebra 78
  – domain 350
augmentation ideal 132, 167, 293
augmented algebra 132
automorphism 27, 80, 211
axiom of choice 10, 60

Baer's criterion 116
balanced mapping 122
basis 105, 398
  – theorem for abelian groups 38
bidual 126
bifunctor 87
bilinear form 249ff.
  – mapping 117, 122
bimodule 83
binary form 249
Binet-Cauchy identity 182
Boolean algebra 70, 134
  – polynomial 71
  – ring 80
Brauer group 152, 296
Burnside problem 28, 178

cancellation 31
cardinal (number) 1ff.
Cartan-Dieudonné theorem 258
Castelnuovo-Zariski theorem 408
casus irreducibilis 247
category 65
Cauchy sequence 275, 314
Cayley's theorem 32, 135, 214
central chain 39
  – simple algebra 150
centralizer 30, 149
centrally primitive idempotent 103
centre 30, 131
chain 8
  – condition 60
character (group) 125
characteristic of field 189
  – – prime ideal 297
  – – ring 80, 104
  – function 6
  – polynomial 153, 175
  – subgroup 46
Chevalley's lemma 333
chief factor, series 36

Chinese remainder theorem 102, 115
class 65
– equation 30
– of nilpotence 40
classical logic 71
Clifford algebra 260
– group 265f.
closed set 378ff.
cofinal 14
cofinite subset 70
I. S. Cohen's theorem 395
coimage, cokernel 85
comaximal ideals 102
comma category 69
commutative diagram 85
– law 25, 53
– ring 79
commutator 39, 42
companion matrix 155
complement 56, 92
complementary graph 16
complete lattice 55
– ordered field 275
– space 314
completely primary ring 104, 187, 386
completion 277, 315
composite of fields 428
composition series 36
compound matrix 182
concrete category 66
conductor 391
cone 279f.
congruent matrices 250
conjugacy class 30
conjugate 30, 42, 199, 213
conjunctive normal form 72
connected graph 19
conorm mapping 370
consistent system 377
constant 171, 415
continuum hypothesis 7
contracted ideal 356
converge 275
co(ntra)variant functor 66
coordinate ring 377
core of a field 280
coset (space) 27ff.
countable 2
cubic equation 238, 245
cubical norm 318
cycle notation 33
cyclic extension 235
– group 27
– module 82
cyclotomic polynomial 219ff.

De Morgan's laws 70
decomposition lemma 63, 361, 365
Dedekind domain 115, 365ff., 390
Dedekind's lemma 206, 231
defining relation 26
degree of field extension 190
– – polynomial 166
– – rational function 405
– – symmetric group 26
Delian problem 194
denominator 335, 354f.

dense embedding 275, 393
– functor 67
density principle 228
dependence relation 397ff.
derivation 171, 415ff.
derivative 204
derived group, series 39
determinant 182
– of a form 252
diagonal argument 8
diamond lattice 57
dicyclic group 50
different 395
digraph 17
dihedral group 26, 50
Dilworth's theorem 18
dimension 140, 249, 387, 403
– index 295
direct power 87
– product 27, 40, 87
– – of rings 101
– sum of modules 87
directed graph 52
– system 64, 202
Dirichlet box principle 2
Dirichlet's theorem 222
discrete absolute value 313
– rank 1 valuation 308
discriminant 231, 263
disjunctive normal form 72
distributive lattice 59, 96
– law 59, 79, 119
divisible module 116
division algebra 133, 150
– ring 80
dominate 332
dual basis lemma 114
– categories 67
– group 125
– homomorphism 71
duality 67f.

edge 15
Einseinheiten 326
Eisenstein polynomial 345
Eisenstein's criterion 200
elementary abelian group 26
embedded component 385
endomorphism 27, 80
– ring 82, 135
endpoint 15
enumerable 2
equipotent 1
equivalence xii, 67
equivalent valuations 311
– absolute values 315
essential extension 129
Euclid's Elements 193

