Barebones Algebra, EH ?! :

Ancient Abstract Algebra

Peter Hoffman

This book contains the bare bones of the theory, covering basic abstract algebra in a fairly solid way at the undergraduate level, but not attempting the comprehensiveness needed for PhD students. That is, it's a textbook, not an encyclopedia. On the other hand, various tedious calculations and mechanical illustrations probably occur less frequently here than they do in the 'average' text. The book is aimed at mathematicians teaching talented and industrious students, and also at especially talented students for self-study. I've had the pleasure of teaching many very good Honours Pure Math classes over the years. Some students in our Mathematics Faculty at Waterloo get their honours degrees without having much demanded of them in the way of basic mathematics, so these classes have often been quite elite. What follows is the endpoint in the evolution of sets of notes for these courses. As a result, the pace here would possibly be considered rather brisk by many, but not by anyone with some interest and talent in mathematics, who is willing to work hard. Math is certainly *fun*; but *not* fun *and* easy for most of us!

The other side of the coin is that being succinct should help the reader to see the forest as a whole without being distracted by too much scrub and deadwood; and should also help the lecturer, if existing, to concentrate on examples, applications and motivation. Not appearing to need 449 pages (or even 149) to arrive at Abel/Galois/Ruffini's famous discoveries also helps to maintain motivation. As we proceed, more of the theory is left for the reader to fill in, especially routine checking which hardly merits the name 'proof'. This approximates day-to-day research in mathematics just as much as work on solving problems. Such problems of course cannot be overemphasized as tools for mastering the material in any mathematics course. Here, an exercise will often be inserted, as an essential step, into the middle of a proof. Involvement as a player, rather than spectator, makes the learning more intense. As students proceed through the theoretical material, they will develop the ability to see when some checking which ought to be done has been left out. There are some other exercises as well, but not as many routine

computations as some texts contain. Another 'novelty' is the inclusion of a large collection of exercises 'in random order' at the end of the book. Researchers (or students writing an exam) cannot argue as follows: "This problem is in Section xyz; therefore, the tools needed are from that section." So these should be useful.

The reader will find it helpful (but essential only for a few examples and for the sections from **46** onwards) to have a working knowledge of linear algebra. You'll need to be familiar with the elementary facts about the standard number systems, Cartesian coordinates, induction, and with simple counting arguments using binomial coefficients.

Sections **3** to **13** are on groups, **14** to **22** on rings, **23** to **34** (excluding **25**) on basic field theory, **35** to **41** on Galois theory (of finite extensions and mostly in characteristic zero), **42** to **46** on modules and canonical forms for similarity, **47** to **50** on linear representations of finite groups, and **51**–**2** on division algebras. The numbering of results and exercises is such as requires no explanation. The extensive index should be helpful to readers who already have some background in abstract algebra. Much of the later material depends on earlier definitions and theorems. The mathematician referred to above will easily see how to omit and permute sections as needed; and will not need gratuitous advice from me concerning how much material can be covered per unit of teaching time—suffice to say that there are about two or three solid one-term courses in what follows, perhaps two for a graduate student filling in gaps in his or her background.

Risking condemnation for **nattering nabobery**[1], we note that if it's not in **boldface**, then it's not a joke, but not conversely.

I wish to thank the students in my algebra classes over the years for helping me to learn algebra better. In particular Ian Goldberg, David Kerr and Mike Mosca, who found some errors in an earlier version of this book, deserve special thanks. Of course, all remaining errors are ones for which I claim priority. I am very grateful also to the following good friends: Debbie Brown, Lois Graham, Carolyn Jackson, Linda Kelly and Annemarie Nittel, for their excellent help with the 'keyboarding'.

[1]Spiro T. Agnew, circa 1971

Basic Notation.

Suppose that B is a set, and that b is a member of B. Such a statement will frequently be shortened to "$b \in B$". As an aid to short-term memory, we'll often correlate notation in this way, with members denoted by the lowercase letter whose corresponding uppercase letter denotes the set, sometimes using subscripts etc. Also we'll try to denote different species of objects in a given discussion using different species of letters: lowercase, uppercase, Greek, etc. But, of course, sets themselves can be members of other sets, so this has some limitations.

A third species in a given discussion might be *maps*, or *functions*, between sets, with notation such as $\gamma : A \to B$. The *domain* of γ is A, and its *codomain* is B. If B were a proper subset of C—this is written $B \subset C$, $B \neq C$: here the symbol $\subset$ includes the possibility that the two sets are equal—and if we enlarged the codomain of the function to C, then the new function, from A to C, must be regarded as being *different* from γ, even though it has the same domain and the same 'formula'. This is usually not done in calculus. But it is necessary for the adjective 'surjective' to be meaningful: the previous map γ is *surjective* if and only if

$$\mathrm{Im}\gamma \ := \ \{ \ b \in B \ : \ \exists a \in A \text{ with } \gamma(a) = b \ \} \ = \ B \ .$$

The display asserts that every element of B has the form $\gamma(a)$ for at least one $a \in A$. We have used " := " to indicate that the notation to its left is being *defined*. So we've defined the *image of* γ, denoted $\mathrm{Im}\gamma$. The real assertion in the display is made by the last " = " sign without the " : " sign. (The latter symbol is called a 'colon', in case you didn't know! A proliferation of symbols " ; " is known as **cancer of the semi-colon**.)

We have also used the quantifier " $\exists$ " ; this should be read: "there exists". Occasionally we'll also use " $\exists!$ " ; this should be read: "there exists a unique". The other quantifier is " $\forall$ " ; it is read: "for all".

Our map γ is *injective* if and only if

$$\forall a_1, a_2 \in A \ , \quad [\ \gamma(a_1) = \gamma(a_2) \implies a_1 = a_2 \ ] \ .$$

This says that distinct members of A have distinct images in B under the function γ. The symbol " $\implies$ " is read: "implies". We say that the function γ is *bijective* if and only if it is both injective and surjective. In this case, γ

has an *inverse*, $\gamma^{-1} : B \to A$. Furthermore, the sets A and B then have the same 'size' or *cardinal number*. When the sets are infinite, this is a definition of great significance; when they are finite, whether it is a definition or an assertion is a question of how you set up your foundations of mathematics. When dealing with (potentially) infinite sets, the phrase *"almost all"* means *"all but finitely many"*.

The standard number systems will be denoted in the usual way:
$\mathbf{Z}$:= the set of all integers (positive and negative) ;
$\mathbf{Q}$:= the set of all rational numbers ;
$\mathbf{R}$:= the set of all real numbers ;
$\mathbf{C}$:= the set of all complex numbers .

The first two of these have *countable* cardinality; that is, there is a bijective function in each case from the set to the set of positive integers. However, neither $\mathbf{R}$ nor $\mathbf{C}$ is countable. Each of these is a subset of the next. Each can be constructed using the previous set, so that all of standard mathematics can in principle be based on logic and a few assumptions about sets. Detailed knowledge of this will not be needed here, but it is assumed that you know a certain amount about these number systems.

In particular, the notorious real number e, which arises basically from seeking functions which equal their own derivatives, satisfies the famous identity of Euler :

$$\mathrm{e}^{i\theta} \;=\; \cos(\theta) \;+\; i\,\sin(\theta)\ .$$

It follows that $\mathrm{e}^{2\pi i/n}$ is a complex number whose n^{th} power equals 1. The only other complex numbers with this property are the powers of $\mathrm{e}^{2\pi i/n}$, of which there are only "n" in total. Notice that we use the ordinary Roman 'e' for this number, leaving the italicized e to be used in many other contexts without causing ambiguity.

Very occasionally in the book, the word(s) "(finite) multiset" occur. Informally, this just means a 'set where some elements can appear more than once'—so it's not really a set. If the reader were to insist, this could be made more formal in at least two ways: as a finite set plus a function from that set to the positive integers (the 'frequency of occurrence function'); or as an equivalence class of finite sequences (where two are equivalent if they are re-arrangements of each other).

We shall often deal with an *equivalence relation* on a set S. Recall that this is any relation, say $\sim$, between some pairs of elements of S which has

the following three properties:
(i) $\forall a,\ a \in S,\ a \sim a$ (reflexivity) ;
(ii) $\forall a,\ b \in S,\ a \sim b \implies b \sim a$ (symmetry) ;
(iii) $\forall a,\ b,\ c \in S,\ (a \sim b \ \ \& \ \ b \sim c) \implies a \sim c$ (transitivity) .
The *equivalence class* of an element a of S is denoted $[a]$ or $[a]_\sim$; it consists of all elements in S which are related to a with respect to the relation $\sim$. Each equivalence class is non-empty; the union of all classes is S; and no two distinct classes have any element in common, i.e. they are disjoint (although, in all but the extreme case where $\sim$ means equality, we'll have elements $a \neq b$ for which $[a] = [b]$). Sometimes the set whose elements are the equivalence classes is denoted as $S \ / \sim$.

An important instance is where $S = \mathbf{Z}$ and $\sim$ is *congruence (mod k)*, for some fixed positive integer k. Two integers a and b are related under this relation precisely when their difference is an integer multiple of k. This is denoted $a \equiv b \pmod{k}$—and when a and b are *not* related with respect to congruence (mod k), this is of course denoted $a \not\equiv b \pmod{k}$. In this example, the equivalence classes are also called *congruence classes*. The infinite set $\mathbf{Z}$ is thus partitioned as the union of finitely many congruence classes, each of which is an infinite set. In fact, the number of congruence classes is "k". In this example, the set of all congruence classes is denoted $\mathbf{Z}_k$, rather than using the cumbersome notation $\mathbf{Z}\ /(\equiv \bmod k)$. Another special fact about this example is that there is a well-defined method for adding, and another for multiplying, two congruence classes. The formulae are

$$[a] + [b] := [a + b] \quad ; \quad [a] \cdot [b] := [a \cdot b] \ .$$

This looks tautological at first, or circular. But it isn't—the operations on the left-hand sides are being defined in terms of the familiar operations applied to ordinary integers on the right-hand sides. The (hopefully temporary) confusion is caused by the *abuse of notation*—we've used the same notation, "+", for two different mathematical objects (and similarly for the notation "$\cdot$").

Of course, there are examples of equivalence relations where equivalence classes have cardinalities different from each other, even with some classes finite and some infinite.

CONTENTS

1. Permutations.

A *permutation* of the set $\{\ 1,\ 2,\ \cdots,\ n\ \}$ is a bijective map from that set to itself. The set of all " $n!$ " such permutations is denoted S_n. If α and β are in S_n , define $\alpha\beta$ (or sometimes $\alpha \cdot \beta$) in S_n to be their composition, $\alpha \circ \beta$. (**Caution:** The opposite convention is sometimes used, $\alpha\beta$ meaning " do α before doing β ".) In this book, $\alpha\beta$ means " do β before doing α ". Let $e \in S_n$ be the identity map. (There is, strictly speaking, a different e for each n, but using the same name for all of them won't lead to any confusion.) Then $\alpha e = \alpha = e\alpha$ for all $\alpha \in S_n$. If $\sigma \in S_n$, let σ^{-1} be the map inverse to σ. Then $\sigma\sigma^{-1} = e = \sigma^{-1}\sigma$.

A given permutation σ is often denoted

$$\begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ \sigma(1) & \sigma(2) & \sigma(3) & \cdots & \sigma(n) \end{pmatrix}.$$

For example, $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 5 & 1 & 4 & 3 & 7 & 6 \end{pmatrix}$ is the permutation $\sigma \in S_7$ defined by $\sigma(1) = 2,\ \sigma(2) = 5,\ \sigma(3) = 1,\ \sigma(4) = 4,\ \sigma(5) = 3,\ \sigma(6) = 7$ and $\sigma(7) = 6$. In this notation, the lower line is a re-arrangement of the sequence $1, 2, 3, \cdots, n$. Some people prefer to think of permutations as re-arrangements.

Given "k" distinct numbers $a_1, a_2, \cdots, a_k$ between 1 and n for $k \geq 2$, the symbol $(a_1 a_2 \cdots a_k)$ denotes the permutation σ defined by specifying its values : $\sigma(a_1) = a_2,\ \sigma(a_2) = a_3, \cdots, \sigma(a_{k-1}) = a_k,\ \sigma(a_k) = a_1$, and $\sigma(x) = x$ for any other x between 1 and n . Such permutations are called *cycles.* The length of the cycle is k . The identity permutation, e, is sometimes considered to be a cycle of length 1, but *not* here. A *transposition* is a cycle of length 2; it interchanges two numbers, leaving everything else fixed. A set of cycles is *disjoint* if no number is moved by more than one of them; i.e. if $(a_1 a_2 \cdots a_k)$ is in the set, then no a_i occurs in any other cycle in the set.

Theorem 1.1. *Any permutation is a product of a disjoint set of cycles.*

Proof. Prove it for all elements of S_n by induction on n. By convention, e is the product of the empty set of cycles. For $n = 1, 2$ or 3, every other element of S_n is a cycle. (Check this !) Suppose that the theorem holds for all elements of S_{n-1}, where $n > 3$. Let $\sigma \in S_n$.

Case 1. If $\sigma(n) = n$, then σ maps $\{ 1, 2, \cdots, n-1 \}$ to itself, so it may be thought of as an element of S_{n-1}. By the inductive hypothesis, it is a product of disjoint cycles. (In this case, none of the cycles will involve n itself).

Case 2. If $\sigma(n) \neq n$, let $a_1 = n$, $a_2 = \sigma(a_1)$, $a_3 = \sigma(a_2)$, $\cdots$. This sequence must eventually have repeats (why?), and the first repeat is a_1 since σ is injective. So let k be the unique integer larger than 1 such that $a_{k+1} = a_1$, but $a_1, a_2, \cdots, a_k$ are all distinct. Define $\bar{\sigma}$ by $\bar{\sigma}(a_i) = a_i$ for all i, and $\bar{\sigma}(x) = \sigma(x)$ if $1 \leq x \leq n$ but $x \neq a_i$ for any i . Then $\bar{\sigma}(n) = n$, so $\bar{\sigma}$ is a product of disjoint cycles, by Case 1. Also $\sigma = \bar{\sigma} \cdot (a_1 a_2 \cdots a_k)$ since they map elements of $\{ 1, 2, \cdots, n \}$ in the same way as each other. No a_i occurs in any cycle in the decomposition of $\bar{\sigma}$ since $\bar{\sigma}(a_i) = a_i$ for all i . After decomposing $\bar{\sigma}$, we see that σ is a product of disjoint cycles.

Exercise 1A. Verify the equation $\sigma = \bar{\sigma} \cdot (a_1 a_2 \cdots a_k)$ from a few lines above.

Exercise 1B. Show that the order of appearance of the cycles in a decomposition as in **1.1** is irrelevant, and furthermore that such a decomposition is unique (i.e. no other disjoint set of cycles will do).

Corollary 1.2. *Any permutation σ is a product of transpositions.* (Intuitively, any re-arrangement can be achieved by a sequence of interchanges of pairs of elements.)

Proof. Since σ is a product of cycles, it is enough to show that any cycle is a product of transpositions. But

$$(a_1 a_2 \cdots a_k) \;=\; (a_1 a_k)(a_1 a_{k-1}) \cdots (a_1 a_2) \; .$$

Exercise 1C. Show that, in the corollary, there is no uniqueness, the order is relevant, and the transpositions won't always be disjoint.

Theorem 1.3. *Given σ, the number of transpositions in products of transpositions which equal σ is either always even or always odd.*

Proof. Let $X = \{ 1, 2, \cdots, n \}$. Define

$$\text{sign}(\sigma) \;:=\; \prod_{\{k,\ell\} \subset X} \left(\frac{\sigma(k) - \sigma(\ell)}{k - \ell}\right)$$

where the product is over all subsets $\{ k, \ell \}$ with two elements.
[The number of such subsets is "$n(n-1)/2$".] Since $\frac{\sigma(k)-\sigma(\ell)}{k-\ell}$ is the same as $\frac{\sigma(\ell)-\sigma(k)}{\ell-k}$, the order chosen for k and ℓ doesn't matter.

Now, for all σ, we have $\mathrm{sign}(\sigma) = \pm 1$:

$$| \mathrm{sign}(\sigma) | = \prod_{\{k,\ell\} \subset X} | \frac{\sigma(k) - \sigma(\ell)}{k - \ell} | = \frac{\prod_{\{i,j\} \subset X} | i - j |}{\prod_{\{k,\ell\} \subset X} | k - \ell |} = 1$$

(making the change of variable $i = \sigma(k),\ j = \sigma(\ell)$ in the numerator).
Exercise 1D. Why can't we use the same argument to show (the false statement) that $\mathrm{sign}(\sigma)$ itself is always $+1$??
We are simply saying that each difference, in one or the other order, of two distinct elements of X occurs once in the numerator and once in the denominator. It follows that $\mathrm{sign}(\sigma) = (-1)^N$, where N is the number of two-element subsets $\{k, \ell\}$ for which $\sigma(k) - \sigma(\ell)$ differs in sign from $k - \ell$.

Thus, for any transposition τ, we have $\mathrm{sign}(\tau) = -1$:
for if $\tau = (ab)$ where $a < b$, the number of subsets which have either the form $\{a, x\}$ with $a < x \leq b$, or the form $\{x, b\}$ with $a \leq x < b$, is exactly "$2b - 2a - 1$", an odd integer.

Finally, $\mathrm{sign}(\alpha\beta) = \mathrm{sign}(\alpha)\mathrm{sign}(\beta)$: for

$$\frac{\mathrm{sign}(\alpha\beta)}{\mathrm{sign}(\beta)} = \prod_{\{k,\ell\} \subset X} \left(\frac{\alpha\beta(k) - \alpha\beta(\ell)}{\beta(k) - \beta(\ell)}\right) = \prod_{\{i,j\} \subset X} \left(\frac{\alpha(i) - \alpha(j)}{i - j}\right) = \mathrm{sign}(\alpha)$$

[making the change of variable $i = \beta(k),\ j = \beta(\ell)$].

It follows easily by induction on m that

$$\mathrm{sign}(\sigma_1\sigma_2 \cdots \sigma_m) = \mathrm{sign}(\sigma_1)\mathrm{sign}(\sigma_2) \cdots \mathrm{sign}(\sigma_m).$$

Now suppose that σ can be written in two ways as a product of transpositions,

$$\sigma = \tau_1\tau_2 \cdots \tau_m \qquad \text{and} \qquad \sigma = \tau'_1\tau'_2 \cdots \tau'_s \ .$$

Then

$$\mathrm{sign}(\sigma) = \mathrm{sign}(\tau_1)\mathrm{sign}(\tau_2) \cdots \mathrm{sign}(\tau_m) = (-1)(-1) \cdots (-1) = (-1)^m \ .$$

But similarly $\text{sign}(\sigma) = (-1)^s$. Thus $(-1)^m = (-1)^s$, so either m and s are both even, or they are both odd, as required.

Corollary to the proof 1.4. *We have* $\text{sign}(\sigma) = (-1)^m$*, where* m *is the number of transpositions in some chosen decomposition of* σ *as a product of transpositions.*

Definition. We say that σ is *even* or *odd* according as σ is a product of an even or an odd number of transpositions. Thus

σ is even $\iff$ $\text{sign}(\sigma) = +1$;

σ is odd $\iff$ $\text{sign}(\sigma) = -1$.

A product of two even, or two odd, permutations is even. A product of an odd times an even, or an even times an odd, is always odd. Note that a cycle of even length is odd, and one of odd length is even.

2. Binary operations.

Definition. A binary operation on a set S is a map $S \times S \to S$.

Notation. Denote the image of (a, b) as $a * b$, or $a \cdot b$, or $a + b$ etc.; and then the operation is denoted $*$ or $\cdot$ or $+$, respectively. The additive notation, $+$, is seldom used if the operation is not *commutative*, as defined below.

Definition. The operation $*$ is

(i) *associative* $\iff$ $(a * b) * c = a * (b * c)$ $\forall a, b, c$ in S ; and

(ii) *commutative* $\iff$ $a * b = b * a$ $\forall a, b \in S$.

It is not difficult to give examples of operations which have neither, either or both of these properties.

Exercise 2A. Give such examples.

Given more than two elements of S, but a finite number, one can multiply them in many different ways using a given operation, either by altering the brackets (which are necessary since, to begin with, one can only multiply *pairs* of elements), or by altering the order. When $*$ is associative, brackets are unnecessary:

Proposition 2.1. *When* $*$ *is associative, all products of* $a_1, a_2, \cdots, a_n$ *in that order with any bracketing are equal.*

See Appendix **ZZZ** after Section **13** for a more precise statement, and proof. As an example when $n = 4$, to show that

$$(a_1 * a_2) * (a_3 * a_4) \;=\; a_1 * [(a_2 * a_3) * a_4] \ ,$$

notice that the left hand side is $a_1 * [a_2 * (a_3 * a_4)]$ by applying associativity to $a_1,\ a_2,\ a_3 * a_4$. But the right hand side also is equal to $a_1 * [a_2 * (a_3 * a_4)]$ by applying associativity to $a_2,\ a_3,\ a_4$, and leaving a_1 alone.

Notation. When $*$ is associative, denote by $a_1 * a_2 * \cdots * a_n$ the element obtained by multiplying $a_1, a_2, \cdots, a_n$ in that order. (We have already used this in Section **1** but there is no logical circularity.)

When $*$ is also commutative, the order doesn't matter:

Proposition 2.2. *If $*$ is both associative and commutative, then*

$$a_{\sigma(1)} * a_{\sigma(2)} * \cdots * a_{\sigma(n)} \;=\; a_1 * a_2 * \cdots * a_n$$

for all $\sigma \in S_n$.

Example. When $n = 3$,

$$a_1 * a_2 * a_3 = a_1 * a_3 * a_2 = a_2 * a_1 * a_3 = a_2 * a_3 * a_1 = a_3 * a_2 * a_1 = a_3 * a_1 * a_2 \ .$$

Proof. Suppose first that $\tau = (ij)$ where $i < j$. Then

$$\begin{aligned} & a_{\tau(1)} * a_{\tau(2)} * \cdots * a_{\tau(n)} \\ &= a_1 * a_2 * \cdots * a_{i-1} * a_j * a_{i+1} * a_{i+2} * \cdots * a_{j-1} * a_i * *a_{j+1} * \cdots * a_n \\ &= a_1 * a_2 * \cdots * a_{i-1} * a_{i+1} * a_{i+2} * \cdots * a_{j-1} * a_i * a_j * a_{j+1} * \cdots * a_n \\ &= a_1 * a_2 * \cdots * a_{i-1} * a_i * a_{i+1} * \cdots * a_{j-1} * a_j * a_{j+1} * \cdots * a_n \ , \end{aligned}$$

applying commutativity first to the pair $[\ a_j\ ,\ a_{i+1} * a_{i+2} * \cdots * a_{j-1} * a_i\ ]$ and then to the pair $[\ a_{i+1} * a_{i+2} * \cdots * a_{j-1}\ ,\ a_i\ ]$.

Now prove the theorem by induction on the minimum number of transpositions needed to decompose σ as a product of transpositions. If that number is zero, then $\sigma = e$ for which the theorem statement is clearly true. If it is one, then σ is a transposition and we've just proved it. For the inductive

step, we can write $\sigma = \bar{\sigma}\tau$, where τ is a transposition and the number for $\bar{\sigma}$ is one less than for σ. Then

$$\begin{aligned} a_{\sigma(1)} * a_{\sigma(2)} * \cdots * a_{\sigma(n)} &= a_{\bar{\sigma}\tau(1)} * a_{\bar{\sigma}\tau(2)} * \cdots * a_{\bar{\sigma}\tau(n)} \\ &= a_{\bar{\sigma}(1)} * a_{\bar{\sigma}(2)} * \cdots * a_{\bar{\sigma}(n)} \quad \text{(by the case just done : let } b_i = a_{\bar{\sigma}(i)}\text{)} \\ &= a_1 * a_2 * \cdots * a_n \qquad\qquad \text{(by the inductive hypothesis).} \end{aligned}$$

I. Elementary Group Theory

Sections **3** to **13** study a type of algebraic object, called a *group*, which is the most important object within non-commutative algebra. Its historical origin was the material which is studied here in the sections on Galois theory (beginning with **35**), and this remains one of the main motivations for studying groups. There are many other motivations within algebra and also, for example, from geometry, topology, physics, chemistry and differential equations, not necessarily entirely separate from each other.

3. Groups.

Definition. A *group* is a set G together with a binary operation $*$ on G, such that the following three axioms hold:

(1) The operation $*$ is associative.

(2) There exists an element $1 \in G$ (called the *identity element*) such that $1 * g = g$ and $g * 1 = g$ for all $g \in G$.

(3) For all $g \in G$, there exists an element $g^{-1} \in G$ (called the *inverse* of g) such that $g^{-1} * g = 1$ and $g * g^{-1} = 1$.

We do not need to assume uniqueness of 1, nor of g^{-1} for fixed g. These facts are proved in the next section. When considering several groups $G,\ H, \cdots$ simultaneously, we often use the notations $1_G,\ 1_H, \cdots$ for their identity elements.

Examples.

A. $GL(\mathbf{R}^n)$, the set of all linear isomorphisms (that is, bijective linear transformations) from $\mathbf{R}^n$ to itself, with composition as operation.

B. $GL(n,\ \mathbf{R})$, the set of all real $n \times n$ *non-singular* (i.e. *invertible*) matrices , with matrix multiplication as operation. The set $\mathbf{R}^\times$ of non-zero real numbers under ordinary multiplication is the case $n = 1$ of this.

C. The set, $\mathbf{C}^\times$, of non-zero complex numbers under multiplication.

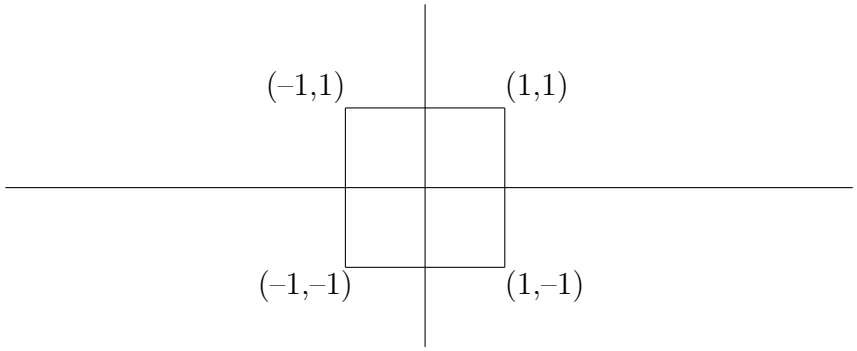
D. $C_k = \{\ 1,\ \rho,\ \rho^2,\ \cdots,\ \rho^{k-1}\ \}$, where $\rho = \mathrm{e}^{2\pi i/k}$, under multiplication.

E. S_n, with the usual product of permutations, namely composition.

F. $O(\mathbf{R}^n)$, the group of orthogonal transformations of $\mathbf{R}^n$.

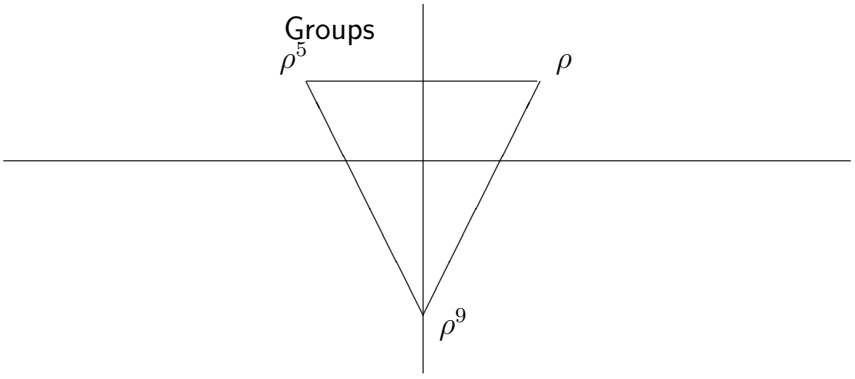
G. $O(n)$, the group of $n \times n$ real orthogonal matrices.

H. D_4, the group of symmetries of the square; i.e. those eight elements of $O(\mathbf{R}^2)$ which map the square below to itself.



I. The set of eight matrices which represent the elements of D_4 with respect to the standard basis for $\mathbf{R}^2$.

J. D_3 , the group of symmetries of an equilateral triangle; to be precise, those six elements of $O(\mathbf{R}^2)$ which map the triangle to itself.



Here $\rho = e^{2\pi i/12}$. (Try to describe the elements of D_3 and D_4—see examples (iii) and (iv) after **4E**.)

K. $\mathbf{R}^n$ is a group under vector addition. The case $n = 1$ is the reals under addition. The case $n = 2$ is the complex numbers under addition. To generalize, any vector space is a group with respect to its addition.

L. The set $\mathbf{Z}$ of integers under addition.

M. The set $\mathbf{Z}_k$ of all "k" of the congruence classes (mod k) under addition (mod k). Another group, this time using instead the operation of multiplication (mod k), is formed by taking the subset, $\mathbf{Z}_k^\times$, of only those congruence classes containing integers *prime* to k. Except when there is no chance of ambiguity, it is important to be quite explicit about what the operation is, when specifying a group.

N. If A is any set, the set G of all bijective maps from A to itself under composition. (When $A = \{\ 1,\ 2,\ \cdots,\ n\ \}$, this G is S_n.)

Exercise 3A. Check all the axioms carefully for each of these examples; also, where relevant, check that the operation is well defined.

Definition. A group G is *commutative* if and only if its operation is commutative.

Examples C., D., K., L., M. are commutative.

Non-commutative examples are A. for $n > 1$; B. for $n > 1$; E. for $n > 2$; F. and G. for $n > 1$; H.; I.; J.; and N. if the set A has three or more elements.

Notation. The operation in a group will from now on be denoted by juxtaposition (i.e. ab is the product of a and b), or by $+$. The latter notation

is used here only when G is commutative, and the word *abelian* is often used instead of 'commutative', especially when this additive notation is used. In the additive notation, the identity element in G is called instead the *zero* of G and denoted 0, and the inverse of g is called instead the *negative* of g and denoted $-g$ (instead of g^{-1}).

Definition. A group G is *finite* if and only if it has finitely many elements. The number of elements in G is called the *order* of G and is denoted $|G|$ here. In the examples, D., E., H., I., J., and M. are finite groups of orders k, $n!$, 8, 8, 6, and k, respectively. The second one in M. has order equal to the number of integers which are between 0 and k and are *prime* to k. This number is denoted $\Phi(k)$, and Φ is called *Euler's function.* (**Euler existed before THE Edmonton,** which juxtaposition should suggest to those with good taste in team sports the pronunciation of his name by Anglo mathematicians.) The other groups are infinite, except for N., which is finite if and only if A is a finite set, and both F. and G. for $n = 1$.

4. Elementary properties of groups.

Proposition 4.1. *The identity element in a group is unique.*

Proof. If $1'$ also denotes an element with the properties of the identity, then $11' = 1'$ since 1 is an identity, but $11' = 1$ since $1'$ is an identity. Hence $1 = 1'$, i.e. they denote the same element, as required.

Proposition 4.2. *The inverse of any group element is unique.*

Proof. If $\hat{g}$ also denotes an inverse for g, then

$$\hat{g} \;=\; \hat{g}1 \;=\; \hat{g}gg^{-1} \;=\; 1g^{-1} \;=\; g^{-1}\ ,$$

as required.

Note. $(g^{-1})^{-1} = g$.

Proposition 4.3. *Let $a,\ b \in G$, a group. Then the equations $ax = b$ and $ya = b$ each have exactly one solution in G.*

Proof. Since $a(a^{-1}b) = 1b = b$, the first equation certainly has one solution, namely $x = a^{-1}b$. If s is any solution, then $as = b$. Multiply both

sides on the left by a^{-1}, yielding $s = a^{-1}b$. Thus there is only one solution. The proof for $ya = b$ is left as an exercise.

Exercise 4A. Do the rest of this proof. Also show that the statement of **4.3** may be used as an axiom to replace (2) and (3) in the definition of the term 'group', as long as we assume that the set G is non-empty. Finally, for *finite* $G \neq \emptyset$, so may the 'cancellation' law: $\forall a, b, c,\ (ab = ac \text{ or } ba = ca) \Rightarrow b = c$; but not for infinite G.

Proposition 4.4. $(a_1 a_2 \cdots a_n)^{-1} = a_n^{-1} a_{n-1}^{-1} \cdots a_1^{-1}$.

Proof. Proceed by induction.
For $n = 2$:
$a_2^{-1} a_1^{-1} a_1 a_2 = a_2^{-1} 1 a_2 = 1$, so $a_2^{-1} a_1^{-1}$ is the inverse of $a_1 a_2$.
Inductive step :

$$\begin{aligned}(a_1 a_2 \cdots a_n)^{-1} &= [(a_1)(a_2 \cdots a_n)]^{-1} \\ &= (a_2 \cdots a_n)^{-1}(a_1)^{-1} \qquad \text{(by the case just proved)} \\ &= a_n^{-1} a_{n-1}^{-1} \cdots a_1^{-1} \qquad \text{(by the inductive hypothesis).}\end{aligned}$$

Definition. For $n \geq 1$ and $g \in G$, define
$g^n := gg \cdots g$ ("n" copies of g) , or, more succinctly :
$g^1 := g$ and $g^n := g^{n-1} g$ (inductive definition).
Define $g^0 := 1$.
Define $g^{-n} := (g^{-1})^n$.

Corollary 4.5. *For all integers* n, *we have* $(g^{-1})^n = (g^n)^{-1}$.

So we could have defined g^{-n} to be $(g^n)^{-1}$ for $n > 0$.

Proof. Let $a_i = g$ for all i in **4.4**. This does it for $n > 0$. Do it as **Exercise 4B**, for $n \leq 0$.

Proposition 4.6. *For all* $i, j \in \mathbf{Z}$,
(i) $g^i g^j = g^{i+j}$ *and*
(ii) $(g^i)^j = g^{ij}$.

Proof. Prove *(ii)* as **Exercise 4C.** As for *(i)*, to prove it for $j \geq 1$, fix i and use induction on j. (**Exercise 4D.** Write this out.) For $j = 0$, it is

trivial, and the case $j < 0$ follows from the case $j > 0$ as follows:

$$g^i g^j \;=\; (g^{-1})^{-i}(g^{-1})^{-j} \;=\; (g^{-1})^{(-i)+(-j)} \;=\; (g^{-1})^{-(i+j)} \;=\; g^{i+j}\ .$$

Definition. An element g has *finite order* if and only if $g^k = 1$ for some $k \geq 1$. In this case, *the order of* g , written $||g||$, is the least such $k \geq 1$. Otherwise, g is said to have *'infinite order'*.

Theorem 4.7. *If* $||g|| = k < \infty$, *then the elements* $1, g, g^2, \cdots, g^{k-1}$ *are all distinct from each other. Every other power of* g *is equal to one of these. In fact*

$$g^i \;=\; g^j \quad\Longleftrightarrow\quad i \equiv j \pmod{k}\ .$$

Proof. First we prove the last statement. If $i \equiv j (\mathrm{mod}\ k)$, then $j = i + sk$ for some $s \in \mathbf{Z}$. Thus

$$g^j \;=\; g^{i+sk} \;=\; g^i g^{sk} \;=\; g^i 1^s \;=\; g^i\ ,$$

proving $\Leftarrow$. Conversely, suppose that $g^i = g^j$. Then $g^{i-j} = g^i g^{-j} = 1$. Divide k into $i - j$ to give quotient m and remainder r with $0 \leq r \leq k - 1$. Then $i - j = mk + r$, so

$$1 \;=\; g^{i-j} \;=\; (g^k)^m g^r \;=\; 1^m g^r \;=\; g^r\ .$$

But $g^s \neq 1$ for $1 \leq s \leq k - 1$, so $r = 0$. Thus $i \equiv j \pmod{k}$, proving $\Rightarrow$. The first statement follows from the last, since no two of the integers $0, 1, 2, \cdots, k-1$ are congruent (mod k). So does the second statement, since every integer is congruent (mod k) to exactly one of the integers $0, 1, \cdots, k-1$.

Exercise 4E. Show that if g has infinite order, then $g^i \neq g^j$ when $i \neq j$. Deduce that every element of a finite group has finite order.

Examples.

(i) $e^{2\pi i/k}$ has order k in C_k. It and its powers, for all k, give all the elements of finite order in $\mathbf{C}^\times$, that is, all the complex roots of unity.

(ii) In S_n, every transposition has order 2.

(iii) In D_4, rotations of 90° and 270° have order 4. Rotation of 180° and the four reflections have order 2.

(iv) In D_3, rotations of 120° and 240° have order 3. The three reflections have order 2.

(v) In $\mathbf{R}^\times$, every element has infinite order except 1 (which has order 1) and -1 (which has order 2). That is, ± 1 are the only roots of unity in $\mathbf{R}^\times$.

(vi) In $\mathbf{Z}$, every element has infinite order except 0.

(vii) In $\mathbf{Z}_k$, the element [1] has order k.

(viii) Let G be the set of all complex k^{th} roots of unity for all $k \geq 1$, under complex multiplication. Then G is an infinite group, but all of its elements have finite order.

(ix) In any group, there is exactly one element of order 1.

(x) If an invertible square matrix has finite order, then all of its eigenvalues must be roots of unity. Is the converse true?

Exercise 4F. Check all the details concerning the above examples.

(xi) **Orders of permutations.**

Theorem 4.8. *If $\sigma = \gamma_1\gamma_2\cdots\gamma_s$ is a product of disjoint cycles of lengths $\ell_1, \ell_2, \cdots, \ell_s$, then the order of σ is the least common multiple of $\ell_1, \ell_2, \cdots, \ell_s$.*

Proof. We must show that

$$\sigma^j(x) = x \quad \forall x \quad \Longleftrightarrow \quad \ell_i \mid j \quad \forall i \ .$$

Suppose that $\sigma^j(x) = x$ for all $x \leq n$. Fix i, and let γ_i be the cycle $(x_1x_2\cdots x_{\ell_i})$. Define k by requiring that $1 \leq k \leq \ell_i$ and that $k \equiv j + 1 (\text{mod } \ell_i)$. Then $(\gamma_i)^j(x_1) = x_k$. Since no x_r is moved by any cycle other

than γ_i, we have $\sigma^j(x_1) = x_k$. Since all the x_r's are distinct, and $\sigma^j(x_1) = x_1$, we have $k = 1$, so $1 \equiv j + 1 \pmod{\ell_i}$. Thus $j \equiv 0 (\mathrm{mod}\ \ell_i)$, i.e. $\ell_i \mid j$, proving $\Rightarrow$.

Conversely, suppose that $\ell_i \mid j \quad \forall i \leq s$. Fix x. If x doesn't occur in any of the cycles, then $\sigma(x) = x$, so $\sigma^j(x) = x$. Otherwise, suppose that x occurs in $\gamma_i = (x_1 x_2 \cdots x_{\ell_i})$, say $x = x_m$. Then $\sigma^j(x) = x_k$, where k is defined by $1 \leq k \leq \gamma_i$ and $k \equiv j + m \pmod{\ell_i}$. But $j \equiv 0 \pmod{\ell_i}$, so $k \equiv m (\mathrm{mod}\ \ell_i)$. Since k and m are both between 1 and ℓ_i , we have $k = m$. Thus $\sigma^j(x) = x_k = x_m = x$, as required, proving $\Leftarrow$.

5. Isomorphisms and multiplication tables.

Definition. Let G and H be groups. We say that G is *isomorphic* to H, and write $G \cong H$, if and only if there is a bijective map $\phi : G \to H$ such that

$$\begin{array}{ccc} \phi(g_1 g_2) & = & \phi(g_1)\phi(g_2) \\ \uparrow & & \uparrow \\ \underline{\text{product in } G} & & \underline{\text{product in } H} \end{array}$$

for all $g_1, g_2 \in G$.

Such a map ϕ is called an *isomorphism* . This means that we have a 1-1 correspondence between elements in the two groups which 'respects' the multiplications.

Exercise 5A. Check details for the following examples :

(i) $GL(\mathbf{R}^n) \cong GL(n,\ \mathbf{R})$, an isomorphism being given by assigning to each linear map its matrix relative to some fixed basis.

(ii) $O(\mathbf{R}^n) \cong O(n)$, the isomorphism given as in (i), making sure this time that the basis is orthonormal.

(iii) $C_k \cong \mathbf{Z}_k$ by mapping $(\mathrm{e}^{2\pi i/k})^j$ to $[j]$.
[When H is additive, the condition on ϕ reads:

$$\phi(g_1 g_2) \ = \ \phi(g_1) + \phi(g_2) \ . \ ]$$

(iv) $\mathbf{Z}_2 \cong S_2$ via that ϕ for which $\phi([0]) = e$ and $\phi([1]) = (12)$.

(v) $S_3 \cong D_3$, although this is not yet obvious.

(vi) $S_4 \ncong D_4$, since $|\ S_4\ | \neq |\ D_4\ |$.

(vii) $\mathbf{R} \cong \mathbf{R}_+$ (:= the positive reals under multiplication) by letting $\phi(x) = \mathrm{e}^x$.

(viii) $\mathbf{R} \ncong \mathbf{Z}$, since $\mathbf{Z}$ is countable and $\mathbf{R}$ is not.

Clearly, isomorphic finite groups must have the same order. In addition, they must have the same 'group theoretic' properties. For example:

Proposition 5.1. *If G is commutative and $G \cong H$, then H is also commutative.*

Proof. Let $h_1, h_2 \in H$. Choose an isomorphism $\phi : G \to H$ and denote by $g_i \in G$ the elements for which $\phi(g_i) = h_i$. Then

$$h_1 h_2 \ = \ \phi(g_1)\phi(g_2) \ = \ \phi(g_1 g_2) \ = \ \phi(g_2 g_1) \ = \ \phi(g_2)\phi(g_1) \ = \ h_2 h_1 \ .$$

For example, $D_3 \ncong C_6$; $D_4 \ncong C_8$; and $S_n \ncong C_{n!}$ for $n > 2$; although in each case the two groups have the same order.

Proposition 5.2. *If $\phi : G \to G'$ is an isomorphism, then*

$$\phi(1) \ = \ 1' \ , \quad \text{and} \quad \phi(g^{-1}) \ = \ \phi(g)^{-1} \text{ for all } g.$$

Proof. $\phi(1) = \phi(1 \cdot 1) = \phi(1)\phi(1)$. Multiply by $\phi(1)^{-1}$ to get the first identity. Now $\phi(g^{-1})\phi(g) = \phi(g^{-1}g) = \phi(1) = 1'$, so $\phi(g^{-1})$ is the inverse of $\phi(g)$, as required, using the following.

Exercise 5B. In any group, if $ab = 1$, then $ba = 1$.

Proposition 5.3. *If $\phi : G \to G'$ is an isomorphism, then $\|\phi(g)\| = \|g\|$ for all $g \in G$.*

Proof. Firstly $g^j = 1 \Rightarrow [\phi(g)]^j = \phi(g^j) = \phi(1) = 1'$. (Prove the second equality as an exercise.) On the other hand,

$$[\phi(g)]^j \ = \ 1' \ \implies \ \phi(1) \ = \ 1' \ = \ \phi(g^j) \ \implies \ g^j \ = \ 1 \ ,$$

since ϕ is injective. Thus $g^j = 1 \Leftrightarrow [\phi(g)]^j = 1'$, and so g and $\phi(g)$ have the same order, the smallest positive j for which both statements hold.

Examples. (i) $\mathbf{Z} \ncong \cup_{k=1}^{\infty} C_k$, although both of them are countable abelian groups.

(ii) Let

$$D_2 = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\}$$

The set D_2 is a group under matrix multiplication (and is isomorphic to the group of symmetries of a suitable two-sided figure). If A and B are any two non-identity elements in D_2, then AB is the third, and we have $A^2 = B^2 = (AB)^2 = I$, so D_2 has no element of order 4. Thus $D_2 \ncong C_4$, though both are commutative of order 4. The group D_2 is called the *Klein 4-group* .

For a *finite* group one can make a multiplication table :

C_4

	1	i	i^2	i^3
1	1	i	i^2	i^3
i	i	i^2	i^3	1
i^2	i^2	i^3	1	i
i^3	i^3	1	i	i^2

D_2

	I	A	B	AB
I	I	A	B	AB
A	A	I	AB	B
B	B	AB	I	A
AB	AB	B	A	I

Pseudotheorem 5.3.5: *Two finite groups are isomorphic if and only if you can name the elements of both in such a way that the corresponding multiplication tables are the same.*

Example. In D_3, let a = rotation of 120°, and let b = reflection in the y-axis. In S_3, let $a = (123)$ and let $b = (12)$. Then both have the following multiplication table :

	1	a	a^2	b	ab	a^2b
1	1	a	a^2	b	ab	a^2b
a	a	a^2	1	ab	a^2b	b
a^2	a^2	1	a	a^2b	b	ab
b	b	a^2b	ab	1	a^2	a
ab	ab	b	a^2b	a	1	a^2
a^2b	a^2b	ab	b	a^2	a	1

This gives an algorithm (but a very impractical one) for determining up to isomorphism all possible groups of any given finite order. However, it is

not even known how many there are, for example, of order 256. (See also Sections **8, 12** and **13**).

Proposition 5.4. *Isomorphism behaves like an equivalence relation. That is, we have the following.*

(i) $G \cong G$ *for all* G, ; *in fact, the identity map* id: $G \to G$ *is an isomorphism.*

(ii) $G \cong H \Rightarrow H \cong G$; *in fact, if* $\phi : G \to H$ *is an isomorphism, then* $\phi^{-1} : H \to G$ *is also an isomorphism.*

(iii) $[G \cong H \ \& \ H \cong K] \Rightarrow G \cong K$; *in fact, if* $\phi : G \to H$ *and* $\psi : H \to K$ *are isomorphisms, then so is* $\psi\phi : G \to K$.

Proof. The maps id, ϕ^{-1} and $\psi\phi$ are certainly bijective. We must show that they map products to products:

$\mathrm{id}(g_1g_2) \ = \ g_1g_2 \ = \ \mathrm{id}(g_1)\mathrm{id}(g_2)$.

$$(\psi\phi)(g_1g_2) \ = \ \psi[\phi(g_1g_2)] \ = \ \psi[\phi(g_1)\phi(g_2)]$$
$$= \ [\psi(\phi(g_1)][(\psi(\phi(g_2)] \ = \ [(\psi\phi)(g_1)][(\psi\phi)(g_2)] \ .$$

Given $h_1, h_2 \in H$, choose $g_i \in G$ with $\phi(g_i) = h_i$. Then

$$\phi^{-1}(h_1h_2) \ = \ \phi^{-1}[\phi(g_1)\phi(g_2)] \ = \ \phi^{-1}[\phi(g_1g_2)]$$
$$= \ g_1g_2 \ = \ \phi^{-1}(h_1)\phi^{-1}(h_2) \ .$$

6. Subgroups and cyclic groups.

Definition. A sub*set* H of a group G is a *subgroup* of G exactly when the following three conditions hold:
(I) $a \in H$ and $b \in H \implies ab \in H$;
(II) $1 \in H$; and
(III) $a \in H \implies a^{-1} \in H$.

Note. By (I), H inherits a multiplication from G, one which is associative. By (II) and (III), H is a group with this multiplication.

Examples. $C_k \subset \mathbf{C}^*$; $O(\mathbf{R}^n) \subset GL(\mathbf{R}^n)$; $O(n) \subset GL(n, \mathbf{R})$;
$D_3 \subset O(\mathbf{R}^2)$; $D_4 \subset O(\mathbf{R}^2)$; $3\mathbf{Z} = \{0, \pm 3, \pm 6, \pm 9, \cdots\} \subset \mathbf{Z} \subset \mathbf{R}$;
$\{[0],\ [3]\} \subset \mathbf{Z}_6$; $\{1,\ -1\} \subset C_6$; $A_n := \{\sigma \ : \ \mathrm{sign}(\sigma) = 1\} \subset S_n$.

Exercise 6A. Prove: i) In each case above, the smaller is a subgroup of the larger. ii) For any group G, the extreme subsets $\{1\}$ and G itself are both subgroups of G.

Proposition 6.1. *Let H be a non-empty subset of a group G.*
Then (H is a subgroup of G) $\iff$ ($[a \in H$ and $b \in H]$ $\Rightarrow$ $a^{-1}b \in H$).

Proof. To prove $\Rightarrow$, note that $a^{-1} \in H$ by (III), so that $a^{-1}b \in H$ by (I). To prove $\Leftarrow$, choose $b \in H$, since H is non-empty. Then $b^{-1}b = 1 \in H$, so (II) holds. If $a \in H$, then $a^{-1}1 = a^{-1} \in H$, so (III) holds. Now if $a \in H$ and $b \in H$, then $a^{-1} \in H$, so $(a^{-1})^{-1}b = ab \in H$, so (I) holds.

Proposition 6.2. *If $\{H_\alpha\}$ is an indexed non-empty collection of subgroups of G, then $\cap_\alpha H_\alpha$ is a subgroup of G.*

Proof. Since $1 \in H_\alpha$ $\forall\alpha$, we have $1 \in \cap_\alpha H_\alpha$, so $\cap_\alpha H_\alpha \neq \emptyset$. Now apply **6.1**. If $a \in \cap_\alpha H_\alpha$ and $b \in \cap_\alpha H_\alpha$, then $\forall\alpha$, $a \in H_\alpha$ and $b \in H_\alpha$. Thus $a^{-1}b \in H_\alpha$ $\forall\alpha$. Thus $a^{-1}b \in \cap_\alpha H_\alpha$, and so $\cap_\alpha H_\alpha$ is a subgroup.

Exercise 6B. In general, a union of subgroups is not a subgroup. What set theoretic condition on a family of subgroups is equivalent to its union being a subgroup?

Definition. Let A be any sub*set* of a group G. The *subgroup of G generated by A* is the smallest subgroup of G which contains A. This definition is somewhat glib (**greatest lower informative bound**), since there is no guarantee beforehand that such a subgroup exists. The following explicit descriptions of this subgroup take care of that problem.

Two descriptions:

$$H_1 \ := \ \bigcap\{ \ H \ : \ A \subset H \ \text{ and } \ H \text{ is a subgroup of } G \ \} \ ;$$

i.e. H_1 is the intersection of all subgroups containing A.

$$H_2 \ := \ \{\prod_{i=1}^{k} a_i \ : \ k \geq 0 \ ; \text{ and } \ \forall i, \text{ either } a_i \in A \text{ or } a_i^{-1} \in A \ \} \ ;$$

i.e. H_2 is the set of all products, $a_1 a_2 \cdots a_k$, each of whose factors is either an element from A or the inverse of an element from A.

Theorem 6.3. *The sets H_1 and H_2 are both subgroups of G and are in fact equal.*

Proof. The set H_1 is a subgroup by **6.2**. The set H_2 is a subgroup using either **6.1** or the definition of the term *subgroup*. Now $H_1 \subset H_2$, since H_2 qualifies as an H in the intersection defining H_1, i.e. $A \subset H_2$ and H_2 is a subgroup. On the other hand, $H_2 \subset H_1$ since each $a_i \in H_1$ and H_1 is closed under multiplication.

Examples.
(i) If A *is* a subgroup, then the subgroup generated by A is A itself.
(ii) The symmetric group S_n is generated by its subset of transpositions.

Exercise 6C. Prove that S_n is generated by $\{\ (12)\ ,\ (123\cdots n)\ \}$, and also by $\{\ (12)\ ,\ (23)\ ,\ \cdots,\ (n-1\ n)\ \}$. Prove that A_n (which is called the *alternating group*) is generated by $\{\ (123)\ ,\ (124)\ ,\ \cdots,\ (12n)\ \}$.

(iii) If $A = \{g\}$, description H_2 shows that the group generated by A is the set, $\{\ g^j\ :\ j \in \mathbf{Z}\ \}$, of all powers of g.

Definition. A group G is a *cyclic group* if and only if G can be generated by one element (more precisely, $\cdots$ by a singleton set).

Theorem 6.4. *If G is finite cyclic, then $G \cong \mathbf{Z}_{|G|}$. If G is infinite cyclic, then $G \cong \mathbf{Z}$.*

Proof. Suppose that G is cyclic and choose a generator g. Define a map $\phi : \mathbf{Z} \to G$ by $\phi(j) = g^j$. Then

$$\phi(j+k) \ = \ g^{j+k} \ = \ g^j g^k \ = \ \phi(j)\phi(k)\ .$$

Also ϕ is surjective since G consists of all the powers of g.

Now suppose further that G is infinite. Then all powers of g are distinct by **4E**, so ϕ is injective. Thus ϕ is an isomorphism and $G \cong \mathbf{Z}$.

On the other hand, suppose that G is finite of order k. By **4.7**,

$$G \ = \ \{\ 1, g, g^2, \cdots, g^{k-1}\ \}\ ,$$

and g has order k. Define $\psi : \mathbf{Z}_k \to G$ by $\psi([j]) = g^j$. By **4.7**, ψ is well

defined and bijective. Also

$$\psi([j]+[k]) \;=\; \psi([j+k]) \;=\; g^{j+k} \;=\; g^j g^k \;=\; \psi([j])\psi([k])\ .$$

Hence ψ is an isomorphism, and $G \cong \mathbf{Z}_k$, as required.

Corollary 6.5. *Up to isomorphism, there is only one cyclic group of each finite order and one infinite cyclic group.*

Note. Any element in a group generates a unique cyclic subgroup of the group. The order of that subgroup is the order of the element.

6^+ Cayley's Theorem.

If $G = \{\ g_1,\ g_2,\ \cdots,\ g_n\ \}$ is a finite group of order n, and $x \in G$, then the sequence $xg_1,\ xg_2,\ \cdots,\ xg_n$ is just $g_{\sigma(1)},\ g_{\sigma(2)},\ \cdots,\ g_{\sigma(n)}$ for some unique $\sigma \in S_n$. If x corresponds to σ in this way, and y to γ, then xy corresponds to $\sigma\gamma$, and 1 to e, and x^{-1} to σ^{-1}. The subset of such σ corresponding to elements of G is a subgroup, H, of S_n , and $G \cong H$. This is **Cayley's theorem** :

Any group of order n is isomorphic to a subgroup of S_n.

Exercise 6D. Verify all these statements to prove the theorem (or prove it in some other way, if you prefer).

7. Cosets and Lagrange's theorem.

One can generalize the number theoretic notion of *congruence*(mod k) as follows.

Definition. Let H be a subgroup of G and let $a, b \in G$. Then we say that *a is congruent to b (*mod H*)*, and write $a \equiv b(\text{mod } H)$, if and only if $a^{-1}b \in H$. [In an additive group, $\cdots \Leftrightarrow b - a \in H$.]

Examples.

(i) $G = \mathbf{Z}$, $H = n\mathbf{Z} = \{\ 0,\ \pm n,\ \pm 2n,\ \pm 3n,\ \cdots\ \}$. Then
$a \equiv b(\text{mod } n\mathbf{Z}) \iff a \equiv b(\text{mod } n)$.

(ii) $G = \mathbf{C}^\times, \quad H = \{\ z \in \mathbf{C}\ :\ |z| = 1\ \}$, the circle group.
Then $a \equiv b (\text{mod } H) \iff |a| = |b|$. (N.B. Of course, here $|\ |$ means 'complex modulus', not 'order'.)

(iii) $G = \mathbf{R}$, $H = \mathbf{Z}$. Then $a \equiv b (\text{mod } \mathbf{Z}) \iff a$ and b have the same decimal expansion to the right of the decimal point.

(iv) $G = S_n$, $H = A_n =$ the group of even permutations. Then $a \equiv b (\text{mod } A_n) \iff a$ and b are either both even or both odd.

(v) If $H = G$, then $a \equiv b (\text{mod } H)$ for all a and b.

(vi) If $H = \{1\}$, then $a \equiv b (\text{mod } H) \iff a = b$.

(vii) If $G = S_3$ and $H = \{e, (12)\}$, then we have three congruences:

$$e \equiv (12) (\text{mod } H)\ ,\ \ (123) \equiv (13) (\text{mod } H)\ , \text{and } (132) \equiv (23) (\text{mod } H)\ ,$$

but no other pair of distinct elements is congruent (mod H) (except by reversing the order of a pair above).

Exercise 7A. Check the details for each example.

Proposition 7.1. *Congruence (mod H) is an equivalence relation on the set G.*

Proof. (i) For all a, we have $a \equiv a (\text{mod } H)$, since $a^{-1}a = 1 \in H$.

(ii) If $a \equiv b (\text{mod } H)$ then $b \equiv a (\text{mod } H)$, since $a^{-1}b \in H$ implies that $b^{-1}a = (a^{-1}b)^{-1} \in H$.

(iii) If $a \equiv b (\text{mod } H)$ and $b \equiv c (\text{mod } H)$, then $a \equiv c (\text{mod } H)$: assuming that both $a^{-1}b \in H$ and $b^{-1}c \in H$ results in $a^{-1}c \in H$, since $a^{-1}c = (a^{-1}b)(b^{-1}c)$.

Proposition 7.2. *The equivalence class, $[g_0]$, of g_0 under congruence (mod H) is $\{\ g_0h\ :\ h \in H\ \}$, the set of 'right multiples' of g_0 by elements of H.*

Proof. $x \in [g_0] \iff g_0 \equiv x (\text{mod } H) \iff g_0^{-1}x \in H \iff \exists h \in H$ such that $x = g_0h \iff x \in \{\ g_0h\ :\ h \in H\ \}$.

Definition. The class in **7.2** is called the *left coset of g_0 (mod H)*, and is denoted g_0H (in additive notation, $g_0 + H$), rather than $[g_0]$.

Corollary 7.3. *If $g_1H \neq g_2H$, then $g_1H \cap g_2H$ is empty. Also*

$$G \;=\; \bigcup_{g_0 \in G} g_0H \ .$$

Proof. This is just the partitioning effect of any equivalence relation. (It may also be proved directly rather easily.)

Proposition 7.4. *If H is finite, then g_0H has the same number of elements as H, for any $g_0 \in G$.*

Proof. Define $\mu : H \to g_0H$ by $\mu(h) = g_0h$. Then μ is bijective (whether H is finite or not), proving the assertion (and more).

Theorem 7.5. (Lagrange) *The order of any subgroup of a finite group divides the order of the group. In fact,*

$$|G| \;=\; |H| \times \text{ number of left cosets of } G(\text{mod } H) \ ,$$

for any subgroup, H, of G.

Proof. By **7.3**, G is the disjoint union of its left cosets (mod H). By **7.4**, all left cosets have the same number of elements, namely $|H|$.

Definition. The number of left cosets of $G(\text{mod } H)$ is called the *index of H in G*, and denoted $[G : H]$. Thus **7.5** may be written $[G : H] \;=\; |G| \,/\, |H|$.

Corollary 7.6. *The order of any element of a finite group divides the order of the group .*

Proof. The order of an element g is the order of the cyclic subgroup H generated by g, so this follows from **7.5**.

Exercise 7B. Define *right coset* and a relation analogous to congruence whose equivalence classes are the right cosets. Deduce from an analogue of **7.5** that the number of right cosets is also the index.

Examples.
(i) C_{12} has the following proper subgroups :

$$\{1\}\ ;\ \{1,\rho^6\}\ ;\ \{1,\rho^4,\rho^8\}\ ;\ \{1,\rho^3,\rho^6,\rho^9\} \text{ and } \{1,\rho^2,\rho^4,\rho^6,\rho^8,\rho^{10}\}\ ,$$

of orders $1, 2, 3, 4$ and 6 respectively, all dividing 12. For example, $\{1,\rho^6\}$ has index 6 in C_{12} , the six cosets being:

$$\{1,\rho^6\}\ ;\ \{\rho,\rho^7\}\ ;\ \{\rho^2,\rho^8\}\ ;\ \{\rho^3,\rho^9\}\ ;\ \{\rho^4,\rho^{10}\} \textit{ and } \{\rho^5,\rho^{11}\}\ .$$

(ii) $\mathbf{Z}_{12}$ has proper subgroups

$$\{[0]\}\ ;\ \{[0],[6]\}\ ;\ \{[0],[4],[8]\}\ ;\ \{[0],[3],[6],[9]\} \text{ and } \{[0],[2],[4],[6],[8],[10]\}.$$

This example is similar to (i) because $C_{12} \cong \mathbf{Z}_{12}$.

(iii) S_3 has proper subgroups

$$\{e\}\ ;\ \{e,(12)\}\ ;\ \{e,(13)\}\ ;\ \{e,(23)\}\ ;\ \text{and } \{e,(123),(132)\}$$

of orders 1,2,2,2 and 3, all dividing 6. For example $\{e,(12)\}$ has index 3 in S_3, the left cosets being

$$\{e,(12)\}\ ;\ \{(123),(13)\} \text{ and } \{(132),(23)\}\ .$$

(iv) D_4 has subgroups : $\{I\}$, of order 1 ;

$$\{I,A^2\}\ ,\ \{I,B\}\ ,\ \{I,AB\}\ ,\ \{I,A^2B\}\ ,\ \{I,A^3B\}\ ,$$

of order 2, all isomorphic to C_2 ;

$$\{I,A,A^2,A^3\} \text{ isomorphic to } C_4 \text{ and}$$

$$\{I,A^2,B,A^2B\} \text{ and } \{I,A^2,AB,A^3B\} \text{ isomorphic to } D_2$$

all of order 4; and itself of order 8.

Corollary 7.7. *Assuming that G is a finite group, for all $g \in G$, we have*

$$g^{|G|} \ = \ 1\ .$$

Proof. By **7.6**, there exists $r \in \mathbf{Z}$ with $|G| = r||g||$. Then

$$g^{|G|} \;=\; g^{||g||r} \;=\; 1^r \;=\; 1\ .$$

Corollary 7.8. (Euler) *$a^{\Phi(k)} \equiv 1 (mod\ k)$ for all integers a prime to k, where $\Phi(k)$ is defined to be the number of integers prime to k between 1 and k.*

Proof. Apply **7.7** with $g = [a]$ and $G = \mathbf{Z}_k^\times$, the group of congruence classes prime to k under multiplication (mod k), noting that $|\mathbf{Z}_k^\times| \;=\; \Phi(k)$.

We've come down from the ethereal abstractions of general group theory to the **terra firma** of number theory. And you've just pronounced the next mathematician's name more or less correctly, except perhaps for emphasizing the second syllable. In particular, Fermat doesn't rhyme with doormat!

Corollary 7.9. (Fermat) *If p is prime, then $a^p \equiv a (mod\ p)$ for all $a \in \mathbf{Z}$.*

Proof. This is trivial if p divides a. If not, let $k = p$ in **7.8**, giving $a^{p-1} \equiv 1 (\mathrm{mod} p)$, so $a^p \equiv a (\mathrm{mod}\ p)$.

Exercise 7C. This last step seems to use some general fact. Assume that $a \equiv b (\mathrm{mod}\ H)$ Given a third element c, which, if either, of the following then hold :

$$ca \equiv cb (\mathrm{mod}\ H) \qquad ; \qquad ac \equiv bc (\mathrm{mod}\ H) \qquad ?$$

8. Direct product ; groups of small order.

We've already been using the Cartesian product of two sets A and B, which is defined to be

$$A \times B \;:=\; \{\ (a,b)\ :\ a \in A \text{ and } b \in B\ \}\ .$$

When both sets have a group structure, so does their product, as follows:

Definition. If G and H are groups, define their direct product to be the set $G \times H$ with the multiplication :

$$\begin{array}{ccc} (g_1, h_1)\ (g_2, h_2) & := & (g_1 g_2\ ,\ h_1 h_2) \\ \uparrow & & \uparrow \quad \uparrow \\ \overline{\text{(multiplication being defined)}} & & \overline{\text{(already known multiplications in } G \text{ and } H\text{)}} \end{array}$$

Proposition 8.1. *The set $G \times H$ is a group with this multiplication. The identity is $(1_G, 1_H)$, where 1_G and 1_H are the identities in G and H respectively. The inverse of (g, h) is (g^{-1}, h^{-1}).*

Proof. (1) To verify associativity, we have

$$\begin{aligned} [(g_1, h_1)\ (g_2, h_2)]\ (g_3, h_3) &= (g_1 g_2\ ,\ h_1 h_2)\ (g_3, h_3) \\ = ((g_1 g_2) g_3\ ,\ (h_1 h_2) h_3) &= (g_1 (g_2 g_3)\ ,\ h_1 (h_2 h_3)) \\ = (g_1, h_1)\ (g_2 g_3\ ,\ h_2 h_3) &= (g_1, h_1)\ [(g_2, h_2)\ (g_3, h_3)]\ . \end{aligned}$$

(2) Next,

$$(1_G, 1_H)(g, h) = (1_G \cdot g, 1_H \cdot h) = (g, h) = (g \cdot 1_G, h \cdot 1_H) = (g, h)(1_G, 1_H)\ .$$

(3) Finally,

$$(g^{-1}, h^{-1})(g, h) = (g^{-1} g, h^{-1} h) = (1_G, 1_H)\ .$$

Similarly, $(g, h)(g^{-1}, h^{-1}) = (1_G, 1_H)$.

Exercise 8A. *Some lines above are unnecessary.* Show that, to check that a given 'set with associative operation'—called a *'semigroup'*— is actually a group, it is only necessary to check that there is a (say) *left* identity element, and then that each element has a *left* inverse. On the other hand, replacing "left" by "right" only once above is not a sufficient condition for 'groupiness'.

Notes. (a) If G_1, G_2 and G_3 are groups, then one might wish to define $G_1 \times G_2 \times G_3$ to be the set of ordered triples (g_1, g_2, g_3), with the obvious

coordinate-wise multiplication. It is almost equal to, and certainly isomorphic to, both $(G_1 \times G_2) \times G_3$ and $G_1 \times (G_2 \times G_3)$. Similarly for any finite sequence, G_1 , G_2 , $\cdots$, G_n , of groups.

(b) Clearly $|G \times H| \ = \ |G| \cdot |H|$, so inductively

$$|G_1 \times G_2 \times \cdots \times G_n| \ = \ |G_1| \cdot |G_2| \cdot \ldots \cdot |G_n| \ .$$

(c) If G and H are both commutative, then so is $G \times H$.

Examples. (i) $C_2 \times C_2 \ \cong \ D_2 \ \ncong \ C_4$.
(ii) $C_2 \times C_3 \ \cong \ C_6 \ \ncong \ S_3 \ \cong \ D_3$.
(iii) $C_2 \times C_4 \ \ncong$ either C_8 or D_4 ; $C_2 \times C_2 \times C_2 \ \ncong$ any of the previous three.
(iv) $\mathbf{R} \times \mathbf{R} \times \cdots \times \mathbf{R} \ = \ \mathbf{R}^n$; $\mathbf{R}^n$ and $\mathbf{R}$ are isomorphic *as groups* (HINT: even as vector spaces over $\mathbf{Q}$), but *not* isomorphic as real vector spaces if n is larger than one.
(v) $\mathbf{R}^\times \times \mathbf{R}^\times \times \cdots \times \mathbf{R}^\times \ \cong$ the subgroup consisting of all the diagonal matrices in $GL(n, \mathbf{R})$.
(vi) $S^1 \times S^1$ looks like a (hollow) torus, where S^1 is the circle group.
(vii) For any G, we have $\{1\} \times G \ \cong \ G$; for any G and H, we have $G \times H \ \cong H \ \times G$.

Exercise 8B. Check the details above.

Theorem 8.2. *If $|G|$ is a prime p, then $G \cong C_p$.*
Thus, for $k = 2, 3, 5, 7, 11, \cdots$, there is only *one* group, up to isomorphism, of order k.

Proof. Since $p > 1$, we can choose $g \in G$ with $g \neq 1$. Then $||g|| > 1$, but $||g||$ divides p by **7.6**, so $||g|| = p$, since p is prime. Thus there are "p" distinct powers of g. Thus $G = \{1, g, g^2, \cdots, g^{p-1}\}$ is the cyclic group generated by g. By **5.4**, $G \cong C_p$.

Proposition 8.3. *Any group of order 4 is isomorphic to exactly one of C_4 or $C_2 \times C_2$.*

Proof. If G has an element g of order 4, then $G = \{1, g, g^2, g^3\}$ is cyclic of order 4, so $G \cong C_4$ by **5.4**. If not, then every non-identity element has

order 2, by **7.6**. Let a and b denote distinct elements of order 2. Then $ab \neq a$ since $b \neq 1$; $ab \neq b$ since $a \neq 1$; and $ab \neq 1$ since $a^{-1} = a \neq b$. Hence $G = \{1, a, b, ab\}$ and $a^2 = b^2 = (ab)^2 = 1$. Also $ba = a^2(ba)b^2 = a(ab)^2b = ab$. This determines the multiplication table of G; and $C_2 \times C_2$ has the same multiplication table if we let a and b be $(-1, 1)$ and $(1, -1)$. (This is the same as the table given for D_2 previously, if we change A to a and B to b.) Thus $G \cong C_2 \times C_2$ (using **5.3.5**).

Exercise 8C. Show that a group whose elements satisfy $g^2 = 1$ for all g is necessarily commutative.

The classification of the remaining cases for orders less than 12 is given below in **8.4** to **8.8**.

Proposition 8.4. *Any group of order 6 is isomorphic to exactly one of C_6 or D_3.*

Proposition 8.5. *Any group of order 10 is isomorphic to exactly one of C_{10} or D_5.*

Proposition 8.6. *Any group of order 9 is isomorphic to exactly one of C_9 or $C_3 \times C_3$.*

The proofs of these are more enlightening using general theorems. They are done in sections **11** and **13**. In **8.4** and **8.5**, the two given groups are not isomorphic because one is commutative and the other isn't. Alternatively we can look at the orders of elements. In **8.6**, the group C_9 has elements of order 9, but $C_3 \times C_3$ doesn't, using:

Proposition 8.7. *If group elements g and h have finite order, then*

$$||(g,h)|| \;=\; \text{LCM}\{\; ||g|| \;,\; ||h|| \;\}\;.$$

Proof. By induction and manipulation, $(g,h)^n \;=\; (g^n, h^n)$. So :

$$(g,h)^n \;=\; (1_G, 1_H) \iff (g^n \;=\; 1_G \quad \text{and} \quad h^n \;=\; 1_H) \iff$$

$$(||g|| \text{ divides } n \;,\; \text{and } ||h\;|| \text{ divides } n) \Longleftrightarrow \text{LCM}\{\; ||g|| \;,\; ||h|| \;\} \text{ divides } n\;.$$

Finally, consider groups of order 8. We have three commutative groups C_8 , $C_4 \times C_2$, and $C_2 \times C_2 \times C_2$ (no pair isomorphic since the maximum orders of elements are 8, 4, and 2 respectively). There is also the non-commutative group D_4. There is a fifth group of order 8, namely Qrn, the quaternion group :

$$Qrn \ = \ \{ \ \pm \mathbf{1} \ , \ \pm i \ , \ \pm j \ , \ \pm k \ \} \ ,$$

often pictured as a subset of $\mathbf{R}^4$, where

$$\mathbf{1} \ = \ (1,0,0,0) \ \ ; \ \ i \ = \ (0,1,0,0) \ \ ; \ \ j \ = \ (0,0,1,0) \ \ ; \ \ k \ = \ (0,0,0,1) \ .$$

The multiplication table is determined by

$$i^2 = j^2 = k^2 = -\mathbf{1} \ ; \ (-\mathbf{1})(i) = -i \ ; \ (-\mathbf{1})(j) = -j \ ; \ (-\mathbf{1})(k) = -k; \text{ and}$$

$$ij \ = \ k \ = \ -ji \ \ ; \ \ jk \ = \ i \ = \ -kj \ \ ; \ \ ki \ = \ j \ = \ -ik \ .$$

One can verify associativity by computing; Qrn is clearly not commutative; and $Qrn \ncong D_4$, since Qrn has only one element, $-\mathbf{1}$, of order 2, whereas D_4 has five such elements (four reflections and rotation of 180°).

Proposition 8.8. *Any group of order 8 is isomorphic to exactly one of* C_8 , $C_4 \times C_2$, $C_2 \times C_2 \times C_2$, D_4 *or* Qrn.
(See sections **11** and **13**).

The problem of classifying finite groups has attracted mathematicians for many years, with much success, but no final solution yet in sight. Summary of our statements:

order	1	2	3	4	5	6	7	8	9	10	11
# of groups	1	1	1	2	1	2	1	5	2	2	1
# of commutative groups	1	1	1	2	1	1	1	3	2	1	1

Definition. Let H and L be subgroups of a group G. We say that G *splits as the internal direct product of* H *and* L if and only if the following map is an isomorphism:

$$\phi : H \times L \longrightarrow G$$
$$(h, l) \mapsto hl \quad ; \quad \text{i.e. } \phi(h, l) = hl \ .$$

Theorem 8.9. *A group G splits as the internal direct product of H and L if and only if the following three conditions hold:*
(i) $H \cup L$ generates G;
(ii) $H \cap L = \{1\}$; and
(iii) if $h \in H$, and $\ell \in L$, then $\ell h = h\ell$.

Note. *(i) ,(ii)*, and *(iii)* are often used as the *definition* of splitting.

Proof. $\Rightarrow$:
(i) If $g \in G$, then $g \in Im\phi$, so $g = h\ell$ for some $h \in H$ and $\ell \in L$. Thus g is in the subgroup generated by $H \cup L$, as required.
(ii) If $g \in H \cap L$, then $(g, g^{-1}) \in H \times L$. But $\phi(g, g^{-1}) = 1$, and ϕ is injective, so $(g, g^{-1}) = (1, 1)$, i.e. $g = 1$, as required.
(iii) We have

$$h\ell \ = \ \phi((h, \ell)) \ = \ \phi((1, \ell)(h, 1)) \ = \ \phi((1, \ell))\phi((h, 1)) \ = \ \ell h \ .$$

$\Leftarrow$: We prove this using *iii), ii)* and *i)* in that order. The function ϕ is a *morphism of groups* (see Section **9**), since
$\phi[(h_1, \ell_1)(h_2, \ell_2)] = \phi(h_1h_2, \ell_1\ell_2) = h_1h_2\ell_1\ell_2$
$$= h_1\ell_1h_2\ell_2 = \phi(h_1, \ell_1)\phi(h_2, \ell_2) \ .$$
It is *injective*, since
$$\phi[(h, \ell)] \ = \ 1 \implies h\ell \ = \ 1 \implies h \ = \ \ell^{-1} \in H \cap L$$
$$\implies h \ = \ \ell^{-1} \ = \ 1 \implies (h, \ell) \ = \ (1, 1) \ ,$$
as required.
It is *surjective* since Imϕ is a subgroup of G containing $H \cup L$ and so it must be all of G, as required. It contains $H \cup L$ because

$$\phi[(h, 1)] \ = \ h \quad \text{and} \quad \phi[(1, \ell)] \ = \ \ell \ .$$

Example. C_{12} is *not* the internal direct product of C_2 and C_3, since $C_2 \cup C_3$ generates only C_6. It is *not* the internal direct product of C_4 and C_6 since $C_4 \cap C_6 = C_2 \neq 1$. It *is* the internal direct product of C_4 and C_3. The

first two statements follow more easily by noting that 12 is equal to neither 2×3 nor 4×6.

9. Morphisms.

Definition. A function $\phi : G \to H$ between groups is a *homomorphism* or a *morphism of groups* if and only if, for all g_1 and g_2 in G, we have

$$\phi(g_1g_2) \;=\; \phi(g_1)\phi(g_2)\ .$$

An *epimorphism* is a surjective morphism. A *monomorphism* is an injective morphism. An *isomorphism* is a bijective morphism.

Thus 'iso.' is the same as 'both mono. and epi.', and agrees with the previous definition.

Exercise 9A. Check all the details in the examples below.

Examples. (i) $\phi : \mathbf{Z} \to \mathbf{Z}_k$, $\phi(n) := [n]$, is an epimorphism.

(ii) $\phi : C_k \to \mathbf{C}^*$, $\phi(x) := x$, is a monomorphism.

(iii) $\phi : H \to G$, $\phi(x) := x$, where H is a subgroup of G, is a monomorphism.

(iv) $\phi : GL(n, \mathbf{R}) \to \mathbf{R}^*$, $\phi(A) := \det A$, is an epimorphism.

(v) $\phi : \mathbf{R}^* \to GL(n, \mathbf{R})$, $\phi(x) := xI$, the diagonal matrix with all x's in the diagonal, is a monomorphism.

(vi) $\phi : S_n \to C_2$, $\phi(\sigma) := \mathrm{sign}(\sigma)$, is an epimorphism for $n \geq 2$, an isomorphism for $n = 2$, and a monomorphism for $n = 1$ and 2.

(vii) If $\phi : G \to H$ and $\psi : H \to J$ are both morphisms, then so is $\psi \circ \phi : G \to J$.

From here until the end of Section 10, ASSUME THAT
$\phi : G \to G'$ is a morphism, and that 1 and $1'$ are the identity elements in groups G and G', respectively.

Proposition 9.1. We have $\phi(1) \;=\; 1'$, and $\phi(g^{-1}) \;=\; \phi(g)^{-1}$ for all $g \in G$.

Proof. This is the same as that for **5.2**; bijectivity was not used in that proof.

Proposition 9.2. $\mathrm{Im}\phi := \{\ \phi(g)\ :\ g \in G\ \}$ *is a subgroup of* G'.

Proof. Now G is not empty, so $\mathrm{Im}\phi$ is not empty. Also

$$\phi(a)^{-1}\phi(b)\ =\ \phi(a^{-1})\phi(b)\ =\ \phi(a^{-1}b) \in \mathrm{Im}\phi\ ,$$

so $\mathrm{Im}\phi$ is a subgroup by **6.1**. ($\mathrm{Im}\phi$ is called the *image of* ϕ.)

Definition. The *kernel of* ϕ is the subset of G defined by

$$\mathrm{Ker}\phi\ :=\ \{\ g \in G\ :\ \phi(g) = 1'\ \}\ .$$

Proposition 9.3. $\mathrm{Ker}\phi$ *is a subgroup of* G.

Proof. We know that $\phi(1) = 1'$, so $\mathrm{Ker}\phi$ isn't empty. Also

$$a, b \in \mathrm{Ker}\phi\ \implies\ \phi(a^{-1}b)\ =\ \phi(a)^{-1}\phi(b)\ =\ 1'^{-1}1'\ =\ 1'\ ,$$

so **6.1** proves it.

Examples. (i) $\phi : \mathbf{Z} \to \mathbf{Z}_k\ , \phi(n) := [n]$, has kernel equal to the subgroup $k\mathbf{Z} = \{\ 0,\ \pm k,\ \pm 2k,\ \cdots\ \}$. (Note that $k\mathbf{Z}$ is *not* coset notation here.)

(ii) $\phi : GL(n, \mathbf{R}) \to \mathbf{R}^*\ , \phi(A) := \det A$, has kernel equal to the group, $SL(n, \mathbf{R})$, of all matrices of determinant $+1$. These groups are called the *general linear group* and the *special linear group*, respectively.

(iii) $\phi : S_n \to C_2\ , \phi(\sigma) := \mathrm{sign}(\sigma)$, has kernel A_n, the *alternating* group (of all even permutations).

(iv) If ϕ is a monomorphism, then $\mathrm{Ker}\phi = \{1\}$ and $\mathrm{Im}\phi \cong G$.

(v) The trivial morphism $\phi : G \to G'$, $\phi(g) := 1'$ for all g, has kernel equal to G.

(vi) $\phi : C_k \to C_\ell$, $\phi(x) := x^{k/d}$, where $d = \mathrm{GCD}\{k, \ell\}$, has kernel equal to $C_{k/d}$, and image C_d.

Exercise 9B. Show that this last ϕ is a well-defined function and a morphism. Also check the other details in these examples.

Proposition 9.4. ϕ *is a monomorphism if and only if* $\mathrm{Ker}\phi = \{1\}$.

Proof. If ϕ is a monomorphism, then $\phi^{-1}(g')$ contains at most one element, for all $g' \in G'$. But then $\mathrm{Ker}\phi = \phi^{-1}(1') = \{1\}$. Conversely suppose that $\mathrm{Ker}\phi = \{1\}$. Then

$$\phi(a) = \phi(b) \implies \phi(a^{-1}b) = \phi(a^{-1})\phi(b) = 1'^{-1}1' = 1'$$

$$\implies a^{-1}b \in \mathrm{Ker}\phi \implies a^{-1}b = 1 \implies a = b\ .$$

Thus ϕ is injective, as required.

10. Normal subgroups and the first isomorphism theorem.

Proposition 10.1. $\phi(a) = \phi(b) \iff a \equiv b(\mathrm{mod}\ \mathrm{Ker}\phi)$.

Proof. We have : $\phi(a) = \phi(b) \iff \phi(a^{-1}b) = 1'$

$$\iff a^{-1}b \in \mathrm{Ker}\phi \iff a \equiv b(\mathrm{mod}\ \mathrm{Ker}\phi)\ .$$

Corollary 10.2. $\phi(a) = \phi(b) \iff aK = bK$, *where* $K = \mathrm{Ker}\phi$.

Proof. Each set gK is an equivalence class under the equivalence relation which is congruence (mod K).

Thus the equivalence relation $\sim$ defined by $[a \sim b \iff \phi(a) = \phi(b)]$ is just congruence (mod Kerϕ). But it's an elementary fact of set theory that the equivalence classes under $\sim$ are in 1-1 correspondence with the image of ϕ, for any function ϕ between sets, no group structure being needed. Let's reprove it in this case:

Definition. If H is any subgroup of a group G, denote by G/H the set of all left cosets of $G(\mathrm{mod}\ H)$. (The number of elements of G/H is then $[G : H]$.)

Proposition 10.3. *If* $K = Ker\phi$, *then* $\psi : G/K \to \mathrm{Im}\phi$, *defined by* $\psi(gK) = \phi(g)$ *is a well-defined bijective set function.*

Proof. To see that ψ is well defined, we must show that the formula is independent of the choice of g within a given coset; i.e. that
$aK = bK \Rightarrow \phi(a) = \phi(b)$. This is half of **10.2**. To show that ψ is injective, we

must show that $\psi(aK) = \psi(bK) \Rightarrow aK = bK$; i.e. $\phi(a) = \phi(b) \Rightarrow aK = bK$. This is just the other half of **10.2**. Finally, ψ is surjective, since

$$(g' \in \mathrm{Im}\phi) \Rightarrow (\exists g \in G \text{ with } \phi(g) = g') \Rightarrow (\exists gK \in G/K \text{ with } \psi(gK) = g') \ .$$

Now $\mathrm{Im}\phi$ is a group, by **9.2**. The natural thing to do is to make explicit the unique binary operation on G/K which converts ψ from being merely a bijective set function into being a group isomorphism. Now

$$\psi(aK)\psi(bK) \ = \ \phi(a)\phi(b) \ = \ \phi(ab) \ = \ \psi(abK) \ .$$

We want the latter to be equal to $\psi[(aK)(bK)]$ for the (as yet undefined) product of aK and bK. But ψ is injective, so we are forced to define:

$$(aK)(bK) \ := \ (ab)K \ .$$

Now we could try to define a binary operation in this way on G/H for *any* subgroup H, but it would **not be well-defined** in general. The subgroup must be **normal in** G:

Definition. A subgroup N of a group G is *normal in* G if and only if, for all g in G and n in N, we have $g^{-1}ng \in N$.

Definition. When N is a normal subgroup of a group G, define (modulo checking that it is well defined) a multiplication on G/N as follows :

$$\begin{array}{ccc} (aN)(bN) & := & (ab)N \\ \uparrow & & \uparrow \\ \text{(multiplication to be defined)} & & \text{(multiplication in } G \text{ already given)} \end{array}$$

Theorem 10.4. *This multiplication is well defined, and G/N becomes a group using it. The identity element is N. The inverse of gN is $g^{-1}N$.*

Definition. G/N is then called *the quotient group 'G modulo N'*.

Proof. For well definition, we must show that the definition is independent of choice of a and b within their respective cosets, i.e. that

$$aN \ = \ \tilde{a}N \ \text{ and } \ bN \ = \ \tilde{b}N \quad \Longrightarrow \quad abN \ = \ \tilde{a}\tilde{b}N \ .$$

The assumptions yield $a^{-1}\tilde{a} \in N$ and $b^{-1}\tilde{b} \in N$, and therefore

$$(ab)^{-1}\tilde{a}\tilde{b} \;=\; [b^{-1}(a^{-1}\tilde{a})b][b^{-1}\tilde{b}] \;\in\; N \;,$$

since the last element is a product of two elements of N (the left-hand factor being in N because N is normal). But $(ab)^{-1}\tilde{a}\tilde{b} \in N$ yields the desired conclusion.

Associativity follows easily:

$$aN\{(bN)(cN)\} \;=\; (aN)(bcN) \;=\; \{a(bc)\}N$$

$$=\; \{(ab)c\}N \;=\; (abN)(cN) \;=\; \{(aN)(bN)\}cN \;.$$

As for the identity element, $(1N)(gN) = (1 \cdot g)N = gN = (gN)(1N)$, as required. For inverses,

$$(g^{-1}N)(gN) \;=\; (g^{-1}g)N \;=\; 1N \;=\; N \;,$$

as required.

Proposition 10.5. *For any morphism $\phi : G \to G'$, the subgroup Kerϕ is normal in G.*

Proof. If $g \in G$ and $n \in \mathrm{Ker}\phi$, then

$$\phi(g^{-1}ng) \;=\; \phi(g^{-1})\phi(n)\phi(g) \;=\; \phi(g)^{-1}1'\phi(g) \;=\; 1' \;,$$

so $g^{-1}ng \in \mathrm{Ker}\phi$.

Theorem 10.6. (1st isomorphism theorem) *If $\phi : G \to G'$ is any morphism, then the map $\psi : G/\mathrm{Ker}\phi \to \mathrm{Im}\phi$, given by setting $\psi(g\mathrm{Ker}\phi) := \phi(g)$, is an isomorphism.*

Proof. The group structure on $G/\mathrm{Ker}\phi$ comes from **10.5** and **10.4**. Also ψ is a well defined bijective set function by **10.3**. It remains only to show that ψ is a morphism. But this was essentially done just after the proof of **10.3**:

$$\psi(aK)\psi(bK) \;=\; \phi(a)\phi(b) \;=\; \phi(ab) \;=\; \psi(abK) \;=\; \psi\{(aK)(bK)\} \;.$$

This last computation had to work because we chose the multiplication in $G/\mathrm{Ker}\phi$ to make it work—see the discussion after **10.3**.

There is one important case where *every* subgroup is normal—when G is abelian:

Proposition 10.7. *If G is commutative, then every subgroup H of G is normal in G.*

Proof. If $g \in G$ and $n \in H$, then $g^{-1}ng = g^{-1}gn = n \in H$.

Examples. (i) $k\mathbf{Z}$ is the kernel of $\phi : \mathbf{Z} \to \mathbf{Z}_k$, $\phi(m) = [m]$, and $\mathrm{Im}\phi = \mathbf{Z}_k$, so $\mathbf{Z}/k\mathbf{Z} \cong \mathbf{Z}_k$ by **10.6.** In fact, the groups $\mathbf{Z}/k\mathbf{Z}$ and $\mathbf{Z}_k$ are *equal*, and the map ψ corresponding to this particular ϕ is the identity map.

(ii) The subgroup $\{e, (12)\}$ is *not* normal in S_3, nor are $\{e, (13)\}$ nor $\{e, (23)\}$. But $A_3 = \{e, (123), (132)\}$ *is* normal—in general A_n is the kernel of $\phi : S_n \to C_2$, $\phi(\sigma) = \mathrm{sign}(\sigma)$, so A_n is normal in S_n. As long as $n > 1$, $\mathrm{Im}\phi = C_2$, so, by **10.6**, $S_n/A_n \cong C_2$. The map ψ, corresponding to this ϕ, maps A_n , the set of even permutations, to $+1$, and the other coset, the set of odd permutations, to -1.

Exercise 10A. Show that a subgroup of index 2 is necessarily normal.

(iii) Let D_n be the subgroup of all the orthogonal transformations which preserve some chosen regular n-gon centred at the origin in $\mathbf{R}^2$. It consists of "n" reflections, and "n" rotations (including the identity element). Define $\phi : D_n \to \mathbf{R}^*$ by $\phi(T) = \det T$. Then $\mathrm{Im}\phi = C_2$ and $\mathrm{Ker}\phi = D_n^+$, the cyclic group of rotations. By **10.6**, $D_n/D_n^+ \cong C_2$. The corresponding ψ maps D_n^+ to $+1$, and the other coset, consisting of reflections, to -1.

(iv) The trivial morphism $\phi : G \to G'$, $\phi(g) = 1'\ \forall g$, has image $\{1'\}$ and has kernel G, so $G/G \cong \{1'\}$, which is rather obvious directly.

(v) The identity map from G to G has image G and has kernel $\{1\}$, so $G/\{1\} \cong G$, another unsurprising consequence of **10.6**.

Exercise 10B. Check, check, check!

Exercise 10C. For both of the following asserted isomorphisms, state the needed hypotheses concerning normality, and then find isomorphisms. Note that when one is looking for an isomorphism whose domain is a quotient group, it's usually easiest to define an epimorphism on the numerator, and prove that its kernel is the denominator. In the second one, A and B are both subgroups of a larger containing group, and $ab = ba$ for all $a \in A$ and

$b \in B$.

$$(G/K)/(H/K) \;\cong\; G/H \quad ; \quad (A \times B)/D \;\cong\; \text{subgroup gen}^{\text{d.}} \text{ by } A \cup B \ ,$$

where $D \;=\; \{\ (x,\ x)\ :\ x \in A \cap B\ \}$.
(One could also take $D \;=\; \{\ (x^{-1},\ x)\ :\ x \in A \cap B\ \}$.)

11. Solubility and a Sylow theorem.

To begin, here is a (perhaps nasty)
Exercise 11A. Find an example of a group G, and k, a positive integer dividing $|G|$, such that G has no subgroup of order k.

There is, however, an important partial converse to Lagrange's theorem. In this section, p and q always denote primes.

Theorem 11.1. (Sylow) *If p is a prime and p^t divides $|G|$, where G is a finite group, then G* has *a subgroup of order p^t.*

Proof. For any non-empty subset A of G, define a subset
$H_A := \{\ g \in G\ :\ gA = A\ \}$, where $gA := \{\ ga\ :\ a \in A\ \}$. Then H_A is a subgroup of G. Furthermore $|H_A| \leq |A|$, since multiplying distinct elements g into some fixed element of A yields distinct answers. Let n_A be the number of distinct subsets gA as g ranges over G. Then $g_1A = g_2A \iff g_1 \equiv g_2 (\text{mod } H_A)$, so $n_A = |G/H_A| = |G|/|H_A|$. Since the binomial coefficient $\left(\begin{array}{c} |G|-1 \\ p^t-1 \end{array}\right)$ isn't divisible by p, and

$$p^t \left(\begin{array}{c} |G| \\ p^t \end{array}\right) \;=\; |G| \left(\begin{array}{c} |G|-1 \\ p^t-1 \end{array}\right) \ ,$$

the largest powers of p dividing $\left(\begin{array}{c} |G| \\ p^t \end{array}\right)$ and $|G|/p^t$ are the same. But $\left(\begin{array}{c} |G| \\ p^t \end{array}\right)$ is the number of subsets with p^t elements, which is a sum of numbers n_A for various A with p^t elements. Hence there exists A_0 with p^t elements such that n_{A_0} is not divisible by any larger power of p than the largest such

power which divides $|G|/p^t$. Then $|H_{A_0}| = |G|/n_{A_0}$ is divisible by p^t , but $|H_{A_0}| \leq |A_0| = p^t$, so $|H_{A_0}| = p^t$. Hence we let the required subgroup be H_{A_0}.

Exercise 11B. Look up the definition of *the action of a group on a set* at the start of Section **47.** See also **47S**. Now check the first two of the following comments about actions.

Above we have G acting on the set of all its subsets with p^t elements, and we construct a subgroup *fixing* a subset. On the other hand, in the proof of **11.4** below, the group acts on *itself* by conjugation.

In sections **47** to **50**, we study the linear actions of a group on vector spaces. In the sections on Galois theory, the groups which arise do so by virtue of their actions on sets of roots of polynomials, and also via operation-preserving actions on algebraic objects called fields.

Corollary 11.2. (Cauchy) *If a prime p divides the order of the finite group G, then G has an element of order p* (a non-identity element of smallest order in a subgroup of order p^t).

Theorem 11.3. *If p is an odd prime, then any group of order $2p$ is isomorphic to exactly one of C_{2p} or D_p (cf.* **8.4** *and* **8.5)**.