Euclidean domain 351, 363, 371
– field 283
Euler function 161, 219, 248
– summation formula 248
– criterion 248
Eulerian graph 23

even Clifford algebra 261
exact functor 111
– sequence 84
exceptional extension 414
exchange lemma 399
– property 397
expanded ideal 356
exponent of a group 28, 441
extension (field) 190
exterior algebra 179, 263
external 40, 87

factor 36
– theorem 34, 69, 85
faithful action 50
– functor 67
– representation 135
Fermat primes 242
Fermat's last theorem 362
fibre product, sum 88
field 80, 189
– of sets 70
filtered algebra 187
final object 68
finite 1, 335
– character 12
– field 223
finitely presented, related 106
Fitting subgroup 48
fixed field 211
flat module 124, 358
forgetful functor 66
formally real field 280
fraction 335, 355
fractional ideal 363f.
Frattini subgroup 46f.
free associative algebra 134, 168
– field extensions 427
– group 69
– module 105f.
Frobenius mapping 203
full subcategory 65
fully invariant 94
function ring 377
functionally complete 73, 225
functor 66ff.
fundamental involution 266
– theorem of algebra 202f., 285

Galois connexion 211ff., 434
– descent 437ff.
– extension, group 211, 432
– field 223
– theory, main theorem 212f., 434
Gauss's lemma 217, 360, 425
Gaussian extension 338
– integer 134, 152
– sum 248
generating set 26
generic point 379
going-up theorem 388
Golod-Shafarevich theorem 177
graded algebra 165, 187
– partially ordered set 77
graph 15
– of a mapping 96
Grassmann algebra 184
greatest 8
ground field 190
group 25ff.
– action 29
– algebra 133
– word 26

Hall's theorem 18
– 3 subgroup lemma 44
Hamel basis 401
Hasse invariant 296
HCF highest common factor 347
height of p-radical element 410, 414
– of prime ideal 389
Hensel's lemma 340f.
henselization 341
hereditary ring 365, 372
Hermitian conjugate 188
Hilbert basis theorem 361
– Nullstellensatz 392ff., 379
– polynomial 175
– series 174ff.
– 'theorem 90' 438f.
homogeneous component 165f.
homomorphism 26, 54, 80, 131, 167, 190
Hopkins' theorem 145f.
hyperbolic pair, plane 267ff., 299
– space 270

IBN invariant basis number 107, 110
ideal 78, 80
– class group 370
– numbers 347, 362
idempotent 80
– law 53
incidence algebra 157
– matrix 23
inclusion-exclusion principle 160
independence property of tensor product 120
index of a subgroup 28
induced subgraph 15
induction 13
inductive ordered set 10
inert subring 217
inf, infimum 51
infinite set 2
initial object 68
injective cogenerator 128
– module 112
inner automorphism 27
– derivation 171
– product space 249
inseparable degree 411

integers 79
integral closure 332
– domain 80
– element 331ff.
– extension 387
– ideal 364
interior multiplication 180
intermediate value property 278
internal direct product 40
– – sum 87
intersection graph 23
invariant chain 36
inverse 25, 65
invertible element 80
– ideal 364
(S-)inverting 335, 355
involution 180, 257
irreducible algebraic set 379
– element 63, 349
– polynomial 193
irredundant decomposition 383
– intersection 143
isolated component 385
isometry 251
isomorphic extensions 208
isomorphism 27, 29, 65, 80
– theorems 35ff., 80
isotropic vector 257
isotypic module 94

Jacobi identity 43
Jacobson radical 48, 142f.
join 52
join-(ir)reducible 63
Jordan-Hölder theorem 36

kernel 27, 80
Klein 4-group 26, 34, 328
– quadric 185
König's lemma 21
Königsberg bridge problem 16
Krasner's lemma 345
Kronecker's theorem 195
Krull dimension 387
– intersection theorem 395
– topology 434
Krull's theorem 90, 351
Krull-Schmidt theorem 96
Kummer extension 441
Kurosh problem 178