Proof. Using **11.2**, choose a, b in G with $||a|| = p$ and $||b|| = 2$. Then

$$G \;=\; \{\; a^i b^j \;:\; 0 \leq i \leq p-1 \;,\; 0 \leq j \leq 1 \;\} \;,$$

since all these products are distinct. Now $ba \neq a^i$ for any i, since $b \neq a^{i-1}$. Hence $ba = a^i b$ for some i. Then

$$a \;=\; b^2 a b^{-2} \;=\; b a^i b^{-1} \;=\; (bab^{-1})^i \;=\; a^{i^2} \;.$$

Hence $i^2 \equiv 1 (\text{mod } p)$, so $i \equiv \pm 1$. When $i \equiv +1$, we get $G \cong C_{2p}$.
When $i \equiv p-1$, we get $G \cong D_p$. (Use **5.3.5**).

Hard Exercise 11C. Classify groups of order pq , where p and q are distinct primes.

Theorem 11.4. *If G is a non-trivial group whose order is a prime power, then G contains a non-identity element x for which $g^{-1}xg = x$ for all $g \in G$.*

Remark and Definitions. The above condition, namely $xg = gx$ for all $g \in G$, says that *x lies in the centre of G*. So the theorem says that *a p-group* has a non-trivial centre. That the centre of a group is always a (clearly normal) subgroup is a routine verification.

Proof. For all $x \in G$, let n_x be the number of elements in the set $\{\ g^{-1}xg\ :\ g \in G\ \}$. Let $H_x = \{\ g \in G\ :\ g^{-1}xg = x\ \}$, a subgroup. Now $g_1^{-1}xg_1 = g_2^{-1}xg_2$ if and only if $g_1^{-1} \equiv g_2^{-1} (\mathrm{mod}\ H_x)$. Thus $n_x = |G|/|H_x|$ is a power of p. But $|G| = p^s$ is a sum of numbers n_x for various x. Thus the number of x with $n_x = 1$ is a multiple of p. It isn't 0 because $n_1 = 1$; hence there are at least "$p - 1$" elements as claimed in the theorem.

Definition. The set containing n_x elements in the above proof is called the *conjugacy class of x in G* .

Exercises 11D. Deduce that a non-trivial group G of prime power order p^s has a subgroup H_1 of order p which is normal in G.
By considering a normal subgroup of order p in G/H_1, show that G has a normal subgroup H_2 of order p^2 if $s \geq 2$.
Continue by induction to construct subgroups H_t normal in G of order p^t for $0 \leq t \leq s$.
Deduce that there is a sequence of subgroups

$$\{1\} = H_0\ \subset\ H_1\ \subset\ H_2\ \subset\ \cdots\ \subset\ H_s = G$$

where $|H_i| = p^i$. Furthermore, each H_i is normal in H_{i+1}, and the corresponding quotient group is cyclic (since its order is p, a prime).

Remark and Definition. Hence a *p-group* is *soluble.*
In our situation just above, each H_i is actually normal in all of G, but that is not required of the tower of subgroups in general for solubility. We only require the existence of a tower where successive group extensions are normal with cyclic quotient group.

Any tower in which successive group extensions are normal will be referred to as a *subnormal series* for the top group. The notation

$$\{1\} = H_0 \ \lhd \ H_1 \ \lhd \ H_2 \ \lhd \ \cdots \ \lhd \ H_s = G$$

will often be used for such a series.

ON WORD USAGE. Wishing to annoy our friends in both Great Britain and the United States, we have adopted *solubility* for groups, and *solvability* for equations (in Section **35**, where the two are mathematically related).

Theorem 11.5. *Any group of order p^2 is isomorphic to exactly one of C_{p^2} or $C_p \times C_p$.* (cf. **8.3** and **8.6**).

Proof. If $G \ncong C_{p^2}$, then every non-identity element has order p. Choose an element $a \neq 1$ with $g^{-1}ag = a$ for all g, by **11.4**. Choose an element $b \notin \{1, a, a^2, \cdots a^{p-1}\}$. Then $G = \{a^i b^j \ : \ 0 \leq i \ , \ j \leq p-1\}$. Since $a^p = b^p = 1$ and $ba = ab$, we have $G \cong C_p \times C_p$. (Use **5.3.5**.)

Exercises 11E. Let G be a non-abelian group of order 8. Show that G has an element a of order 4. Choosing an element $b \notin \{1, a, a^2, a^3\}$, show that $G = \{a^i b^j : 0 \leq i \leq 3 \ , \ 0 \leq j \leq 1\}$. Show that $ba = a^3 b$. Show that either $b^2 = 1$ and $G \cong D_4$; or else $b^2 = a^2$ and $G \cong Qrn$. Combined with the theory of abelian groups in Section **13**, this proves **8.8**. Try to classify groups of orders p^3 and $p^2 q$. First look up the other Sylow theorems in **Artin**, p. 206.

Exercises 11F. Much later we shall need the fact that the image of a soluble group under a morphism is also soluble. (Equivalently, a quotient of a soluble group is soluble.) Here is a sequence of exercises to establish that.

Let H be a normal subgroup of a group G, and ϕ a morphism with domain G.

(A) Prove that $\phi(H)$ is normal in $\phi(G)$.

(B) Show how ϕ induces an epimorphism

$$\theta : G/H \ \longrightarrow \ \phi(G)/\phi(H)$$

$$gH \ \mapsto \ \phi(g)\phi(H) \ .$$

(C) Show that the image of a cyclic group under a morphism is itself cyclic. (More generally, the image of a generating set is a generating set for the image.)

(D) By applying ϕ to each term of a *subnormal series*,

$$\{1\} = H_0 \lhd H_1 \lhd H_2 \lhd \cdots \lhd H_s = G \ ,$$

deduce that if G is soluble, then so is $\phi(G)$.

Exercise 11G. Prove that if N, a normal subgroup of G, and G/N are both soluble, then so is G.

Exercise 11H. Prove that any subgroup of a soluble group is also soluble.

REMARKS CONCERNING GENERATORS AND RELATIONS. We have had several examples of specific groups in which we found generators and relations. In such circumstances, one knows when 'all' relations have been found : a list of all the group elements (with no redundancies) presumably exists in some form, and to say that enough relations have been found is substantiated by giving a method to take an arbitrary 'word' (that is, an iterated product) in the generators and their inverses, and reduce it to one of the group elements in the given list. As we've seen above, this way of thinking about a group is useful in studying classification results.

But there is a more subtle question which we'll leave to your later studies. (It is much better studied in connection with the fundamental group in algebraic topology and with actions of groups on graphs). Suppose given not a group, but just some symbols to be used as generators, and a set of relations in 'group words' involving these symbols. Is there then a group with exactly these generators and relations? The answer is yes. But even when the set of relations is empty, it's a bit of work to give a correct exposition constructing the group. Such a group is called the *free group* on the set of generators. (It is infinite cyclic if there is only one generator, but highly non-abelian if more than one.) But once *it* has been constructed, the answer to the general case comes easily : one constructs the free group on new, different symbols which are in 1-1 correspondence with the given generators, and then factors this free group by the smallest *normal* subgroup containing the words in the new symbols which correspond to the desired relations. The group defined by a given specification of generators and relations is unique up to a unique isomorphism, in a suitable sense.

Challenging Exercise 11I. In the symmetric group S_n, denote the transposition $(i\ i+1)$ as τ_i for $0 < i < n$. Prove that S_n is given by

these "$n-1$" generators and the relations

$$\tau_i^2 \;=\; e \quad ; \quad (\tau_i\tau_{i+1})^3 \;=\; e \quad ; \quad \tau_j\tau_i \;=\; \tau_i\tau_j \text{ if } j > i+1 \; .$$

Above one takes all values of i for which the left-hand sides are meaningful.

12. S_n is not soluble.

More precisely, this is true for all $n > 4$; we have towers

$$S_2 \;\rhd\; \{e\} \; ;$$

$$S_3 \;\rhd\; A_3 \;\rhd\; \{e\} \; ;$$

$$S_4 \;\rhd\; A_4 \;\rhd\; \{e,\ (12)(34),\ (13)(24),\ (14)(23)\} \;\rhd\; \{e,\ (12)(34)\} \;\rhd\; \{e\} \; .$$

Exercise 12A. Show that, in every case, the smaller group *is* normal in the larger (as implicitly claimed by using the symbols $\rhd$), and the corresponding quotient group is cyclic.

Theorem 12.1. *The group S_n is not soluble for $n \geq 5$.*

This theorem will be deduced easily from the next one.

Theorem 12.2. *For $n \geq 5$, the group A_n is* simple, *i.e. has no normal subgroups other than the two extreme ones.*

Remarks. The smallest non-abelian simple group is A_5, of order 60. It follows easily, using the next section, that every group of smaller order is soluble. Several other finite and infinite families of finite simple groups were discovered a long time ago. Between 1960 and 1975, a few new, individual, such groups were added to earlier ones (and collectively called *sporadic* simple groups). By 1980, many mathematicians were convinced that a proof had been constructed that this was the complete list of finite simple groups (up to isomorphism), although apparently no individual mathematician had completely checked the (several thousand page) proof by 1993. This rumoured theorem is important because *all* finite groups are built up in towers where the successive quotients are finite simple groups. About half of the proof of classification involves representation theory, whose rudiments are given here in sections **47** to **50**. (We'll go a small distance to demonstrate the flavour of

this by proving the famous $(p,\ q)$–theorem of Burnside : a finite non-abelian simple group has order divisible by at least three distinct primes. This comes close to precluding any non-abelian group of order less than 60 from being simple.) 'Simple' refers not to the innards of such a group, but to its external relation to other groups: it cannot be regarded in any non-trivial way as being 'built up' out of a normal subgroup and the corresponding quotient group.

Proof of 12.2. Let H be a non-trivial normal subgroup of A_n. We'll show that $H = A_n$ by a sequence of reductions.

Firstly, it suffices to show that each 3-cycle is in H , since the set of 3-cycles $(12k)$, for $3 \leq k \leq n$, generates A_n (Ex. **6C**).

Since, for three distinct positive integers a, b and c not exceeding n, there is an even permutation

$$\gamma \ = \ \begin{pmatrix} 1 & 2 & 3 & \cdots \\ a & b & c & \cdots \end{pmatrix} ,$$

(write such a permutation down and then, if necessary to make it even, interchange the two entries at the right hand end of the lower line—recall that $n \geq 5$), and since

$$\gamma^{-1}(abc)\gamma \ = \ (123) ,$$

it suffices to show that H contains at least one 3-cycle, that is, some element which fixes exactly "$n-3$" integers.

Define

$m = m_{_H} := \max\{\ \ell < n \ : \ \exists \beta \in H$ with β fixing exactly "ℓ" integers $\}$.

It remains only to show that $m = n-3$; that is, $m \geq n-3$, since $m = n-1$ or $n-2$ are clearly impossible. Pick some $\beta \in H$ which fixes "m" integers, and assume, for a contradiction, that $m < n-3$, so that β moves at least 4 integers. We'll divide into three cases, in each instance producing an even permutation, α, and thence a non-identity element, $\beta^{-1}\alpha^{-1}\beta\alpha$ (clearly in H), which fixes more than "m" integers, contradicting the definition of m.

(I) Suppose that β is a cycle, $(abcde\cdots)$, of odd length at least five. Let $\alpha = (abc)$. Then $\beta^{-1}\alpha^{-1}\beta\alpha$ fixes a as well as all numbers which β fixes, but moves b (to c), as required.

(II) Suppose that β is a product of a disjoint set of an even number of transpositions. Let (ab) and (cd) be two of them. Let $\alpha = (cde)$ for an integer e differing from a, b, c and d. Then $\beta\alpha^{-1}\beta\alpha$ fixes all integers which

β fixes except possibly e, but also fixes a and b. Furthermore it moves c to $\beta(e) \neq c$, producing the required contradiction.

(III) Alternatively, the set of cycles in β must contain two cycles, δ and ϵ, of lengths at least three and at least two, respectively. Write $\delta = (uvw\cdots)$ and $\epsilon = (xy\cdots)$. Let $\alpha = (yxw)$. Then $\beta^{-1}\alpha^{-1}\beta\alpha$ again fixes everything that β fixes and fixes u as well, but moves v (to $\epsilon^{-1}(x)$), producing the required contradiction.

Proof of Theorem 12.1. If S_n were soluble, a tower showing this would need to have at least one intermediate group, since S_n is not cyclic. We'll show that S_n has only three normal subgroups, the obvious two, and A_n . Since A_n is simple and not cyclic, this leaves no possibilities for such a tower. So let H be a proper non-trivial normal subgroup of S_n . By **12.2**, $H \cap A_n$ is either A_n or $\{e\}$.

Exercise 12B. Show that if B is normal in C, then $B \cap A$ is normal in A, for any subgroup A of C (as just used).

In the first case, since no group fits between A_n and S_n, we have $H = A_n$, as required. In the second case, since the product of two odd permutations is even, H must have order 2. But none of the elements of order 2 in S_n is fixed by all conjugations, such elements being products of disjoint transpositions, by **4.8**. So the second case cannot occur.

This theorem will be used in Section **35** (as the input from group theory) in the proof that there can be no formula, involving only the operations of arithmetic and n^{th} roots, for solving the general polynomial equation of degree 5 or greater.

13. Finitely generated abelian groups.

The theorem below gives a complete, non-redundant list, up to isomorphism, of all those abelian groups which have finite sets of generators. The list consists of certain direct products of cyclic groups, so that the structure of these groups is especially simple. In Section **43** ahead, a more general theorem is presented. It is proved in slightly more detail than we give here, and depends on only a few facts about rings from sections **14** and **17**.

Definitions. When groups G and H are abelian and additive notation is used, the group $G \times H$ is denoted $G \bigoplus H$ and is called the *direct sum* of G and H. *Internal* direct sums are defined as before for internal direct products. A group G is *finitely generated* if and only if at least one of the finite subsets of G generates G. Obviously any *finite* group is finitely generated.

Lemma 13.1 *An abelian group is the internal direct sum of subgroups* $G_1\ ,\ \cdots\ ,\ G_r$ *if and only if each element can be written uniquely in the form* $g_1 + g_2 + \cdots + g_r$ *with each* $g_i \in G_i$.

Exercise 13A. Prove this.

Theorem 13.2. *If G is a finitely generated abelian group, then there is exactly one sequence* $\{\ k_1,\ k_2,\ \cdots,\ k_r\ \}$, *where for each i, the integer k_i divides k_{i+1} and either $k_i = 0$ or $k_i > 1$, such that*

$$G \ \cong\ \mathbf{Z}_{k_1} \oplus \mathbf{Z}_{k_2} \oplus \cdots \oplus \mathbf{Z}_{k_r} \ =:\ \bigoplus_{i=1}^{r} \mathbf{Z}_{k_i}\ .$$

(Here $\mathbf{Z}_0$ means $\mathbf{Z}$.)

Note. The sequence of integers $\{\ k_1,\ \cdots,\ k_r\ \}$ is called the sequence of *invariant factors* of G. Any zeros must come at the end, and they will occur if and only if G is infinite. If $|G| = 8$ for example, the possible sequences are $\{8\}$, $\{2, 4\}$, and $\{2, 2, 2\}$ (cf. **8.8**). By its very definition, a cyclic group is finitely generated. As an easy exercise, show that a direct product of finitely many finitely generated groups is also finitely generated. Thus **13.2** does what is claimed in the first sentence of this section.

Proof. We prove the *existence* of $\{\ k_1,\ \cdots,\ k_r\ \}$ with the required properties by induction on $r :=$ the minimum number of elements which can

generate G. For $r = 0$,

$$G \;=\; \{0\} \;=\; \bigoplus_{i=1}^{0} \mathbf{Z}_{k_i} \;.$$

(The reader who dislikes starting with $r = 0$, and/or taking a direct sum indexed by the empty set, should exclude the zero group from the statement of the theorem, and modify/simplify the inductive step just ahead by setting $r = 1$, in order to obtain a proof of an initial case with $r = 1$.)

For the inductive step: Let

$$P \;:=\; \{\; m \in \mathbf{Z} \;:\; \exists \{g_1, \cdots g_r\} \text{ generating } G, \text{ and}$$

$$\text{integers } m_2, \cdots, m_r \text{ with } mg_1 + \sum_{i=2}^{r} m_i g_i = 0\} \;.$$

First assume that $P = \{0\}$. Choose $\{g_1, \cdots, g_r\}$ generating G. Then each element can be written as $m_1 g_1 + \cdots + m_r g_r$ (uniquely, since $P = \{0\}$), so G is the direct sum of the cyclic groups generated by the g_i (use **13.1**). Since $mg_i \neq 0$ for all $m \neq 0$ (because $P = \{0\}$), each of these cyclic groups is isomorphic to $\mathbf{Z}$. So let $k_1 = k_2 = \cdots = k_r = 0$, and we're done.

If $P \neq \{0\}$, then P has positive elements. Let

$$k_1 \;:=\; \min\{\; m \in P \;:\; m > 0 \;\} \;.$$

Note that $k_1 \neq 1$, since G cannot be generated by $\{g_2, \cdots, g_r\}$.

Now consider all possible choices of integers $\{\ell_2, \ell_3, \cdots, \ell_r\}$ and of generators $\{g_1, \cdots, g_r\}$ such that $k_1 g_1 + \sum_{i=2}^{r} \ell_i g_i = 0$.

(a) *For all such choices, $k_1 \mid \ell_j$ for all $j > 1$:*
Divide k_1 into ℓ_j, giving $\ell_j = sk_1 + R$ with $0 \leq R < k_1$. Then we have a relation

$$Rg_j + k_1(g_1 + sg_j) + \sum_{i \neq 1, j} \ell_i g_i \;=\; 0,$$

which by minimality of k_1 implies that $R = 0$, since the set

$$\{\; g_j,\; g_1 + sg_j,\; g_2,\; \cdots,\; g_{j-1},\; g_{j+1},\; \cdots,\; g_r \;\}$$

also generates G. Thus $\ell_j = sk_1$ is a multiple of k_1, as required.

(b) *For all such choices, if* $\sum_{i=1}^{r} m_i g_i = 0$, *then* $k_1 \mid m_1$:
Write $m_1 = qk_1 + R$ with $0 \le R < k_1$. Then we have a relation

$$Rg_1 + \sum_{i=2}^{r} (m_i - q\ell_i) g_i \;=\; 0,$$

so again $R = 0$.

Now make a fixed such choice, and define

$$\bar{g_1} \;:=\; g_1 + \sum_{i=2}^{r} (\ell_i / k_1) g_i \ ,$$

using a). Then $k_1 \bar{g_1} = 0$. Let H be the cyclic group generated by $\bar{g_1}$. Then $H \cong \mathbf{Z}_{k_1}$, since, by choice of k_1, the element $\bar{g_1}$ cannot have order less than k_1. Let J be the group generated by $\{g_2, \cdots, g_r\}$. By the inductive hypothesis, J is the internal direct sum of cyclic groups generated by (say) $\{\bar{g_2}, \bar{g_3}, \cdots, \bar{g_r}\}$, and isomorphic to (say) $\mathbf{Z}_{k_2} \;\oplus\; \cdots \;\oplus\; \mathbf{Z}_{k_r}$, where $k_2 \mid k_3 \mid k_4 \cdots$.

It remains to show that G is the internal direct sum of H and J, and that $k_1 \mid k_2$. The set $H \cup J$ generates G, since $\{\bar{g_1}, g_2, \cdots, g_r\}$ generates G. If $x \in H \cap J$, let

$$x \;=\; m_1 \bar{g_1} \;=\; \sum_{i=2}^{r} (-m_i) g_i \ .$$

Then $\sum_{i=1}^{r} m_i g_i \;=\; 0$, so $k_1 \mid m_1$ by (b). Hence $x \;=\; m_1 \bar{g_1} \;=\; 0$. Thus $H \cap J = \{0\}$, so $G \cong H \oplus J$. Finally we have

$$\sum_{i=1}^{r} k_i \bar{g_i} \;=\; \sum_{i=1}^{r} 0 \;=\; 0 \ ,$$

so $k_1 \mid k_2$ by (a).

To prove uniqueness, suppose that

$$G \;=\; \bigoplus_{i=1}^{r} \mathbf{Z}_{k_i} \;\;\cong\;\; \bigoplus_{i=1}^{s} \mathbf{Z}_{\ell_i} \;=\; H \ ,$$

where $k_i \mid k_{i+1}$ and $\ell_i \mid \ell_{i+1}$, all these integers being non-negative. By adding 1's at the beginning of the shorter sequence, we can assume that $r = s$. We shall show that $k_i = \ell_i$ for all i. If not, for a contradiction, let $j =$

$\min\{i : k_i \neq \ell_i\}$. Let p be a prime and α an integer such that $p^{\alpha+1} \mid k_j$, but $p^{\alpha+1}$ does not divide ℓ_j (interchanging the k's and ℓ's if necessary). Let $p^\beta G := \{\ p^\beta x \ : \ x \in G\ \}$. Since $G \cong H$, we have

$$p^\alpha G/p^{\alpha+1}G \ \cong \ p^\alpha H/p^{\alpha+1}H \ .$$

But the left hand side has at least p^{r-j+1} elements since

$$p^\alpha G/p^{\alpha+1}G \ \cong \ \bigoplus_{i=1}^{r} p^\alpha \mathbf{Z}_{k_i}/p^{\alpha+1}\mathbf{Z}_{k_i}$$

and $p^\alpha \mathbf{Z}_{k_i}/p^{\alpha+1}\mathbf{Z}_{k_i}$ has p elements as long as $p^{\alpha+1} \mid k_i$, which holds for $j \leq i \leq r$. However, the right hand side has at most p^{r-j} elements, because $p^{\alpha+1}$ does not divide ℓ_i for $1 \leq i \leq j$, and so $p^\alpha \mathbf{Z}_{\ell_i}/p^{\alpha+1}\mathbf{Z}_{\ell_i}$ is a trivial group for these i. From this contradiction, we conclude that $k_i = \ell_i$ for all i, as required.

Definitions. The *torsion subgroup* of G is its set of elements of finite order. G is *torsion-free* if and only if its torsion subgroup is $\{0\}$. A *basis* for G is an indexed set $\{g_\alpha\} \subset G$ such that each element of G can be expressed uniquely as a finite integral combination $\sum_\alpha k_\alpha g_\alpha$, where the k_α are integers, zero for all but finitely many α. Then G is a *free abelian group* if and only if it has a basis.

If G has a finite basis $\{g_1, g_2, \cdots, g_r\}$, then clearly

$$G \ \cong \ \mathbf{Z}^r \ := \ \mathbf{Z} \oplus \mathbf{Z} \oplus \ \cdots \ \oplus \mathbf{Z} \quad (\text{“}r\text{” copies of } \mathbf{Z}) \ .$$

A free abelian group is evidently torsion-free. The group of rational numbers, $\mathbf{Q}$, is torsion free, but is *not* free abelian. (CAUTION. There is a definition of an important idea in general, not-necessarily-abelian group theory, that of *free group*. See the remarks after **11H**. The only [free abelian] group which is an abelian [free group] is the case of rank $r = 1$ [defined below], i.e. isomorphic to $\mathbf{Z}$.) The group $\mathbf{Q}/\mathbf{Z}$ is a torsion abelian group (i.e. it coincides with its torsion subgroup) but is not finitely generated, nor is it a direct product or direct sum of cyclic groups, even in any sense in which one uses infinitely many factors (an idea which the reader may wish to investigate!)

Note. The above definitions make sense for any abelian group G. In the following corollaries, *G is assumed also to be finitely generated.* Some are

true more generally, and these and others can be proved independently of **13.2**.

Corollary 13.3. *G is torsion-free if and only if G is free.*

Corollary 13.4. *The torsion subgroup is finite.*

Corollary 13.5. *If G is free, then any two bases have the same number of elements (called the* **rank** *of G).*

Corollary 13.6. *G is the internal direct sum of its torsion subgroup and a free abelian group whose rank is unique.*

Definition. The *p-component* of G is the subgroup

$$\{\ g \in G\ :\ \exists \alpha \text{ with } p^\alpha g = 0\ \}\ .$$

Corollary 13.7. *The p-component of G is isomorphic to*

$$\mathbf{Z}_{p^{t_1}} \oplus \mathbf{Z}_{p^{t_2}} \oplus \cdots \oplus \mathbf{Z}_{p^{t_\ell}}$$

for a unique sequence $1 \le t_1 \le t_2 \le \cdots \le t_\ell$.

Theorem 13.8. *The torsion subgroup of G is the internal direct sum of the p-components of G for those primes p such that the p-component is non-zero.*

(By **13.4** there are finitely many such primes.)

Proof. Given g of order $k = \prod_{i=1}^t p_i^{\alpha_i}$, where $p_1, \cdots, p_t$ are the primes referred to, define k_j to be $\prod_{i \neq j} p_i^{\alpha_i}$. Since $\mathrm{GCD}\{k_1, \cdots, k_t\} = 1$, we have $\sum s_i k_i = 1$ for some integers s_i . Then $g = \sum_{i=1}^t s_i k_i g$ is a decomposition, as required, since $k_i g$ has order $p_i^{\alpha_i}$.

Uniqueness of the decomposition follows easily from the fact that 0 is the only element whose order is both a power of p_i and prime to p_i .

The p-primary decomposition is often more convenient than the invariant factor decomposition, but passing from either to the other, and listing all possible abelian groups of some given finite order, are easy tasks. See **44D** and the paragraphs following it.

We now give a precise definition of *'the abelian group with generators* $x_1, \cdots, x_n$ *and relations* $r_j := \sum_{i=1}^n \alpha_{ji} x_i \ = 0$ *for* $1 \le j \le k$ *'*, and show how to decompose such a group.

Let

$$F \ := \ \mathbf{Z}^n \ := \ \mathbf{Z} \oplus \mathbf{Z} \oplus \ \cdots \ \oplus \mathbf{Z} \qquad (\text{“}n\text{” copies of } \mathbf{Z}) \ ,$$

and let $\bar{x}_i := (0, 0, \cdots, 0, 1, 0, \cdots, 0) \in F$ (with 1 in the i^{th} place). Let R be the subgroup generated by $\{\bar{r}_1, \cdots, \bar{r}_k\}$ where $\bar{r}_j := \sum \alpha_{ji} \bar{x}_i$. Define $G := F/R$ and define $x_i := \bar{x}_i + R$.

To decompose, let A be the $(k \times n)$-matrix (α_{ji}). Using *integer* row and column operations—again we refer to Section **44** for more details—reduce A to a diagonal matrix

$$D = \begin{pmatrix} d_1 & & & \\ & \ddots & & 0 \\ 0 & & d_\ell & \\ & & & 0 \end{pmatrix} ,$$

where $d_i \mid d_{i+1}$ for all i. Then $D = PAQ$ where P and Q are integer matrices with integer matrix inverses. They are found by applying the row and column operations respectively to identity matrices of the correct sizes. For $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, define $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} := Q^{-1}\mathbf{x}$. If we write the relations $A\mathbf{x} = \mathbf{0}$, an equivalent system of relations is $(PAQ)(Q^{-1}\mathbf{x}) = \mathbf{0}$, i.e. $D\mathbf{y} = \mathbf{0}$, i.e. $d_i y_i = 0$ for $1 \le i \le \min(k, n)$. Define $d_{k+1} = d_{k+2} = \cdots = d_n = 0$ if $k < n$. Since $\mathbf{x} = Q\mathbf{y}$, the set $\{y_1, \cdots, y_n\}$ also generates G. Let $t = \min\{i : d_i \neq 1\}$—(except if $d_i = 1$ for all i where we'd have $G = \{0\}$ and could forget about this calculation!) Then $y_1 = y_2 = \cdots = y_{t-1} = 0$, and G is the direct sum of the cyclic groups generated by $y_t, y_{t+1}, \cdots, y_n$ and has invariant factors $d_t, d_{t+1}, \cdots, d_n$.

Exercise 13B. Decompose in both ways the abelian group generated by $\{x_1, x_2, x_3, x_4\}$ with the following relations. Also find generators for the direct summands.

(a)

$$8x_1 + 8x_4 \ = \ 10x_3$$

$$\begin{aligned} 10x_1 &= 350x_4 \\ 16x_1 + 90x_3 + 16x_4 &= 20x_2 \\ 8x_1 &= 10x_2 + 352x_4 \end{aligned}$$

(b)

$$\begin{aligned} 2x_1 - 5x_2 + 37x_3 + 39x_4 &= 0 \\ x_1 + x_3 + 2x_4 &= 0 \\ x_1 - 5x_2 - 4x_3 - 3x_4 &= 0 \end{aligned}$$

(c)

$$\begin{aligned} 4x_1 + 3x_2 &= 6x_3 \\ x_1 + 2x_2 &= 2x_3 \\ 5x_1 + 5x_2 + x_4 &= 7x_3 \\ 2x_3 &= 0 \\ x_1 + x_2 &= 2x_3 \end{aligned}$$

Exercise 13C. Prove that a non-zero finite abelian group cannot be decomposed as a direct sum of non-trivial groups if and only if it is cyclic of prime power order. (We then refer to it as being *indecomposable.*) Deduce the **Krull-Schmidt Theorem** for finite *abelian* groups: The decomposition of such a group into a direct product of indecomposable factors is *unique* up to isomorphism of the factors and the order in which they appear. (This theorem holds for *arbitrary* finite groups.) The *existence* of such a decomposition for any finite group is quite easy to see.

Exercise 13D. Given distinct finite abelian groups $H \subset G$, prove that there is no intermediate group between G and H if and only if G/H is cyclic of prime order (i.e. is an abelian simple group). Deduce the **Jordan-Holder Theorem** for finite *abelian* groups : a *composition series*

$$\{1\} = H_0 \lhd H_1 \lhd H_2 \lhd \cdots \lhd H_s = G$$

for G (i.e. each inclusion is the inclusion of a proper normal subgroup, and no new groups can be 'inserted' in between, preserving the normality of each extension) is *unique* in the following sense: the sequence of quotient groups, H_{i+1}/H_i, is unique up to isomorphism and re-ordering. (This theorem also

is true for arbitrary finite groups—that's why we bothered to write the word "normal" above.) The *existence* of such a composition series for any finite group is quite easy to see.

For example, the composition factors for $\mathbf{Z}_6$ are $\{ \mathbf{Z}_2 , \mathbf{Z}_3 \}$, and can be made to appear in either order. For S_4 they are $\{ \mathbf{Z}_2 , \mathbf{Z}_2 , \mathbf{Z}_3 , \mathbf{Z}_2 \}$; see the composition series given just before **12.1**. Although they aren't forbidden by Jordan and Holder from appearing in a different order, they actually cannot in this case, because S_4 has no normal subgroup of order 8, and A_4 has no subgroup at all of order 6 (cf. **11A**). A finite soluble group (see **11D**) is one whose composition quotients are cyclic of prime order. This wasn't quite our definition, but can be deduced. In the other direction, for solubility , it suffices to have a tower in which the successive quotients are abelian.

Exercise 13E. Prove this. (Use the structure theorem— first show that if a normal subgroup N can be 'inserted' into a group Q, and if $H \lhd G$ with $G/H \cong Q$, then a group K can be 'inserted' as follows: $H \lhd K \lhd G$ with $K/H \cong N$ and $G/K \cong Q/N$.)

A non-soluble group is one whose composition quotients include at least one non-abelian simple group.

If you reacted as intended, your proofs in **13C** and **13D** used (at least parts of) the structure theory for finite abelian groups. On the one hand, this illustrates the power of having such a strong grip on what all the finite abelian groups look like. On the other hand, the proofs of these results at their correct level of generality (no assumption of 'abelianity') are much more elegant.

Exercise 13F. Prove that any subgroup of a finitely generated abelian group is itself finitely generated. More delicately, for any subset S of a finitely generated abelian group, there is a *finite* subset T of S, such that T and S generate the same subgroup.

APPENDIX ZZZ. The generalized associative law.

This boring section gives a rigorous formulation of **2.1**, the generalized associative law. The word 'bracketing' needs to be eliminated. Since **2.1** says that we can multiply sequences of more than two elements unambiguously, the correct formulation should involve the notion of an n-fold product map P_n from sequences of length n yielding elements in S.

Proposition. *Suppose that $*$ is an associative binary operation on a set S. Then there is exactly one sequence of maps $P_n : S^n \to S$ for $n \geq 1$ having the following properties.*

(i) $P_1(s) = s$ *for all* s ;

(ii) $P_n(s_1, s_2, \cdots, s_n) = P_i(s_1, s_2, \cdots, s_i) * P_{n-i}(s_{i+1}, s_{i+2}, \cdots, s_n)$ *for all* $n \geq 1$ *and* $1 \leq i \leq n-1$.

Proof. (*Existence*): Define inductively $P_1(s) = s$, and, for $n \geq 2$,

$$P_n(s_1, s_2, \cdots, s_n) = s_1 * P_{n-1}(s_2, \cdots, s_n) .$$

So (i) holds by definition. We prove (ii) by induction on n. For $n = 1$, there are no values of i with $1 \leq i \leq n-1$, so the condition holds. Suppose that $n \geq 2$ and the condition (ii) holds for $n-1$. Prove it for n in two cases:

If $i = 1$, the needed identity is immediate from the definition.

If $2 \leq i \leq n-1$, then

$$P_i(s_1, s_2, \cdots, s_i) * P_{n-i}(s_{i+1}, s_{i+2}, \cdots, s_n)$$

$$= \{s_1 * P_{i-1}(s_2, \cdots, s_i)\} * P_{n-i}(s_{i+1}, s_{i+2}, \cdots, s_n) \text{ [by def}^{\text{n.}}, \text{ since } i \geq 2]$$

$$= s_1 * \{P_{i-1}(s_2, \cdots, s_i) * P_{n-i}(s_{i+1}, s_{i+2}, \cdots, s_n)\} \quad \text{[by associativity]}$$

$$= s_1 * P_{n-1}(s_2, \cdots, s_n) \quad \text{[by the inductive hypothesis]}$$

$$= P_n(s_1, s_2, \cdots, s_n) \quad \text{[by definition]} .$$

(*Uniqueness*): Suppose that $Q_n : S^n \to S$ is another such sequence. We prove that $Q_n = P_n$ by induction on n. For $n = 1$, we have that $Q_1(s) = s = P_1(s)$ for all s, so $Q_1 = P_1$.

Suppose that $n \geq 2$ and $Q_{n-1} = P_{n-1}$. Then

$$Q_n(s_1, \cdots, s_n) = Q_1(s_1) * Q_{n-1}(s_2, \cdots, s_n) \quad \text{[by (ii) with } i = 1]$$

$$= s_1 * Q_{n-1}(s_2, \cdots, s_n) \quad \text{[by (i)]}$$

$$= s_1 * P_{n-1}(s_2, \cdots, s_n) \quad \text{[by inductive hypothesis]}$$

$$= P_n(s_1, \cdots, s_n) \quad \text{[by definition]} .$$

Since this holds for all $(s_1, \cdots, s_n)$, we get $Q_n = P_n$.

Remark. Of course we normally use the notation $s_1 * s_2 * \cdots * s_n$, or even $s_1 s_2 \cdots s_n$, rather than $P_n(s_1, \cdots, s_n)$.

II. Commutative Rings.

Sections **14** to **22** study the most basic objects of commutative algebra. This will probably look somewhat more familiar than did groups, since the idea of *ring* is based on the standard number systems and polynomials, and, in the non-commutative case, on square matrices.

14. Rings.

Definitions. A **Z**-*algebra* (or *associative algebra over* **Z**) is an ordered triple $(R\ ,\ +\ ,\ \cdot)$, where R is a set, and both $+$ and $\cdot$ are binary operations on R, with the following properties.
(1) $(R\ ,\ +)$ is an abelian group.
(2) $a(bc) = (ab)c$ for all a, b, c in R. (This, of course, is associativity.)
(3) $a(b+c) = (ab) + (ac)$ and
$(b+c)a = (ba) + (ca)$ for all a, b, c in R. (This is *distributivity*.)
Above, and henceforth, we write the multiplication using juxtaposition. Also we continue with the conventions from kindergarten that multiplication is done before addition etc., so that use of brackets can be minimized. For example, the brackets on the right hand sides of the distributive laws are unnecessary.

An element 1 is an *identity element* if and only if $a1 = a = 1a$ for all $a \in R$. If R has a 1, then we call R a **ring**. If R is a ring, then an element u is *invertible* (or is a *unit*) if and only if there is an element v (in R, of course) with $uv = 1 = vu$. Then v is the *inverse* of u. A *division ring* (or *skewfield*) is a ring in which every non-zero element is invertible. (Sometimes $1 \neq 0$ is also assumed in a division ring, as we do in Sections **51** and **52**. This merely excludes the one example $R = \{0\}$.) We say that R is *commutative* if and only if $ab = ba$ for all $a, b \in R$. A *field* is a commutative division ring in which $1 \neq 0$.

Remarks. It has become customary to require a ring to have an identity element. The reader should be aware, when reading books on the subject, that some do not make this requirement. See Section **52** for the general definition of an (associative) R-*algebra*. It is convenient that this, with $R = \mathbf{Z}$, is a ready-made term for what used to be called a ring. We shall try to stick to the words *identity element* and *invertible element*, since *unit* is sometimes used for each of them.

Exercises 14A. $0x = 0 = x0$ for all x ; $1 = 0 \Longrightarrow R = \{0\}$; $-(ab) = (-a)(b) = (a)(-b)$; $(-a)(-b) = ab$;—ad infinitum—; in any ring.

Examples. The standard number systems $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{C}$ are commutative rings, all but $\mathbf{Z}$ being fields. Given a commutative ring R, one has:
(1) $R[x]$, the ring of polynomials with coefficients in R—see Section **16**;
(2) $M_n(R)$ or $R^{n\times n}$, the ring of $n \times n$ matrices with entries from R. This ring is almost never commutative. Its invertibles are the matrices with inverses, sometimes called 'non-singular'.

An example of a non-commutative division ring is not so easy to find. (There aren't any finite ones; see Section **51**.) The classic example is the *quaternions* . This is $\mathbf{R}^4$ with vector addition and the following multiplication. Let $\{\mathbf{1},\ i,\ j,\ k\}$ be the standard ordered basis for $\mathbf{R}^4$, and multiply them using the multiplication table for the quaternion group, Qrn, in Section **8**. Then the distributive law , together with the extra requirement, involving scalar multiplication,

$$\alpha\cdot(vw) \ = \ (v)(\alpha\cdot w) \ = \ (\alpha\cdot v)(w) \tag{!!}$$

for all $\alpha \in \mathbf{R}$ and $v, w \in \mathbf{R}^4$, determines the product of any two vectors. This is the best way to remember the multiplication, though one can write a horrible formula

$$(a_1,\ a_2,\ a_3,\ a_4)(b_1,\ b_2,\ b_3,\ b_4) \ = \ (f_1,\ f_2,\ f_3,\ f_4)$$

where each f_i is a function of 8 variables $a_1,\ a_2,\ \cdots,\ b_4$. Proving associativity and distributivity may be done by a lengthy computation.

Associativity may also verified by checking it for products of the basic elements $i,\ j$ and k, and then using distributivity to verify it in general. The following is easier. Regard a quaternion $(a_1,\ a_2,\ a_3,\ a_4)$ as a pair of pairs $((a_1,\ a_2),(a_3,\ a_4))$, i.e. as a pair of complex numbers, $(c_1,\ c_2)$. Letting $\bar{c} =$ complex conjugate of $c \in \mathbf{C}$, the multiplication becomes

$$(c_1,\ c_2)(d_1,\ d_2) \ = \ (c_1d_1 - c_2\bar{d_2},\ c_1d_2 + c_2\bar{d_1}) \ .$$

Now to verify associativity:

$$[(c_1,\ c_2)(d_1,\ d_2)](e_1,\ e_2) \ = \ (c_1d_1 - c_2\bar{d_2}\ ,\ c_1d_2 + c_2\bar{d_1})(e_1,\ e_2)$$

$= (c_1d_1e_1 - c_2\bar{d_2}e_1 - c_1d_2\bar{e_2} - c_2\bar{d_1}\bar{e_2}\ ,\ c_1d_1e_2 - c_2\bar{d_2}e_2 + c_1d_2\bar{e_1} + c_2\bar{d_1}\bar{e_1}).$

On the other hand,

$$(c_1,\ c_2)[(d_1,\ d_2)(e_1,\ e_2)] \ = \ (c_1,\ c_2)(d_1e_1 - d_2\bar{e_2}\ ,\ d_1e_2 + d_2\bar{e_1})$$

$= (c_1d_1e_1 - c_1d_2\bar{e_2} - c_2\bar{d_1}\bar{e_2} - c_2\bar{d_2}e_1\ ,\ c_1d_1e_2 + c_1d_2\bar{e_1} + c_2\bar{d_1}\bar{e_1} - c_2\bar{d_2}e_2).$

Comparing answers gives the result.

As for distributivity:

$$(c_1,\ c_2)[(d_1,\ d_2) + (e_1,\ e_2)] \ = \ (c_1,\ c_2)(d_1 + e_1,\ d_2 + e_2)$$

$$= \ (c_1d_1 + c_1e_1 - c_2\bar{d_2} - c_2\bar{e_2}\ ,\ c_1d_2 + c_1e_2 + c_2\bar{d_1} + c_2\bar{e_1})\ ,$$

whereas

$$(c_1,\ c_2)(d_1,\ d_2) + (c_1,\ c_2)(e_1,\ e_2)$$

$$= \ (c_1d_1 - c_2\bar{d_2}\ ,\ c_1d_2 + c_2\bar{d_1}) + (c_1e_1 - c_2\bar{e_2}\ ,\ c_1e_2 + c_2\bar{e_1})$$

$$= \ (c_1d_1 - c_2\bar{d_2} + c_1e_1 - c_2\bar{e_2}\ ,\ c_1d_2 + c_2\bar{d_1} + c_1e_2 + c_2\bar{e_1}),$$

as required.

The other distributive law may be verified similarly, or else by the following trick, using the *quaternionic conjugate (a,b,c* and d are reals):

$$\overline{a + bi + cj + dk} \ := \ a - bi - cj - dk\ .$$

Exercise 14B. Prove the identities

$$\bar{\bar{h}} = h \ \ ; \ \ \overline{h_1 + h_2} = \overline{h_1} + \overline{h_2} \ \ ; \ \ \text{and} \ \ \overline{h_1\ h_2} = \overline{h_2}\ \overline{h_1}\ .$$

(*Note the order reversal.*)

Then, using the distributive law already verified,

$$(h_1\ +\ h_2)h_3 \ = \ \overline{\overline{(h_1\ +\ h_2)h_3}} \ = \ \overline{\overline{h_3}\ \overline{(h_1\ +\ h_2)}} \ =$$

$$\overline{\overline{h_3}\ (\overline{h_1}\ +\ \overline{h_2})} \ = \ \overline{\overline{h_3}\ \overline{h_1}\ +\ \overline{h_3}\ \overline{h_2}} \ = \ \overline{\overline{h_3}\ \overline{h_1}} + \overline{\overline{h_3}\ \overline{h_2}} \ =$$

$$\overline{\overline{h_1}\ \overline{h_3}} \ + \ \overline{\overline{h_2}\ \overline{h_3}} \ = \ h_1h_3 \ + \ h_2h_3 \ .$$

The formula $h\bar{h} = ||h||^2$, where $||h||$ here means the length of h as a vector in $\mathbf{R}^4$, immediately leads to the existence of inverses:

$$h^{-1} \;=\; \bar{h} \;/\; ||h||^2 \quad \text{for} \;\; h \neq 0 \; .$$

Here we are identifying each real number with the corresponding multiple of **1**, the quaternionic identity element. The quaternions are unique in the sense that no other $\mathbf{R}^n$ has a multiplication making it into a non-commutative division ring satisfying the condition above labeled as (!!). See Section **52**.

Let's discuss the analogues, for rings, of group morphisms, etc.

Definitions. Let R and R' be rings, whose identity elements are 1 and $1'$ respectively. A map $\phi : R \to R'$ between rings is a *morphism of rings* if and only if, for all x and y in R,

$$\phi(x+y) \;=\; \phi(x) + \phi(y) \quad ; \quad \phi(xy) \;=\; \phi(x)\phi(y) \quad ; \quad \text{and } \phi(1) \;=\; 1' \; .$$

Exercise 14C. None of these can be deduced from the other two.

'Epi-', 'mono-' and 'iso-' are used as prefixes for 'morphism' just as in group theory. An additive subgroup $I \subset R$ is a *left* (resp. *right*) (resp. *two-sided) ideal* if and only if [$r \in R$ and $x \in I$] implies that [$rx \in I$ (resp. $xr \in I$) (resp. both $rx \in I$ and $xr \in I$)].

An additive subgroup $T \subset R$ is a *subring* of R if and only if it is also closed under multiplication and contains the identity element of R.

Proposition 14.1. *Let ϕ be a ring morphism from R to R'. Then*
(i) Imϕ is a subring of R' ;
(ii) Kerϕ is a two-sided ideal in R .

Proof. *(i)* The map ϕ is, in particular, a morphism of additive groups, so Imϕ is an additive subgroup of R'. Also $\phi(1) = 1' \in \text{Im}\phi$. Finally, if x' and y' are in Imϕ, let $x' = \phi(x)$ and $y' = \phi(y)$. Then $x'y' = \phi(x)\phi(y) = \phi(xy) \in \text{Im}\phi$, as required.

(ii) Kerϕ is certainly an additive subgroup of R, since ϕ is a morphism of abelian groups. Also, if both $r \in R$ and $x \in \text{Ker}\phi$ then $\phi(rx) = r\phi(x) = r0 = 0$, so $rx \in \text{Ker}\phi$; and similarly for xr.

Definition. If I is a **two-sided** ideal in R, define a multiplication on the additive quotient group R/I as follows : $(x+I)(y+I) := xy + I$.

Theorem 14.2. *The abelian group R/I is a ring with this multiplication, which is, in particular, well defined.*

Proof. To show that the multiplication is well defined, we must show that if $x+I=\tilde{x}+I$ and $y+I=\tilde{y}+I$, then $xy+I=\tilde{x}\tilde{y}+I$. But $\tilde{x}=x+r$ for some $r \in I$, and $\tilde{y}=y+s$ for some $s \in I$. Thus $\tilde{x}\tilde{y}=xy+xs+ry+rs$. But $xs \in I$ since I is a right ideal, and $ry \in I$ since I is a left ideal, and $rs \in I$ for both reasons. Thus the multiplication is well defined.
To prove associativity,

$$\begin{aligned}[(x+I)(y+I)](z+I) &= (xy+I)(z+I) = (xy)z+I \\ &= x(yz)+I = (x+I)(yz+I) = (x+I)[(y+I)(z+I)] \ .\end{aligned}$$

Distributivity is just as easy. The identity element is $1+I$.