Lagrange interpolation formula 104, 225, 395
Lagrange resolvent 236
Lagrange's theorem 28
Laplace expansion 183
lattice 52
Laurent polynomial 133, 166
law of quadratic reciprocity 248
LCM least common multiple 347
least element 8
left, right exact 111
– – inverse 109
left-normed product 43
Legendre polynomial 290
– symbol 247
Leibniz's formula 173
length of a chain 61, 387
– – – lattice 62
Levitzki's theorem 146
Lie algebra 43
lie over 388
limit 275
– ordinal 13
line complex, coordinates 185
linearly (in)dependent 105
– disjoint 148, 418
local ring 335, 355, 396
localization 355
locally cyclic group 59, 436
– finite group 28
– – partially ordered set 157
– nilpotent 382
loop 15
lower bound 8
– central series 40
– segment 9
– semimodular lattice 77
Lüroth's theorem 407

Mac Lane's criterion 423
marriage theorem 19
Maschke's theorem 162
matrix representation 134
– ring 97
matroid 445
maximal 8
maximum condition 60, 89
maxterm 72
meet 52
meet-(ir)reducible 63, 383
minimal element 8
– generating set 105
– polynomial 192
minimum condition 61, 89
minterm 72
Möbius function 158, 248
– inversion formula 158f.
modular lattice 55
– law 28, 55
module 81ff.
monic polynomial 134
monoid 30
– algebra 133, 168
Morita equivalence 100, 135f.
morphism 65
multiplication 79, 147
– table 133
multiplicative function 161
– representatives 326
– set 335, 351
multivector 179

Nakayama's lemma 144f., 395
natural duality 71
– homomorphism 34
– irrationality 233f.
– transformation 66

negation 71
negative 273
neutral element 25
Newton-Fourier rule 324
nilideal 144
nilpotence class 40
nilpotent group, chain 39
– ideal 141

nilradical 353
Noether normalization lemma 391
Noether's equations 438
– problem 404
Noetherian induction 60
– module 89
– ring 89, 361
non-defective form 302
non-generator 46
norm 153, 230
– on Clifford algebra 266
normal basis theorem 439
– chain 36
– closure 199
– equation 215
– extension 198, 413
– subgroup 27
normalized valuation 309
normalizer 30
normed vector space 317
null sequence 275, 314
Nullstellensatz 392ff.

object 65
one 79
opposite category 66
– ring 82
orbit 29
order 332
order of an element 26
– of a group 28
order set 273
order-isomorphism 9, 54, 274
order-preserving mapping 54
order-type 9ff.
ordered ring 272
ordinal (number) 11
orthogonal basis 254
– group,transformation 256
– idempotents 103
– sum, complement 252f.
– vectors, space 251
orthogonality relations 127
Ostrowski's theorems 319f.
outer derivation 171

p-adic integer, valuation 308, 311, 315, 369
p-dependent 446
p-group 30
p-radical extension 410
parallelogram law 35, 85
partially ordered monoid 63
path 17
pentagon lattice 59

perfect closure 411
– field 203
– group 39
permutation 26
– group 32
perspective intervals 56
PID principal ideal domain 80, 90, 372
place 334
Plücker coordinates 185
Poincaré series 174
– 's theorem 32
point 15
polynomial 133, 166
positive order set 273ff.
positive-definite 251, 285
power of the continuum 7
power set 6
preordering xi
primary decomposition 383
– submodule 374, 382
prime avoidance lemma 384
– element 196, 310, 349
– ideal 196, 351
– subfield 189
primitive element 223f., 228
– n-th root of 1 219
– permutation group 50
– polynomial 217
principal ideal 78, 80, 90
– valuation 308
– – ring 311
principle of domination 322
– – inclusion-exclusion 160
product formula 191ff.
profinite group 434f.
projective intervals 56
– module 112ff.
projective-free ring 136
proper orthogonal 256
pullback action 86, 357
– diagram 88
purely inseparable 156, 410
– transcendental 404
pushout 88
Pythagorean field 298

quadratic extension 214
– form, space 249
– residue 247
quadrature of circle 194
quartic equation 238, 245
quasicompactness 380
quasi-Galois extension 413
quasi-inverse 143
quasiprimary ideal 386
quaternion algebra 148, 263f.
– group 26, 50
quintic equation 238f., 245
quiver 17
quotient group 28