Example. The subgroup $n\mathbf{Z}$ is a two-sided ideal in $\mathbf{Z}$, so we have that $\mathbf{Z}_n = \mathbf{Z}/n\mathbf{Z}$ is now 'officially known' to be a ring. The ring $\mathbf{Z}$ is unusual in that all of its additive subgroups happen to be ideals.

Theorem 14.3. (First isomorphism theorem for rings) *If $\phi : R \to R'$ is a ring morphism, then the isomorphism of additive groups $\psi : R/\mathrm{Ker}\phi \to \mathrm{Im}\phi$ (given by the first isomorphism theorem for groups) is actually a ring isomorphism.*

Proof. The only extra assertions, beyond the theorem for groups, are that ψ behaves with respect to multiplication,

$$\begin{aligned}\psi[(x+\mathrm{Ker}\phi)(y+\mathrm{Ker}\phi)] &= \psi(xy+\mathrm{Ker}\phi) \\ &= \phi(xy) = \phi(x)\phi(y) = \psi(x+\mathrm{Ker}\phi)\psi(y+\mathrm{Ker}\phi) \ ,\end{aligned}$$

and that

$$\psi(1+\mathrm{Ker}\phi) = \phi(1) = 1' \ .$$

Definition. An *integral domain* is a commutative ring (of course, with 1) such that:
(i) if $x \neq 0$ and $y \neq 0$, then $xy \neq 0$ (i.e. the only *zero divisor* is 0);
(ii) $1 \neq 0$ (equivalently $R \neq \{0\}$).

Caution. Some authors don't require commutativity, or don't require the ring to have a 1, or allow the zero ring to be an integral domain.

Proposition 14.4. *If F is a field, and R is a subring of F, then R is an integral domain.*

Proof. Certainly R is a commutative ring, since it inherits commutativity from F. Furthermore $1 \neq 0$ in a field F. It also inherits the property of having no non-zero zero divisors: if $x \neq 0$ and $xy = 0$, then

$$y \;=\; (x^{-1}x)y \;=\; x^{-1}(xy) \;=\; x^{-1}0 \;=\; 0\ .$$

Corollary. *Any field is an integral domain;* $\mathbf{Z}$, $\mathbf{Q}$, $\mathbf{R}$ *and* $\mathbf{C}$ *are integral domains* (facts which you presumably don't find new or surprising).

Note. This is a phoney proof for $\mathbf{Z}$, since, in the construction of the number systems, one needs to verify the integral domain property of $\mathbf{Z}$ before constructing $\mathbf{Q}$ and verifying that $\mathbf{Q}$ is a field. A converse of **14.4** will be proved in **18.1**, mimicking the construction of $\mathbf{Q}$, and showing that, up to isomorphism, *all* integral domains arise as in **14.4**.

15. Structure of $\mathbf{Z}_n^\times$.

The set $R^\times$ of invertibles in a ring R is easily seen to be a group under multiplication. Thus $\mathbf{Z}_k^\times$ is a commutative group of order

$$\Phi(k) := \text{the number of integers between 1 and } k \text{ which are prime to } k\ ,$$

since $[\ell]_{\mathrm{mod}k}$ has an inverse in $\mathbf{Z}_k$ if and only if ℓ is prime to k.

Exercise 15A. Check these assertions.

Exercise 15B. If R and S are rings, make the abelian group $R \times S$ into a ring by using the multiplication

$$(r_1\ ,\ s_1)(r_2\ ,\ s_2) \;:=\; (r_1r_2\ ,\ s_1s_2)\ .$$

We have $1_{R\times S} \;=\; (1_R,\ 1_S)$. Furthermore, $(R \times S)^\times \;=\; R^\times \times S^\times$ since $(r, s)^{-1} = (r^{-1}, s^{-1})$. Verify these statements.

Proposition 15.1. *If k and ℓ are relatively prime, then, as rings,*

$$\mathbf{Z}_{k\ell} \;\cong\; \mathbf{Z}_k \times \mathbf{Z}_\ell\ .$$

Exercise 15C. The map $\mathbf{Z}_{k\ell} \to \mathbf{Z}_k \times \mathbf{Z}_\ell$ sending $[x]_{k\ell}$ to $([x]_k\ ,\ [x]_\ell)$ is a ring isomorphism. Complete the proof as an exercise.

Corollary. *If* $\mathrm{GCD}\{k,\ell\} = 1$, *then* $\mathbf{Z}_{k\ell}^\times \cong \mathbf{Z}_k^\times \times \mathbf{Z}_\ell^\times$, *so that* $\Phi(k\ell) = \Phi(k)\Phi(\ell)$. *Extending to any number of factors, for* distinct *primes* $p_1, \cdots, p_t$, *we have*

$$(\mathbf{Z}_{p_1^{\alpha_1}\cdots p_t^{\alpha_t}})^\times \quad \cong \quad \mathbf{Z}_{p_1^{\alpha_1}}^\times \times \cdots \times \mathbf{Z}_{p_t^{\alpha_t}}^\times$$

and

$$\Phi(p_1^{\alpha_1}\cdots p_t^{\alpha_t}) \;=\; \Phi(p_1^{\alpha_1})\cdots\Phi(p_t^{\alpha_t}) \;=\; \prod_i \; (p_i^{\alpha_i} - p_i^{\alpha_i - 1}) \; .$$

We can write the last formula as

$$\Phi(k) \quad = \quad k \; \prod_{p|k} \; (\; 1 - \frac{1}{p} \;) \; ,$$

product over all primes p dividing k.

Lemma 15.2 *For all integers* y, *all non-negative* i *and all positive* α, *we have :*

(i)

$$y^{2^i} \equiv \pm 1 (\mathrm{mod}\ 2^\alpha) \quad \Longleftrightarrow \quad y^{2^{i+1}} \equiv 1 (\mathrm{mod}\ 2^{\alpha+1}) \; ;$$

and

(ii) if p *is an odd prime,*

$$y^{p^i} \equiv 1 (\mathrm{mod}\ p^\alpha) \quad \Longleftrightarrow \quad y^{p^{i+1}} \equiv 1 (\mathrm{mod}\ p^{\alpha+1}) \; .$$

Exercise 15D. Give the proof. *Hint.* For $\Leftarrow$ in *(ii)*, show that : $z^p \equiv 1 (\mathrm{mod}\ p^{\alpha+1}) \Rightarrow z \equiv 1 (\mathrm{mod}\ p^j)$ for $1 \leq j \leq \alpha$, by induction on j.

The theorem below, when combined with the observations above, shows how to write the finite commutative group $\mathbf{Z}_n^\times$ as a direct product of cyclic groups. We already know from Theorem **13.2** (written in multiplicative notation) that such a decomposition must exist.

Theorem 15.3. *(i) For* $\alpha > 1$, *we have*

$$\mathbf{Z}_{2^\alpha}^\times \quad \cong \quad C_2 \; \times \; C_{2^{\alpha-2}} \; .$$

(ii) If p is an odd prime, then

$$\mathbf{Z}_{p^\alpha}^\times \quad \cong \quad C_{p^{\alpha-1}(p-1)} \ .$$

Proof. (ii) For $\alpha = 1$, the group $\mathbf{Z}_p^\times$ is cyclic, since if r is its largest invariant factor, then the polynomial $x^r - 1$ (over $\mathbf{Z}_p$) has all $p-1$ elements of the group at issue as roots. This is impossible for $r < p-1$, so $r = p-1$, as required. (See Sections **16** and **30** for a generalization of this idea, and more details.)

For $\alpha > 1$, it suffices to find group elements $[x]$ of order $p-1$, and $[y]$ of order $p^{\alpha-1}$, since then, by coprimeness, $[xy]$ will have order $p^{\alpha-1}(p-1)$. So let $x = z^{p^{\alpha-1}}$, where $[z]_p$ is a generator for $\mathbf{Z}_p^\times$, which exists by the previous paragraph. For $\alpha = 2$, the group $\mathbf{Z}_{p^2}^\times$ certainly has an element $[y]_{p^2}$ of order p , since p divides the order of the group. (Use either **13.2** or **11.2**.) By (ii) of the previous lemma , $[y]_{p^\alpha}$ will have order $p^{\alpha-1}$ in $\mathbf{Z}_{p^\alpha}^\times$, using induction on α.

(i) For $\alpha = 2$, this is clear, since the group at issue has order 2. For $\alpha = 3$, the group has four elements, all of order divisible by 2, as required, since

$$1^2 \equiv 3^2 \equiv 5^2 \equiv 7^2 \equiv 1 (\text{mod } 8) \ .$$

Using (i) of the previous lemma and induction on α, we see that the element of maximum order in $\mathbf{Z}_{2^\alpha}^\times$ has order $2^{\alpha-2}$ for $\alpha \geq 3$. Thus $\mathbf{Z}_{2^\alpha}^\times$ has invariant factors $\{2, 2^{\alpha-2}\}$ for $\alpha \geq 3$, as required.

Note. We have used several well known facts about $\mathbf{Z}$. These are proved in a more general context in Section **17**.

Exercise 15E. Write the group of invertibles in the ring of integers mod 2160 as a product of cyclic groups, in both invariant factor form and p-primary form.

Exercise 15F. Fix positive integers $m > n$. Find a formula for the exact power of p which divides

$$\text{GCD}\{ \ k^m - k^n \ : \ k \geq 2 \ \} \ ,$$

for each prime p.

EXCEPT WHERE NOTED EXPLICITLY, ALL RINGS
IN THE REMAINING SECTIONS ARE **COMMUTATIVE**.

For a good treatment of the 'non-commutative theory', see **Lam**(1991). See also the last two sections, **51** and **52**, of this book.

16. Polynomials and ring extensions.

An *extension* of a ring R is any ring S which contains R as a subring. Particularly later in field theory, the extension will be referred to with the notation $S \supset R$ (a noun), since R is almost as important to keep in mind as is S. The notation $R \subset S$ should be read as a phrase: 'R is a subring of S', or 'R, which is a subring of S', or 'S is an extension of R'.

Definitions. Given a ring extension $S \supset R$ and $s \in S$, we say that s *is algebraic over* R if and only if, for some $m > 0$ and some $r_0,\ r_1,\ \ldots,\ r_m$ in R with $r_m \neq 0$, we have

$$r_0 \ + \ r_1 s \ + \ r_2 s^2 \ + \ \cdots \ + \ r_m s^m \ = \ 0 \ .$$

Write the left-hand side as $\Sigma r_i s^i$. (It can be thought of as an infinite sum, but with $r_i = 0$ for all $i > m$.) If $\Sigma r_i s^i = 0$ with $r_i \in R$ implies that all $r_i = 0$, then we say that s *is transcendental over* R. So 'transcendental' means the same as 'not algebraic'. In either case, define a subset $R[s]$ of S by

$$R[s] \ := \ \{\ \Sigma r_i s^i \ : \ r_i \ \in \ R,\ r_i = 0 \text{ for almost all } i\ \}.$$

Theorem 16.1. *The set $R[s]$ is a subring of S, and is an extension of R. It coincides with the intersection of the collection of all subrings of S which contain $R \cup \{s\}$.*

Definition. The ring $R[s]$ is called the *subring generated by* $R \cup \{s\}$ or *the subring generated by s over R.*

Exercise 16A. Show that the intersection of any collection of subrings of a given ring S is a subring of S.

Proof of 16.1. Applying inductively proved general associative, commutative and distributive laws for any (finite) number of ring elements, we find

$$\Sigma r_i s^i \ \pm \ \Sigma r'_i s^i \ = \ \Sigma (r_i \pm r'_i) s^i$$

and

$$(\Sigma r_i s^i)(\Sigma r'_i s^i) \ = \ \Sigma (\Sigma_{j=0}^{i} r_j r'_{i-j}) s^i,$$

so that $R[s]$ is closed under $+$, $-$ and $\cdot$. Taking $m = 0$ (that is, $r_i = 0$ for all $i > 0$) shows that $R \subset R[s]$. In particular, $1 \in R[s]$. The first assertion is thus established. The intersection T in the second assertion is a subring of S by **16A**. It clearly contains $R \cup \{s\}$, so by closure, $R[s] \subset T$. But that $T \subset R[s]$ is obvious, since $R[s]$ is one of the rings in the intersection defining T.

Remark. You may have noted the analogy with the two descriptions given in **6.3** of the subgroup generated by a subset of a group.

Exercise 16B. Write out all the details concerning the first sentence of the above proof.

Proposition 16.2. *Continuing with the same notation, elements of $R[s]$ are* uniquely *expressible in the form $\Sigma r_i s^i$ if and only if s is transcendental over R.*

Examples of non-uniqueness. Let $R = \mathbf{R}$, $S = \mathbf{C}$ and $s = i$, where $i^2 = -1$. Then

$$0 \ + \ 0i \ + \ 0i^2 \ = \ 1 \ + \ 0i \ + \ 1i^2 \ .$$

Here $R[s] = S$. Alternatively, let $R = \mathbf{Q}$, $S = \mathbf{R}$ and $s = \sqrt{2}$. Then

$$0 \ + \ 0\sqrt{2} \ + \ 0(\sqrt{2})^2 \ = \ 2 \ + \ 0\sqrt{2} \ + \ (-1)(\sqrt{2})^2 \ .$$

Here $R[s] \neq S$. In both of these examples, elements of $R[s]$ are uniquely expressible in the form $r_0 + r_1 s$; whereas, if $\sqrt[3]{2}$ had been used instead of $\sqrt{2}$ for s, we'd get unique expressions $r_0 + r_1 s + r_2 s^2$. In **24.2**, this is explained in general, at least when s is algebraic and both R and S are fields.

Proof of 16.2. Assume that s is transcendental over R and that $\Sigma r_i s^i = \Sigma r'_i s^i$. Then $\Sigma(r_i - r'_i)s^i = 0$, so $r_i - r'_i = 0$ for all i, by transcendence, as required. Conversely, if $\Sigma r_i s^i = 0 = \Sigma 0 s^i$, then, by uniqueness, $r_i = 0$ for all i, proving transcendence of s.

Remarks. The reader will be familiar with polynomials $\Sigma r_i x^i$, at least when $R = \mathbf{R}$. The object x is often called a 'variable' or 'indeterminate', without any attempt being made to define the words. Actually, in kindergarten algebra, the above polynomial would usually be thought of as the function from $\mathbf{R}$ to $\mathbf{R}$ which maps each real number t to the real $\Sigma a_i t^i$. Below we'll see that it is important to think of a polynomial as an *expression*,

rather than as a *function*, since, for certain R such as $\mathbf{Z}_p$, the two rings (of polynomials and of polynomial functions) are *not* in 1-1 correspondence. We want two polynomials to differ if any coefficients differ, so **16.2** motivates the following.

Definition. *'The' polynomial ring in one variable over* R is $R[x]$ for any x which is transcendental over R.

Notation. To avoid having to say explicitly that we are considering the polynomial ring, we shall never place $[x]$ to the right of R except when x is transcendental over R. This applies to the letter x as well as to x', x_1, x_2 etc..., but not to other letters. This depends on having a ring R none of whose extensions involved in the discussion has an element algebraic over R and already named x. For example, x certainly isn't transcendental either over $R[x]$, or over $R[x^3]$. There is nothing to stop the reader from placing $[x]$ to the right of either of these, but I shall refrain from doing so. If x *is* transcendental over R, then a shorter notation for the object denoted by doing so is just $R[x]$.

It is fairly obvious that $R[x]$ is unique up to isomorphism, and not quite so obvious that it always exists:

Theorem 16.3. i) **(Uniqueness)** *If s and t are both transcendental over R, then $R[s] \cong R[t]$.*

ii) **(Existence)** *For any R, there exists a polynomial ring $R[x]$. That is, there exists an extension ring containing an element which is transcendental over R.*

Proof i) It is a routine verification to check that the function $\Sigma r_i s^i \mapsto \Sigma r_i t^i$ is an isomorphism of rings, as required.

Exercise 16C. Do it!

ii) **Preliminary Remarks.** Many mathematicians, including perhaps the reader, will feel that this next proof is hardly necessary. They say: "Just take an abstract symbol x and operate with it according to the usual rules of algebra, including the rule that distinct polynomial expressions are to be regarded by definition as denoting different objects." The author would disagree with this only in pointing out that, although following these rules of algebra has not yet led to a contradiction, this fact alone is no guarantee that

such a contradiction won't occur tomorrow. The following proof gives this guarantee: such a contradiction would be caused by something other than 'rules for polynomials'. This is analogous to 'rules for complex numbers', where i is regarded as an abstract symbol satisfying only laws which follow from commutativity and the law $i^2 = -1$. A construction of $\mathbf{C}$ (say, as $\mathbf{R} \times \mathbf{R}$) is similarly needed simply to guarantee *relative consistency.*

To proceed to the proof, define

$$S \ := \ \{ \ (r_0, \ r_1, \ r_2, \ \cdots \ ) \ : \ r_i \ \in \ R, \ r_i = 0 \text{ for almost all } i \ \} \ .$$

Thus, S is a subset of the set of infinite sequences from R, the latter, strictly speaking, being the set of functions from $\{0, 1, 2, \ldots\}$ to R. Define operations on S as follows:

$$(r_0, \ r_1, \ \cdots \ ) + (r'_0, \ r'_1, \ \cdots \ ) \ := \ (r_0 + r'_0, \ r_1 + r'_1, \ \cdots \ ) \ ;$$

$$(r_0, \ r_1, \ \cdots \ ) \ (r'_0, \ r'_1, \ \cdots \ ) \ := \ (r_0 r'_0, \ r_0 r'_1 + r_1 r'_0, \ \cdots \ , \Sigma_{j=0}^{i} r_j r'_{i-j}, \ \cdots \ ) \ .$$

(The last summation gives the entry in the $(i+1)^{\text{th}}$ slot, which is indexed by i.) Note that both right-hand sides are in S : the terms are in R, and all but finitely many are zero. Verification of the ring axioms is now routine; note that the identity element of S is $(1, \ 0, \ 0, \ 0, \ \cdots \ )$. **Exercise 16D.** Do this verification.

We do have a problem: strictly speaking, R is *not* a subring of S. Define $\phi : R \to S$ by $\phi(r) = (r, \ 0, \ 0, \ \cdots \ )$. Another routine verification shows that ϕ is an injective morphism of rings. Thus S has a subring $\phi(R)$ isomorphic to R. If we 'identify' R with $\phi(R)$, then S becomes an extension of R, as required.

Aside: Remarks on 'Identification'. The last assertion perhaps smells a bit fishy. But most readers will already have seen something similar: for example, after constructing $\mathbf{Q}$ from $\mathbf{Z}$, one identifies $a \ \in \ \mathbf{Z}$ with $a/1 \ \in \mathbf{Q}$; after constructing $\mathbf{C}$ as $\mathbf{R} \times \mathbf{R}$, one identifies $\mathbf{R}$ with the subfield $\mathbf{R} \times \{0\}$ of $\mathbf{C}$. These relatively innocuous sleights-of-hand can be accomplished in at least three different ways: Suppose that $\phi : R \to S$ is an injective ring morphism. The arrow-theorist's approach is to generalize the meaning of the term *ring extension* to mean any ordered triple $(R, \ S, \ \phi)$ as above. The conspicuous consumer's approach is to throw away the old copy

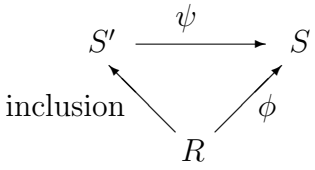
of R **(it's last year's model !!)**. Let the symbol R now stand for $\phi(R)$. Finally, wanting R itself to be the subring, the careful plodder's approach is to remove $\phi(R)$ from the set S and to replace it by R. That is, define

$$S' \ := \ R \cup T$$

where T is disjoint from R and has the same cardinality as $S \setminus \phi(R)$. Then pick any bijection $\psi : S' \to S$ which agrees on R with ϕ. Define operations on S' by

$$a \ +_{S'} \ b \ := \ \psi^{-1}[\psi(a) \ +_S \ \psi(b)]$$
$$a \ \cdot_{S'} \ b \ := \ \psi^{-1}[\psi(a) \ \cdot_S \ \psi(b)]$$

Then S' is a ring, R is a subring of S', and the diagram



commutes, with ψ being an isomorphism of rings. Any ring theoretic property of $R \xrightarrow{\phi} S$ will carry over to $R \hookrightarrow S'$. Henceforth we shall not dwell at all on this business of identification. There are no algebraic pitfalls associated with it, and the set theoretic pitfalls are few and obvious. **End of Aside.**

To complete the proof, let $x := (0,\ 1,\ 0,\ 0,\ \cdots\ ) \in S$. By induction, $x^n = (0,\ 0,\ \cdots,\ 0,\ 1,\ 0,\ 0,\ \cdots\ )$, with 1 in the $(n+1)^{\text{th}}$ slot. Now, using the identification of $a \in R$ with $(a,\ 0,\ 0,\ \cdots\ ) \in S$, we get

$$\sum a_i x^i \ = \ (a_0,\ a_1,\ a_2,\ \cdots\ ) \ .$$

Thus, if $\sum a_i x^i$ is zero in S, then $a_i = 0$ in R for all i, as required.

Exercise 16E. Write out the details of the above paragraph.

Definition. Let $Map(R, R)$ be the set of *all* functions $\phi : R \to R$. (ϕ needn't be a morphism.) Define operations by

$$(\phi_1 + \phi_2)(b) \ := \ \phi_1(b) + \phi_2(b) \ \ ; \ \ (\phi_1\phi_2)(b) \ := \ \phi_1(b)\phi_2(b) \ .$$

Exercise 16F. Verify that $Map(R, R)$ becomes a commutative ring. Check that your proof generalizes to $Map(X, R)$ for any set X.

Caution. The multiplication in $Map(R, R)$ is *not* composition, and the multiplicative identity is *not* the identity function: it is the constant function sending a to 1 for all a.

Definition. Define $\mu : R[x] \to Map(R, R)$ by 'substitution':

$$[\ \mu(\sum_{i=0}^{n} a_i x^i)\ ](b) \ := \ \sum_{i=0}^{n} a_i b^i \ .$$

Then μ is a morphism of rings, and Imμ is the *ring of polynomial functions on R.*

Example. Let $R = \mathbf{Z}_3$. Then $1 + x + x^2 \neq 1 + x^2 + x^3$ in $\mathbf{Z}_3[x]$. But $\mu(1+x+x^2) = \ \mu(1+x^2+x^3)$. It is the function from $\mathbf{Z}_3$ to itself which sends 0 to 1, 1 to 0, and 2 to 1. This example will be explained more generally in **16G**.

Conclusion. The morphism μ is **not a monomorphism** in general. This is why we don't define a polynomial to be a polynomial function.

Theorem 16.4. *If R is an infinite field, then μ is a monomorphism.*
So in this important and familiar case, there is a 1-1 correspondence between polynomials and polynomial functions. The proof is at the end of this section. The theorem is true more generally for any infinite integral domain R, as we shall see later.

Definition. If $f \ = \ \sum_{i=0}^{n} a_i x^i$ is a non-zero polynomial, i.e. if $a_i \neq 0$ for at least one i, define the *degree* of f by

$$\deg(\sum a_i x^i) \ = \ \max\{i \ : \ a_i \neq 0\} \ .$$

Proposition 16.5. *Whenever all three degrees are defined, we have*

$$\deg(f + g) \ \leq \ \max\{\deg(f), \deg(g)\}$$

The proof is trivial.

Theorem 16.6. *Assume that R is an integral domain. Then*
(i) $R[x]$ is an integral domain;

(ii) if $f \neq 0 \neq g$ *, then* $\deg(fg) = \deg(f) + \deg(g)$;
(iii) f *is invertible in* $R[x]$ *if and only if* f *is a constant which is invertible in* R*.*

Proof. If $f = \sum_{i=0}^{n} a_i x^i$ with $a_n \neq 0$, and $g = \sum_{i=0}^{m} b_i x^i$ with $b_m \neq 0$, then $a_n b_m$ is the coefficient of x^{n+m} in fg. But $a_n b_m \neq 0$ since R is an integral domain. Thus $fg \neq 0$, proving (i). Also if $i + j > m + n$, then either $i > n$ or $j > m$, so $a_i b_j = 0$. Thus the coefficient of x^k in fg is zero if $k > m + n$. Hence $\deg(fg) = m + n$. Finally, if $fg = 1$, then $0 = \deg 1 = \deg(fg) = \deg f + \deg g = m + n$. But $m \geq 0$ and $n \geq 0$. Hence $n = 0$ and $m = 0$, and so f is a constant, whose inverse in R is the constant g. Hence f is a constant which is invertible in R. Conversely, if $ab = 1$ in R, the same is clearly true in $R[x]$, as required.

Assume for the rest of this section that F is a field.

Theorem 16.7. (The Division Algorithm) *Suppose given polynomials $f \in F[x]$ and $d \in F[x]$, with $d \neq 0$. Then there is exactly one pair $(q, r) \in F[x] \times F[x]$ with the properties:*
(i) $f = qd + r$ (q is the quotient and r the remainder); and
(ii) either $r = 0$ or $\deg(r) < \deg(d)$.

Proof. (Uniqueness) Suppose that (q_1, r_1) and (q_2, r_2) both satisfy *(i)* and *(ii)*. Then $q_1 d + r_1 = f = q_2 d + r_2$, so $r_1 - r_2 = (q_2 - q_1)d$. If $r_1 \neq r_2$, then neither is 0 (check degrees in the last equation). Also

$$\max(\deg r_1, \deg r_2) \geq \deg(r_1 - r_2) = \deg d + \deg(q_2 - q_1) \geq \deg d .$$

But $\deg d < \deg r_1$ and $\deg d < \deg r_2$ by *(ii)*, giving a contradiction. Thus $r_1 = r_2$, so $(q_2 - q_1)d = 0$. But $d \neq 0$, so $q_1 = q_2$, proving uniqueness.

(Existence) For fixed d, proceed by induction on $\deg f$. If $f = 0$ or if $\deg f < \deg d$, let $q = 0$ and let $r = f$. Then *(i)* and *(ii)* are satisfied. If $\deg f = \deg d$, suppose that $f = \sum_{i=0}^{n} a_i x^i$ and $d = \sum_{i=0}^{n} b_i x^i$ with $b_n \neq 0 \neq a_n$. Let $q = b_n^{-1} a_n$. If $n = 0$, let $r = 0$. If $n > 0$, let $r = \sum_{i=0}^{n-1} (a_i - b_n^{-1} a_n b_i) x^i$. Having proved existence for all f with $\deg f < m$, where $m > \deg d$, suppose that $f = \sum_{i=0}^{m} a_i x^i$ where $a_m \neq 0$. Then $f = a_0 + x f_1$ where $f_1 = \sum_{i=0}^{m-1} a_{i+1} x^i$. Since $\deg f_1 = m - 1$, we have $f_1 = q_1 d + r_1$, where $r_1 = 0$ or $\deg r_1 < \deg d$ by the inductive hypothesis. Also $a_0 = q_0 d + r_0$, where $r_0 = 0$ or $\deg r_0 < \deg d$. Thus

$$f = a_0 + x f_1 = (q_0 + x q_1) d + (r_0 + x r_1) .$$

Also $r_0 + x r_1 = q_2 d + r_2$ by the first part of the proof. So let

$$q = q_0 + x q_1 + q_2 \qquad \text{and} \qquad r = r_2 .$$

Definition. If $f \in F[x]$ and $b \in F$, define

$$f(b) := \sum_{i=1}^{n} a_i b^i = [\mu(f)](b) ,$$

where $f = \sum_{i=1}^{n} a_i x^i$. We say that b *is a root* of f (or a *zero* of f) if and only if $f(b) = 0$. Since μ is a morphism,

$$(f + g)(b) = f(b) + g(b) \qquad \text{and} \qquad (fg)(b) = f(b) g(b) .$$

Remainder Theorem 16.8. *If $f \in F[x]$ and $a \in F$, then there exists q with $f = (x-a)q + f(a)$; i.e. the remainder on dividing f by $(x-a)$ is $f(a)$.*

Proof. We have $f = (x-a)q + r$, where either $r = 0$ or else deg$r$ <deg$(x-a) = 1$. Thus r is a constant. But

$$f(a) \;=\; (a-a)q(a) + r(a) \;=\; r(a)\ ,$$

so r is the constant $f(a)$, as required.

Corollary 16.9. *An element $a \in F$ is a root of f if and only if there exists a polynomial $q \in F[x]$ with $f = (x-a)q$, (that is, $\cdots$if and only if $x - a$ is a* divisor *of f in $F[x]$.)*

Theorem 16.10. *If* deg$(f) = k$, *then f has at most "k" roots.*

Proof. When $k = 1$, $f = a_0 + a_1 x$, and f has exactly one root, namely $-a_1^{-1}a_0$. Proceeding inductively, suppose that this theorem is true for all f with deg$f < n$. For a contradiction, let deg$f = n$, and suppose that f has at least "$n+1$" distinct roots, namely $b_0,\ b_1,\ \cdots,\ b_n$. Then $f = (x - b_0)g$ by the Remainder theorem, and deg$g = n - 1$. But $0 = f(b_i) = (b_i - b_0)g(b_i)$. Since $b_i \neq b_0$ for $i > 0$, $g(b_i) = 0$ for $1 \leq i \leq n$. Thus g has "n" distinct roots $b_1, \cdots, b_n$, contradicting the inductive hypothesis.

Proof of Theorem 16.4. $[\mu(f) = 0] \implies [f(a) = 0\ \forall a \in R] \implies$ [f has infinitely many roots] $\implies$ $[f = 0]$ by **16.10**. Thus Ker$\mu = \{0\}$, as required.

Exercise 16G. i) Prove that the kernel of $\mu : \mathbf{Z}_p[x] \to Map(\mathbf{Z}_p,\ \mathbf{Z}_p)$ is the ideal I consisting of all multiples of $x^p - x$, (a *principal* ideal—see the next section).
ii) Thus the ring of polynomial functions over $\mathbf{Z}_p$ is isomorphic to the ring $\mathbf{Z}_p[x]/I$.
iii) Prove that the latter ring has exactly " $p^{\,p}$ " elements.
iv) Deduce that every function $\mathbf{Z}_p \to \mathbf{Z}_p$ is a polynomial function.

Exercise 16H. If R is a non-zero *finite* commutative ring, then the substitution map $\mu : R[x] \to Map(R,\ R)$ is not injective—give a simple cardinality argument to prove this, rather than an example.

17. Principal ideals & unique factorization.

Let R denote a commutative ring.

Definition. For $a \in R$, the *principal ideal generated by* a is the set

$$aR \ := \ \{ \ ar \ : \ r \in R \ \} \ .$$

Exercise 17A. Prove that this set aR *is* an ideal..

Definition. A *principal ideal domain* ("PID") is an integral domain in which every ideal is a principal ideal.

Theorem 17.1. *The rings* $\mathbf{Z}$*, and* $F[x]$*, when* F *is a field, are* PID*'s.*

Proof. We already know that they are integral domains, so let I be an ideal in $F[x]$ (resp. in $\mathbf{Z}$). If $I = \{0\}$, then I is the principal ideal generated by 0. If $I \neq \{0\}$, choose a non-zero $a \in I$ such that no non- zero element of I has degree (resp. absolute value) smaller than that of a. Then we'll prove that I is the principal ideal generated by a, as required. Since $a \in I$, we have $ar \in I$ for all $r \in F[x]$ (resp. all $r \in \mathbf{Z}$), so $aF[x] \subset I$ (resp. $a\mathbf{Z} \subset I$). To reverse the inclusion, let $b \in I$, and divide a into b to obtain $b = qa + r$ where either $r = 0$ or deg$r$ <deg$a$ (resp. $|r| < |a|$). Since $b \in I$ and $qa \in I$, we have $r = b - qa \in I$. But a was a non-zero element of I with smallest possible degree (resp. absolute value), so we must have $r = 0$. Thus $b = qa$, and so $b \in aF[x]$ (resp. $b \in a\mathbf{Z}$), as required.

Definition. In our commutative ring, we say that a *divides* b, and write $a \mid b$, if and only if there exists $c \in R$ with $b = ac$. We say that a and b are *associates*, and write $a \sim b$, if and only if there exists an invertible $u \in R$ with $b = au$.

Proposition 17.2. *Assume that* R *is an integral domain. Then*
(i) for all a*, we have* $a \mid a$ *;*
(ii) $a \mid b$ *and* $b \mid c$ *implies that* $a \mid c$ *;*
(iii) $a \mid b \iff bR \subset aR$ *;*
(iv) $a \sim b \iff$ (both $a \mid b$ and $b \mid a$) $\iff aR = bR$ *;*
(v) $\sim$ *is an equivalence relation ;*
(vi) 0 *is an associate of only itself ;*
(vii) the invertibles are the associates of 1 .

Proof. *(i)* $a = a1$.
(ii) $b = ar$ and $c = bs$ implies that $c = ars$.
(iii) $(a \mid b) \iff (\exists c,\ b = ac) \iff (b \in aR) \iff (bR \subset aR)$.
(iv) $(a \sim b\ ) \implies (\exists$ invertible u with $b = au) \implies$
$(b = au \quad \& \quad a = bu^{-1}) \implies (a \mid b \quad \& \quad b \mid a) \implies$
$(bR \subset aR \quad \& \quad aR \subset bR)$, by (iii) $\implies (aR = bR) \implies$
$(\exists\ c,\ d$ with $a = bc$ and $b = ad) \implies (a = b = 0$ or else $cd = 1)$, since $b = bcd$, and cancellation of non-zero elements is valid in an integral domain, $\implies a \sim b$, since in the first case, $0 \sim 0$, and in the second case c is a unit whose inverse is d.
(v) Clearly "$a \sim b \iff aR = bR$ " defines an equivalence relation.
(vi) $0 \sim a \iff 0R = aR \iff a \in \{0\} \iff a = 0$.
(vii) $1 \sim a \iff 1R = aR \iff 1 \in aR$
$\iff \exists b$ with $ab = 1 \iff a$ is invertible.

Note. We only used the integral domain property once.

Definition. An element d is a *greatest common divisor* for a and b, written $d = \mathrm{GCD}\{a, b\}$, if and only if
(i) $d \mid a$ and $d \mid b$, and
(ii) if $c \mid a$ and $c \mid b$, then $c \mid d$.

Definition. An element m is a *least common multiple* for a and b, written $m = \mathrm{LCM}\{a, b\}$, if and only if
(i) $a \mid m$ and $b \mid m$, and
(ii) if $a \mid n$ and $b \mid n$, then $m \mid n$.

Note. GCD's and LCM's don't always exist.

Proposition 17.3. *Assume that R is an integral domain. Then any two* GCD*'s for $\{a, b\}$ are associates. Similarly for* LCM's.

Proof. If d_1 and d_2 are both GCD's for a and b , then $d_1 \mid d_2$, since d_1 divides any common divisor by (ii) of the definition. Symmetrically, $d_2 \mid d_1$. So by **17.2**(iv), $d_1 \sim d_2$. The proof for LCM's is similar.

Exercise 17B. Show that if $\mathrm{GCD}\{a, b\}$ and $\mathrm{LCM}\{a, b\}$ both exist, then their product is ab, up to associates. Use this to investigate under what conditions the existence of one implies the existence of the other.

Theorem 17.4. *Assume that R is a principal ideal domain. Then, if $a, b \in R$,*
(i) the set $I = \{\ ra + sb\ :\ r \in R,\ s \in R\ \}$ is an ideal in R, and is the principal ideal generated by some GCD *for $\{a, b\}$. In particular,* GCD$\{a, b\}$ *exists and has the form $ra + sb$ for some $r, s \in R$. Also,*
(ii) $aR \cap bR$ is an ideal in R and is the principal ideal generated by LCM$\{a, b\}$.

Proof. (i) An easy computation shows that I is an ideal. Since R is a PID, we can choose d such that $I = dR$. Now $a \in I$, so $d \mid a$. Similarly, $d \mid b$. Thus d is a common divisor. Suppose that $c \mid a$ and $c \mid b$. Then $c \mid (ra + sb)$ for any r and s, i.e. c divides any element of I. In particular $c \mid d$. Thus d is a greatest common divisor.

(ii) Any intersection of ideals is an ideal.

Exercise 17C. Prove this, including the case of an infinite collection of ideals.

Thus we may choose m such that $aR \cap bR = mR$. Then $m \in mR \subset aR$, so $a \mid m$. Symmetrically, $b \mid m$. Thus m is a common multiple of a and b. Now suppose that $a \mid n$ and $b \mid n$. Then $n \in aR$ and $n \in bR$. Hence $n \in aR \cap bR = mR$. So $m \mid n$, and m is a least common multiple.

Note. In $\mathbf{Z}$ and $F[x]$, use the *Euclidean algorithm* to compute GCD's, LCM's, and $\{r,\ s\}$ as in the theorem. That algorithm is the same in $F[x]$ as in $\mathbf{Z}$.

Definition. A factorization $a = bc$ is *trivial* if and only if one of b, c is invertible and the other is an associate of a. An element $a \in R$ is *irreducible* if and only if it is not zero, not invertible, and it admits only trivial factorizations. Thus in an integral domain, a non-zero non-invertible is irreducible if and only if all of its divisors are either invertibles or associates of itself.

Example. In $F[x]$ the invertibles are the non-zero constants, so $f \sim g$ if and only if either is a non-zero constant multiple of the other. A polynomial is irreducible if and only it has positive degree and it is not the product of two polynomials of strictly smaller degree. In particular, every polynomial of degree 1 is irreducible.

Definition. A *unique factorization domain* (sometimes called *Gaussian domain* or UFD—not to be confused with **unidentified flying domain**) is an integral domain R such that the following hold:

(1) Any $a \in R$ which is non-zero and not invertible is a product of irreducibles; i.e. there exists $s \geq 1$ and irreducibles $p_1,\ p_2,\ \cdots,\ p_s$ (not necessarily distinct) such that $a = \prod_{i=1}^{s} p_i$.

(2) Furthermore, any other such factorization of a may be obtained by replacing each p_i by an associate and re-indexing (or re-ordering). That is, if $\prod_{i=1}^{s} p_i = \prod_{j=1}^{r} q_j$, where each p_i and q_j is irreducible, then $r = s$ and there exists a permutation σ of $\{\ 1,\ 2,\ \cdots,\ s\ \}$ such that $p_i \sim q_{\sigma(i)}$ for $1 \leq i \leq s$.

Note. See Exercises **17D** for important non-UFD's.

Theorem 17.5. *For any field F, the ring $F[x]$ is a* UFD. *In other words:* Any element of $F[x]$ with positive degree is a product of irreducible polynomials. Furthermore, any two such factorizations of a polynomial may be obtained from one another by multiplying the factors by non-zero constants and re-ordering.

Proof. *Existence.* We prove by induction on $\deg f$ that $f = \prod_{i=1}^{s} p_i$ for some $s \geq 1$ and irreducibles p_i. If $\deg f = 1$, then f is already irreducible so let $s = 1$ and $p_1 = f$. Now suppose that $\deg f = n > 1$, and that any polynomial of degree less than n can be so factored. If f is irreducible, again let $s = 1$ and $p_1 = f$. If not, then $f = gh$, where g and h both have degree less than n. By the inductive hypothesis, $g = \prod_{i=1}^{t} p_i$ and $h = \prod_{i=t+1}^{s} p_i$ where each p_i is irreducible. Multiplying these together completes the proof of existence.

Uniqueness. To do this we first need a lemma :

Lemma 17.6. *In any* PID, *if p is irreducible and $p \mid g_1 g_2$, then either $p \mid g_1$ or $p \mid g_2$.*

Proof. Let $d = \mathrm{GCD}\{p, g_1\}$. Then $d \mid p$ and p is irreducible, so either $d \sim p$ or d is invertible. If $d \sim p$, then $p \mid g_1$ since $d \mid g_1$. If d is invertible, we may take $d = 1$. Then there exist elements r and s with $1 = pr + g_1 s$. Hence $g_2 = (pr + g_1 s)g_2 = (rg_2)(p) + (s)(g_1 g_2)$. But $p \mid (rg_2)(p)$, and $p \mid (s)(g_1 g_2)$ since $p \mid g_1 g_2$, so $p \mid g_2$ as required.

Proof of Uniqueness. It follows easily from **17.6**, by induction on s, that if p is irreducible and $p \mid \prod_{i=1}^{s} g_i$, then $p \mid g_i$ for some i.

We prove by induction on $\deg f$ that, if $f = \prod_{i=1}^{s} p_i \sim \prod_{j=1}^{r} q_j$ for irreducibles p_i and q_j, then $r = s$ and $q_{\sigma(i)} \sim p_i$ for some permutation σ of

$\{ 1, \cdots, s \}$. If deg $f = 1$, because $\deg(\prod_{i=1}^{s} p_i) \geq s$, we must have $s = 1$ and $f = p_1$. Similarly $r = 1$ and $f \sim q_1$. So we have $p_1 \sim q_1$, as required.

Now suppose that uniqueness has been proved for polynomials of degree less than n, and that deg $f = n$, and, as above, that $f = \prod_{i=1}^{s} p_i \sim \prod_{j=1}^{r} q_j$. Then $p_1 \mid \prod_{j=1}^{r} q_j$, so $p_1 \mid q_{j_1}$ for some j_1 . Since q_{j_1} is also irreducible we have $p_1 \sim q_{j_1}$. Thus $\prod_{i=2}^{s} p_i \sim \prod_{j=1}^{j_1-1} q_j \prod_{j=j_1+1}^{r} q_j = g$, say. Since deg $g < n$, we may apply our inductive hypothesis, and conclude that $s-1 = r-1$ and that $p_i \sim q_{j_i}$ for $2 \leq i \leq s$ and some rearrangement $\{j_2, \cdots, j_s\}$ of $\{1, \cdots, s\} \setminus \{j_1\}$. This completes the proof.

Comments. (1) When $\prod_{i=1}^{s} p_i = \prod_{j=1}^{s} q_j$ (not merely $\sim$), let u_i be the units such that $p_i = u_i q_{\sigma(i)}$. Then we must have $\prod_{i=1}^{s} u_i = 1$.

(2) Just as in **17.1**, we could go through the last proof, replacing $F[x]$ by $\mathbf{Z}$, polynomial by integer, and degree by absolute value. This would give a proof of unique factorization in $\mathbf{Z}$—the 'fundamental theorem of arithmetic'.

(3) In fact, any PID is a UFD. This of course implies the result for $F[x]$ and for $\mathbf{Z}$. But there are plenty of important UFD's which are not PID's. For example, $\mathbf{Z}[x]$, and $F[x_1, x_2, \cdots]$ (polynomials in several variables) for any field F, are UFD's. But the proof requires more work than above. See Section **20**.

(4) In **17.1** and **17.5**, we may replace $F[x]$ by any **Euclideanizable domain**, as defined below. Then replacing deg by δ, these results and their proofs generalise immediately. (Scholars may object to the word used above, but presumably not to the distinction between the adjectives Euclidean and Euclideanizable, any more than they would object to teaching students the distinction between metric and metrizable spaces). Thus any such domain which admits at least one δ as below is a PID. And in the generalisation of **17.5** we would have partly proved (3); but not wholly, since one can find PID's which are not Euclideanizable.

Definition. A *Euclidean domain* is a pair (R, δ), where R is an integral domain, and $\delta : R \setminus \{0\} \to \mathbf{N}$ is a function satisfying :

(i) $\delta(ab) \geq \delta(a)$ for all non-zero a and b ;

(ii) for all a and non-zero b in R, there is a pair (q, r) from R such that $a = qb + r$ and either $r = 0$ or $\delta(r) < \delta(b)$—(Uniqueness of (q, r) isn't assumed.)

An integral domain R is *Euclideanizable* iff $\exists\delta$ with $(R,\ \delta)$ being a Euclidean domain.

The standard examples are then the ones we've been concentrating on: $\mathbf{Z}$ with δ being the absolute value, and $F[x]$ (for a *field* F) with δ being the degree.

Exercises 17D. (A) Write the proofs in detail of the first claim in paragraph (4) just above.

(B) Now try to generalize to arbitrary PID's, i.e. prove the first statement in (3). You will probably be led, for example, to showing that a PID cannot have an infinite, strictly increasing sequence of ideals.

(C) Show that $(R,\ \delta)$ is a Euclidean domain (called the *Gaussian integers*), where $R\ =\ \{\ a+b\sqrt{-1}\ :\ a,\ b\in\mathbf{Z}\ \}$, and δ is the restriction of the complex modulus. Show also that the remainder is not always unique in this example.

(D) Prove that, changing $\sqrt{-1}$ to $\sqrt{-5}$ in the domain above, the complex modulus continues to satisfy (i) in the definition above, and can be used to show that:

a) ±1 are the only invertibles;

b) elements such as 2, 3 and $1\pm\sqrt{-5}$ are irreducible in this domain; and

c) the existence half of unique factorization holds.

But show that $2\cdot3=(1+\sqrt{-5})(1-\sqrt{-5})$ is an instance where the uniqueness half fails [and so (ii) certainly fails for the complex modulus restricted to this domain, and for any other function satisfying (i)].

Exercise 17E. Make a careful construction of a ring which is exactly like $F[x]$, except that arbitrary non-negative *rational* powers of x are used, only finitely many in a given element. Show that you've produced an integral domain in which the *existence* half of unique factorization fails—note that $x=x^{1/2}x^{1/2}=x^{1/4}x^{1/4}x^{1/4}x^{1/4}=\cdots$.

(5) There is a nice treatment of a theorem of Fermat, that any integer prime which is congruent to 1(mod 4) is a sum of two integer squares, using the Euclidean domain in (C) just above. This is an example in which the Euclidean function δ allows for more than one (quotient, remainder) pair in the 'division algorithm', (ii). See **Herstein**, p. 113.

The following facts, concerning $\mathbf{R}$ and $\mathbf{C}$, are probably already known to

the reader.

(i) Every non-constant polynomial in $\mathbf{C}[x]$ has a root in $\mathbf{C}$.

(ii) The only irreducibles in $\mathbf{C}[x]$ are the linear polynomials $ax + b$.

(iii) The irreducibles in $\mathbf{R}[x]$ are the linears and the quadratics $ax^2 + bx + c$ for which $b^2 < 4ac$.

Statement (i) is often called the **fundamental theorem of algebra**; see Appendix **A** and Section **41** for two proofs. Statements (i), (ii) and (iii) are easily shown to be logically equivalent. The unique factorization theorem takes a more specific form for $\mathbf{R}[x]$ and $\mathbf{C}[x]$, using (ii) and (iii) respectively. Later (in **20A** and **31.4**), we'll see that $\mathbf{Q}[x]$, as well as $\mathbf{Z}_p[x]$ for primes p, have irreducibles of *every* positive degree.

18. The field of fractions.

Definition. Let R be an integral domain. Define Q_R, the *field of fractions* of R as follows: as a set, it is the set of equivalence classes

$$Q_R \ := \ (\ R \times [R \setminus \{0\}] \ ) \ / \ \sim \quad ,$$

where $(a_1, b_1) \sim (a_2, b_2)$ if and only if $a_1 b_2 = a_2 b_1$.

Exercise 18A. Before reading further, check that this is an equivalence relation.

Denote the equivalence class of (a, b) as a/b. The operations are defined by:

$$(a/b) + (c/d) \ := \ (ad + bc)/(bd) \qquad \text{and} \qquad (a/b)(c/d) \ := \ (ac)/(bd) \ .$$

You may remember these formulae from kindergarten.

Note. If b and d are both non-zero, then so is bd, so denominators remain in $R \setminus \{0\}$.

Theorem 18.1. *The triple $(Q_R, +, \cdot)$ is well defined and is a field. The zero is $0/b$ for any $b \neq 0$. The negative of a/b is $(-a)/b$. The identity is a/a for any $a \neq 0$. If $a \neq 0 \neq b$, the inverse of a/b is b/a. Finally, the map $\phi : R \to Q_R$ given by $\phi(a) = a/1$ is a monomorphism of rings.*

Proof. It is not hard to verify that $\sim$ is an equivalence relation. For example, the relations $(a_1, b_1) \sim (a_2, b_2)$ and $(a_2, b_2) \sim (a_3, b_3)$ imply that

$a_1b_2 = a_2b_1$ and $a_2b_3 = a_3b_2$. Therefore $(a_1b_2)(a_2b_3) = (a_2b_1)(a_3b_2)$. Now if $a_2 \neq 0$, then $a_2b_2 \neq 0$, so we can cancel a_2b_2 and obtain that $a_1b_3 = a_3b_1$. If $a_2 = 0$, then $a_1b_2 = 0 = a_3b_2$, so $a_1 = 0 = a_3$, and again $a_1b_3 = a_3b_1$. Thus, in both cases, $(a_1, b_1) \sim (a_3, b_3)$. This proves the transitivity of $\sim$. The other two properties are much easier to prove.

To show that the operations are well defined amounts to showing that if

$$(a_1, b_1) \sim (a_2, b_2) \qquad \text{and} \qquad (c_1, d_1) \sim (c_2, d_2) \ ,$$

then

$$(a_1d_1 + b_1c_1, b_1d_1) \ \sim \ (a_2d_2 + b_2c_2, b_2d_2) \quad \text{and} \quad (a_1c_1, b_1d_1) \ \sim \ (a_2c_2, b_2d_2) \ .$$

This is an easy computation. The verifications of the associative, commutative and distributive laws are straightforward calculations, using the corresponding laws in R. For example, to prove distributivity:

$$(a/b + c/d)(e/f) \ = \ [(ad+bc)/bd][e/f] \ = \ (ad+bc)e/bdf \ = \ (ade+bce)/bdf.$$

But

$$(a/b)(e/f) + (c/d)(e/f) \ = \ ae/bf \ + \ ce/df \ = \ (aedf + cebf)/bfdf.$$

The two answers are the same, since

$$(aedf + cebf, bfdf) \ \sim \ (ade + bce, bdf) \ .$$

We have $0/b + c/d = (0d + bc)/bd = bc/bd = c/d$, so $0/b$ is the zero. Also, $(-a)/b + a/b = [(-a)(b) + ab]/b \ = 0/b$, so $(-a)/b$ is the negative of a/b. Furthermore, $(a/a)(c/d) = ac/ad = c/d$, so a/a is the identity for any $a \neq 0$. Also, $(a/b)(b/a) = ab/ba = 1/1$, the identity, so a/b and b/a are inverse, for non-zero a and b.

Finally, the zero and identity in Q_R are distinct, since $0/1 = 1/1$ implies that $0 = 1$ in R, which is false. Thus Q_R is a field.

The map ϕ is a morphism by a trivial calculation. To show that it is injective, suppose that $\phi(a)$ is zero. Then $a/1 = 0/1$ in Q_R, so $(a, 1) \sim (0, 1)$, or $a \cdot 1 = 1 \cdot 0$. Hence $a = 0$. This shows that $\mathrm{Ker}\phi = \{0\}$, as required.

Exercise 18B. Fill in all the details in the proof above.

Note. This theorem shows that any integral domain is isomorphic to a subring of some field. Conversely, any subring R of a field is an integral domain. See **14.4**.

Examples. (1) $\mathrm{Q}_{\mathbf{Z}} = \mathbf{Q}$, and **18.1** reduces to the usual construction of the field of rational numbers, starting from the integers.

(2) If F is a field, $Q_{F[x]}$ is denoted $F(x)$←(*curly* brackets!!), and is called *the field of rational functions with coefficients in* F. We shall identify a polynomial f with the rational function $f/1$ (just as we identify integers with rationals which can be written in the form $a/1$).

19. Prime & maximal ideals.

Let I be an ideal in R, a commutative ring (with 1 , of course).

Definition. We say that I is *prime in* R if and only if
(1) $I \neq R$; and
(2) $ab \in I$ implies that either $a \in I$ or $b \in I$.
We say that I is *maximal in* R if and only if
(1) $I \neq R$; and
(2) if J is an ideal with $I \subset J \subset R$, then either $J = I$ or $J = R$.

The following statements may easily be verified directly, but they are also trivial corollaries of the next theorem:
$\{0\}$ *is prime* $\iff$ *R is an integral domain.*
$\{0\}$ *is maximal* $\iff$ *R is a field.*
Any maximal ideal is a prime ideal.

Exercise 19A. Prove the statements above.

Theorem 19.1. *(a) I is prime* $\iff$ *R/I is an integral domain.*
(b) I is maximal $\iff$ *R/I is a field.*

Proof. (a)⇒: Assume that I is prime. Then :
$(a+I)(b+I) = 0+I \Rightarrow ab+I = I \Rightarrow ab \in I \Rightarrow (a \in I$ or $b \in I) \Rightarrow (a+I =$ $0+I$ or $b+I = 0+I)$. Also, $I \neq R$, so R/I is not the zero ring.

(a)⇐: Assume that R/I is an integral domain. Then :
$ab \in I \Rightarrow (a+I)(b+I) = 0+I \Rightarrow (a+I = 0+I$ or $b+I = 0+I) \Rightarrow (a \in I$ or $b \in I)$. Also, since R/I is not the zero ring, $I \neq R$.

(b)⇒: Assume that I is maximal. Then R/I is not the zero ring, since $I \neq R$. If $a + I \neq 0 + I$, then $a \notin I$. It remains to construct an inverse for $a + I$. Let $J = \{\ x + ay\ :\ x \in I,\ y \in R\ \}$. Then J is an ideal, and $J \neq I$ since $a \in J$, and $I \subset J \subset R$. Thus $J = R$. Choose $x_0 \in I$ and y_0 with $x_0 + ay_0 = 1$. Then

$$(a + I)(y_0 + I) \ = \ (1 - x_0) + I \ = \ 1 + I\ ,$$

so $y_0 + I$ is an inverse for $a + I$.

(b)⇐: Assume that R/I is a field. Then R/I is not the zero ring, so $I \neq R$. Suppose that J is an ideal, $I \subset J \subset R$, and $J \neq I$. Choose $x \in J \setminus I$. Then there exists $y \in R$ with $(x + I)(y + I) = 1 + I$. Thus $1 - xy \in I$, so $1 - xy \in J$. But $xy \in J$ since $x \in J$. Thus $1 \in J$, and $J = R$, as required.

The following can be generalized quite a bit, but is sufficient for our purposes.

Theorem 19.2. *Let R be a principal ideal domain with $d \in R$. Then the following hold.*
(a) The ideal dR is prime if and only if either $d = 0$ or d is irreducible.
(b) If R is a field, then dR is maximal if and only if $d = 0$.
(c) If R is not a field, then dR is maximal if and only if d is irreducible. Thus prime and maximal ideals are nearly the same in a PID.

Proof. (a) Assume that dR is prime and $d \neq 0$. Then d is not invertible since $dR \neq R$. Suppose that $d = ab$. Then $ab \in dR$ so either $a \in dR$ or $b \in dR$. If $a \in dR$, let $a = dc$. Then $d = dcb$, so $cb = 1$, and so b is invertible. Symmetrically, if $b \in dR$, then a is invertible. Thus d admits only trivial factorizations, i.e. d is irreducible. Conversely, if d is irreducible and $ab \in dR$, then d divides ab, so either d divides a or d divides b (by **17.6**). Thus either $a \in dR$ or $b \in dR$, as required. If $d = 0$, then $dR = \{0\}$ is prime, since R is an integral domain.

(b) If R is a field, then $\{0\}$ and R are the only ideals, $\{0\}$ is maximal, and $dR = \{0\}$ if and only if $d = 0$.

(c) Assume that R is not a field and dR is maximal. Then dR is prime, so by (a), either $d = 0$ or d is irreducible. But the first possibility is ruled out since $\{0\}$ is not maximal. Conversely, suppose that d is irreducible, J is an ideal and $dR \subset J \subset R$. Then $J = aR$ for some a, so a divides d. Thus, either $a \sim d$, which implies that $J = dR$, or a is invertible, which implies

that $J = R$. Finally, $dR \neq R$ since d is not invertible. Thus dR is maximal, as required.

Theorem 19.3. *Given a ring morphism $\phi : F[x] \to K$, where F and K are fields, there are two possibilities:*
(1) ϕ is a monomorphism; or
(2) Imϕ is a subfield of K isomorphic to $F[x]/dF[x]$ for some irreducible $d \in F[x]$.

Proof. Suppose that ϕ is not injective. Now $F[x]$ is a PID, so Ker$\phi = dF[x]$ for some $d \in F[x]$. By the first isomorphism theorem, Im$\phi \cong F[x]/dF[x]$. Thus $F[x]/dF[x]$ is an integral domain, since Imϕ is a subring of the field K. Thus $dF[x]$ is a prime ideal in $F[x]$. Hence d is irreducible, since $d \neq 0$ (for $d = 0 \Leftrightarrow \phi$ is a monomorphism). Therefore $F[x]/dF[x]$ is a field and so Imϕ is a field.

Remarks. a) It is immediate from Zorn's Lemma (see **Artin**, p. 588) that every non-zero ring *has* at least one maximal ideal. We won't need this fact later except in one alternative proof.

b) There is an exceptionally important example connecting algebra and topology. Let X be a compact Hausdorff space. The continuous functions form a subring, $\mathbf{C}(X)$, of the ring of all functions from X to $\mathbf{C}$, where the operations are defined by just adding and multiplying values of the functions. For each $x \in X$, the set of continuous functions f for which $f(x) = 0$ is an ideal, I_x, in $\mathbf{C}(X)$. This ideal is maximal, and it can be proved that every maximal ideal has this form. Thus there is a 1-1 correspondence between the maximal ideals in $\mathbf{C}(X)$ and the points of the space X. One can go further and even define, purely algebraically, the topology on the set of maximal ideals which makes the above correspondence into a homeomorphism. This theme of 'algebraizing topology and geometry' has become very important in recent years, especially the generalization to the non-commutative situation.

A purely algebraic analogue of the above correspondence is the evident fact that the maximal ideals, $(x - z)\mathbf{C}[x]$, in the ring $\mathbf{C}[x]$ are in 1-1 correspondence with the 'points', z, in $\mathbf{C}$. It is a non-trivial theorem (**Hilbert's Nullstellensatz**) that this generalizes to a 1-1 correspondence between the set of maximal ideals in the ring of polynomials in n variables and the set of points in complex n-space. This is a fundamental connection between algebra and geometry in a subject called **algebraic geometry**, discussed briefly

at the end of Section **21**.

Exercise 19B. Prove that I_x above is a maximal ideal. (Take X to be the closed interval $[0,\ 1]$ if you haven't studied topology.)

20. Gauss' theorem.

Let R be a UFD. We'll prove :

Theorem 20.1. (Gauss) *$R[x]$ is also a* UFD.

It follows that polynomials in several variables over a UFD also form a UFD; see the next section, especially **21.3**. Coefficients in **Z** or any field are especially important examples.

Definition. The non-zero polynomial $\sum_{i=0}^{n} a_i x^i \in R[x]$ is *primitive in* $R[x]$ if and only if $\mathrm{GCD}\{a_0, \cdots, a_n\} = 1$.

Gauss' Lemma 20.2. *If f and g are primitive in $R[x]$, then so is the polynomial fg.*

Proof. For a contradiction, suppose that p divides all coefficients of fg, where p is an irreducible in R. Let

$$f \ = \ \sum a_i x^i \qquad ; \qquad g \ = \ \sum b_j x^j \ \ ;$$

$k = \min\{\ i\ :\ p \text{ does not divide } a_i\ \}$; $\ell = \min\{\ j\ :\ p \text{ does not divide } b_j\ \}$.

But p divides each of : $\sum_{i+j=k+l} a_i b_j$; a_i for all $i < k$; and b_j for all $j < \ell$. Thus $p \mid a_k b_\ell$, so p divides either a_k or b_ℓ, contradicting the definitions of k and ℓ.

Theorem 20.3. *Let $f \in R[x]$.*

(i) If f is constant, then (f is irreducible in $R[x] \iff f$ is irreducible in R).

(ii) If deg $f > 0$, then (f is irreducible in $R[x] \iff f$ is irreducible in $Q_R[x]$ and primitive in $R[x]$).

Proof. (i) Clearly $f = gh$ is a non-trivial factorization in $R[x]$ if and only if it is a non-trivial factorization in R.

(ii)$\Leftarrow$: If $f = gh$ in $R[x]$, then g and h cannot both have positive degree since f is irreducible in $Q_R[x]$. We may assume that g is a non-zero constant.

Then g must be invertible in R, since f is primitive in $R[x]$. Hence $f = gh$ is a trivial factorization in $R[x]$.

ii)$\Rightarrow$: If f were not primitive in $R[x]$, it would have a non-trivial factorization $f = gh$, where g is the GCD of the coefficients of f. Supposing, for a contradiction, that f is not irreducible in $Q_R[x]$, we would have a non-trivial factorization $f = gh$. But for all non-zero $g \in Q_R[x]$, there exists a $\lambda \in Q_R$ with λg primitive in $R[x]$. Hence exists a $\mu \in Q_R$, such that $\mu f = \bar{g}\bar{h}$, where $\bar{g}$ and $\bar{h}$ are primitive in $R[x]$. Since, by Gauss' lemma, $\bar{g}\bar{h}$ is primitive in $R[x]$, the element μ must be invertible in R, so $f = \mu^{-1}\bar{g}\bar{h}$ is a non-trivial factorization in $R[x]$, contradicting the hypothesis.

Proof of Gauss' Theorem 20.1.

Existence. Factor f as $p_1 p_2 \cdots p_t$ in $Q_R[x]$. Multiplying each p_i by an element in Q_R, we obtain $f = \mu \bar{p}_1 \bar{p}_2 \cdots \bar{p}_t$, where each $\bar{p}_i$ is primitive and irreducible in $R[x]$, and $\mu \in Q_R$. Since $\bar{p}_1 \bar{p}_2 \cdots \bar{p}_t$ is primitive, actually $\mu \in R$. Now factorize $\mu = \mu_1 \mu_2 \cdots \mu_s$ in R. Then $f = \mu_1 \mu_2 \cdots \mu_s \bar{p}_1 \bar{p}_2 \cdots \bar{p}_t$ is a factorization in $R[x]$.

Uniqueness. Suppose that

$$\mu_1 \mu_2 \cdots \mu_s \bar{p}_1 \bar{p}_2 \cdots \bar{p}_t \quad = \quad \nu_1 \nu_2 \cdots \nu_r \bar{q}_1 \bar{q}_2 \cdots \bar{q}_v \ ,$$

where : μ_i and ν_j are irreducibles in R; $\bar{p}_i$ and $\bar{q}_j$ are primitive irreducibles in $R[x]$. We must have $\mu_1 \mu_2 \cdots \mu_s \sim \nu_1 \nu_2 \cdots \nu_r$ since these are the GCD's of the coefficients. Hence $r = s$ and $\mu_i \sim \nu_i$ for all i after re-arranging, since R is a UFD. Also $t = v$, and $\bar{p}_i \sim \bar{q}_i$ for all i (after re-arranging) in $Q_R[x]$—and therefore in $R[x]$—, since $Q_R[x]$ is a UFD.

Exercise 20A. (Eisenstein's Criterion). Let p be a prime. Show that a polynomial with integer coefficients which has all but the top coefficient divisible by p, and the bottom one not divisible by p^2, is irreducible in $\mathbf{Z}[x]$—and therefore in $\mathbf{Q}[x]$. (Argue as in the proof of Gauss' Lemma **20.2**.)

21. Polynomials in several variables and ring extensions.

Let k be a positive integer. The case $k = 1$ of what follows has already been given in the first part of Section **16.** Let $S \supset R$ be a ring extension, and let $(s_1, \cdots, s_k)$ be a sequence from S. Initially we don't preclude the possibility that $s_i = s_j$ when $i \neq j$.

Proposition 21.1. *The ring* $R[s_1][\ s_2]\cdots[s_k]$ *is*

$\{\ \sum r_{i_1,\cdots,i_k} s_1^{i_1} s_2^{i_2}\cdots s_k^{i_k}\ :\ r_{i_1,\cdots,i_k}$ is in R and is almost always zero. $\}$.

It coincides with the intersection, T*, of all subrings of* S *which contain* $R \cup \{s_1,\cdots,s_k\}$*. It is independent of the order of* $(s_1,\cdots,s_k)$*.*

Definition. Denote this ring as $R[s_1,\cdots,s_k]$, and call it the ring *generated by* $R\cup\{s_1,\cdots,s_k\}$, or the ring *generated by* $\{s_1,\ldots,s_k\}$ *over* R*.*

Sketch Proof of 21.1. The first assertion follows by induction on k from the definition of $R[s]$ before **16.1**. The third assertion clearly follows from the second, which is proved by the same argument as in **16.1**: Because T is a ring containing $R\cup\{s_1,\ldots,s_k\}$, it contains all elements as in the first assertion. Because $R[s_1,\cdots,s_k]$ is one of the rings in the intersection defining T, it contains T.

Definition. The sequence $(s_1,\ldots,s_k)$ is *algebraically independent over* R if and only if $\sum r_I s_1^{i_1}\cdots s_k^{i_k} = 0$ implies that all r_I are zero, where the $r_I \in R$. Otherwise it is *algebraically dependent over* R*.*

Remarks. Clearly (s_1) is algebraically dependent over R if and only if s_1 is algebraic over R. Any sequence with $s_i = s_j$ for some $i \neq j$ is evidently algebraically dependent. The algebraic dependence of a *sequence* of *distinct* elements depends only on the *set* of those elements. Thus we shall refer to the algebraic dependence and independence of sets. An *infinite* set will, by definition, be algebraically independent when all of its finite subsets are. That case will seldom occur here.

Exercise 21A. Prove that $(s_1,\ldots,s_k)$ is algebraically independent over R if and only if, for all i, the element s_i is transcendental over the subring $R[s_1,\cdots,s_{i-1}]$.

Definition. *'The' polynomial ring over* R *in "*k*" variables* is the ring $R[x_1,\cdots,x_k]$ for any $(x_1,\cdots,x_k)$ which is algebraically independent over R. As a convention which saves words, we shall never place the string $[x_1,\cdots,x_k]$ after R unless $(x_1,\cdots,x_k)$ is algebraically independent over R (but this applies only to the letter x).

Remarks. By obvious inductions on k from the case $k=1$ done in **16.3**, we get the following.

1) *$R[x_1, \cdots, x_k]$ is unique up to isomorphism.* More precisely, if $(s_1, \cdots, s_k)$ and $(t_1, \cdots, t_k)$ are both algebraically independent over R, then there is a unique isomorphism

$$R[s_1, \cdots, s_k] \longrightarrow R[t_1, \cdots, t_k]$$

which fixes each $r \in R$ and maps each s_i to t_i. See also the extension principle in k variables just ahead.

2) *For any R, there does exist a polynomial ring $R[x_1, \cdots, x_k]$.* For the inductive step, use **16.3**ii) to choose an x_k which is transcendental over $R[x_1, \cdots, x_{k-1}]$. The construction in the proof of **16.3** would then give a ring $R[x_1, x_2]$ consisting of sequences whose terms are sequences whose terms are in R ! Evidently this is not usually a good way to think of a polynomial in two variables; and it gets worse for larger k.

Also by induction k, **16.5** and **20.1** give

Proposition 21.2. *If R is an* ID*, then so is $R[x_1, \ldots, x_k]$. In this case , the invertibles in the latter ring are simply the invertibles in its subring R of 'constants'.*

Theorem 21.3. *If R is a* UFD*, then so is $R[x_1, \ldots, x_k]$.*

Examples. Thus $\mathbf{Z}[x_1, \ldots, x_k]$ and (when F is a field) $F[x_1, \ldots, x_k]$ are UFD's.

In the above two results, we cannot change ID/UFD to PID. In fact $R[x_1, \ldots, x_k]$ is *never* a PID if $k > 1$. For example, the ideal

$$\{\ s_1x_1 + s_2x_2\ :\ s_i \in R[x_1, \cdots, x_k]\ \}\ ,$$

generated by x_1 and x_2, is not principal. In $\mathbf{Z}[x]$, the ideal

$$\{\ \sum a_i x^i\ :\ a_0 \text{ is even}\ \}\ ,$$

generated by 2 and x, is not principal.

Exercise 21B. Take this to its evident conclusion: Prove in detail that if $k > 0$, and R is an ID such that $R[x_1, \cdots, x_k]$ is a PID, then $k = 1$ and R is a field.

Let $S \supset R$ be a ring extension.

Morphism Extension Principle (1 variable). Given $a \in S$, there is a unique ring morphism $\phi : R[x] \longrightarrow S$ which

i) sends elements of R to themselves in S, and

ii) sends x to a.

In fact $\phi(f) = f(a)$. Certainly this ϕ has properties i) and ii). Furthermore, any ring morphism satisfying i) and ii) must send $\sum b_i x^i$ to $\sum b_i a^i$, i.e. send f to $f(a)$. Thus it coincides with our given ϕ. This extension property is the basic property which is *not* always shared by the ring of polynomial *functions*. The image of ϕ is clearly equal to $R[a]$. The map ϕ is injective if and only if $f(a) \neq 0$ whenever $f \neq 0$, i.e. if and only if a is transcendental over R. Thus a is transcendental over R if and only if ϕ determines an isomorphism $R[x] \to R[a]$.

Morphism Extension Principle (k variables). Given a sequence $(a_1, \cdots, a_k) \subset S$, there is a unique ring morphism
$\phi : R[x_1, \cdots, x_k] \longrightarrow S$ which

i) sends elements of R to themselves in S, and

ii) sends each x_i to a_i.

In fact $\phi(f) = f(a_1, \cdots, a_k)$.
This follows directly as in the case $k = 1$ just above, and also by induction on k using the case $k = 1$. The image of ϕ is $R[a_1, \cdots, a_k]$. The map ϕ is injective if and only if $(a_1 \cdots, a_k)$ is algebraically independent over R, and in this case, $R[a_1, \cdots, a_k] \cong R[x_1, \cdots, x_k]$.

The glories of algebraic geometry.

Not much will be done in this book with polynomials in more than one variable. The subject which studies the solution sets in 'n-space' of systems consisting of one or more polynomial equations in n variables is known as *algebraic geometry*. It has probably produced and motivated the most important body of pure mathematics in this century. Its influence pervades algebra, topology, and parts of analysis. Its applications range from number theory to particle physics to computer graphics to mathematical logic. See **Artin**, pp. 373–379, for a brief introduction to algebraic geometry.

22. Symmetric polynomials.

Besides the uniqueness of the polynomial ring, here's another application of the extension principle in several variables. It gives the ring morphism from $R[x_1, \cdots, x_k]$ to itself which permutes the variables according to some given permutation, γ, in the symmetric group S_k . Simply take S to be $R[x_1, \cdots, x_k]$ and take each a_i to be $x_{\gamma(i)}$. Call the resulting map ϕ_γ. It sends $f(x_1, \cdots, x_k)$ to $f(x_{\gamma(1)}, \cdots, x_{\gamma(k)})$. Then ϕ_γ is an isomorphism, whose inverse, $(\phi_\gamma)^{-1}$, is $\phi_{\gamma^{-1}}$, since first re-arranging the variables, then putting them back in place gives the identity map. More precisely, the identity map of the polynomial ring, and $\phi_{\gamma^{-1}} \circ \phi_\gamma$, and $\phi_\gamma \circ \phi_{\gamma^{-1}}$, are all morphisms sending constants to themselves and each x_i to itself. By the uniqueness part of the extension principle, they all must be the same, as required. The fact that re-arranging the variables gives an isomorphism is a more detailed expression of the fact that the order of the variables is immaterial.

Definition. Define $SymmR[x_1, \cdots, x_n]$, the set of *symmetric polynomials*, by: $f \in SymmR[x_1, \cdots, x_k]$ if and only if $\phi_\gamma(f) = f \ \forall \gamma \in S_k$.

Proposition 22.1. *The set $SymmR[x_1 \cdots, x_k]$ is a subring of $R[x_1, \cdots, x_k]$.*

Proof. We have $\phi_\gamma(1) = 1$. If $\phi_\gamma(f) = f$ and $\phi_\gamma(g) = g$, then $\phi_\gamma(f \pm g) = f \pm g$, and $\phi_\gamma(fg) = fg$.

Definition. Define the i^{th} *elementary symmetric polynomial*, $e_i \in R[x_1 \cdots, x_k]$, to be the coefficient of t^i in the element

$$\prod_{j=1}^{k} (1 + x_j t) \ \in \ R[x_1 \cdots, x_n][t] \ ,$$

where t is transcendental over $R[x_1 \cdots, x_n]$.

Proposition 22.2. *We have $e_i \in SymmR[x_1 \cdots, x_k]$;*
$e_0 = 1$; $e_i = 0$ for $i > k$; and, for $1 \leq i \leq k$,

$$e_i \ = \sum_{1 \leq j_1 < \cdots < j_i \leq k} x_{j_1} x_{j_2} \cdots x_{j_i} \ .$$

Note. We're assuming that k is 'fixed'. The polynomial e_i, of course, depends on k.

Proof. To show that e_i is symmetric, note that for all $\gamma \in S_k$, we have

$$\prod_{j=1}^{k}(1 + x_{\gamma(j)}t) \;=\; \prod_{j=1}^{k}(1 + x_j t) \ .$$

Note also that the right-hand side has 1 as coefficient of t^0, and has degree k as a polynomial in t, so $e_0 = 1$ and $e_i = 0$ for $i > k$.

Exercise 22A. Prove the last formula in **22.2** by induction on k, or otherwise.

The following is a fundamental fact about symmetric polynomials, and is often phrased: Any symmetric polynomial can be expressed uniquely as a polynomial in the elementary symmetric polynomials.

Theorem 22.3. *The set* $\{\ e_1,\ e_2,\ \cdots,\ e_k\ \}$ *is algebraically independent over* R*, and generates* $SymmR[x_1, \cdots, x_k]$*. Thus*

$$SymmR[x_1,\ \cdots,\ x_k] \;=\; R[e_1,\ \cdots,\ e_k] \ .$$

(So the subring, $SymmR[x_1,\ \cdots,\ x_k]$, is actually isomorphic to its extension ring, $R[x_1,\ \cdots,\ x_k]$.)

Proof. Proceed by induction on k. When $k = 1$, it is clear, since $e_1(x_1) = x_1$ is transcendental over R, and $SymmR[x_1] = R[x_1]$. For the inductive step, let $\bar{e}_1, \cdots, \bar{e}_{k-1}$ denote the ESF's in $x_1, \cdots, x_{k-1}$.

Algebraic independence: Suppose, for a contradiction, that each f_i is a polynomial in "$k-1$" variables, and that $f = \sum_{i=0}^{n} f_i t^i$ is a non-zero polynomial of least degree n (in the last variable t) such that $f(e_1, \cdots, e_{k-1})(e_k) = 0$. Now $f_0 \neq 0$, since otherwise we could factor out a copy of e_k from the relation, contradicting the minimality of n.

Exercise 22B. Justify this, without assuming that R is an ID.

On the other hand, let $\phi : R[x_1, \cdots, x_k] \to R[x_1, \cdots, x_{k-1}]$ be the ring morphism obtained by 'setting x_k equal to zero'. Then $\phi(e_i) = \bar{e}_i$ for $1 \leq i \leq k-1$, and $\phi(e_k) = 0$. Applying ϕ to the relation $\sum f_i(e_1, \cdots, e_{k-1})(e_k)^i = 0$ yields $f_0(\bar{e}_1, \cdots, \bar{e}_{k-1}) = 0$. By the inductive hypothesis, $\{\ \bar{e}_1,\ \cdots,\ \bar{e}_{k-1}\ \}$ is algebraically independent, so $f_0 = 0$, giving the required contradiction.

Generation: We must show that $SymmR[x_1, \cdots, x_k] \subset R[e_1, \cdots, e_k]$. Define the total degree of a non-zero polynomial in $R[x_1, \cdots, x_k]$ to be the maximum sum of exponents of x_i's which occurs in a monomial with non-zero coefficient in the polynomial. Suppose, for a contradiction, that $g \in SymmR[x_1, \cdots, x_k]$ is a non-constant polynomial of least total degree not lying in $R[e_1, \cdots, e_k]$. Then $\phi(g) \in SymmR[x_1, \cdots, x_{k-1}]$, so by the inductive hypothesis, $\phi(g) = f_0(\bar{e}_1, \cdots, \bar{e}_{k-1})$ for some f_0. Regarding f_0 as a polynomial in k variables (constant with respect to the last variable), let $g_1 = g - f_0(e_1, \cdots, e_k) \in SymmR[x_1, \cdots, x_k]$. For $1 \leq j \leq k$, let $\psi_j : R[x_1, \cdots, x_k] \to R[x_1, \cdots, x_k]$ be the map obtained by setting x_j equal to zero. Then $\psi_k(g_1) = 0$ since $\phi(g_1) = 0$. Since g_1 is symmetric, we have $\psi_j(g_1) = 0$ for all j. But, if

$$g_1 \;=\; \sum a_{i_1,\cdots,i_k} x_1^{i_1} \cdots x_k^{i_k} \;,$$

then

$$\psi_j(g_1) \;=\; \sum_{i_j=0} a_{i_1,\cdots,i_k} x_1^{i_1} \cdots x_k^{i_k} \;.$$

Thus $a_{i_1,\cdots,i_k} = 0$ if $i_j = 0$ for some j, i.e. every non-zero term in g_1 involves each x_j to a positive power. Thus $g_1 = x_1 x_2 \cdots x_k g_2 = e_k g_2$ for some g_2. Now g_2 is symmetric. Also

$$\text{totaldeg.} f_0(e_1, \cdots, e_k) \;=\; \text{tot.deg.} f_0(\bar{e}_1, \cdots, \bar{e}_{k-1}) \;=\; \text{tot.deg.}\phi(g) \;.$$

But $\text{tot.deg.}\phi(g) \leq \text{tot.deg.}(g)$. Thus

$$\text{tot.deg.} g_1 \;\leq\; \max\{\text{tot.deg.} g, \text{tot.deg.} f_0(e_1, \cdots, e_k)\} \;\leq\; \text{tot.deg.} g \;.$$

So $\text{tot.deg.} g_2 = \text{tot.deg.}(g_1) - k <$ tot.deg.g. By the minimality of tot.deg.g, we get $g_2 = f_2(e_1, \cdots, e_k)$ for some f_2. Thus

$$\begin{aligned} g \;&=\; g_1 + f_0(e_1, \cdots, e_k) \;=\; e_k g_2 + f_0(e_1, \cdots, e_k) \\ &=\; e_k f_2(e_1, \cdots, e_k) + f_0(e_1, \cdots, e_k) \;. \end{aligned}$$

So $g \in R[e_1, \cdots, e_k]$, completing the proof.

Theorem 22.4. *If F is a field, and $f \in F[t]$ is a monic polynomial, in the trancendental t, of degree n with roots $a_1,\ a_2,\ \cdots,\ a_n$ (repeats possible) in F, then the coefficient of t^i in f is $(-1)^{n-i} e_{n-i}(a_1, \cdots, a_n)$.*

Proof. Let $\phi : F[x_1, \cdots, x_n][t] \to F[t]$ be the ring morphism defined by $\phi(\sum f_i t^i) = \sum f_i(-a_1, \ -a_2, \ \cdots, \ -a_n)t^i$. Then, working in the field of fractions of $F[x_1, \cdots, x_n][t]$, we get

$$f \ = \ \prod_{i=1}^{n}(t - a_i) \ = \ \phi[\prod_{i=1}^{n}(t + x_i)] = \phi[t^n \prod_{i=1}^{n}(1 + t^{-1}x_i)]$$

$$= \ \phi[t^n \sum_{i=0}^{n} e_{n-i}(x_1, \cdots, x_n)(t^{-1})^{n-i}] \ = \ \sum_{i=0}^{n} e_{n-i}(-a_1, \cdots, -a_n)t^i$$

$$= \ \sum_{i=0}^{n}(-1)^{n-i}e_{n-i}(a_1, \cdots, a_n)t^i \ ,$$

as required.

Note. This gives a formula for getting the coefficients, given the roots. Going the other way is a different kettle of fish, as the next 60 or so pages will demonstrate.

Exercise 22C. Prove that **22.4** holds more generally with coefficients in any commutative ring.

Remark. There is a method for making sense of a 'limit' of the rings $SymmR[x_1, \cdots, x_k]$, as the number of variables, k, tends to infinity. This produces a very important ring, which is often called the *ring of symmetric functions*, although its elements resemble functions even less than in the case of finitely many variables. See the first chapter of **Macdonald** for a veritable feast of information on this ring and its connections to classical algebra and combinatorics.

Exercise 22D. In $R[x_1, \cdots, x_k]$, define $\Delta \ := \ \prod_{i<j} \ (x_i - x_j)$. Let $\gamma \in S_k$. Prove that

$$\phi_\gamma(\Delta) = \text{sign}(\gamma)\Delta \ .$$

In many books, the proof that the *sign* function is well defined uses this 'partially symmetric' polynomial Δ (also called an *alternating* polynomial, or an *alternating* function, because the subgroup which fixes it is A_k), instead of using the closely related direct formula for the *sign* which we used in **1.3**.

III. Basic Field Theory.

Sections 23 to **34** give the material which is crucial to the Galois theory in the following group of sections, where we prove the famous results about non-solvability of polynomial equations. The sections here are essentially that portion of elementary field theory which doesn't use anything at all substantial about groups. We give a review without proofs of the linear algebra needed, as well as the application of field theory to the classical problems concerning construction of figures using only straight-edge and compass (Section **27**). One exception here is part of Gauss' theorem concerning which regular n-gons can be constructed. He did it without Galois theory, but for us it's easier to delay the last step and use the Galois correspondence. The second notable result here is the complete classification of finite fields, up to isomorphism, done in Section **31**. This is much in contrast to the case of finite groups, where classification still seems to lie far in the future. (Perhaps one of you will do it?)

23. Prime fields and characteristic.

Definitions. Let $b \in F$, a field. If $n \in \mathbf{Z}$, define $n \cdot b$ as follows:

$$n \cdot b \ := \ b + b + \cdots + b \qquad (\text{“}n\text{”times}) \qquad \text{if} \ \ n > 0;$$

$$0 \cdot b \ := \ 0 \quad ; \quad n \cdot b \ := \ -\left[(-n) \cdot (b)\right] \ = \ (-n) \cdot (-b) \qquad \text{if} \ \ n < 0 \ .$$

The *characteristic* of F is zero, $\mathrm{ch}(F) := 0$, if and only if $n \cdot 1 \neq 0$ for all $n > 0$. If $\mathrm{ch}(F) \neq 0$, define

$$\mathrm{ch}(F) \ := \ \min\{ \ n \ : \ n \cdot 1 = 0 \ , \ n > 0 \ \} \ .$$

The *prime subfield* of F, denoted P, is the intersection of all subfields of F. Clearly P is a subfield of every subfield of F.

Theorem 23.1. *Assume that $ch(F) = 0$. Then $n \cdot b \neq 0$ for any $b \neq 0$ and $n \neq 0$. In this case, $P \cong \mathbf{Q}$.*

Theorem 23.2. *Assume that $ch(F) = p > 0$. Then p is a prime integer. If $b \neq 0$, then $n \cdot b = 0$ if and only if p divides n. In this case, $P \cong \mathbf{Z}_p$.*

Proof of both. Define $\phi : \mathbf{Z} \to F$ by $\phi(n) := n \cdot 1$. Then ϕ is a ring morphism, so Imϕ is an integral domain, by **14.1** and **14.4**. Thus Kerϕ is a prime ideal in $\mathbf{Z}$ by **19.1**a). Therefore, by **19.2**a):
either Ker $\phi = \{0\}$, i.e. $n \cdot 1 = 0$ implies that $n = 0$, and so ch$(F) = 0$;
or else Ker$\phi = p\mathbf{Z}$ for some prime p.
In the latter case, $n \cdot 1 = 0 \iff p$ divides n.
In both cases, if $b \neq 0$, then $n \cdot b = 0 \iff n \cdot 1 = (n \cdot b)(b^{-1}) = 0$.

So if ch$(F) = 0$, then $n \cdot b = 0 \iff n = 0$.

And if ch$(F) = p$, then $n \cdot b = 0 \iff p$ divides n. So by the first isomorphism theorem, Im$\phi \cong \mathbf{Z}_p$. Thus Imϕ is a sub*field* of F, so P is contained in Imϕ. But $\mathbf{Z}_p$ has no proper subfields, so $P =$Imϕ, and $P \cong \mathbf{Z}_p$, as required.

Finally, if ch$(F) = 0$, define $\psi : \mathbf{Q} \to F$ by $\psi(m/n) := \phi(n)^{-1}\phi(m)$. Then ψ is well defined, partly because $\phi(n) \neq 0$ if $n \neq 0$. It is a ring morphism between fields, so is injective. Thus Im$\psi \cong \mathbf{Q}$, so Imψ is a subfield of F and therefore contains P. The reverse inclusion follows because $\mathbf{Q}$ contains no proper subfields, and therefore neither does any field isomorphic to $\mathbf{Q}$ such as Imψ. More directly we can check this last point by noting that, for integers m and n, we have $1 \in P$, and therefore $m \cdot 1 = \phi(m) \in P$. Thus $\phi(n)^{-1} \in P$, and so $\phi(n)^{-1}\phi(m) \in P$. Hence $P =$ Imψ, and so $P \cong \mathbf{Q}$, as required.

Note. Since any field has only the two extreme ideals, and a morphism between fields, $F \to F'$, cannot have F as kernel (since it maps 1 to 1), it has $\{0\}$ as kernel, so is injective. We'll call it a *field map*. It provides an isomorphism of F with a subfield of F'.

Exercises 23A. Show that ch$(F) = 0$ implies that F is infinite. Is the converse true? Prove that, if ch$(F) \neq$ ch(F'), then there are no field maps $F \to F'$.

Tedious(?) Exercise 23B. Rewrite the proof of **23.1/23.2** avoiding the use of the results on prime and maximal ideals in Section **19**.

24. Simple extensions.

Definition. Let $a \in K$, an extension field of F. The *field generated by F and a*, denoted $F(a)$, is the intersection of all those subfields of K which contain $F \cup \{a\}$.

Note that $F[a] \subset F(a)$, but possibly $F[a]$ is not a field, in which case it is a proper subring of $F(a)$.

If $K = F(a)$ for some $a \in K$, then K will be called a *simple extension* of F.

Theorem 24.1. *Suppose that a is transcendental over F. Then there is a unique isomorphism $\psi : F(x) \to F(a)$ for which both $\psi(b) = b$ for all $b \in F$ and $\psi(x) = a$. In this case, $F[a] \neq F(a)$.*
(Recall that $F(x)$ is the field of rational functions over F, i.e. the field of fractions of the polynomial ring—so there is no conflict of notation between this earlier use of $F(x)$ and the use introduced in this section.)

Definition. In this case, $F(a)$ is called a *simple transcendental extension* of F.

Theorem 24.2. *Suppose that a is algebraic over F. Then there is a unique monic irreducible $f_0 \in F[x]$ such that there exists an isomorphism*

$$\psi : F[x]/f_0F[x] \longrightarrow F(a)$$

for which both $\psi(b + f_0F[x]) = b$ for all $b \in F$ and $\psi(x + f_0F[x]) = a$. In this case, ψ is also unique, and $F[a] = F(a)$.

Definition. In this case, $F(a)$ is called a *simple algebraic extension* of F. Also f_0 is called *the minimal polynomial* of a over F. It is the monic polynomial of least degree such that $f_0(a) = 0$, as we shall see in the proof. Its degree is called also the *degree of a over F*.

Proof of both. *Uniqueness of ψ*: In **24.1**, the conditions on ψ imply that

$$\psi(\sum b_i x^i / \sum c_i x^i) \;=\; \psi(\sum b_i x^i)\psi(\sum c_i x^i)^{-1}$$
$$=\; (\sum \psi(b_i)\psi(x)^i)(\sum \psi(c_i)\psi(x)^i)^{-1} \;=\; (\sum b_i a^i)(\sum c_i a^i)^{-1}\ ,$$

so ψ is unique. In **24.2** we have

$$\psi(\sum b_i x^i + f_0F[x]) \;=\; \sum \psi(b_i + f_0F[x])\psi(x + f_0F[x])^i \;=\; \sum b_i a^i\ ,$$

so ψ is unique. [Using these formulae, we could now verify the properties, but we'll proceed differently.]

Existence. (Notice the analogies between this proof and the proof of **23.1/23.2**.) Let $\phi : F[x] \to K$ be the morphism defined by setting $\phi(f) := f(a)$. Then $F[a] = \mathrm{Im}\phi$, which is an integral domain, by **14.1** and **14.2**. Thus $\mathrm{Ker}\phi$ is a prime ideal in $F[x]$, by **19.1**a). Hence $\mathrm{Ker}\phi$ is either $\{0\}$ or $f_0F[x]$ for some irreducible f_0 in $F[x]$, by **19.2**a). Now

$$\phi \text{ is injective } \iff [f(a) = 0 \Rightarrow f = 0] \iff a \text{ is transcendental } .$$

If a is transcendental, let $\psi : F(x) \to F(a)$ be $\psi(f/g) = \phi(g)^{-1}\phi(f)$. Then ψ is well-defined and is a non-zero morphism between fields, so is injective. Because $\mathrm{Im}\psi$ contains $F \cup \{a\}$ and is a subfield of $F(a)$, it is clear that $\mathrm{Im}\psi = F(a)$. Thus ψ is also surjective, as required. Since $F[a] = \psi(F[x])$ and $F[x] \neq F(x)$, we have $F[a] \neq F(a)$.

Now assume that a is algebraic, and so $\mathrm{Ker}\phi$ is a prime ideal generated by a monic irreducible f_0. By the first isomorphism theorem, there exists an isomorphism ψ from $F[x]/f_0F[x]$ to $\mathrm{Im}\phi$ with the given properties. Now $\mathrm{Im}\phi$ is a subfield of K containing $F \cup \{a\}$, so $F(a) \subset \mathrm{Im}\phi$. But $\mathrm{Im}\phi = F[a] \subset F(a)$. Thus $\mathrm{Im}\phi = F(a) = F[a]$. Finally any monic irreducible f such that $f(a) = 0$ must have $\mathrm{GCD}\{f, f_0\}$ of positive degree (since a is a root of $\mathrm{GCD}\{f, f_0\} = rf + sf_0$ for some r, s), so $f = f_0$. Thus f_0 is unique.

Tedious(?) Exercise 24A. Rewrite the proof of **24.1/24.2** avoiding the use of the results on prime and maximal ideals in Section **19**.

25. Review of vector spaces and linear maps.

Let F be any field. An *F-vector space* is a set V together with
i) an addition $+$ on V such that $(V, +)$ is an abelian group;
ii) a scalar multiplication $F \times V \to V$, $(\alpha, v) \mapsto \alpha \cdot v$;
such that for all α, β in F and v, v_1, v_2 in V, we have

$$(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v \quad ; \quad \alpha \cdot (v_1 + v_2) = \alpha \cdot v_1 + \alpha \cdot v_2 \ ;$$

$$\alpha \cdot (\beta \cdot v) = (\alpha\beta) \cdot v \quad ; \quad 1 \cdot v = v \ .$$

A *linear combination* from a set $S \subset V$ is any element of the form

$$\sum_{s \in S} \alpha_s \cdot s \ ,$$

where $\{\ s \in S \mid \alpha_s \neq 0\ \}$ is **finite**.

S *generates* $V \iff$ every element is a linear combination from S ;

$$S \text{ is } \textit{linearly independent} \iff [\sum_{s\in S} \alpha_s \cdot s = 0 \Rightarrow \alpha_s = 0\ \forall s\ ]\ ;$$

S is a *basis* for $V \iff S$ is linearly independent and generates V .

Theorem 25.1. *(i) Any generating set contains a basis.*
(ii) Any linearly independent set is contained in some basis.
(iii) Any two bases for V 'are in 1-1 correspondence'.

Note. Using either *(i)*, since V generates V, or *(ii)*, since the empty set is linearly independent, we see that V *has* a basis. The number of elements in any basis, well defined by *(iii)*, is the *dimension* of V, denoted $\dim V$, and is either a non-negative integer, or ∞ (or rather an infinite cardinal, for the sophisticated).

Definition. A map $\phi : V \to W$ between F-vector spaces is *linear* if and only if

$$\phi(v_1 + v_2) \ = \ \phi(v_1) + \phi(v_2) \qquad \text{and} \qquad \phi(\alpha \cdot v) \ = \ \alpha \cdot \phi(v)$$

for all $\alpha \in F$ and $v_1, v_2, v \in V$.

Proposition 25.2. *Assume that ϕ is linear. Then*

$$\phi(\sum_{s\in S} \alpha_s \cdot s) \ = \ \sum_{s\in S} \alpha_s \cdot \phi(s).$$

If ϕ is surjective and S generates V, then $\phi(S)$ generates W. If ϕ is injective and S is linearly independent, then $\phi(S)$ is linearly independent. If ϕ is bijective and S is a basis for V, then $\phi(S)$ is a basis for W.