R-module 81f.
Rabinowitsch trick 393

radical of inner product space 251
– – ideal 353
– – ring 141ff., 353
– extension 235
ramification index 337, 370
ramified 370
Ramsey number, theorem 20
rank of free algebra 173
– – – module 106
– – quadratic form 254
real closed field 282
reciprocal equation 247
recursive definition 13
reduced ring 104, 421
reducible algebraic set 379
refinement 36, 61
reflexion 257
regular field extension 425
– mapping 396
– part of quadratic space254
– permutation group 34, 237
– representaton 134, 253
– quadratic form, space 251
relation, relator 26
relative complement 56
relatively algebraically closed 405
represent 254
residue class field 310, 356
– degree 337
resolvent 246
retraction 94
ring 79, 170, 347
root 192
– tower 238
rotation 256
ruler-and-compass construction 194

saturated set 352
Schreier refinement theorem 37
Schröder-Bernstein theorem 4, 77
Schur's lemma 137ff.
section 94
self-regular 427
semi-Artinian module 129
semidirect product 41
semigroup 30
semilinear transformation 437
semisimple module 91
– ring 137
separable algebra 421
– closure 411
– degree 409
– element, polynomial 204f.
– extension 205, 432
separably generated 423
separating transcendence basis 423
set 1
signature of a form 285
similar algebras 152
– quadratic forms 294
simple extension 192, 228
– group 34
– module 91
– ring 137
– transcendental extension 405
simplicity of Alt_5 239f.
singular form 251
skeleton 68
skew field 80
– symmetric matrix 298
small category 65
socle 94
soluble (by radicals) 238
– group 39
solution 376
source 65
spanning (set) 398
– relation 400
special orthogonal group 256
Speiser's theorem 438

Sperner's lemma 22
spin group, representation 267
spinor kernel, norm 266f.
split exact 85
split inner product space 270
– quadratic space 293
splitting field 197, 200
– extension 430
stabilizer 29
standard involution 180
Steinitz number 435
– criterion 228
stochastic matrix, algebra 162
strongly regular ring 104
structure constants 133
Sturm sequence, theorem 288ff.
subring 80
sup, supremum 51
superalgebra 168
supernatural number 435
support 358
Sylow subgroup theorems 37
Sylvester-Franke theorem 186
Sylvester's law of inertia 285
symbol homomorphism 305
symbolic power 385
symmetric difference 81
– functions 214
– group 26
symmetry 257
symplectic basis, group 299f.
– space 298

target 65
tensor algebra 169f.
– product 117ff.
theorem of the primitive element 228
tiled ring 99
torsion (sub)module 90, 373
– element 90
– group 26
torsion-free 90, 376
totally ordered set 8
– positive element 282
trace 153, 230, 285
transcendence basis, degree 404

transcendental extension 192, 404
transduction 187
transfinite induction 13, 61
– number 2
transitive 29
transitivity formulae 154, 230
transposition 33
transversal 28
tree 19
triangle inequality 273, 312
triangular matrix ring 99
trisection of an angle 194
trivial absolute value 313
– group 26
– ring 79
– valuation 308
type component 94

UFD, unique factorization domain 217, 349, 359, 395
UGN unbounded generating number 110
ultrametric inequality 313
unary operator 70
uniformizer 310
unit 80
unit-element 25, 79
unital algebra 132
universal mapping property 67
– quadratic form 255
upper bound 8
– – property 278
– central series 40

valency 17
valuation (ring) 308ff.
Vandermonde determinant, matrix 232, 238
variety 379
versor 265
vertex 15

weakly finite ring 107
Wedderburn structure theorems 137ff.
– nilpotence theorem 156
– theorem on finite fields 226
well-ordered set 8
width of ordered set 17
Witt (Grothendieck) ring 291ff.
– group 271
– identity 43
– index, decomposition 270, 293
– invariant 296
– ring 270, 294
– cancellation theorem 269
– chain equivalence theorem 293
– extension theorem 271

Yoneda's lemma 78

Zariski topology 380, 393
Zassenhaus lemma 37, 56
Zech logarithm 224
zero 192
– element 25, 79
zerodivisor 80, 381
Zorn's lemma 10