Definition. $V \cong W$ as F-vector spaces if and only if there is a bijective linear $\phi : V \to W$.

Corollary. $V \cong W \implies \dim V = \dim W$ *as (*possibly infinite*) cardinals.*

Theorem 25.3. *Conversely, dim* $V = \dim W < \infty \implies V \cong W$.

Note. We won't need it here, but **25.3** holds without requiring finite dimensionality, as long as we interpret dimension as a cardinal. A countable dimensional vector space will *not* be isomorphic to a vector space whose dimension is some uncountable cardinal. As a $\mathbf{Q}$-vector space, $\mathbf{R}$ has uncountable dimension.

26. The degree of an extension.

Let F be any field.

Theorem 26.1. *If K is an extension of F, then the addition in K, together with the scalar multiplication $\mu : F \times K \to K$, $\mu(b, a) := ba$, gives K the structure of a vector space with scalars in F.*

Proof. Clearly K is an abelian group under addition. Furthermore, the laws given in the second part of the definition of *vector space* all hold, since K is a ring, and those laws are part of the definition of *ring*.

Definitions. The *degree of K over F,* which is denoted as $[K : F]$, is the dimension of K as a vector space over F. Possibly $[K : F] = \infty$. Here we won't need to distinguish between different infinite cardinals. In this case, K is an *infinite extension of F* (which is saying much more than just that K is infinite as a set). Otherwise, i.e. if $[K : F]$ is finite, we say that K is a *finite extension of F* (even though K is *not* finite *as a set* unless F is, in this case).

Remark. It is clear now what the additive group structure of a field is; i.e. as a group under $+$, we have $K \cong P^{[K:P]} \cong \mathbf{Q}^n$ or ${\mathbf{Z}_p}^n$ for some (possibly infinite) cardinal n. Once the characteristic and the degree over the prime field are known, any additional information about a field necessarily involves its multiplication.

Theorem 26.2. *Let $a \in K$, an extension of F. Then the following hold.*
(i) $[F(a) : F] = \infty \iff$ *a is transcendental over F.*
(ii) $[F(a) : F]$ *is finite* $\iff$ *a is algebraic over F. In this case, the degree $[F(a) : F]$ is also the degree of the minimal polynomial of a, i.e. the degree of a over F, and* $\{\ 1,\ a,\ a^2,\ \cdots,\ a^{n-1}\ \}$ *is a basis for $F(a)$ as an F-vector space, where $n := [F(a) : F]$.*

Proof. We prove the implications $\Longleftarrow$, since, for example, *(i)*$\Rightarrow$ is the same as *(ii)*$\Leftarrow$.

(i) $\{a$ is transcendental over $F\} \implies \{F(a) \cong F(x)$ by a ring isomorphism mapping all elements of F to themselves $\}$
$\implies \{F(a) \cong F(x)$ by an isomorphism of F-vector spaces $\}$
$\implies$ $F(a)$ and $F(x)$ have the same dimension over F,
i.e. $[F(a) : F] = [F(x) : F] = \infty$. The last equality follows from the fact that $F(x)$ contains a subspace $F[x]$ which is easily seen to be infinite dimensional.

Challenge. Can you actually write down a basis over F for the vector space $F(x)$?? At least, can you show that its dimension is the smallest infinite cardinal, i.e. 'countable'? Think about partial fractions! While you're at it, why not formulate and prove a theorem on partial fraction decomposition in $F(x)$, drawing on your experience learning methods of integration?

(ii) $\{a$ is algebraic over $F\} \implies \{F(a) \cong F[x]/f_0F[x]$ by an isomorphism of F-vector spaces, where f_0 is the minimal polynomial of $a\} \implies$

$$[F(a) : F] \ = \ [(F[x]/f_0F[x]) : F] \ = \ \deg f_0\ .$$

If $n := \deg f_0$, then

$$\{\ 1 + f_0F[x],\ x + f_0F[x],\ x^2 + f_0F[x],\ \cdots\cdots,\ x^{n-1} + f_0F[x]\ \}$$

is a basis for $F[x]/f_0F[x]$. Thus the image of that set under the isomorphism, namely $\{\ 1,\ a,\ \cdots,\ a^{n-1}\ \}$, is a basis for $F(a)$.

Theorem 26.3. *Given an iterated field extension $F \subset E \subset K$, we have*

$$[K : F] \;=\; [K : E]\,[E : F]\;.$$

In fact, if $\{a_\lambda\}_{\lambda\in L}$ is an indexed basis for E as an F-vector space, and $\{b_\mu\}_{\mu\in M}$ is an indexed basis for K as an E-vector space, then $\{a_\lambda b_\mu\}_{(\lambda,\mu)\in L\times M}$ is an indexed basis for K as an F-vector space.

Proof. The first sentence follows from the second, since $[E : F]$ is the cardinality of L, and $[K : E] = \mathrm{card}M$, so $[K : F]$ would be, as required, $\mathrm{card}(L \times M) = (\mathrm{card}L)(\mathrm{card}M)$.

(i) *Every element of K is a linear combination of $\{a_\lambda b_\mu\}$ with coefficients in F:* Let $c \in K$. Since $\{b_\mu\}$ is a basis for K over E, there exist $e_\mu \in E$ such that $c = \sum_{\mu\in M} e_\mu b_\mu$. Since $\{a_\lambda\}$ is a basis for E over F, for all μ there exist $f_{\mu\lambda} \in F$ such that $e_\mu = \sum_{\lambda\in L} f_{\mu\lambda}a_\lambda$. Then

$$c \;=\; \sum_{(\lambda,\mu)\in L\times M} f_{\mu\lambda}a_\lambda b_\mu\;,$$

as required.

A field full of e_μ's will excite the naturalist even more than the algebraist—to say nothing of ν's.

(ii) *The set $\{a_\lambda b_\mu\}$ is linearly independent over F*: Suppose that

$$\sum_{\mu\in M}\,(\sum_{\lambda\in L} f_{\mu\lambda}a_\lambda)b_\mu \;=\; 0\;.$$

Since $\sum f_{\mu\lambda}a_\lambda \in E$, and $\{b_\mu\}$ is linearly independent over E, it follows that $\sum f_{\mu\lambda}a_\lambda = 0$ for all $\mu \in M$. Since $\{a_\lambda\}$ is linearly independent over F, it now follows that $f_{\mu\lambda} = 0$ for all μ and λ, as required.

Corollary 26.4. *If $F \subset E \subset K$ are field extensions, then (K is a finite extension of F) $\iff$ (both K is a finite extension of E, and E is a finite extension of F) . In this case, both $[E : F]$ and $[K : E]$ are divisors of $[K : F]$.*

Corollary 26.5. *If $a \in K$, a finite extension of F, then a is algebraic over F, and the degree of a over F divides $[K : F]$.*

Proof. Let $E = F(a)$ in **26.4** and apply **26.2**.

Definition. If $\{a_1, \cdots, a_n\} \subset K \supset F$, define $F(a_1, \cdots, a_n)$ to be the intersection of all subfields of K which contain $F \cup \{a_1, \cdots, a_n\}$. It is called the *field generated by* $\{a_1, \cdots, a_n\}$ *over* F, and is in fact the set

$$\{\ g(a_1, \cdots, a_n)^{-1} f(a_1, \cdots, a_n)\ :\ f, g \in F[x_1, \cdots, x_n]\ ;\ g(a_1, \cdots, a_n) \neq 0\ \}\ .$$

It is easily seen that

$$F(a_1, \cdots, a_i)(a_{i+1}, \cdots, a_n)\ =\ F(a_1, \cdots\cdots, a_n)$$

for $1 \leq i \leq n$.

Theorem 26.6. *Let K be a finite extension of F. Then there exists a positive integer n and a set $\{\ a_1,\ \cdots,\ a_n\ \} \subset K$ such that*

$$F\ \subset\ F(a_1)\ \subset\ F(a_1, a_2)\ \subset\ \bullet\ \bullet\ \bullet\ \subset\ F(a_1, \cdots, a_n) = K\ .$$

Only the last equality is at issue, but we want to emphasize the tower of fields.

Note. In a sense, this determines the structure of finite extensions, since at each stage $F(a_1, \cdots, a_i) = F(a_1, \cdots, a_{i-1})(a_i)$ is a simple algebraic extension of the previous stage $F(a_1, \cdots, a_{i-1})$; and the structure of simple algebraic extensions was analysed in **24.2** and **26.2**. The element a_i is even algebraic over F, by **26.5**.

Proof. Proceed by induction on $[K : F]$, simultaneously for all K and F. If $[K : F] = 1$, any n and a_i will do, since $K = F$, and each inclusion is actually equality. For the inductive step, let $[K : F] > 1$, so that $K \neq F$. Choose $a_1 \in\ K \setminus F$. Then

$$[K : F]\ =\ [K : F(a_1)][F(a_1) : F]\ ,$$

and $a_1 \notin F$, so $F(a_1) \neq F$. Thus $[F(a_1) : F] > 1$. It follows that $[K : F(a_1)] < [K : F]$, and so we may apply the inductive hypothesis to the extension K of $F(a_1)$. Thus there exist $a_2, \cdots, a_n$ such that $K = F(a_1)(a_2, \cdots, a_n) = F(a_1, \cdots, a_n)$, as required.

Exercise 26A. Suppose that $a \in K \supset E \supset F$. Show that

$$[E(a) : E] \ \leq \ [F(a) : F] \ .$$

Give an example where the inequality is strict.

The following calculations will probably need at least **26A** and the fact that $x^n - k$ is irreducible in $\mathbf{Q}[x]$ for $n > 0$ when the integer k is divisible by some prime but not by its square. This fact follows immediately from **Eisenstein's criterion** given in **20A**.

Exercise 26B. Calculate $[\mathbf{Q}(\sqrt{2},\ \sqrt[3]{5}) \ : \ \mathbf{Q}]$.

Exercise 26C. Calculate $[\mathbf{Q}(\sqrt{2},\ \sqrt{5}) \ : \ \mathbf{Q}]$.

Exercise 26D. Calculate $[\mathbf{Q}(\sqrt[4]{2},\ \sqrt[3]{2}) \ : \ \mathbf{Q}]$.

Exercise 26E. Calculate $[\mathbf{Q}(\sqrt[4]{2},\ \sqrt[6]{2}) \ : \ \mathbf{Q}]$.

Exercise 26F. Calculate $[\mathbf{Q}(\sqrt[4]{5},\ \sqrt[6]{7}) \ : \ \mathbf{Q}]$.

Exercise 26G. Calculate $[\mathbf{Q}(\sqrt[3]{2},\ e^{2\pi i/3}) \ : \ \mathbf{Q}]$. Show that the bigger field is also the field generated over $\mathbf{Q}$ by all the complex roots of $x^3 - 2$.

Exercise 26H. Calculate $[\mathbf{Q}(\sqrt{2},\ e^{2\pi i/3}) \ : \ \mathbf{Q}]$.

Exercise 26I. Calculate $[\mathbf{Q}(e^{2\pi i/8}) \ : \ \mathbf{Q}]$, and $[K : \mathbf{Q}]$ where K is the field generated over $\mathbf{Q}$ by all the complex roots of $x^8 - 1$.

Exercise 26J. Calculate $[\mathbf{Q}(e^{2\pi i/12}) \ : \ \mathbf{Q}]$, and $[L : \mathbf{Q}]$ where L is the field generated over $\mathbf{Q}$ by all the complex roots of $x^{12} - 1$.

27. Straight-edge & compass constructions.

The reader may prefer to skip this section, and return to it only after Section **33**, or even **39**, since we have left the proofs of a couple of later results here to those sections. We have included this section here to make it clear that the proofs of such things as the impossibility of angle trisection and of construction of certain other figures depends only on the elementary field theory that we've developed so far.

The main work will be to consider the set of all points which can be produced in finitely many steps starting from two initial points, where each

step is the use of a straight-edge to draw a line or of a compass to draw a circle. New points are produced as the intersections of these curves. We want to give an algebraic characterization of the coordinates of such 'constructible points'.

We'll assume that the reader has already seen in kindergarten geometry how to make certain basic constructions such as perpendicular bisectors, angle bisection, and getting the line parallel to a given line through a given point

All points, lines and circles below are in the plane $\mathbf{R}^2$. Starting from two points, define inductively sets of *constructible* points $\mathcal{P}$, lines $\mathcal{L}$, and circles $\mathcal{C}$, as follows.

$\mathcal{P} := \cup_{k\geq 0} \mathcal{P}_k \subset \mathbf{R}^2$; $\mathcal{L} := \cup_{k\geq 0} \mathcal{L}_k$; $\mathcal{C} := \cup_{k\geq 0} \mathcal{C}_k$; where :

$\mathcal{L}_0 :=$ empty set,

$\mathcal{L}_k := \mathcal{L}_{k-1} \cup [$ set of lines joining pairs of points in $\mathcal{P}_{k-1}$] ;

$\mathcal{C}_0 :=$ empty set,

$\mathcal{C}_k := \mathcal{C}_{k-1} \cup [$ circles centred in $\mathcal{P}_{k-1}$ and radius a distance between a pair of points in $\mathcal{P}_{k-1}$] ;

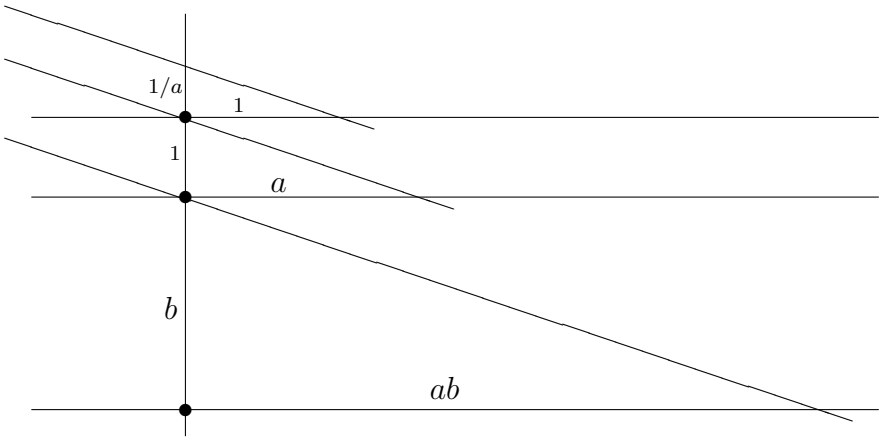
$\mathcal{P}_0 := \{ (0,0) , (1,0) \}$,

$\mathcal{P}_k := \mathcal{P}_{k-1} \cup$ [intersection points of lines/circles in $\mathcal{C}_k \cup \mathcal{L}_k$].

Let $Quad \subset \mathbf{R}$ be the set of all the coordinates of all the points in $\mathcal{P}$. It is easily seen that $\alpha \in Quad$ if and only if $|\alpha|$ is a distance between points in $\mathcal{P}$.

Proposition 27.1. *Quad is a subfield of* $\mathbf{R}$.

Proof. Closure under the four field operations follows in general quite easily from the case of positive reals. Given such positive a and b in $Quad$, ponder the following diagram to see closure under multiplication and inversion (addition and subtraction are very easy) :



Definition. Let F_k be the field generated by all the coordinates of all the points in $\mathcal{P}_k$.

Proposition 27.2. $[F_k : \mathbf{Q}]$ *is a power of 2.*

Proof. Since

$$[F_k : \mathbf{Q}] \quad = \quad [F_k : F_{k-1}][F_{k-1} : F_{k-2}] \cdots [F_2 : F_1][F_1 : \mathbf{Q}] \ ,$$

it suffices to show that each $[F_{i+1} : F_i]$ is a power of 2 (where $F_0 = \mathbf{Q}$ when $i = 0$). If $a_1,\ a_2,\ \cdots,\ a_r$ are the coordinates of all the points in $\mathcal{P}_{i+1}$, then $F_{i+1} = F_i(a_1, a_2, \cdots, a_r)$, so we think of getting to F_{i+1} from F_i by a tower as in **26.6**. But, by **26A**,

$$[F_i(a_1, a_2, \cdots, a_j) : F_i(a_1, a_2, \cdots, a_{j-1})] \ \leq \ [F_i(a_j) : F_i] \ ,$$

so it suffices to show that $[F_i(a_j) : F_i]$= 1 or 2 for all i and j. But this is clear, since either $a_j \in F_i$, or else a_j is a solution to one of : a pair of linear, a linear and a quadratic, or a pair of quadratic equations with coefficients in F_i, the latter case being reducible to a linear and quadratic. This follows from the kindergarten procedures for finding intersection points of lines and circles with one another.

Main Theorem 27.3. *If $\alpha \in Quad$, then the degree, $[\mathbf{Q}(\alpha) : \mathbf{Q}]$, is a power of 2.*

Proof. If $\alpha \in Quad = \bigcup_{k \geq 0} F_k$, then $\alpha \in F_k$ for some k. This gives the existence of a k with $\mathbf{Q}(\alpha) \subset F_k$. Hence $[\mathbf{Q}(\alpha) : \mathbf{Q}]$ divides $[F_k : \mathbf{Q}]$, which is a power of 2, as required.

Applications. 1. General angle trisection is impossible, since cosine$(\pi/9)$ has degree 3 over $\mathbf{Q}$; its minimal polynomial is $x^3 - \frac{3}{4}x - \frac{1}{8}$. [Use trigonometric identities, e.g. for $\cos(3\theta)$.]

2. Hence a regular 9-gon cannot be constructed. Gauss completely determined for which n the regular n-gon is constructible; see sections **33** and **39** later in this book.

3. 'Doubling the cube' is impossible, since $[\mathbf{Q}(\sqrt[3]{2}) : \mathbf{Q}] = 3$.

4. See the comment after **A7** below concerning the impossibility of 'squaring the circle'.

Better Theorem 27.4. *A real number α is in Quad if and only if there exist fields*

$$K_0 = \mathbf{Q} \ \subset \ K_1 \ \subset \ K_2 \ \subset \ \cdots \ \subset \ K_r = \mathbf{Q}(\alpha)$$

such that $[K_{i+1} : K_i] = 2$ for all i. But there exist α with $[\mathbf{Q}(\alpha) : \mathbf{Q}]$ equal to power of 2 and $\alpha \notin Quad$; for example, when α is equal to either real root of $x^4 - 2x - 2$.

Proof of the first statement. In the proof of **27.2**, we saw how to fit a tower of fields between F_{i+1} and F_i, in which each extension has degree 2. Given $\alpha \in Quad$, choose k with $\alpha \in F_k$ and juxtapose all these towers for $0 \le i < k$. Now intersect $\mathbf{Q}(\alpha)$ with each field in the last tower.

Exercise 27A. Show that each extension in this new tower has degree at most 2.

After removing any repeats, this produces a tower as required.

Conversely, given a tower as in the statement, we can see that α is in *Quad* by showing that if K is a subfield of *Quad*, then so is any degree 2 extension L of K which consists of real numbers.

Exercise 27B. i) Show that, except in characteristic 2, any field extension of degree 2 has the form $F(\sqrt{\beta}) \supset F$, i.e. the form $F(\alpha) \supset F$ where $\alpha^2 \in F$.

ii) Show that the exclusion of characteristic 2 is needed.

iii) Show that the analogue for degree 3 and $\sqrt[3]{\beta}$ is false.

To proceed, **27B**i) shows that L has the form $K(\sqrt{\beta})$, where $\beta \in K$ is positive. We need only show that $\sqrt{\beta} \in Quad$, which is clear for any positive $\beta \in Quad$ by drawing a circle of diameter $\beta+1$ (since the chord perpendicular to a diameter, through a point on the diameter of distance 1 from the circular

boundary, has length $2\sqrt{\beta}$).

Exercise 27C. Prove this last statement.

The proof of the second statement now amounts to showing that there are no fields at all which fit between $\mathbf{Q}$ and $\mathbf{Q}(\alpha)$, since the latter has degree 4 over the former, the polynomial in the statement being irreducible by Eisenstein's criterion (given at the end of Section **20**). This proof is done in detail at the end of Section **39**.

28. Roots and splitting fields.

Let F be any field.

Theorem 28.1. *If non-constant $g \in F[x]$, then there exists a finite extension K of F, with $[K : F] \leq \deg(g)$, such that g has a root in K.*

Proof. Let p be an irreducible factor of g. Let $K = F[x]/pF[x]$, where F is identified with the subfield $\{b + pF[x] \ : \ b \in F \ \} \subset K$. Then we have $[K : F] = \deg(p) \leq \deg(g)$. Let $a = x + pF[x]$. Then

$$p(a) \ = \ p(x) + pF[x] \ = \ 0 + pF[x] \ ,$$

so $p(a)$ is zero in K. Thus $g(a) = 0$ in K, and a is a root of g.

Note. Because of choice for p, the field containing a root of g is not unique. However, if g were irreducible, $\deg(g)$ divides $[K : F]$ by **26.5**, so $[K : F] = \deg(g)$ for any K as in **28.1.** Even stronger:

Theorem 28.2. *If $g \in F[x]$ is irreducible, and $\hat{a}, \tilde{a}$ are roots of g in extensions $\hat{K}, \tilde{K}$ of F, then there exists a unique isomorphism ψ from $F(\hat{a})$ to $F(\tilde{a})$ such that ψ maps all elements of F to themselves, and maps $\hat{a}$ to $\tilde{a}$.* (See also **28.5**.)

Proof. The polynomial g is (up to associates) the minimal polynomial of $\hat{a}$ over F, so by **24.2** there is a unique isomorphism

$$\hat{\psi} : F[x]/gF[x] \ \longrightarrow \ F(\hat{a})$$

with the properties $\hat{\psi}(x + gF[x]) = \hat{a}$ and $\hat{\psi}(b) = b$ for all $b \in F$. Similarly, there exists a unique

$$\tilde{\psi} : F[x]/gF[x] \ \longrightarrow \ F(\tilde{a})$$

(using **24.2**). Let $\psi = \tilde{\psi} \circ (\hat{\psi})^{-1}$. If ψ_1 and ψ_2 both had the properties of ψ, then $\psi_1 \circ \hat{\psi}$ and $\psi_2 \circ \hat{\psi}$ would both have the properties of $\tilde{\psi}$, so they agree; and therefore $\psi_1 = \psi_2$.

Theorem 28.3. *Let non-constant $g \in F[x]$. Then there is a finite extension K of F such that g is a product of linear polynomials in $K[x]$, i.e. 'K contains all the roots of g'.*

Proof. Proceed by induction on $\deg(g)$. If $\deg(g) = 1$, let $K = F$. For the inductive step, suppose that $\deg(g) > 1$. By **28.1** , there is a finite extension E of F in which g has a root a. Then $g = (x - a)g_1$ for some $g_1 \in E[x]$. By the inductive hypothesis, since $\deg(g_1)$ is less than $\deg(g)$, there exists a finite extension K of E such that g_1 is a product of linear polynomials in $K[x]$. Then K is a finite extension of F by **26.4** and g 'splits' in $K[x]$.

Definition. If K is an extension of F and $g \in F[x]$ is a non-constant, say *g splits over K* if and only if g is a product of linears in $K[x]$. The field K is a *splitting field for g over F* if and only if
(1) g splits over K, and
(2) g doesn't split over E if $F \subset E \subset K$ and $E \neq K$.

Theorem 28.4. *Any non-constant $g \in F[x]$ has a splitting field over F which is a finite extension of F.*

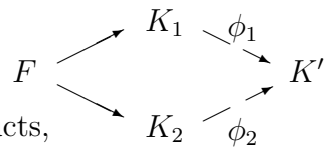
Proof. By **28.3**, construct a finite extension K' of F such that g splits over K'. Then $g = b(x - a_1)(x - a_2) \cdots (x - a_n)$, where b and each a_i are in K'. Now $b \in F$, since it is the leading coefficient of g. Thus, if $F \subset K'' \subset K'$, the polynomial g splits over K'' if and only if K'' contains a_i for all i, using unique factorization in $K'[x]$. Now let K either be the intersection of all such K'' or else be $F(a_1, \cdots, a_n)$; the two coincide by the usual arguments (see **6.3** and **16.1**) about generating an algebraic object either by combining its elements or by intersecting certain subobjects.

Remarks. This argument shows that *inside* any extension K' of F such that $g(x)$ splits in $K'[x]$, there is a *unique* splitting field K over F for $g(x)$. This uniqueness is strong in that K is not just unique up to isomorphism. But it is weak in that we must remain within a fixed extension K'. Below in **28.9**

we shall prove the very important fact that K is unique up to isomorphism, independently of any containing field K'. The principles **28.5** to **28.8** just below will play important roles later in Galois theory, as well as being useful here in proving the uniqueness of splitting fields.

Before giving the 'official uniqueness of splitting fields proof', here is a sketch of a different approach. Given two splitting fields $K_1 \supset F$ and $K_2 \supset F$ for $g(x) \in F[x]$, one can find a field K' and two field maps $\phi_i : K_i \to K'$ which agree on F:

(THE DIAGRAM "COMMUTES")



[For the cognoscenti who know tensor products, take $K' = (K_1 \otimes_F K_2)/I$ for a maximal ideal I. See the exercise below.]
Then both $\phi_1(K_1)$ and $\phi_2(K_2)$ are essentially splitting fields for $g(x)$ over F [modulo the fact that $F \to K'$ isn't, strictly speaking, an extension in the sense of our definition]. Both are subfields of K', so $\phi_1(K_1) = \phi_2(K_2)$ by the 'strong' uniqueness discussed above. Thus $\phi_2^{-1}\phi_1$ defines an isomorphism, as required, from K_1 to K_2.

Extended Exercise 28A. Learn about the tensor product, $\bigotimes$, e.g. in **Greub** or **Atiyah-Macdonald**—the basics are in Appendix $\bigotimes$ after Section **50**—and how $R = K_1 \bigotimes_F K_2$ can be made into a commutative ring (even an *F-algebra*—see Section **52**) into which K_1 and K_2 embed. Alternatively, check that the following messy version works: Pick F-bases $\mathcal{B}' = \{b'_1, b'_2, \cdots, b'_u\}$ and $\mathcal{B}'' = \{b''_1, b''_2, \cdots, b''_t\}$ for K_1 and K_2 respectively. Write

$$\left.\begin{array}{rcl} b'_p b'_q & = & \Sigma_{i=1}^{u} \;\; a'_i(p,q) b'_i \\ b''_r b''_s & = & \Sigma_{j=1}^{t} \; a''_j(r,s) b''_j \end{array}\right\} \quad \text{with } a'_i(p,q) \; , a''_j(r,s) \text{ in } F \; .$$

Define R to be the F-vector space of dimension ut with basis $\mathcal{B}' \times \mathcal{B}''$. Define the multiplication, say $*$, on R by

$$\left[\Sigma_{(p,r)} \quad c_{p,r}(b'_p, b''_r) \right] * \left[\Sigma_{(q,s)} \; d_{q,s}(b'_q, b''_s)\right]$$
$$:= \quad \Sigma_{(i,j)} \; \left[\Sigma_{(p,q,r,s)} \; a'_i(p,q) a''_j(r,s) c_{p,r} d_{q,s}\right] (b'_i, b''_j) \; .$$

The embeddings from K_1 and K_2 into R will be determined by F-linearity and by specifying $b'_i \mapsto (b'_i\ ,\ 1)$ and $b''_j \mapsto (1\ ,\ b''_j)$.

The first of the following extension principles is essentially a rehash of **28.2**.

Definition. Given a field map $\phi : F \to F^*$ and $g(x) = \Sigma\ a_i x^i$ in $F[x]$, *the polynomial* $g^*(x) \in F^*[x]$ *which corresponds to* $g(x)$ is defined to be $g^*(x) := \Sigma\phi(a_i)x^i$. We are actually dealing with a ring morphism $F[x] \to F^*[x]$, but wish to avoid excessive notation.

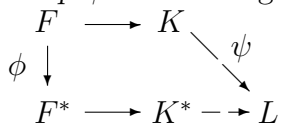
28.5. Simple Extension Principle. *Given $\phi : F \to F^*$, a field map, assume that $g(x)$ is irreducible in $F[x]$, and that $g^*(x) \in F^*[x]$ corresponds to $g(x)$, using ϕ. Let α and α^* be roots $g(x)$ and $g^*(x)$ respectively, in extensions of F and F^*. Then there is a unique field map $\theta : F(\alpha) \to F^*(\alpha^*)$ which agrees with ϕ on F, and such that $\theta(\alpha) = \alpha^*$:*

$$\begin{array}{ccc} F & \longrightarrow & F(\alpha) \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\theta} \\ F^* & \longrightarrow & F^*(\alpha^*) \end{array}$$

Proof. Let $g(x)$ have degree m, and define

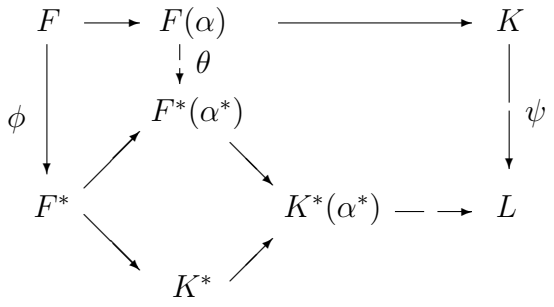
$\theta(\beta_0 + \beta_1\alpha + \cdots + \beta_{m-1}\alpha^{m-1}) := \phi(\beta_0) + \phi(\beta_1)\alpha^* + \cdots + \phi(\beta_{m-1})(\alpha^*)^{m-1}$
with $\beta_i \in F$; i.e. $\theta(f(\alpha)) = f^*(\alpha^*)$. The conditions on θ force this formula, proving uniqueness. Essentially the proof of **28.2** shows that θ is a morphism of rings: multiplication in $F(\alpha)$ [resp. $F^*(\alpha^*)$] is given by reducing modulo $g(\alpha)$ [resp. modulo $h(\alpha^*)$, for some irreducible factor $h(x)$ of $g^*(x)$]. (Possibly $h(x) = g^*(x)$.)

28.6. General Extension Principle. *Given a field map $\phi : F \to F^*$, an extension $K^* \supset F^*$, and a finite extension $K \supset F$, there exists a finite extension $L \supset K^*$ and a field map $\psi : K \to L$ agreeing with ϕ on F :*



Remark. This principle, rather than $\bigotimes$, could have been used in the above sketch alternative proof of the uniqueness of splitting fields.

Proof. If $[K : F] = 1$, then $K = F$, so let $L = K^*$ and $\psi(\beta) = \phi(\beta)$ for all β. Proceed by induction on $[K : F]$. Assume that $[K : F] > 1$. Choose any $\alpha \in K \setminus F$ and let $g(x)$ be its minimal polynomial over F. Let $g(x)$ correspond using ϕ to $g^*(x) \in F^*[x]$. Let α^* be a root of $g^*(x)$ in some extension of K^*. Consider the diagram



First define θ using the simple extension principle. Now

$$[K : F(\alpha)] \;=\; [K : F] \;/\; [F(\alpha) : F] \;<\; [K : F] \;,$$

so we may apply the inductive hypothesis to the upper right-hand part of the diagram to obtain L and ψ. Note that $L \supset K^*$ is finite, since both $L \supset K^*(\alpha^*)$ and $K^*(\alpha^*) \supset K^*$ are. That ψ agrees with ϕ on F follows easily, since the 'diamond' portion of the diagram commutes.

28.7. Splitting Extension Principle. *Given a field map $\phi : F \to F^*$, $g(x) \in F[x]$, a splitting field $K \supset F$ for $g(x)$, and a finite extension $K^* \supset F^*$ such that the corresponding $g^*(x)$ splits in $K^*[x]$, there exists a field map*
$\eta : K \to K^*$ agreeing with ϕ on F :

$$\begin{array}{ccc} F & \longrightarrow & K \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\eta} \\ F^* & \longrightarrow & K^* \end{array}$$

Proof. First apply the General E. P. using exactly its notation and diagram. If $\delta \in K$ and $g(\delta) = 0$, then $g^*(\psi(\delta)) = 0$ by an easy calculation. Thus $\psi(K) \subset K^*$, and we may take $\eta(\omega) = \psi(\omega)$ for all $\omega \in K$ (so η and ψ differ only with respect to their codomains).

Exercise 28B. Let $\phi : F \to F^*$ be an isomorphism of fields. It is intuitively clear that, since ϕ could be used to 'identify' F with F^*, results relating vector spaces over a single field have corresponding results relating vector spaces over the two fields. For example: We could call a function, say $\eta : V \to V^*$ from an F–vector space to an F^*–vector space, '*ϕ–linear*' if and only if it is a group morphism satisfying $\eta(fv) = \phi(f)\eta(v)$. Show that the existence of an injective such map implies that $\dim_F(V) \leq \dim_{F^*}(V^*)$. What would the condition be on η for the opposite inequality?

28.8. Splitting Field Uniqueness Principle. *Given an isomorphism $\phi : F \to F^*$ of fields, and splitting fields $K \supset F$ and $K^* \supset F^*$ respectively for $g(x) \in F[x]$ and $g^*(x) \in F^*[x]$ which correspond using ϕ, there exists an isomorphism $\eta : K \to K^*$ agreeing with ϕ on F.*
(Use the same diagram as in **28.7**.)

Proof. Pick any η using **28.7**. Since η is a field map, it is injective, yielding $[K : F] \leq [K^* : F^*]$ by **28B**. But our hypotheses are symmetric, so the reverse inequality holds, and so $[K : F] = [K^* : F^*]$. Thus η is surjective, as required, by the 'linear pigeon-hole principle'—a linear map between vector spaces of the same finite dimension is injective $\Leftrightarrow$ it is surjective—and similarly for a ϕ–linear map as in **28B**.

Corollary 28.9. Taking $F = F^*$ and ϕ to be the identity map, it follows that *a splitting field over F for $g(x) \in F[x]$ is unique up to an isomorphism fixing each element of F.*

Remarks. i) Surjectivity of η also follows from the observation that $\eta(K)$ is a splitting field for $g^*(x)$ over $\phi(F)$ in **28.7**.
ii) The next 50 pages or so of this book depend for their significance largely upon the fact that the isomorphism in **28.9** is almost never unique—when $K = K^*$, the set of all such automorphisms of K is called the Galois group (under composition).

Where to find your roots?

Many readers will already be aware that any polynomial in $\mathbf{C}[x]$ has all of its roots in $\mathbf{C}$; see the following appendix. Most of classical algebra is concerned with polynomial equations over $\mathbf{C}$ and its subfields. Therefore it is not an unreasonable tactic for the reader to imagine that most of the upcoming work (except for finite fields in Section **31**) takes place within $\mathbf{C}$, or even within subfields of $\mathbf{C}$ which are finite extensions of $\mathbf{Q}$. The following appendix also indicates how, replacing $\mathbf{Q}$ by an arbitrary 'base field' F, one may find a (usually non-finite) extension of F (namely, its *algebraic closure*) in which all the action can take place. However, the material of the present section—the existence and uniqueness of splitting fields—is all that is really needed. It really is needed, even when only working over $\mathbf{Q}$; and it is definitely more elementary than any proofs of either the 'fundamental theorem of algebra' or the existence of algebraic closures.

Appendix A. Algebraic closure and the fundamental theorem of (19^{th} century) algebra.

As indicated in the last paragraph, study of the material in this appendix is not essential for the understanding of subsequent sections, which only use it for a few examples. The student should however, read at least the statements of **A1** and **A2**.

Definition. A field F is *algebraically closed* if and only if any one, and therefore all, of the following hold.

Theorem A1. *Given a field F, the following are equivalent.*
(i) Every irreducible in $F[x]$ is linear.
(ii) Every non-constant in $F[x]$ has a root in F.

(iii) Every non-constant in $F[x]$ splits in $F[x]$.
(iv) If $K \supset F$ is an algebraic extension, then $K = F$.
(v) If $K \supset F$ is a finite extension, then $K = F$.
(vi) For any simple extension $F(\alpha) \supset F$, either $\alpha \in F$ or α is transcendental over F.

The proof, which is quite easy, will be left as an exercise.

Theorem A2. *The field* **C** *is algebraically closed.*

This is commonly known as the **fundamental theorem of algebra**, but that name should be qualified as in the appendix title. It was first proved (in several different ways) by Gauss. All proofs depend on some ideas of a topological nature. (Note that $\mathbf{Q}(\sqrt{-1}\,)$, for example, is definitely *not* algebraically closed). Most proofs also depend on one or another piece of mathematical machinery deeper than what we have so far given. For example, courses in one variable complex analysis often give a proof using the theory of complex line integrals. In Section **41** ahead is a proof in which the *machinery* is algebraic (Galois theory), and the 'topology' is reduced to the familiar facts that

1) an odd degree real polynomial has a real root, and

2) a positive real has a positive real square root.

At the end of this appendix is sketched the 'best' proof, in the sense that it uses a single, intuitively appealing topological idea which directly explains why the theorem is true.

Proposition A3. *If $E \supset F$ and $K \supset E$ are both algebraic extensions, then so is $K \supset F$.*

Proof. Let $\alpha \in K$ have minimal polynomial $\Sigma_0^n\, a_i x^i$ over E. Since each a_i is algebraic over F, the extension $L := F(a_0, a_1, \ldots, a_{n-1}) \supset F$ is finite. Also $L(\alpha) \supset L$ is finite. Thus $L(\alpha) \supset F$ is finite, and so α is algebraic over F.

Theorem A4. *Given an extension $K \supset F$, let*

$$L \;=\; \{\; \alpha \in K \;:\; \alpha \;\; \textit{is algebraic over} \;\; F \;\}\;.$$

Then
(i) *L is a subfield of K; and*
(ii) *if K is algebraically closed, then so is L.*

Proof. i) If α and β are in L, then $F(\alpha, \beta) \supset F$ is finite, so algebraic. But $F(\alpha, \beta)$ contains the elements $\alpha - \beta,\ \alpha\beta$, and α/β if $\beta \neq 0$. Thus each of them is in L.

ii) Let $g(x)$ be a non-constant in $L[x]$. It has a root α in K. Then $L \supset F$ and $L(\alpha) \supset L$ are algebraic. By **A3**, so is $L(\alpha) \supset F$. Since α is algebraic over F, the definition of L yields that $\alpha \in L$.

Corollary A5. *The set* **A** *of all algebraic numbers* (all complex numbers which are algebraic over **Q**) *is an algebraically closed field.*

Remark. Thus $\mathbf{A} \supset \mathbf{Q}$ is a non-finite algebraic extension. The straight-edge and compass constructible numbers $Quad \supset \mathbf{Q}$ gave us one earlier.

Theorem A6. *Let $K \supset F$ be algebraic. Then $|K| = |F|$ unless F is a finite field, in which case K is either finite or countable.*

Sketch Proof. Write $F[x] = \bigcup_{r=0}^{\infty} P_r$, where P_r consists of zero and all polynomials of degree at most r. Thus

$$|P_r| \ = \ |F^{r+1}| \ = \ |F|^{r+1} \quad (= |F| \ \text{ if } \quad F \ \text{ is infinite}) \ .$$

Each non-zero element of P_r has at most r roots in K. We can write $K = \bigcup_{r=0}^{\infty} K_r$, where K_r consists of those elements in K which are algebraic over F of degree at most r. By the above facts about $|P_r|$, we see that $|K_r| = |F|$ if F is infinite, and K_r is finite if F is finite. The proof is completed by noting that a countable union of sets with a fixed infinite cardinality has that same cardinality, and a countable union of finite sets is at most countable.

Corollary A7. (Cantor) *The field* **A** *is countable, and so the set,* $\mathbf{C} \setminus \mathbf{A}$*, of transcendental numbers is uncountable (in particular, non-empty!).*

Proving that special numbers (such as π and e) are transcendental is much harder than Cantor's indirect proof above of the existence of transcendental numbers. The fact that $\sqrt{\pi}$ is not algebraic immediately implies that 'the circle cannot be squared' by straight-edge and compass. This was one of the famous geometry problems from antiquity: *Construct a square of the same area as a given disc.*

Exercise AA. (i) Prove that if F is countable, then so is $F[x]$.
(ii) Deduce that for any $n > 0$, there exists a subset $\{\alpha_1, \ldots, \alpha_n\}$ of $\mathbf{C}$ which is algebraically independent over $\mathbf{Q}$.

ASIDE. Here is a simple method, due to Liouville, for proving that certain specially designed numbers are transcendental. This preceded Cantor's method by only a few decades. Gauss apparently had no proof of the existence of transcendentals—but one never knows !

Theorem A8. *If z is a real algebraic number of degree $d > 1$, then there is a number $M > 0$ such that $|z - \frac{p}{q}| \geq M/q^d$ for all integers p and q with $q > 0$.* The number M depends on z but not on p and q.
(Hence, in the sense just described, algebraic numbers are *harder* to approximate by rationals than are transcendental numbers!)

Proof. Multiplying to rid the minimal polynomial of denominators, we get a polynomial $f(x) = \Sigma_{i=0}^{d}\, a_i x^i$ with each $a_i \in \mathbf{Z}$, satisfying $f(z) = 0$ and $f(p/q) \neq 0$ for all $p/q \in \mathbf{Q}$. By factoring $(x-y)$ out of each $(x^i - y^i)$ we can write

$$f(x) - f(y) \;=\; (x-y)h(x,y)\ ,$$

where $h(x,y)$ is a polynomial of two variables. Since $h(z,y)$ is a continuous function of y, it is bounded on the interval $z-1 \leq y \leq z+1$; i.e. $\exists B > 0$ with $|h(z,y)| \leq B$ for all y such that $|z-y| \leq 1$. Now

$$|f(p/q)| \;=\; |q^{-d}(\Sigma a_i q^{d-i} p^i)| \;=\; |q^{-d} \cdot (\text{a nonzero integer})| \;\geq\; q^{-d}.$$

Thus, if $|z - \frac{p}{q}| \leq 1$, we have

$$q^{-d} \;\leq\; |f(p/q)| \;=\; |f(z) - f(p/q)| \;=\; |z - \frac{p}{q}|\ |h(z,p/q)| \;\leq\; |z - \frac{p}{q}| \cdot B\ .$$

Hence $|z - \frac{p}{q}| \geq 1/(Bq^d)$ whenever $|z - \frac{p}{q}| \leq 1$. Taking

$$M \;=\; \min\{\ 1\ ,\ 1/B\ \}$$

gives the required result.
The *result* isn't interesting if z isn't real, but does the *proof* depend on it being real? Is there a generalization to approximating by numbers with both real and imaginary parts being rational?

Application. Let $z = \Sigma_{n=1}^{\infty} 10^{-n!} = .1100010\cdots010\cdots010\cdots\cdots$.
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 6^{\underline{th}} \quad 24^{\underline{th}} \quad 120^{\underline{th}} \quad \cdots$

For a contradiction, assume that z is algebraic of degree d. Then $d > 1$ because $z \notin \mathbf{Q}$ (since the decimal expansion isn't periodic). Choose M as in the theorem. But now for a given s, consider the rational

$$\sum_{n=1}^{s} 10^{-n!} \;=\; p/10^{s!} \;=\; p/q \;.$$

We get

$$M/(10^{s!})^d \;\leq\; |z - \frac{p}{q}| \;<\; \frac{1}{10^{(s+1)!-1}} \;.$$

The last inequality is clear since $z - \frac{p}{q} = .00\cdots01\cdots$ (whatever)$\cdots$.
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (s+1)!^{\underline{th}}$ place

Thus $0 < M < 10^{-[(s+1)!-s!d-1]}$. Letting $s \to \infty$, we get $0 < M \leq 0$, a contradiction. Hence z is a transcendental number.

Exercise AB. Write down other transcendental numbers using this method. Write a 'formula' for uncountably many of them.

END OF ASIDE.

Theorem A9. *Let F be a field and let S be any subset of non-constants in $F[x]$. Then there is an extension $M \supset F$ such that all members of S split in $M[x]$.*

The proof is trivially reduced to the largest case, $S = F[x] \setminus F$, for which we sketch a proof at the end of this appendix.

Definition. Let $K \supset F$ be an extension, and let S be a subset of $F[x] \setminus F$. We say that K is a *splitting field for* $(F,\ S)$ if and only if **A9** holds for $M = K$, but fails for $M = E$ if $F \subset E \subset K$ and $E \neq K$.

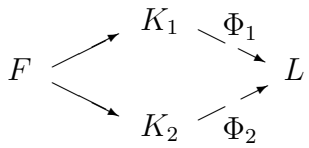
Exercise AC. Show that K is algebraically closed if and only if it is the splitting field for $(K,\ K[x] \setminus K)$.

Theorem A10. *Every pair $(F,\ S)$ has a splitting field K, which is unique up to an isomorphism fixing each element of F.*

Remark. This improves upon the last section only when S is infinite: if S is finite, a splitting field for (F, S) is easily seen to be the same as a

splitting field over F for the polynomial which is the product of all members of S.

Sketch Proof. Using **A9**, let $M \supset F$ be an extension such that all members of S split in $M[x]$. Let R be the set of all roots in M of all members of S. Clearly $F(R)$, the field generated by $F \cup R$, is a splitting field for (F, S) and is the *only one* inside M. The latter "bang-on uniqueness", and the fact that any diagram



can be completed, with a field L and field maps Φ_i , to be a commutative diagram, gives uniqueness: we have $K_i \cong \Phi_i(K_i)$, and $\Phi_1(K_1) = \Phi_2(K_2)$ by "bang-on uniqueness."

Remark. The last sentence is the same as the alternative proof of uniqueness for splitting fields in the previous section. A proof by extending maps is more awkward here, since it needs some sort of transfinite induction. Note however that the existence of L in the proof depends in general on the axiom of choice. For example, if L is constructed as $(K_1 \bigotimes_F K_2)/I$, that dependence is on the existence of the maximal ideal I.

Definition. Given an extension $K \supset F$, we say that K is 'the' *algebraic closure of* F if and only if any one, and so all four, of the following hold.

Theorem A11. *Given $K \supset F$, the following are equivalent.*
(i) K is algebraically closed and $K \supset F$ is algebraic.
(ii) K is algebraically closed, but L isn't for $F \subset L \subset K$ and $L \neq K$.
(iii) K is a splitting field for all of $F[x]$ over F (i.e. for $(F,\ F[x] \setminus F)$ in previous notation).
(iv) $K \supset F$ is algebraic, and, for any algebraic extension $E \supset F$, there exists a field morphism $\phi : E \to K$ which is the 'identity' on F.

Remarks and examples. (I) By *iii)* and **A10**, each F has an algebraic closure which is unique up to an isomorphism fixing elements of F.

(II) By *i)* and **A5**, the algebraic closure of $\mathbf{Q}$ is $\mathbf{A}$ (not $\mathbf{C}$!!).

(III) For each prime p, all the finite fields $\mathbf{F}_{p^n}$ in Section **31** ahead have the same algebraic closure: Any algebraic closure $\mathbf{K}_p$ for $\mathbf{F}_{p^n}$ is also an algebraic closure for its prime subfield $\mathbf{F}_p$, by *i)*, since $\mathbf{K}_p$ is an algebraic extension of $\mathbf{F}_p$ by **A3**. In fact $\mathbf{K}_p$ is simply the *union* of all the finite fields of characteristic

p, in the following sense. Firstly, for each n, the field $\mathbf{K}_p$ contains exactly one subfield of order p^n, the splitting field of $x^{p^n} - x$ over $\mathbf{F}_p$. The union of all these finite subfields is easily seen to be a subfield E of $\mathbf{K}_p$. To see that $E = \mathbf{K}_p$, note that any root in $\mathbf{K}_p$ of a polynomial in $\mathbf{F}_p[x]$ lies in some finite extension of $\mathbf{F}_p$, so in E.

Exercise AD. Find a proper infinite subfield of $\mathbf{K}_p$.

(IV) It is mildly surprising that building a minimal K over F so that all of $F[x]$ splits in $K[x]$ results in all of $K[x]$ splitting in $K[x]$.

Sketch Proof of A11. i) $\Rightarrow$ ii): Choose $\alpha \in K \setminus L$ with minimal polynomial $g(x)$ over F. Then $g(x)$ does not split in $L[x]$.

ii) $\Rightarrow$ i): By **A4**, the set L, of those elements of K which are algebraic over F, is an algebraically closed field, so $L = K$.

i) $\Rightarrow$ iii): Any $g(x)$ in $F[x]$ splits in $K[x]$ since any $g(x)$ in $K[x]$ does. Suppose that $F \subset E \subset K$ and $E \neq K$. Let $\alpha \in K \setminus E$. Since $K \supset F$ is algebraic, α is algebraic over F. Let $g(x)$ be its minimal polynomial over F. Then $g(x)$ does not split in $E[x]$.

iii) $\Rightarrow$ i): Certainly $K \supset F$ is algebraic. Suppose given a simple algebraic extension $K(\alpha) \supset K$. By **A3**, $K(\alpha) \supset F$ is algebraic. Thus α is algebraic over F, so $\alpha \in K$, as required.

Exercise AE. Prove the equivalence of *(iv)* with the remaining conditions in the theorem.

Outline of a deduction of A9 from Zorn's lemma. Choose any set T containing F such that T is infinite and $|T| > |F|$. Consider the collection $\mathcal{C}$ of all fields K such that K is an algebraic extension of F and, as a set, K is a subset of T. Then $\mathcal{C}$ is a set. By the cardinality result **A6** and by **A3**, if $K \in \mathcal{C}$ and $L \supset K$ is an algebraic extension, one can construct $L' \supset K$ with $L' \in \mathcal{C}$ such that $L' \cong L$ by an isomorphism fixing elements of K. Given K and K' in $\mathcal{C}$, define $K \leq K'$ to mean that K' is an extension of K (*more than just $K \subset K'$ as sets*). Suppose that $\mathcal{L}$ is a non-empty subset of $\mathcal{C}$ which is linearly ordered by the relation $\leq$. Then $\cup\mathcal{L}$ is easily seen to be in $\mathcal{C}$. By Zorn's lemma (see **Artin**, p.588), $\mathcal{C}$ has a maximal member M. Suppose for a contradiction that non-constant $g(x) \in F[x]$ does not split in $M[x]$. Choose an algebraic extension $L \supset M$ such that $g(x)$ splits in $L[x]$. We may assume that $L \in \mathcal{C}$ by the earlier remark. This contradicts the maximality of M.

Outline of the 'homotopy' proof of A2. Dividing by the leading coefficient, and noting that zero is a root if the bottom coefficient is zero, it remains to show that any polynomial

$$g(x) \;=\; x^n + a_{n-1}x^{n-1} + \cdots + a_0 \;\in\; \mathbf{C}[x]$$

has a root in $\mathbf{C}$ when $a_0 \neq 0$ and $n > 0$. For a contradiction, suppose that it doesn't. Then for each real $r > 0$, we get a continuous function

$$\omega_r \;:\; S^1 \;\longrightarrow\; \mathbf{C} \setminus \{0\}$$

by $\omega_r(z) := g(rz)$, where S^1 is the unit circle $\{z : |z| = 1\}$. Values are taken in the *punctured* plane $\mathbf{C} \setminus \{0\}$ by our assumption—this is crucial here! Picture each ω_r as a parameterized closed loop; that is, as a particle moving continuously in $\mathbf{C} \setminus \{0\}$, as the parameter z starts from 1 and moves around the circle S^1. By varying r continuously on the positive line from any r_0 to any r_1, it is intuitively clear that ω_{r_0} and ω_{r_1} are *homotopic* : i.e. they can be continuously deformed from one to the other (without passing through the origin in $\mathbf{C}$!) For r_0 very close to zero, it is clear from continuity that $\omega_{r_0}(z)$ stays close to a_0 for all z, and so the loop ω_{r_0} does not 'wrap around' the origin at all; in fact, it can be continuously deformed to the *constant loop* which just stays at a_0. But for r_1 sufficiently large, we shall show that ω_{r_1} wraps around the origin "n" times in a counterclockwise direction; it could be deformed to the loop $z \mapsto (r_1 z)^n$ which travels n times around the circle of radius r_1^n at constant speed. This is the contradiction to the fact that ω_{r_0} is homotopic to ω_{r_1} which proves the theorem. (It is also the point which requires the most work when setting up the machinery to make this into a complete proof.)

To prove the claim about ω_{r_1}, it suffices to show that, for all $z \;\in\; S^1$,

$$|\omega_{r_1}(z) - (r_1 z)^n| \;\leq\; r_1^n/2$$

by choosing r_1 suitably: For then, $\omega_{r_1}(z)$ is always within the 'moving disc' of radius $r_1^n/2$ centered at $(r_1 z)^n$. So the 'ω_{r_1}-particle' gets dragged around the origin "n" times, being trapped in the moving disc, however crazily it moves around within that disc. In fact, a homotopy converting ω_{r_1} into $z \mapsto (r_1 z)^n$ is definable by continuously shrinking the moving disc to its center. You should think of the particle as a dog attached to a leash of length $r_1^n/2$, whose

master walks "n" times around the origin, along the circle of radius r_1^n centred at the origin.

To prove the required inequality, if $r_1 \geq 1$, then

$$\begin{aligned} |\omega_{r_1}(z) - (r_1 z)^n| &= |a_{n-1}(r_1 z)^{n-1} + a_{n-2}(r_1 z)^{n-2} + \cdots + a_0| \\ &\leq |a_{n-1}(r_1 z)^{n-1}| + \cdots\cdots\cdots + |a_0| \\ &= |a_{n-1}| r_1^{n-1} + |a_{n-2}| r_1^{n-2} + \cdots + |a_0| \\ &\leq nM r_1^{n-1} \end{aligned}$$

for $M = \max\{\, |a_i| : 0 \leq i < n \,\}$. But if we choose $r_1 \geq 2Mn$, then

$$Mn r_1^{n-1} \leq r_1^n/2 \ ,$$

as required. Thus the job is done by choosing

$$r_1 := \max\{\, 1 \ , \ 2n|a_{n-1}| \ , \ \cdots, \ 2n|a_0| \,\} \ .$$

29. Repeated roots and the formal derivative.

Definition. For any field F, define $D : F[x] \to F[x]$ by

$$D(\sum_{i=0}^{n} a_i x^i) = \sum_{i=0}^{n-1} (i+1) a_{i+1} x^i \ .$$

The operator D is called the *formal derivative* for obvious reasons.

Exercise 29A. Prove the identities
i) $D(f+g) = D(f) + D(g)$;
ii) $D(fg) = fD(g) + D(f)g$;
iii) $D(f^n) = nf^{n-1}D(f)$.

Proposition 29.1. *In $F[x]$, we have that $(x-a)^2$ divides f if and only if $(x-a)$ divides both f and $D(f)$.*

Proof. $\Rightarrow$: If $f = (x-a)^2 g$, then

$$D(f) \;=\; (x-a)[2g + (x-a)D(g)] \; .$$

$\Leftarrow$: If $f = (x-a)h$, then $D(f) = (x-a)D(h) + h$. Since $(x-a)$ divides $D(f)$, we have $(x-a)$ dividing h. Thus $(x-a)^2$ divides f.

Definition. a is a *repeated root* of f if and only if $(x-a)^2$ divides f.

Theorem 29.2. *Let $f \in F[x]$. Then f has no repeated root in any extension of F if and only if* $\mathrm{GCD}\{f, D(f)\} = 1$ *in $F[x]$.*
(Note that one of the conditions involves arbitrary extensions of F, whereas the other takes place entirely within $F[x]$.)

Corollary 29.3. *When the characteristic is 0, an irreducible in $F[x]$ can have no repeated root in any extension of F.*

More generally, an irreducible g could have a repeated root only if Dg were 0, since otherwise, because $\deg(Dg) < \deg(g)$, we get that the required GCD is 1. In the case of characteristic zero, Dg is certainly non-zero (and has degree one less than that of g) since

$$D(x^n) \;=\; nx^{n-1} \;\neq\; 0 \qquad \text{for all } n > 0 \; .$$

Proof of 29.2: $\Leftarrow$: There exist s, t in $F[x]$ with

$$sf + tD(f) = 1 \; .$$

Since $(x-a)$ does not divide 1, it cannot divide both f and $D(f)$ in any extension $K[x] \supset F[x]$ for any $a \in K$. Thus a is not a repeated root, i.e. not a root of order greater than 1.

$\Rightarrow$: If $\mathrm{GCD}\{f, D(f)\} = g$, and g is not a non-zero constant, then F has an extension K in which g has a root a. Then $(x-a)$ divides both f and Df in $K[x]$, so a is a repeated root of f.

Example. $x^t - 1$ has no repeated roots in K if t is prime to $\mathrm{ch}(K)$, or if $\mathrm{ch}(K) = 0$.

30. Finite subgroups of $F^\times$.

Recall that $F^\times$ is the group $F \setminus \{0\}$ under multiplication.

Theorem 30.1. *If F is any field, then $F^\times$ has exactly one subgroup of order t for each t for which $x^t - 1$ splits into distinct linear factors in $F[x]$, and no other finite subgroups. For each such t, this subgroup is cyclic.*

Proof. If G is a subgroup of order t, clearly each of its elements is a root of $x^t - 1$. Thus G consists of the " t " distinct roots, so G is the *only* subgroup of order t, and $x^t - 1 = \prod_{a \in G}(x - a)$. To show that G is cyclic, suppose that

$$G \quad \cong \quad C_{t_1} \times C_{t_2} \times \cdots \times C_{t_r}$$

(using **13.2**), where $1 < t_1 \mid t_2 \mid \cdots \mid t_r$, and $t = \prod_1^r t_i$. Then $a^{t_r} = 1$ for every $a \in G$, so $x^{t_r} - 1$ has " t " roots in G. Hence $t_r \geq t$, which implies that $t_r = t,\ r = 1$, and G is cyclic.

Definition. The generators of such a group G are called *primitive* t^{th} roots of unity. They are those t^{th} roots of unity which are *not* s^{th} roots of unity for any $s < t$. For each such t, there are exactly " $\Phi(t)$ " primitive t^{th} roots of unity, where Φ is Euler's function.

Remark. A special case is that, for a finite field F, the group $F^\times$ is cyclic. For $\mathbf{Z}_p$, we already used this to start the inductive calculation of the group $\mathbf{Z}_{p^n}^\times$ in Section **15**. On the other hand, this observation is the crucial one which we use in the next section to obtain the classification of finite fields.

31. The structure of finite fields.

Proposition 31.1. *A finite field F has " p^n " elements for some prime p and some $n \geq 1$.*

Proof. Let P denote the prime subfield of F. Then $[F : P] = n$, for some $n < \infty$. Thus $F \cong P^n$ as a P-vector space. But P has " p " elements where $p = \text{ch}(F)$, so $|F| = |P^n| = p^n$.

Lemma 31.2. *Let F have characteristic p and prime subfield P. Then F has " p^n " elements if and only if F is the splitting field of $x^{p^n} - x$ over P.*

Proof. $\Rightarrow$: If $a \in F^\times$, then $a^{p^n-1} = 1$, since $F^\times$ is a group of order $p^n - 1$. Hence $a^{p^n} = a$ for *all* $a \in F$, so F contains " p^n " roots of $x^{p^n} - x$, and is the splitting field.

$\Leftarrow$: Let R be the set of roots in F of $x^{p^n} - x$. Since

$$(a \pm b)^{p^n} \;=\; a^{p^n} \pm b^{p^n}$$

in any field of characteristic p, the set R is closed under addition and subtraction. Closure under multiplication and division is trivial, so R is a subfield. Hence $R = F$. Since

$$\mathrm{GCD}\{\; x^{p^n} - x \;,\; D(x^{p^n} - x) \;\} \;=\; \mathrm{GCD}\{\; x^{p^n} - x \;,\; -1 \;\} \;=\; 1 \;,$$

the polynomial $x^{p^n} - x$ has no repeated roots, so $|F| = p^n$.

Theorem 31.3. *Given $(p,\ n)$, there is exactly one field of order p^n up to isomorphism, namely the splitting field, $\mathbf{F}_{p^n}$, of $x^{p^n} - x$ over $\mathbf{Z}_p$.*

Proof. The field $\mathbf{F}_{p^n}$ has " p^n " elements by the lemma. If K is any field with " p^n " elements and prime field P, then any isomorphism $P \to \mathbf{Z}_p$ extends to an isomorphism $K \to \mathbf{F}_{p^n}$ by **28.8**, the uniqueness of splitting fields, since K is the splitting field of $x^{p^n} - x$ over P by the lemma.

Note. The group $\mathbf{F}_{p^n}^\times$ is cyclic of order $p^n - 1$ by the previous section, and has primitive t^{th} roots of unity for the divisors t of $p^n - 1$, and only for these t. We have

$$x^{p^n} - x \;=\; \prod_{a \in \mathbf{F}_{p^n}} (x - a) \;.$$

Theorem 31.4. *For all $n \geq 1$, the ring $\mathbf{Z}_p[x]$ contains irreducible polynomials f of degree n, and for any such f,*

$$(\; \mathbf{Z}_p[x] \;/\; f\mathbf{Z}_p[x] \;) \;\cong\; \mathbf{F}_{p^n} \;.$$

Any extension in the realm of finite fields is simple.

Proof. Let θ be any generator of the cyclic group $\mathbf{F}_{p^n}^\times$. Then we have $\mathbf{F}_{p^n} = \mathbf{Z}_p(\theta)$ since the powers of θ fill out $\mathbf{F}_{p^n}^\times$. So θ has degree n over $\mathbf{Z}_p$, and its minimal polynomial is irreducible of degree n. Now $\mathbf{Z}_p[x]/f\mathbf{Z}_p[x]$ is a field with " p^n " elements, if f is irreducible of degree n, so it is isomorphic

to $\mathbf{F}_{p^n}$. We have observed that $\mathbf{F}_{p^n}$ is a simple extension of $\mathbf{F}_p$. Thus any finite field is a simple extension of its prime subfield, and therefore of each of its subfields.

Note. This shows that $\mathbf{Z}[x]$ and $\mathbf{Q}[x]$ have plenty of irreducibles of every degree (as does Eisenstein's criterion in **20A** more directly).

Exercise 31A. Show that $\mathbf{F}_{p^n}$ has a subfield of order p^r if and only if r divides n.

Exercise 31B. Show that $\mathbf{F}_{p^n}$ does not have two subfields of the same order.

Exercise 31C. Show that

$$x^{p^n} - x \;=\; \prod_{r|n} \prod g(x) \quad ,$$

where the inside product is over the set of all monic irreducibles $g(x)$ of degree r in $\mathbf{Z}_p[x]$. After reading the next section, deduce that the cardinality of the latter set is

$$r^{-1} \sum_{s|r} \mu(s) p^{r/s} \ .$$

Exercise 31D. Is every function from $\mathbf{F}_{p^n}$ to itself necessarily a polynomial function (cf. **16G**)?

32. Moebius inversion.

Summations in this section are over all *positive* divisors of some given positive integer. For positive integers k, define $\mu(k)$ inductively by

$$\mu(1) \;=\; 1 \qquad ; \qquad \sum_{r|k} \mu(r) \;=\; 0 \qquad \text{if } k > 1 \ .$$

Exercise 32A. Show that μ is given explicitly by

$$\mu(p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}) \;=\; \begin{cases} (-1)^t & \text{if all } \ \alpha_i = 1 \ ; \\ 0 & \text{otherwise} \ ; \end{cases}$$

where the p_i's are distinct primes and each $\alpha_i > 0$.

Theorem 32.1. (Moebius Inversion Formula) *If $\{a_n\}_{n\geq 1}$ is a sequence from an abelian group, then*

$$b_k \;=\; \sum_{r|k} a_r \quad\Longrightarrow\quad a_k \;=\; \sum_{r|k} \mu(k/r) b_r \;.$$

Proof. We have

$$\sum_{r|k} \mu(k/r) b_r \;=\; \sum_{r|k} [\mu(k/r) \sum_{s|r} a_s] \;=\; \sum_{s|k} [a_s \sum_{s|r|k} \mu(k/r)]$$

(letting $\ell = k/r$)

$$= \sum_{s|k} [\; a_s \sum_{\ell \;|\; k/s} \mu(\ell) \;] \;=\; a_k \cdot 1 \;+ \sum_{s|k \;,\; s<k} a_s \cdot 0 \;=\; a_k \;.$$

Example. The identity, $k = \sum_{r|k} \Phi(r)$, can be obtained by counting the elements of C_k : each element generates the subgroup C_r for some unique r dividing k, and C_r has " $\Phi(r)$ " distinct choices of generator. By Moebius inversion,

$$\Phi(k) \;=\; \sum_{r|k} \mu(k/r) r \;.$$

Exercise 32B. Use this to reprove that

$$\Phi(\prod_i p_i^{\alpha_i}) \;=\; \prod_i (p_i^{\alpha_i} - p_i^{\alpha_i - 1})$$

33. Cyclotomic polynomials.

Definition. Suppose that F is an extension of $\mathbf{Q}$ which contains primitive k^{th} roots of unity. Define the k^{th} *cyclotomic polynomial* by the formula $c_k(x) := \prod(x - \zeta)$, product over the " $\Phi(k)$ " distinct primitive k^{th} roots of unity ζ. The polynomial c_k is monic of degree $\Phi(k)$. Since the roots of $x^k - 1$ are the primitive r^{th} roots of unity for all r dividing k, we get

$$x^k - 1 \;=\; \prod_{r|k} c_r(x) \qquad (*)_k$$

Using $(*)_k$ and induction on k, we see that:

(i) $(*)_k$ determines c_k ;

(ii) $c_k(x) \in \mathbf{Z}[x]$ for all k, since we have $\{f$ primitive in $\mathbf{Z}[x]$ and $fg \in \mathbf{Z}[x]\} \Rightarrow g \in \mathbf{Z}[x]$;

(iii) $c_k(x)$ does not depend on choice of F. (We could take F to be $\mathbf{C}$, or to be any splitting field for $x^k - 1$, since $x^k - 1$ has no repeated roots.)

Applying Moebius inversion to $(*)_k$ in the commutative group $[\mathbf{Z}(x)]^*$ under multiplication, we get

$$c_k(x) \;=\; \prod_{r|k} (x^r - 1)^{\mu(k/r)}$$

in $\mathbf{Z}(x)$.

Theorem 33.1. (Gauss) *The polynomial $c_k(x)$ is irreducible in $\mathbf{Q}[x]$. For primitive k^{th} roots of unity ζ, the extension $\mathbf{Q}(\zeta)$ of $\mathbf{Q}$ has degree $\Phi(k)$. The minimal polynomial over $\mathbf{Q}$ of any primitive k^{th} root of unity is $c_k(x)$.*

Proof. The three statements are clearly equivalent.

To prove that $c_k(x)$ is irreducible, it suffices to show that all primitive k^{th} roots of unity have the *same* minimal polynomial over $\mathbf{Q}$, since this polynomial divides $c_k(x)$, and would have degree at least $\Phi(k)$, so it would coincide with $c_k(x)$. If ζ is any primitive k^{th} root of unity, then any other such root can be written as $\zeta^{p_1 p_2 \cdots p_r}$ for (not necessarily distinct) primes p_i which don't divide k. Hence we need only show that ζ and ζ^p have the same minimal polynomial over $\mathbf{Q}$, if p is a prime which does not divide k. Choose multiples $f(x)$ and $g(x)$ of the minimal polynomials of ζ and ζ^p respectively, such that f and g are primitive in $\mathbf{Z}[x]$. Since ζ is a root of $g(x^p)$, it is

clear that $f(x)$ divides $g(x^p)$ in $\mathbf{Q}[x]$. Now $f(x)$ is primitive in $\mathbf{Z}[x]$, so $f(x)$ divides $g(x^p)$ in $\mathbf{Z}[x]$.

For a contradiction, suppose that $f(x)$ and $g(x)$ are *not* associates in $\mathbf{Q}[x]$. Then $f(x)$ and $g(x)$ are relatively prime in $\mathbf{Q}[x]$ and both divide (x^k-1), so $f(x)g(x)$ divides (x^k-1) in $\mathbf{Z}[x]$. Let $\bar{f}(x)$ and $\bar{g}(x)$ in $\mathbf{Z}_p[x]$ be f and g with coefficients reduced mod p. Then $\bar{g}(x^p) = \bar{g}(x)^p$, so $\bar{f}(x)$ divides $\bar{g}(x)^p$ in $\mathbf{Z}_p[x]$. Thus $\bar{f}(x)$ and $\bar{g}(x)$ have a common root θ in some extension of $\mathbf{Z}_p$. But $\bar{f}(x)\bar{g}(x)$ divides (x^k-1) in $\mathbf{Z}_p[x]$, and so θ is a repeated root of (x^k-1). This is a contradiction, since GCD$\{\ x^k-1\ ,\ kx^{k-1}\ \} = 1$ in $\mathbf{Z}_p[x]$ when p doesn't divide k.

Hence $f(x)$ and $g(x)$ *are* associates in $\mathbf{Q}[x]$, so that ζ and ζ^p have the same minimal polynomial.

Now we can prove most of Gauss' theorem, alluded to earlier in Section **27**, determining which regular n-gons are constructible by straight-edge and compass. The proof below uses only the weaker version of **33.1** in which we assume that k is a prime. See Exercise **33B** below for an easier proof of that weaker version.

Theorem 33.2. (Gauss) *The regular n-gon is constructible if and only if n factors as $2^\alpha p_1 p_2 \cdots p_t$, where the p_i are distinct primes of the form $2^{2^\beta}+1$.*

First half of the Proof. (*See Section* **39** *for the second half.*) It is easily seen that the set $\mathcal{P}$ of constructible points, thought of as a subset of $\mathbf{C}$, is in fact the field $Quad(\sqrt{-1})$. The main theorem on *Quad* extends immediately to imply that elements of $\mathcal{P}$ are necessarily algebraic of degree equal to a power of 2. It is also clear that the regular n-gon is constructible if and only if $e^{2\pi i/n} \in \mathcal{P}$.

But if p is an odd prime dividing n, then $p-1$ divides $\Phi(n)$; and if p^2 divides n, then p divides $\Phi(n)$. Therefore, if $\Phi(n)$ is a power of 2, then the only odd primes which can divide n must be of the form $2^\gamma+1$; and the squares of these primes cannot divide n. Primes of this form are called *Fermat primes*. It is elementary to show that they must have the form $2^{2^\beta}+1$. (Use the factorization of $x^{odd}+1$.) This proves the theorem in one direction.

In the second half of the proof, we'll show that $e^{2\pi i/p} \in \mathcal{P}$ for each Fermat prime p. By angle bisection (or by a simple, purely algebraic argument), we see that $e^{2\pi i/2^\alpha} \in \mathcal{P}$ for all $\alpha > 0$. Thus it remains only to show that

$e^{2\pi i/ab} \in \mathcal{P}$ whenever $e^{2\pi i/a} \in \mathcal{P}$ and $e^{2\pi i/b} \in \mathcal{P}$ with GCD$\{a,\ b\} = 1$. But this is immediate from the fact that $\mathcal{P}$ is a field, and that $sa + tb = 1$ for some integers s and t—write

$$e^{2\pi i/ab} \ = \ (e^{2\pi i/a})^t (e^{2\pi i/b})^s \ .$$

Exercise 33A. Show that $0 \leq \beta \leq 4$ do give primes above, but $\beta = 5$ doesn't. (No other values of β are known which give primes, and many are known which don't.)

Exercise 33B. Show that for primes p, the cyclotomic polynomial $c_p(x)$ is equal to

$$\sum_0^{p-1} x^i \ = \ (x^p - 1)/(x - 1) \ .$$

Deduce that $c_p(x + 1)$ has all but the top coefficient divisible by p, and the bottom one not divisible by p^2. Prove that an integer polynomial with these divisibility properties is irreducible in $\mathbf{Z}[x]$—**Eisenstein's Criterion**, given also at the end of **20**. (Argue as in the proof of Gauss' Lemma **20.1**.) Conclude that $c_p(x)$ is irreducible in $\mathbf{Q}[x]$. (This is somewhat simpler than the given proof of **33.1**, don't you think?)

34. Primitive elements exist in characteristic zero.

A simple extension may be written $F(\gamma) \supset F$ for various choices of γ, each of which is called a *primitive element* for the extension. For example, $\sqrt{2} + \sqrt{3}$ is a primitive element for $\mathbf{Q}(\sqrt{2},\ \sqrt{3}) \supset \mathbf{Q}$: We have

$$(\sqrt{2} + \sqrt{3})^3 - 9(\sqrt{2} + \sqrt{3}) \ = \ 2\sqrt{2}$$

yielding $\sqrt{2} \in \mathbf{Q}(\sqrt{2} + \sqrt{3})$. Thus

$$\sqrt{3} \ = \ (\sqrt{2} + \sqrt{3}) - \sqrt{2} \ \in \ \mathbf{Q}(\sqrt{2} + \sqrt{3}),$$

giving $\mathbf{Q}(\sqrt{2},\ \sqrt{3}) \subset \mathbf{Q}(\sqrt{2} + \sqrt{3})$, the opposite inclusion being obvious.

Theorem 34.1. *In characteristic zero, any finite extension $K \supset F$ is simple.*

Remarks. At first glance the theorem seems both surprising and perhaps also not especially useful. [For most purposes, the notation $\mathbf{Q}(\sqrt{2},\ \sqrt{3})$ says more about what it denotes than does $\mathbf{Q}(\sqrt{2}+\sqrt{3})$.] But later we'll use the theorem several times in crucial situations. It becomes less surprising several sections hence when, as a byproduct of the fundamental theorem of Galois theory, we see the (perhaps even more surprising) fact that any finite extension $K \supset F$ in characteristic zero admits only *finitely many* intermediate fields E; that is, fields with $F \subset E \subset K$. (When $[K : F] > 2$, there are certainly infinitely many F-subspaces V with $F \subset V \subset K$.) Thus with infinitely many $\gamma \in K \setminus F$ (when $K \neq F$), it seems reasonable that each of the intermediate fields, including K itself, should be $F(\gamma)$ for infinitely many γ.

First, here is an example to show that the assumption of characteristic zero is not made merely because we can't think of a proof without it. The example necessarily involves infinite fields of characteristic p since we already know that any extension involving finite fields is simple. Let $K = \mathbf{F}_p(\alpha, \beta)$ where $\{\alpha, \beta\}$ is algebraically independent over $\mathbf{F}_p$, and let $F = \mathbf{F}_p(\alpha^p, \beta^p)$. Then $[K : F] = p^2$, an F-basis for K being

$$\{\ \alpha^i\beta^j\ :\ 0 \le i, j < p\ \}\ .$$

(Check this!) All $\gamma \in K$ have the form

$$\gamma\ =\ \sum u_{ij}\alpha^i\beta^j \Big/ \sum v_{ij}\alpha^i\beta^j \quad ;\quad u_{ij}\ ,\ v_{ij}\ \in\ \mathbf{F}_p\ .$$

Since $(s+t)^p\ =\ s^p + t^p$ in characteristic p,

$$\gamma^p\ =\ \sum u_{ij}\alpha^{pi}\beta^{pj} \Big/ \sum v_{ij}\alpha^{pi}\beta^{pj}\ \in\ F\ .$$

(We used that $u^p = u$ in $\mathbf{F}_p$, but this is irrelevant; any field of characteristic p could replace $\mathbf{F}_p$ in this example.) Thus γ is algebraic of degree at most p over F (in fact, of degree p or 1), and so $K \neq F(\gamma)$ for any γ.

Exercise 34A. In this example, show that there are infinitely many E with $F \subset E \subset K$. HINT: Consider $F(\alpha + \theta\beta)$ for $\theta \in F$.

Challenge. Is there a logical connection between simplicity and finiteness of the set of intermediate fields? See Appendix **B**.

Proof of 34.1. By **26.6**, $K = F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ for some $\alpha_i \in K$. It therefore suffices to prove the theorem for finite extensions of the form $F(\alpha, \beta) \supset F$. [Then taking $(\alpha, \beta) = (\alpha_{n-1}, \alpha_n)$ and changing F to $F(\alpha_1, \cdots, \alpha_{n-2})$ yields $K = F(\alpha_1, \ldots, \alpha_{n-2}, \gamma)$ for some γ; now take $(\alpha, \beta) = (\alpha_{n-2}, \gamma)$ to decrease the number of generators to $n-2$; etc.]

It suffices to find $\lambda \in F$ such that, if $\gamma = \alpha + \lambda\beta$, then $\beta \in F(\gamma)$. [For then $\alpha = \gamma - \lambda\beta \in F(\gamma)$, so $F(\alpha, \beta) \subset F(\gamma)$; and the reverse inclusion is obvious.]

Let α and β have minimal polynomials $a(x)$ and $b(x)$, respectively, over F. Choose any $\lambda \in F$ which disagrees with $(\beta - \tilde{\beta})^{-1}(\tilde{\alpha} - \alpha)$, for all roots $\tilde{\alpha}$ of $a(x)$, and all roots $\tilde{\beta} \neq \beta$ of $b(x)$ [in some splitting field for $a(x)b(x)$ over F]. The choice is possible, since F is infinite. Let

$$h(x) \;=\; a(\gamma - \lambda x) \;\in\; F(\gamma)[x] \;.$$

Then

$$h(\beta) \;=\; a(\gamma - \lambda\beta) \;=\; a(\alpha) \;=\; 0 \;,$$

and, for all $\tilde{\beta}$ as above,

$$h(\tilde{\beta}) \;=\; a(\gamma - \lambda\tilde{\beta}) \;\neq\; 0 \;,$$

since

$$\gamma - \lambda\tilde{\beta} \;=\; \alpha + \lambda(\beta - \tilde{\beta}) \;\neq\; \tilde{\alpha}$$

for any root $\tilde{\alpha}$ of $a(x)$. Thus $h(x)$ and $b(x)$ have β as a common root, but no other common roots in any extension of $F(\gamma)$. The minimal polynomial, $m(x)$, of β over $F(\gamma)$ therefore divides both $h(x)$ and $b(x)$. Being irreducible, $m(x)$ has distinct roots, since the characteristic is zero. These roots will all be common to $h(x)$ and $b(x)$, and so $m(x)$ has only one root [in fact $m(x) = x - \beta$], and $\beta \in F(\gamma)$, as required.

IV. Galois Theory

The main theoretical content here is in sections **35** and **38**. In the former, we introduce the Galois group, and derive enough theory to give a group theoretic condition which is *necessary* for a polynomial equation in one variable to be solvable by radicals and field operations: namely, the condition that this Galois group of the splitting extension for the polynomial is soluble. This is used in the following two sections: firstly to prove that, when dealing with polynomials of degree greater than 4, no formula of that form can exist for any base field; and then to exhibit a specific rational polynomial whose roots can't be expressed in terms of its coefficients using only $+$, $-$, $\times$, $\div$ and $\sqrt[n]{\ }$. Then in **38** we give the 'complete story', a 1-1 correspondence between intermediate fields of a given extension and subgroups of the corresponding Galois group. Mostly we stick to the case of characteristic zero, but in such a way that tacking on the extra subtleties for general characteristic goes quite smoothly and efficiently. (See Appendix **B**.) Other applications of the Galois correspondence include completing the proof concerning the constructibility of regular n-gons, and giving another proof of the fundamental theorem of (19^{th} century) algebra. In Appendix **C**, it's shown that solubility is also *sufficient* for solvability.

35. The Galois group.

Galois' big idea was, in essence, to introduce groups into field theory, as follows.

Definition. For each field extension $K \supset F$, let $\mathrm{Aut}_F(K)$ denote the set of those field isomorphisms $\theta : K \to K$ for which $\theta(a) = a$ for all $a \in F$. Such a θ is called an *automorphism of K fixing elements of F*.

Note. When F is the prime subfield, the condition $\theta(a) = a$ is easily deducible, since $\theta(1) = 1$.

Exercise 35A. Prove that the condition $\theta(a) = a$ is equivalent to F-linearity of θ.

Proposition 35.1. *Using composition as operation, the set $Aut_F K$ becomes a group.*

Proof. This is a routine verification. Check that $\mathrm{Aut}_F K$ contains all three of : the inverse of such a θ, the composition of two such θ, and the identity map of K.

Proposition 35.2. *If $h(x) \in F[x]$ and $\theta \in Aut_F K$, then θ maps to itself the set of those roots of $h(x)$ which happen to be in K.*

Proof. If $h(x) = \Sigma a_i x^i$, then for any $b \in K$,

$$\theta(h(b)) \;=\; \sum \theta(a_i b^i) \;=\; \sum \theta(a_i)\theta(b)^i \;=\; h(\theta(b)) \;,$$

since $\theta(a_i) = a_i$. Thus $h(b) = 0$ implies that $h(\theta(b)) = 0$.

Proposition 35.3. *Each element of $Aut_F(F(b_1, \ldots, b_k))$ is uniquely determined by its values on $b_1, \ldots, b_k$; i.e. if $\theta(b_i) = \tilde{\theta}(b_i)$ for all i, then the elements θ and $\tilde{\theta}$ are equal.*

Proof. For any 'scalars' $a_I = a_{i_1,\cdots,i_k} \in F$, almost all zero, let

$$c \;=\; \sum_I a_I b_1^{i_1} \cdots b_k^{i_k} \;.$$

Then

$$\theta(c) \;=\; \sum_I \; a_I \theta(b_1)^{i_1} \cdots \theta(b_k)^{i_k} \;=\; \sum_I \; a_I \tilde{\theta}(b_1)^{i_1} \cdots \tilde{\theta}(b_k)^{i_k} \;=\; \tilde{\theta}(c) \;.$$

Now every element of $F(b_1, \ldots, b_k)$ can be written as $c^{-1}d$ for elements c and d as above. Then

$$\theta(c^{-1}d) \;=\; \theta(c)^{-1}\theta(d) \;=\; \tilde{\theta}(c)^{-1}\tilde{\theta}(d) \;=\; \tilde{\theta}(c^{-1}d) \;.$$

Thus $\theta = \tilde{\theta}$.

Remark. It is immediate from the last two results that $Aut_F(K)$ is a *finite* group when the extension $K \supset F$ is finite—for K is then generated over F by finitely many algebraic elements, and there are only finitely many possible elements to be images of each of them, by **35.2**, so **35.3** allows only finitely many possibilities for automorphisms.

Definition. For each $g(x) \in F[x]$, the *Galois group of* $g(x)$ *over* F is defined to be the group $\mathrm{Aut}_F(K)$ for the splitting field K of $g(x)$ over F. This is a finite group.

Exercise 35B. Show that the Galois group of $g(x)$ over F is independent of the choice of splitting field. More precisely, any choice of an isomorphism between two splitting extensions induces in a natural way an isomorphism between the corresponding two automorphism groups.

The following is now immediate from **35.2** and **35.3**, since the splitting field is generated over F by the roots of $g(x)$.

Theorem 35.4. *Each element of the Galois group of* $g(x)$ *restricts to a self-bijection of the set of roots of* $g(x)$*, and is completely determined by this permutation of the roots.*

Thus the Galois group becomes identified with a subgroup of S_k, the symmetric group, once a list, $b_1, \ldots, b_k$, of all distinct roots of $g(x)$ is given. (So $k \leq \deg g$, possibly strictly when $g(x)$ is reducible or when the characteristic is not zero—recall that an irreducible polynomial in characteristic zero has no repeated roots).

Example. Let $F = \mathbf{Q}$, $g(x) = x^3 - 2$ and $\omega = \mathrm{e}^{2\pi i/3}$. Then the splitting field of $x^3 - 2$ over $\mathbf{Q}$ is $K = \mathbf{Q}(\omega \ , \ \sqrt[3]{2})$. Furthermore, $[K : \mathbf{Q}] = 6$, a $\mathbf{Q}$-basis for K being

$$\{ \ 1 \ , \ \omega \ , \ \sqrt[3]{2} \ , \ \sqrt[3]{2}\omega \ , \ \sqrt[3]{4} \ , \ \sqrt[3]{4}\omega \ \} \ .$$

Complex conjugation fixes 1 , $\sqrt[3]{2}$ and $\sqrt[3]{4}$, and, in the other three basis elements, replaces ω by $\omega^2 = -1 - \omega$. So it maps K to itself. This gives an element of $\mathrm{Aut}_{\mathbf{Q}}(K)$ which acts as a 2-cycle

$$\{ \ \sqrt[3]{2}\omega \leftrightarrow \sqrt[3]{2}\omega^2 \text{ and } \sqrt[3]{2} \hookleftarrow \ \}$$

on the set of roots of $x^3 - 2$. To show that $\mathrm{Aut}_{\mathbf{Q}}(K)$ is as large as **35.4** allows (i.e. is isomorphic to S_3), it remains only to see that the 2-cycle $\{\sqrt[3]{2} \leftrightarrow \sqrt[3]{2}\omega \text{ and } \sqrt[3]{2}\omega^2 \hookleftarrow \ \}$ can be realized as the restriction of an automorphism θ of K (because S_3 is generated, for example, by $\{ \ (12), \ (23) \ \}$).

Such a θ would have to fix 1 (of course) and to interchange ω and ω^2, since

$$\theta(\omega) = \theta\left((\sqrt[3]{2})^{-1}(\sqrt[3]{2}\omega)\right) = \theta(\sqrt[3]{2}\,)^{-1}\ \theta(\sqrt[3]{2}\omega) = (\sqrt[3]{2}\omega)^{-1}(\sqrt[3]{2}) = \omega^2\ .$$

Since $\omega^2 = -\omega - 1$ and $\sqrt[3]{4}$ maps to $\sqrt[3]{4}\omega^2$, it follows that θ is the $\mathbf{Q}$-linear bijection

$$a_0 + a_1\omega + (b_0 + b_1\omega)(\sqrt[3]{2}) + (c_0 + c_1\omega)(\sqrt[3]{4}) \ \mapsto$$
$$[(a_0 - a_1) - a_1\omega] + [b_1 + b_0\omega](\sqrt[3]{2}) + [-c_0 + (c_1 - c_0)\omega)](\sqrt[3]{4})\ .$$

A slightly tedious calculation shows that this map is indeed 'homomorphic' with respect to multiplication. Alternatively S_3 is also generated by any set { 2-cycle, 3-cycle }. So instead of worrying about θ, one may realize a 3-cycle as the restriction of an automorphism ϕ as follows. Note that $x^3 - 2$ is irreducible also in $\mathbf{Q}(\omega)[x]$. Now take ϕ to be the unique $\mathbf{Q}(\omega)$-linear isomorphism

$$\mathbf{Q}(\omega)\ (\sqrt[3]{2}) \ \longrightarrow\ \mathbf{Q}(\omega)(\sqrt[3]{2}\omega)$$

mapping $\sqrt[3]{2}$ to $\sqrt[3]{2}\omega$, as given by the simple extension principle.

Exercise 35C. Carry out all of the details for the existence of θ and ϕ immediately above.

This example shows that the Galois group is in general not abelian. A famous problem, still unsolved in 1994, asks whether *every* finite group occurs (up to isomorphism) as the Galois group over $\mathbf{Q}$ of at least one $g(x) \in \mathbf{Q}[x]$.

Every finite *abelian* group does occur (see **38D**III). Quite a few of them occur as follows:
The Galois group of $x^n - 1$ *over* $\mathbf{Q}$ *is isomorphic to* $\mathbf{Z}_n^\times$*, the group, of order* $\Phi(n)$*, of invertibles in the ring* $\mathbf{Z}_n$ (see **35D** below).
The structure of $\mathbf{Z}_n^\times$ as a product of finite cyclic groups was determined earlier in Section **15**.

To see how both (abelian) groups of order 4 occur, here are details for the cases $n = 5$ and $n = 12$. We have $\Phi(5) = 4 = \Phi(12)$.

Let $n = 5$ and $\omega = \mathrm{e}^{2\pi i/5}$, so that $K = \mathbf{Q}(\omega)$ has basis $\{\ 1,\ \omega,\ \omega^2,\ \omega^3\ \}$ over $\mathbf{Q}$. By **35.3**, each element of $\mathrm{Aut}_{\mathbf{Q}}(\mathbf{Q}(\omega))$ is determined by its value on ω. This value lies in $\{\ \omega,\ \omega^2,\ \omega^3,\ \omega^4\ \}$, which is the set of roots of $1 + x +$

$x^2 + x^3 + x^4 = c_5(x)$, of which K is also the splitting field. Thus the Galois group here has at most four elements. If there were a field automorphism θ with $\theta(\omega) = \omega^2$, then

$$\begin{array}{lllllll} \theta(\omega^2) & = & (\theta(\omega))^2 & = & (\omega^2)^2 & = & \omega^4 \quad = \quad -1-\omega-\omega^2-\omega^3 \ ; \\ \theta(\omega^3) & = & (\omega^2)^3 & = & \omega^6 & = & \omega \ ; \\ \theta(\omega^4) & = & (\omega^2)^4 & = & \omega^3 \ . & & \end{array}$$

As in the last example, the values of $\theta(\omega^i)$ for $0 \le i \le 3$ determine a $\mathbf{Q}$-linear map which may be checked directly to be in $\mathrm{Aut}_{\mathbf{Q}}(K)$. In this instance, a much simpler argument for the existence of θ is that, by the simple extension principle, it is the unique isomorphism from $\mathbf{Q}(\omega)$ to $\mathbf{Q}(\omega^2)$ [which happens to equal $\mathbf{Q}(\omega)$], mapping ω to ω^2, using the fact that, over $\mathbf{Q}$, the numbers ω and ω^2 have the same minimal polynomial, $c_5(x)$. The above calculations show that, on the set of roots of $c_5(x)$, the restriction of θ is the 4-cycle

$$(\ \omega \ \mapsto \ \omega^2 \ \mapsto \ \omega^4 \ \mapsto \ \omega^3 \ \mapsto \ \omega \ ) \ .$$

And so the Galois group is $\{ \ e, \ \theta, \ \theta^2, \ \theta^3 \ \}$, which is cyclic of order four. It is therefore isomorphic to $\mathbf{Z}_5^\times$, as was claimed. It follows that θ^3 acts as another 4-cycle, whereas θ^2 acts as a product of two 2-cycles $\{\omega \ \leftrightarrow \omega^4 \ , \ \omega^2 \ \leftrightarrow \ \omega^3\}$.

Now let $n = 12$ and $\omega = \mathrm{e}^{2\pi i/12}$, so that $K = \mathbf{Q}(\omega)$ has $\mathbf{Q}$-basis $\{ \ 1, \ \omega, \ \omega^2, \ \omega^3 \ \}$. This time the Galois group (of both $c_{12}(x) = x^4 - x^2 + 1$ and $x^{12} - 1$) should be thought of as a permutation group on the set $\{ \ \omega, \ \omega^5, \ \omega^7, \ \omega^{11} \ \}$ of all the roots of $c_{12}(x)$. It has order four again, with elements determined by

$$\phi(\omega) \ = \ \omega^5 \quad ; \quad \psi(\omega) \ = \ \omega^7 \quad ; \quad (\phi \circ \psi)(\omega) \ = \ \phi(\omega^7) \ = \ (\omega^5)^7 \ = \ \omega^{11} \ .$$

This time each of $\phi, \ \psi$ and $\phi \circ \psi$ acts as a product of two disjoint transpositions: for example, ϕ is $\{\omega \leftrightarrow \omega^5 \ ; \ \omega^7 \leftrightarrow \omega^{11}\}$. Thus we have

$$\phi^2 \ = \ \psi^2 \ = \ e \qquad ; \qquad \phi \circ \psi \ = \ \psi \circ \phi \ ,$$

and

$$\mathrm{Aut}_{\mathbf{Q}}(K) \ = \ \{ \ e, \ \phi, \ \psi, \ \phi \circ \psi \ \} \ \cong \ C_2 \times C_2 \ \cong \ \mathbf{Z}_{12}^\times \ .$$

Exercise 35D. Let F be any field and let ω be a primitive n^{th} root of unity in some extension field. Define a map

$$\Gamma \ : \ \mathrm{Aut}_F(F(\omega)) \ \longrightarrow \ \mathbf{Z}_n^\times$$

by $\Gamma(\theta) = [k]_n$ when $\theta(\omega) = \omega^k$. Using the routine established in the above examples, prove that Γ is well-defined; a morphism of groups; and injective. When $F = \mathbf{Q}$, use Gauss' theorem **33.1** on the irreducibility of $c_n(x)$ in $\mathbf{Q}[x]$ to prove that Γ is also surjective.

Doing this simple exercise establishes the previous claim, and also the following, whose proof we include since the result is crucial later.

Proposition 35.5. *For any F and any root, ω, of unity in an extension field, the group $Aut_F(F(\omega))$ is abelian; i.e. the Galois groups over all fields of all the polynomials $x^n - 1$ are abelian.*

Proof. For two elements ϕ and ψ, let $\phi(\omega) = \omega^k$ and $\psi(\omega) = \omega^\ell$. Then

$$(\phi \circ \psi)(\omega) \;=\; \phi(\omega^\ell) \;=\; \phi(\omega)^\ell \;=\; (\omega^k)^\ell \;=\; (\omega^\ell)^k \;=\; (\psi \circ \phi)(\omega)\ .$$

Since group elements here are determined by their effects on ω, we have $\phi \circ \psi = \psi \circ \phi$, as required.

The reader may have noticed that, in all of the examples so far, we have $|\mathrm{Aut}_F(K)| = [K : F]$. This is always true when K is a splitting field over F. It can be helpful when trying to calculate a Galois group (usually not an easy task), but we'll delay its proof to the section after next. *It is not true for general K.* For example $|\mathrm{Aut}_{\mathbf{Q}}(\mathbf{Q}(\sqrt[3]{2}\ ))| = 1$, since any group element must map $\sqrt[3]{2}$ to itself [neither of the other roots of $x^3 - 2$ being in $\mathbf{Q}(\sqrt[3]{2})$], and since group elements here are determined by their effects on $\sqrt[3]{2}$.

Another family of examples where abelian Galois groups occur is for $x^n - \lambda$ over E, where E contains both λ and a primitive n^{th} root, ω, of unity. The splitting field is $E(\mu)$ for any μ with $\mu^n = \lambda$, since the list of roots of $x^n - \lambda$ is then $\mu,\ \omega\mu,\ \omega^2\mu,\ \cdots,\ \omega^{n-1}\mu$.

Proposition 35.6. *With the above data, the group $Aut_E(E(\mu))$ is abelian.*

Proof. Any element ϕ of the Galois group permutes the above roots, and is determined by its effect on μ since $\phi(\omega^k\mu) = \omega^k\phi(\mu)$ (or, more simply,

since μ generates $E(\mu)$ over E). For elements ϕ and ψ, let $\phi(\mu) = \omega^i\mu$ and $\psi(\mu) = \omega^j\mu$. Then

$$(\phi \circ \psi)(\mu) \;=\; \phi(\omega)^j\phi(\mu) \;=\; \omega^j\omega^i\mu \;=\; \omega^i\omega^j\mu \;=\; (\psi \circ \phi)(\mu)\ ,$$

so $\phi \circ \psi = \psi \circ \phi$, as required.

Exercise 35E. By considering the least $i > 0$ for which there exists ϕ with $\phi(\mu) = \omega^i\mu$, show that the Galois group in **35.6** is actually cyclic.

Note how the first example (the splitting field of $x^3 - 2$ over $\mathbf{Q}$) is built up in two steps corresponding to the two previous propositions:

$$F \;=\; \mathbf{Q} \quad ; \quad \omega \;=\; \mathrm{e}^{2\pi i/3} \quad ; \quad E \;=\; \mathbf{Q}(\omega) \quad ; \quad \mu \;=\; \sqrt[3]{2}\ ;$$

$$\mathbf{Q} \;\subset\; \mathbf{Q}(\omega) \;\subset\; \mathbf{Q}(\omega,\ \mu) = K\ .$$

This tower yields a subnormal series with abelian quotients,

$$S_3 \;\rhd\; A_3 \;\rhd\; \{e\},$$

when $L \mapsto \mathrm{Aut}_L(K)$ is applied.

Let us return to the theory to generalize this, giving a minimal route to understanding the famous assertion of Ruffini, first proved in all detail by Abel: *equations of degree greater than four cannot in general be solved using only 'radicals'* (n^{th} roots) *and field operations.*

Proposition 35.7. *A tower*

$$F_0 \;\subset\; F_1 \;\subset\; F_2 \;\subset\; \cdots \;\subset\; F_n$$

of field extensions yields a decreasing sequence

$$Aut_{F_0}(F_n) \;\supset\; Aut_{F_1}(F_n) \;\supset\; \cdots \;\supset\; Aut_{F_n}(F_n) = \{e\}$$

of groups.

Proof. This has been stated in the form in which it usually arises, but the content is just the case $n = 2$: if $F_0 \subset F_1 \subset F_2$, then $\mathrm{Aut}_{F_1}(F_2)$ is a subgroup of $\mathrm{Aut}_{F_0}(F_2)$, and $\mathrm{Aut}_{F_2}(F_2)$ is the trivial group. These are obvious from the definitions.

N.B. From here onwards, except where the contrary is mentioned explicitly, all fields will have **CHARACTERISTIC ZERO** and all **EXTENSIONS ARE FINITE**.

The proof of the next theorem shows that the existence of primitive elements can be useful. This theorem would definitely need to be modified if we were allowing infinite extensions and/or arbitrary characteristic. It singles out a class of extensions named because of the connection with normal subgroups given in **35.9***ii)a)* ahead.

Definition. A *normal extension* is one for which any one (and therefore all) of the conditions in the following theorem hold.

Theorem 35.8. *Given $K \supset F$, a finite extension in characteristic zero, the following conditions are equivalent.*

i) For all $\alpha \in K \setminus F$, there exists $\theta \in Aut_F(K)$ with $\theta(\alpha) \neq \alpha$.

ii) For each irreducible $g(x) \in F[x]$, either all the roots of $g(x)$ are in K, or none of them are.

iii) For all α such that $K = F(\alpha)$, the splitting field over F of the minimal polynomial of α over F is K.

iv) The field K is the splitting field over F of at least one polynomial $g(x) \in F[x]$.

v) If $F \subset K \subset L$, then all F-linear field maps $\phi : K \to L$ satisfy $\phi(K) = K$.

Remarks. a) It follows from *v)* that for all $\psi \in \mathrm{Aut}_F(L)$, we have $\psi(K) = K$. Does this imply *v)*?

b) In *v)*, it is not important whether $L \supset K$ is restricted to finite extensions or not: the unrestricted *v)* will be deduced, and the restricted *v)* will

be assumed, in the following proof.

Proof. *i) ⇒ ii)*: Assume that $g(x)$ has a root in K and let $\alpha_1, \cdots, \alpha_r$ be all such (distinct) roots which are in K. Let

$$h(x) \;=\; (x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_r),$$

so that $h(x)$ divides $g(x)$ in $K[x]$. It suffices to prove that $h(x) \in F[x]$, since irreducibility there of $g(x)$ then implies that $h(x)$ and $g(x)$ are associates, as required. (Recall that $g(x) = h(x)k(x)$ with $g(x)$ and $h(x)$ in $F[x]$ implies that $k(x) \in F[x]$.) But the coefficients, $\pm e_i(\alpha_1, \cdots, \alpha_r)$, of $h(x)$, where e_i is the i^{th} elementary symmetric polynomial (see **22**), are fixed by all $\theta \in Aut_F(K)$, since θ permutes $\{\alpha, \cdots, \alpha_r\}$. By *i)*, these coefficients must be in F, as required.

ii) ⇒iii): If $g(x)$ denotes the minimal polynomial of α over F, then by *ii)*, all of its roots are in K, which is generated over F by one root, and so K is the splitting field of $g(x)$.

iii) ⇒iv): This is immediate from the existence of primitive elements, proved in the previous section.

iv) ⇒v): The set of roots in K of any $h(x) \in F[x]$ is mapped by ϕ to itself, by the argument proving **35.2**. Now choose any $h(x)$ such that K is its splitting field. Then ϕ maps a set of generators over F for K to itself, and therefore maps all of K to itself.

v) ⇒ ii) ⇒i): The proof of **35.8** is concluded by two arguments using the isomorphism extension properties.

Assuming *v)*, suppose given an irreducible $g(x)$ in $F[x]$ with a root, α, in K. Let $\beta \in M$ be any root of $g(x)$ in any extension M of K. We must prove that $\beta \in K$. Construct a commutative diagram where maps ϕ_1 and ϕ send α to β :

$$\begin{array}{ccccc} F & \longrightarrow & F(\alpha) & \longrightarrow & K \\ {\scriptstyle id}\downarrow & & \downarrow{\scriptstyle \phi_1} & & \searrow^{\phi} \\ F & \longrightarrow & F(\beta) & \longrightarrow & M \dashrightarrow L \end{array}$$

The simple extension principle, **28.5**, gives ϕ_1 uniquely. The general extension principle, **28.6**, gives a finite extension $L \supset M$ and a field map ϕ. Since $K \subset M$, we have $K \subset L$, and thus $\phi(K) = K$ by *v)*. Since $\alpha \in K$ and $\phi(\alpha) = \beta$, we get $\beta \in K$, as required.

Assuming *ii)* and given $\alpha \in K \setminus F$, let $g(x) \in F[x]$ be the minimal polynomial of α over F. Choose another root $\beta \neq \alpha$. This can be done since $\alpha \notin F$, so the degree of $g(x)$ is not 1 , and its roots are distinct by **29.4** because the characteristic is zero. By *ii)*, the element β is also in K. Now construct a commutative ladder of isomorphisms mapping α to β :

$$\begin{array}{ccccc} F & \longrightarrow & F(\alpha) & \longrightarrow & K \\ {\scriptstyle id}\downarrow & & \downarrow{\scriptstyle \theta_1} & & \downarrow{\scriptstyle \theta} \\ F & \longrightarrow & F(\beta) & \longrightarrow & K \end{array}$$

This clearly suffices to prove *i)*, since $\theta(\alpha) = \beta \neq \alpha$. The existence of θ_1 is clear from the simple extension principle, **28.5**. As for θ, note that since *ii)* implies *iv)*, K is a splitting field for some $h(x) \in F[x]$ over F. Therefore it is a splitting field for the same polynomial $h(x)$ over $F(\alpha)$ and also over $F(\beta)$. So the existence of θ is immediate from **28.8**, the splitting field uniqueness principle.

Exercise 35F. Prove that $K \supset F$ is a normal extension if and only if there exists an extension $L \supset K$ for which *v)* holds with $L \supset F$ normal.

Remark. This completes the proof of **35.8**. But here is a proof that K in the last paragraph is a splitting field [i.e. a proof that *ii)* implies *iv)*] which doesn't use the existence of primitive elements. Write K as $F(\alpha_1, \cdots, \alpha_\ell)$, and let $h_i(x)$ be the minimal polynomial of α_i over F. Let $h(x) = h_1(x) \cdots h_\ell(x)$. By *ii)*, K contains all roots of each $h_i(x)$, and so, of $h(x)$. It is generated over F by (a subset of) these roots, completing the proof.

This, and examination of the proof of **35.8**, shows that, without the assumption of characteristic zero, the conditions in **35.8** satisfy

$$i) \implies ii) \Leftrightarrow iv) \Leftrightarrow v) \implies iii) .$$

In arbitrary characteristic, the words *normal extension* usually mean *ii), iv), v)*, whereas *Galois extension* refers to the stronger condition *i)*. See the remarks beginning Section **38**. Now let's return to the world of characteristic zero.

Remark. It is an unfortunate fact that when $E \supset F$ and $K \supset E$ are both normal, it may happen that $K \supset F$ is not normal.

Exercise 35G. Show that $\mathbf{Q} \subset \mathbf{Q}(\sqrt{2}) \subset \mathbf{Q}(\sqrt[4]{2})$ is an example illustrating the remark above, first showing that any extension of degree 2 is normal.

Exercise 35H. The above example shows that there must be something wrong with the following 'proof' that
$\{E \supset F$ and $K \supset E$ both normal implies that $K \supset F$ is normal$\}$
Find the error :
Let $\alpha \in K \setminus F$. By **35.8***i)*, we need only find $\theta \in Aut_F(K)$ with $\theta(\alpha) \neq \alpha$. If $\alpha \in K \setminus E$, then **35.8***i)* yields $\theta \in Aut_E(K)$ with $\theta(\alpha) \neq \alpha$, as required, since $K \supset E$ is normal. If $\alpha \in E \setminus F$, choose $\phi \in Aut_F(E)$ with $\phi(\alpha) \neq \alpha$. Since K is a splitting field over E [by **35.8***iv)*], use the uniqueness of splitting fields to extend ϕ to an automorphism θ from K to itself, as required !!?

Proposition 35.9. *Given a 'short' tower $F \subset E \subset K$ of finite extensions in characteristic zero, we have the following.*

i) If $K \supset F$ is normal, then so is $K \supset E$.

ii) If $K \supset F$ and $E \supset F$ are normal, then

a) $Aut_E(K)$ is a normal subgroup of $Aut_F(K)$, and

b) restricting defines an isomorphism of groups

$$Aut_F(K)/Aut_E(K) \longrightarrow Aut_F(E) \ .$$

Proof. *i)* This is immediate from either *iv)* or *v)* of **35.8**.

ii) If $\theta \in Aut_F(K)$, then θ maps E to E by **35.8***v)*, and so restriction defines a function

$$Aut_F(K) \longrightarrow Aut_F(E) \ ,$$

which is evidently a morphism of groups. Its kernel is by definition $Aut_E(K)$, proving *a)*. The function is surjective, as required for *b)*, since **28.8** allows any $\phi \in Aut_F(E)$ to be extended to an automorphism of K [which is a splitting field over E by part *i)*].

Recall from Section **11** that a finite group G is *soluble* if and only if there exists a subnormal series

$$G = G_0 \ \rhd \ G_1 \ \rhd \ G_2 \ \rhd \ \cdots \ G_{\ell-1} \ \rhd \ G_\ell = \{e\}$$

such that each G_i/G_{i+1} is cyclic. But it suffices to replace *cyclic* by *abelian* as noted in **13F**, using the structure theorem for finite abelian groups.

It is now easy to see that, *for all F of characteristic zero, all $\lambda \in F$ and all $n \geq 1$, the Galois group over F of $x^n - \lambda$ is soluble :* its splitting field K tops a tower

$$F \ \subset \ F(\omega) \ \subset \ F(\omega, \ \mu) = K \ ,$$

where $\mu^n = \lambda$ and ω is a primitive n^{th} root of unity. Now $F(\omega)$ is the splitting field over F of $x^n - 1$, so **35.9***ii)* applies, giving

$$Aut_F(K) \ \rhd \ Aut_{F(\omega)}(K) \ \rhd \ \{e\}$$

and

$$Aut_F(K)/Aut_{F(\omega)}(K) \ \cong \ Aut_F(F(\omega)) \ .$$

The latter group and $Aut_{F(\omega)}(K)$ are both abelian, by **35.5** and **35.6** respectively, and so $Aut_F(K)$ is soluble (of 'length' ≤ 2), as required.

We wish to generalize this to equations which can be solved using only various n^{th} roots and the field operations. Before proceeding, here is a small technicality needed later.

Definition. Given an (algebraic, but we're assuming finite, remember?) extension $K \supset F$ and elements $\alpha, \ \beta \in K$, we say that α and β are *conjugate over F* if and only if they have the same minimal polynomial over F.

Lemma 35.10. *i) With notation as above, if α and β are conjugates, then so are α^n and β^n for each $n \geq 1$.*

ii) Given $F \subset E \subset K$, with α and β in K being conjugate over F, suppose that $E \supset F$ is normal and $\alpha^n \in E$ for some $n \geq 1$. Then $\beta^n \in E$.

Proof. *i)* Let $f(x)$ be the minimal polynomial of α (or β) over F. Let $g(x)$ be the minimal polynomial of α^n over F, and let $h(x) = g(x^n)$. Then $h(\alpha) = g(\alpha^n) = 0$, so $f(x)$ divides $h(x)$ in $F[x]$. Thus

$$g(\beta^n) = h(\beta) = 0 \ ,$$

and so $g(x)$ is also the minimal polynomial of β^n, as required.

ii) Since $E \supset F$ is normal, and $g(x)$ has root $\alpha^n \in E$, we see that β^n is also in E by **35.8***ii)*.

Definition. Let F_0 be a field of characteristic zero. A polynomial $g(x)$ in $F_0[x]$ is said to be *solvable by radicals over* F_0 if and only if there is a tower of field extensions

$$F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_m ,$$

for some m, such that

i) $g(x)$ splits in F_m—equivalently, the splitting field of $g(x)$ over F_0 is a subfield of F_m ;

ii) for each $i > 0$, we have $F_i = F_{i-1}(\mu_i)$ for some $\mu_i \in F_i$ such that $\mu_i^{n_i} \in F_{i-1}$ for some $n_i > 0$—equivalently, each F_i is generated over F_{i-1} by a root of $x^n - \lambda$, where $n > 0$ and $\lambda \in F_{i-1}$ both depend on i.

Another formulation of this definition : $F_0(\mu_1, \cdots, \mu_m)$ contains a splitting field over F_0 for $g(x)$, for some elements μ_i , where μ_i has a positive power in $F_0(\mu_1, \cdots, \mu_{i-1})$ for each i.

This definition is the mathematically usable version of the statement that all the roots of $g(x)$ can be 'expressed using only n^{th} roots for various n, field operations, and elements of F_0'. Subsuming the example after the reminder of the definition of *soluble group*, next comes the main theorem used to show that some equations $g(x) = 0$ of degree greater than four are not susceptible to such 'radical solutions'. The following two sections will go into detail on this. The converse of the next theorem is also true, its proof using the fundamental theorem three sections ahead. The converse has more 'positive' applications (**neutralize nattering nabobery!**). For example, Galois theory can be used to give a systematic derivation of the 'radical formulae' for solving cubics and quartics. See **Stewart**, pp.161–164; **Goldstein**, pp.316–322; **Artin**, pp.560–565; or **Appendix C** of this book.

Note that the phrase "over F_0" in the definition above is very important. If F_0 is $\mathbf{C}$ or $\mathbf{R}$, it follows from the fundamental theorem of (19^{th} century) algebra that *every* $g(x) \in F_0[x]$ is solvable by radicals over F_0. But if $g(x) \in \mathbf{Q}[x]$, it may very well not be solvable by radicals over $\mathbf{Q}$, as we shall see in Section **37**.

Theorem 35.11. *If $g(x) \in F_0[x]$ is solvable by radicals over F_0, then the Galois group of $g(x)$ over F_0 is a soluble group.*

Remark. Consistent, but not mutually consistent, usage of 'soluble' and 'solvable' will be found in British and U.S.ish (American?) texts, respectively.

Proof. Let K be the splitting field of $g(x)$ over F_0. To prove, as required, that $Aut_{F_0}(K)$ is soluble, let μ_i , n_i and

$$F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_m$$

be as in the last definition. We'll produce a new tower

$$F_0 \subset L_0 \subset L_1 \subset \cdots \subset L_r = L \ .$$

ASIDE. The first tower has two possible defects for the purposes of the proof:

i) Possibly F_{i-1} contains no primitive n_i^{th} root of unity. If it does, then by **35.6**, the extension $F_i \supset F_{i-1}$ is normal and $Aut_{F_{i-1}}(F_i)$ is abelian.

ii) Possibly $F_m \supset F_0$ is not normal. If it is, then

$$Aut_{F_0}(K) \cong Aut_{F_0}(F_m)/Aut_K(F_m)$$

by **35.9***ii)*.

If neither defect is present, then the desired solubility of $Aut_{F_0}(K)$ follows from that of $Aut_{F_0}(F_m)$ by ii). The latter follows from i) since, by **35.9***ii)*,

$$Aut_{F_{i-1}}(F_m)/Aut_{F_i}(F_m) \cong Aut_{F_{i-1}}(F_i) \ .$$

These arguments are used at the end of the proof below.

There are a number of texts in use where the present material seems to be done in a quite efficient manner, even relative to what is here, just above and below. Since this book prides itself on being a "bare bones", succinct presentation, the following explanation must be given. There are at least two such texts where one or both of points i) and ii) above are completely missed; a third, better known, text in which both i) and ii) are explicitly built into the *definition* of solvability by radicals; and a very famous text in which the author has decided to assume that every field under consideration contains all needed roots of unity—this gets around i), but leaves one with insufficient theory to see that some quintics in $\mathbf{Q}[x]$ are not solvable by radicals. As for ii), adjusting the tower to make the top field normal over the bottom is left as an exercise.

Definition. Any extension $F_m \supset F_0$ as in ii) of the last definition is called a *radical extension*. (This bears some analogy to the *iterated quadratic extensions* which occurred in the study of geometrical constructions. The existence of numbers relevant to such constructions and not in any iterated quadratic extension proves non-constructibility results. The existence of algebraic numbers not in any radical extension will prove the non-existence of 'radical formulae' for solving polynomial equations.)

Lemma 35.12. *If $F_m \supset F_0$ is any radical extension, then there is an extension $L \supset F_m$ such that $L \supset F_0$ is both radical and normal. Furthermore, this last extension can be chosen so that it has a radical tower whose successive extensions all have abelian Galois groups.*

Proof. Let $g_i(x)$ be the minimal polynomial of μ_i over F_0. Let n be $n_1 \cdots n_m$ (or any common multiple of $\{ n_1, \ \cdots, \ n_m \}$). Let L be a splitting field of $(x^n - 1)g_1(x) \cdots g_m(x)$ over F_0 which contains F_m. In other words, L is the field generated over F_0 by ω, a primitive n^{th} root of unity, together all the μ_i and all their conjugates over F_0. List the roots $\nu_1, \ \cdots, \ \nu_r$ of the product $g_1(x) \cdots g_m(x)$, in such a way that μ_1 and all its conjugates occur first, then μ_2 and all of its conjugates, etc.

Now define $L_0 := F_0(\omega)$, and, inductively,

$$L_j \ := \ L_{j-1}(\nu_j) \ = \ F_0(\omega, \ \nu_1, \ \nu_2, \ \cdots, \ \nu_j) \ .$$

This gives the promised tower, with $L \supset F_0$ certainly normal.

Now L_0 is the splitting field of $x^n - 1$ over F_0, so $L_0 \supset F_0$ is normal and, by **35.5**, the group $Aut_{F_0}(L_0)$ is abelian.

Fix $j > 0$, and let i be such that ν_j is μ_i or one of its conjugates over F_0. For some $k < j$, the field L_k is generated over F_0 by ω plus all of the μ_t and their conjugates over F_0 for $t < i$, i.e. it is the splitting field over F_0 for the product $(x^n - 1)g_1(x) \cdots g_{i-1}(x)$. Since $\mu_i^{n_i} \in F_{i-1} \subset L_k$, we have by **35.10** that $\nu_j^{n_i} \in L_k \subset L_{j-1}$. Thus, since L_{j-1} contains a primitive n_i^{th} root of unity (namely, some power of ω), it follows that L_j is the splitting field over L_{j-1} for $x^{n_i} - \nu_j^{n_i}$. By **35.6**, the group $Aut_{L_{j-1}}(L_j)$ is abelian.

Continuation of the proof of 35.11. The proof is now completed (as in the aside) by using **35.9** several times, as follows. The $g(x)$-splitting field K is a subfield of L, since $F_m \subset L$. The short tower $F_0 \subset L_0 \subset L$ of normal extensions gives an abelian quotient

$$Aut_{F_0}(L)/Aut_{L_0}(L) \ \cong \ Aut_{F_0}(L_0) \ .$$

Each short tower $L_{j-1} \subset L_j \subset L$ of normal extensions gives an abelian quotient

$$Aut_{L_{j-1}}(L)/Aut_{L_j}(L) \ \cong \ Aut_{L_{j-1}}(L_j) \ .$$

Since the subnormal series

$$Aut_{F_0}(L) \ \rhd \ Aut_{L_0}(L) \ \rhd \ Aut_{L_1}(L) \ \rhd \ \cdots \rhd Aut_{L_m}(L) = \{e\}$$

has abelian quotients, it is immediate that $Aut_{F_0}(L)$ is soluble. The short tower $F_0 \subset K \subset L$ of normal extensions gives

$$Aut_{F_0}(L)/Aut_K(L) \ \cong \ Aut_{F_0}(K) \ .$$

By **11F**, the group $Aut_{F_0}(K)$ is soluble, since it is the image of a soluble group under a morphism of groups. This is what we had to prove.

Are you convinced that we have captured what you thought it meant for an equation to be solvable by radicals? The only objection that I can think of is that we are requiring *all* the roots, rather than just *one* root, to lie in a radical extension. Let's say that the phrase '$g(x)$ *has a solution by radicals*' means the latter. Then **35.12** gives an easy solution to the following.

Exercise 35I. Show that :
i) an irreducible polynomial has a solution by radicals if and only if it is solvable by radicals;
ii) a polynomial has a solution by radicals if and only if at least one of its irreducible factors is solvable by radicals;
iii) a polynomial is solvable by radicals if and only if all of its irreducible factors are solvable by radicals.

36. The general equation of degree n.

Given any field F of characteristic 0, and an integer $n > 0$, form the field $F_0 = F(a_0, a_1, \cdots, a_{n-1})$, where $\{\ a_0,\ \cdots,\ a_{n-1}\ \}$ is algebraically independent over F, and let

$$g(x)\ =\ x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\ \in\ F_0[x]\ .$$

One refers to $g(x) = 0$ as *the general equation of degree n over F*. We are thinking of the a_i as 'formal symbols', so to speak. Let K be a splitting field for $g(x)$ over F_0. Of course, $F_0 \supset F$ and $K \supset F$ are *not* finite extensions. Write

$$(*) \qquad g(x)\ =\ (x - r_1)(x - r_2)\cdots(x - r_n)$$

in $K[x]$, so that $K = F_0(r_1, \cdots, r_n)$.

Theorem 36.1. *The Galois group of $g(x)$ over F_0 is isomorphic to the symmetric group S_n .*

Remark. Since S_n is not soluble for $n \geq 5$ by **12.1**, it follows that $g(x) = 0$ *is not solvable by radicals for* $n \geq 5$. Informally, this means that there can be no *formula*, involving only $+$, $-$, $\times$, $\div$, k^{th} roots and the coefficients, for solving all polynomial equations of a given degree (5 or greater), even for cleverly chosen F. The first accepted proof of this famous result was found by Abel for degree 5. It put an end to 2,500 years of attempts to find such formulae (except by a few cranks). Ruffini had earlier produced a proof which turned out to also be substantially complete and correct.

Proof. Multiplying out the right hand side of $(*)$ gives

$$(**) \qquad a_i \;=\; \pm\, e_{n-i}(r_1, \cdots, r_n)$$

by **22.4**, so $a_i \in F(r_1, \cdots, r_n)$, yielding

$$K \;=\; F(a_0, \cdots, a_{i-1}, r_1, \cdots, r_n) \;=\; F(r_1, \cdots, r_n)\ .$$

The proof is completed by showing that $\{\ r_1,\ \cdots,\ r_n\ \}$ is algebraically independent over F, since then, permutations $\sigma \in S_n$ give automorphisms of K by

$$g(r_1, \cdots, r_n)/h(r_1, \cdots, r_n) \;\;\mapsto\;\; g(r_{\sigma(1)}, \cdots, r_{\sigma(n)})/h(r_{\sigma(1)}, \cdots, r_{\sigma(n)})\ .$$

Since these permute the set $\{r_1, \ldots, r_n\}$, they fix each a_i by $(**)$, and so fix all elements of F_0, as required.

Suppose then that $h(r_1, \cdots, r_n) = 0$, where the polynomial $h(x_1, \cdots, x_n) \in F[x_1, \cdots, x_n]$ must be shown to be zero. Let

$$s(x_1, \cdots, x_n) \;=\; \prod_{\sigma \in S_n} h(x_{\sigma(1)}, \cdots, x_{\sigma(n)}) \;\in\; \mathrm{Symm}F[x_1, \cdots, x_n]\ .$$

Using **22.3**, there is a unique $s^\bullet(x_1, \cdots, x_n)$ in $F[x_1, \cdots, x_n]$ with

$$s(x_1, \cdots, x_n) \;=\; s^\bullet(e_1(\mathbf{x}), e_2(\mathbf{x}), \cdots, e_n(\mathbf{x}))\ .$$

$$\text{Then } 0 \;=\; s(r_1, \cdots, r_n) \qquad (\text{since } \; h(\mathbf{r}) = 0)$$

$$= \; s^\bullet(e_1(\mathbf{r}), \cdots, e_n(\mathbf{r})) \;=\; s^\bullet(\pm a_{n-1}, \cdots, \pm a_0)\ .$$

Since the coefficients of $s^\bullet$ are in F, and $\{ \ \pm a_0, \ \cdots, \ \pm a_{n-1} \ \}$ is algebraically independent over F, we get $s^\bullet(x_1, \cdots, x_n) = 0$ in $F[x_1, \cdots, x_n]$. Thus $s(x_1, \cdots, x_n) = 0$, and so $h(x_1, \cdots, x_n) = 0$, as required.

Exercise 36A. Show that if some of the elements in an algebraically independent set are replaced by their negatives, the set remains algebraically independent (as used just above).

37. Radically unsolvable over Q.

The non-existence of a general formula leaves open the possibility that, over some fields, every equation might be solvable by radicals, but involving such a plethora of special cases that they couldn't be welded together into a 'single formula'. Over any algebraically closed field, such as **C** or the algebraic numbers, this is exactly what happens, since there's no need to build any tower at all. And over **R**, one has $\mathbf{C} = \mathbf{R}(i)$ with $i^2 = -1 \in \mathbf{R}$, so again, any $g(x) \in \mathbf{R}[x]$ is solvable by radicals over **R**.

Exercise 37A. Let $B \supset A$ be a radical extension. Show that a polynomial in $A[x]$ being solvable by radicals over A and over B are equivalent.

But over **Q** and many other fields, one can write down explicit polynomials $g(x)$ which are not solvable by radicals. Below is one of degree 5 over **Q**. First here is a result which we could have proved earlier but didn't need then.

Theorem 37.1. *If $K \supset F$ is normal, then*

$$|\mathrm{Aut}_F(K)| \ = \ [K : F] \ .$$

Note. The proof applies to *finite normal extensions* in *characteristic zero*, or more generally, to any *simple extension which is the splitting field of some polynomial in $F[x]$ which has no repeated root.*

Exercise 37B. For the field K in the previous section, find an F_0-basis (which must have "$n!$" elements by **37.1**).

Proof. Choose a primitive element α, so that $K = F(\alpha)$. Let the minimal polynomial $p(x)$ over F of α have degree $[K : F] = n$, say. Let $\alpha = \alpha_1, \ \alpha_2, \ \cdots, \ \alpha_n$ be the roots of $p(x)$, which are distinct by **29.4**, and in K by **35.8** ii). It is clear, by considering degrees, that $K = F(\alpha_i)$ for all

i. Given $\theta \in Aut_F(K)$, we have $\theta(\alpha) = \alpha_i$ for some i, and this determines θ, by **35.2** and **35.3**, so

$$|Aut_F(K)| \ \leq \ n \ .$$

But for each i, the simple extension principle yields an isomorphism from $K = F(\alpha)$ to $K = F(\alpha_i)$ sending α to α_i , since α and α_i have the same minimal polynomial $p(x)$, so we have equality above.

Theorem 37.2. *Let $g(x) \in \mathbf{Q}[x]$ be an irreducible of degree 5 with exactly two roots in $\mathbf{C} \setminus \mathbf{R}$ [and therefore three real roots]. Then the Galois group of $g(x)$ over $\mathbf{Q}$ is isomorphic to S_5.*

Remark. Since S_5 is not soluble, such a $g(x)$ is therefore not solvable by radicals over $\mathbf{Q}$. For example, $x^5 - 4x + 2$ will fill the bill; elementary calculus shows that it has exactly three real roots, and Eisenstein's criterion (**20A**) with $p = 2$ gives irreducibility.

Proof. Let $\{\ \alpha,\ \bar{\alpha},\ \alpha_1,\ \alpha_2,\ \alpha_3\ \}$ be the roots of $g(x)$, with the α_i real. Complex conjugation maps the splitting field, K, of $g(x)$ to itself [by **35.8v)**, taking L to be $\mathbf{C}$, and K to be the splitting field within $\mathbf{C}$ of $g(x)$]. This gives an element of $Aut_{\mathbf{Q}}(K)$ which acts as a 2-cycle, $\alpha \leftrightarrow \bar{\alpha}$, on the set of roots of $g(x)$. Now $|Aut_{\mathbf{Q}}(K)| = [K : \mathbf{Q}]$ by the previous theorem, and so it is divisible by $[\mathbf{Q}(\alpha) : \mathbf{Q}] = 5$, which happens to be a prime. Thus $Aut_{\mathbf{Q}}(K)$ contains an element of order 5 by the Cauchy theorem **11.2**. Only cycles can have prime order in S_n [see **4.8**], so $Aut_{\mathbf{Q}}(K)$ also has an element which acts as a 5-cycle on the set of roots. But S_5 is generated by any $\{$ 2-cycle, 5-cycle $\}$ by **6C**. Thus $Aut_{\mathbf{Q}}(K)$ is the full permutation group of the roots of $g(x)$, as required.

38. The Galois correspondence.

Passing from fields to their automorphism groups has an inverse, obtained by forming the field of elements fixed by each member of a given group. The precise version of this in the theorem below has many applications to polynomial equations and elsewhere in field theory. For example, it follows immediately that there are only finitely many intermediate fields for any finite extension in characteristic zero, since a finite group certainly has only finitely many subgroups.

In other books, the theorem below is often proved *before* the application to solvability by radicals, since it is needed for the converse to **35.11**. We have already proved at least half of it.

The theorem has important analogues both beyond field theory, and within field theory to non-finite extensions and to characteristic p. As for the latter, the reader might wish to check that our proof below applies to finite extensions *with no restriction on characteristic*, as long as the assumption of normality is taken to mean condition *i)* in **35.8**. This should be fairly credible in view of the remark after the proof of **35.8**, indicating that *i)* implies all the other conditions. Note however that **37.1** then would need a new proof. The phrase *Galois extension* is often used in connection with **35.8***i)*, rather than 'normal extension'. That the equivalent **35.8** conditions *ii), iv)* and *v)* do not imply *i)* without an additional assumption (namely *separability* : each element in K is the root of a polynomial in $F[x]$ which has no repeated roots—see Appendix **B**) can be seen by means of the following example.

If K is a field of characteristic p, then the map $\alpha \mapsto \alpha^p$ is a field map $K \to K$, and therefore injective, but not necessarily surjective if K is infinite. The example in Section **34** shows this, but, more simply, take

$$K = \mathbf{Z}_p(\beta) \ \supset \ \mathbf{Z}_p(\beta^p) = F,$$

where β is transcendental over $\mathbf{Z}_p$. Then $K \supset F$ satisfies iv) in **35.8** since K is the splitting field of $x^p - \beta^p$ over F. [Note that, in $K[x]$, we have $x^p - \beta^p = (x - \beta)^p$.] But the extension does *not* satisfy **35.8***i)* since, for any $\theta \in Aut_F(K)$, we have $\beta^p = \theta(\beta^p) = \theta(\beta)^p$, so $\theta(\beta) = \beta$, since $x \mapsto x^p$ is injective. Thus θ is the identity map, and **35.8***i)* fails for *every* $\alpha \in K \setminus F$.

Less relevant here, *iii)* does not imply { *ii), iv), v)* } in **35.8** for characteristic p : one can find a finite extension which is neither simple nor a splitting field—for example,

$$\mathbf{Z}_3(\alpha, \beta) \ \supset \ \mathbf{Z}_3(\alpha^3, \beta^3)$$

where $\{ \ \alpha, \ \beta \ \}$ is algebraically independent over $\mathbf{Z}_3$.

Exercise 38A. Complete the details of this example.

To avoid the trouble of going back through our arguments carefully, we'll just state the theorem in characteristic zero.

Fundamental Theorem 38.1. The Galois Correspondence. *Let $K \supset F$ be a (finite) normal extension in characteristic zero. Denote by $\mathcal{INT}(F,\ K)$ the set of fields E with $F \subset E \subset K$. Denote by $\mathcal{SG}(F,\ K)$ the set of subgroups of $Aut_F(K)$. Define*

$$\mathcal{G} \ = \ \mathcal{G}_F^K \ : \ \mathcal{INT}(F,\ K) \ \longrightarrow \ \mathcal{SG}(F,\ K)$$

by $\mathcal{G}(E) \ := \ Aut_E(K)$. Define

$$\mathcal{F} \ = \ \mathcal{F}_F^K \ : \ \mathcal{SG}(F,\ K) \ \longrightarrow \ \mathcal{INT}(F,\ K)$$

by

$$\mathcal{F}(\Gamma) \ := \ \{\ \alpha \in K \ : \ \theta\alpha = \alpha \ \text{ for all } \ \theta \in \Gamma\ \}$$

(called the 'fixed field of Γ'*). Then:*
i) the maps $\mathcal{G}$ and $\mathcal{F}$ are mutually inverse bijections;
ii) they reverse inclusions : $E \ \subset \ E' \quad \Longleftrightarrow \quad \mathcal{G}(E) \ \supset \ \mathcal{G}(E')$;
iii) we have $|\mathcal{G}(E)| \ = \ [K : E]$;
iv) the extension $E \supset F$ is normal if and only if $\mathcal{G}(E)$ is a normal subgroup of $\mathcal{G}(F)$, in which case

$$Aut_F(E) \ \cong \ \mathcal{G}(F)/\mathcal{G}(E) \ .$$

Remark. It follows from iii) that

$$[E : F] \ = \ |\mathcal{G}(F)|/|\mathcal{G}(E)| \ .$$

Superficially more generally, the 'relative distances' in the towers

$$F \ \subset \ E_1 \ \subset \ E_2 \ \subset \ K$$

and

$$\mathcal{G}(F) \ \supset \ \mathcal{G}(E_1) \ \supset \ \mathcal{G}(E_2) \ \supset \ \mathcal{G}(K) = \{e\}$$

are related as one would expect:

$$[E_2 : E_1] \ = \ |\mathcal{G}(E_1)|/|\mathcal{G}(E_2)| \ .$$

To see this, just write $[E_2 : E_1]$ as $[K : E_1]/[K : E_2]$, and apply *iii)*. Alternatively, apply the Galois correspondence of $K \supset E_1$. In general

$$[E_2 : E_1] \ \geq \ |Aut_{E_1}(E_2)| \ ,$$

by the first half of the argument in **37.1**.

Exercise 38B. Deduce that, for a tower of fields $F \subset E_1 \subset E_2 \subset K$, with $K \supset F$ as in the theorem, we have :

$E_2 \supset E_1$ is normal $\Leftrightarrow$ $Aut_{E_2}(K)$ is a normal subgroup of $Aut_{E_1}(K)$.

Proof of the Galois correspondence theorem. To begin, it is a routine verification to check that $\mathcal{F}$ is well-defined, i.e. that $\mathcal{F}(\Gamma)$ is a subfield of K containing F.

(i) That $\mathcal{F}(\mathcal{G}(E)) = E$ has already been proved: $\mathcal{F}(\mathcal{G}(E)) \supset E$ is trivial, and the reverse inclusion is immediate from condition *i)* of **35.8** for normal extensions, $K \supset E$ being such an extension by **35.9***i)*. The non-obvious inclusion for showing that $\mathcal{G}(\mathcal{F}(\Gamma)) = \Gamma$ is $\mathcal{G}(\mathcal{F}(\Gamma)) \subset \Gamma$. Let $E = \mathcal{F}(\Gamma)$ and $[K : E] = n$. Then

$$|\mathcal{G}(\mathcal{F}(\Gamma))| \;=\; |Aut_E(K)| \;=\; n$$

by **37.1**, so it suffices to prove that $|\Gamma| \geq n$. Write $K = E(\alpha)$ and define

$$g(x) \;=\; \prod_{\theta \in \Gamma} \{x - \theta(\alpha)\} \ ,$$

a polynomial in $K[x]$ which has degree equal to $|\Gamma|$. Then, for any $\theta_0 \in \Gamma$,

$$g(x) \;=\; \prod_{\theta \in \Gamma} \{x - \theta_0\theta(\alpha)\} \ ,$$

so θ_0 fixes the coefficients of $g(x)$. Since $E = \mathcal{F}(\Gamma)$, this implies that $g(x) \in E[x]$. But $g(\alpha) = 0$, so $g(x)$ is divisible by the minimal polynomial of α over E. The latter has degree n, so degree$(g(x)) \geq n$, as required.

(ii) This is obvious.

(iii) This is **37.1**, once again using that $K \supset E$ is normal.

(iv) In view of **35.9***ii)*, it remains only to prove that if $E \supset F$ is not a normal extension (where E is a subfield of K), then $Aut_E(K)$ is not a normal subgroup of $Aut_F(K)$. The contrapositive form of the proof in **35.8**, that *ii)* implies *i)*, proves the existence of $\theta \in Aut_F(K)$ with $\theta(E) \neq E$: { Choose an irreducible $g(x) \in F[x]$ with roots $\alpha \in E$ and $\beta \notin E$ [since **35.8***ii)* fails for the extension $E \supset F$]; then construct

$$\begin{array}{ccccc} F & \supset & F(\alpha) & \supset & K \\ || & & \downarrow \psi & & \downarrow \theta \\ F & \supset & F(\beta) & \supset & K \end{array}$$

where ψ and θ are isomorphisms mapping α to β.} Then, by i) of the present theorem, $Aut_{\theta(E)}(K) \neq Aut_E(K)$, so the following lemma completes the proof (as well as reproving the converse, **35.9***ii)*a), and describing how conjugacy of subgroups in $Aut_F(K)$ behaves.)

Lemma 38.2. *For any normal extension $K \supset F$, intermediate field E, and $\theta \in Aut_F(K)$,*

$$Aut_{\theta(E)}(K) \;=\; \theta \circ Aut_E(K) \circ \theta^{-1} \;.$$

Proof. It is a routine verification that $\theta \circ \phi \circ \theta^{-1} \in Aut_{\theta(E)}(K)$ for any $\phi \in Aut_E(K)$. The reverse inclusion follows either from the fact that the two groups have the same order (by earlier parts of the fundamental theorem), or else from the just proved inclusion, replacing θ by θ^{-1} and E by $\theta(E)$.

Each of the Galois groups which we calculated earlier gives an illustration of the fundamental theorem. See also Section **40** for another example.

$x^2 + 1$ over $\mathbf{R}$:

$$\begin{array}{cc} \mathbf{C} & \{e\} \\ \uparrow & \downarrow \\ \mathbf{R} & \{e,\ \text{complex conjugation}\} \end{array}$$

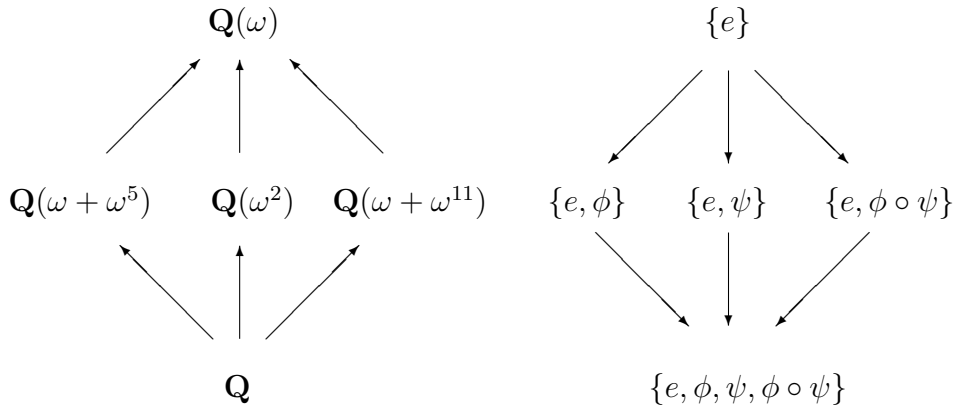
More generally, the Galois correspondence for any quadratic (i.e. degree 2) extension takes this form. In fact this holds for any normal extension of prime degree p (except that the group is cyclic of order p).

$x^5 - 1$ over $\mathbf{Q}$: (Here $\omega = \mathrm{e}^{2\pi i/5}$)

$$\begin{array}{cc} \mathbf{Q}(\omega) & \{e\} \\ \uparrow & \downarrow \\ \mathbf{Q}(\omega^2 + \omega^3) & \{\theta^2, e\} \\ \uparrow & \downarrow \\ \mathbf{Q} & \{\theta, \theta^2, \theta^3, e\} \end{array}$$

This example and the next one show the two possibilities which can occur for normal extensions of degree 4 (since there are only two groups of order 4, up to isomorphism).

$x^{12} - 1$ over $\mathbf{Q}$: (Here $\omega = \mathrm{e}^{2\pi i/12}$.).



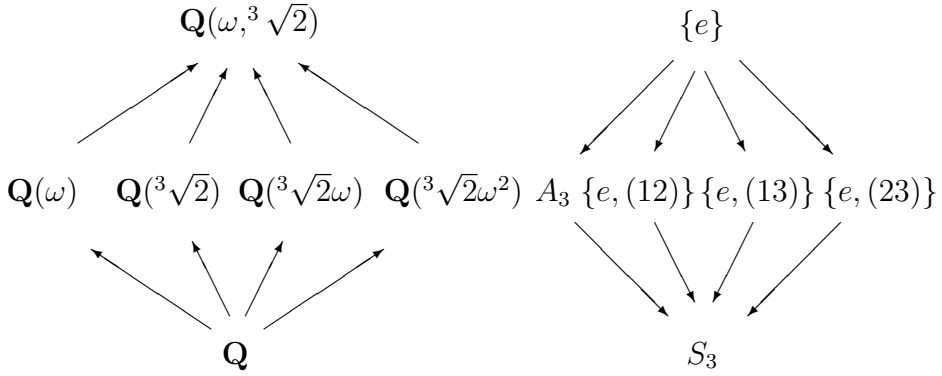
Here we have

$$\omega + \omega^5 \;=\; -(\omega^7 + \omega^{11}) \;=\; \omega^3 \;=\; i \;\;;$$

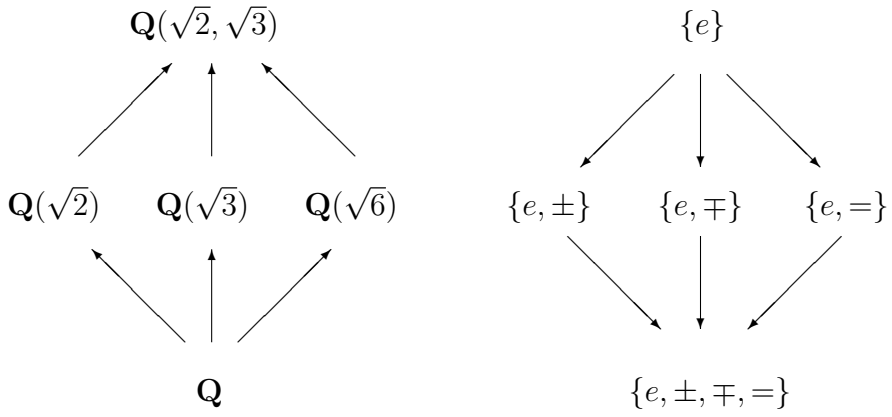
$$\omega + \omega^{11} \;=\; \omega + \bar{\omega} \;=\; 2Re(\omega) \;=\; \sqrt{3} \;=\; -(\omega^7 + \omega^5) \;\;;$$

$$\mathbf{Q}(\omega^2) \;=\; \mathbf{Q}(\omega^4) \;\;.$$

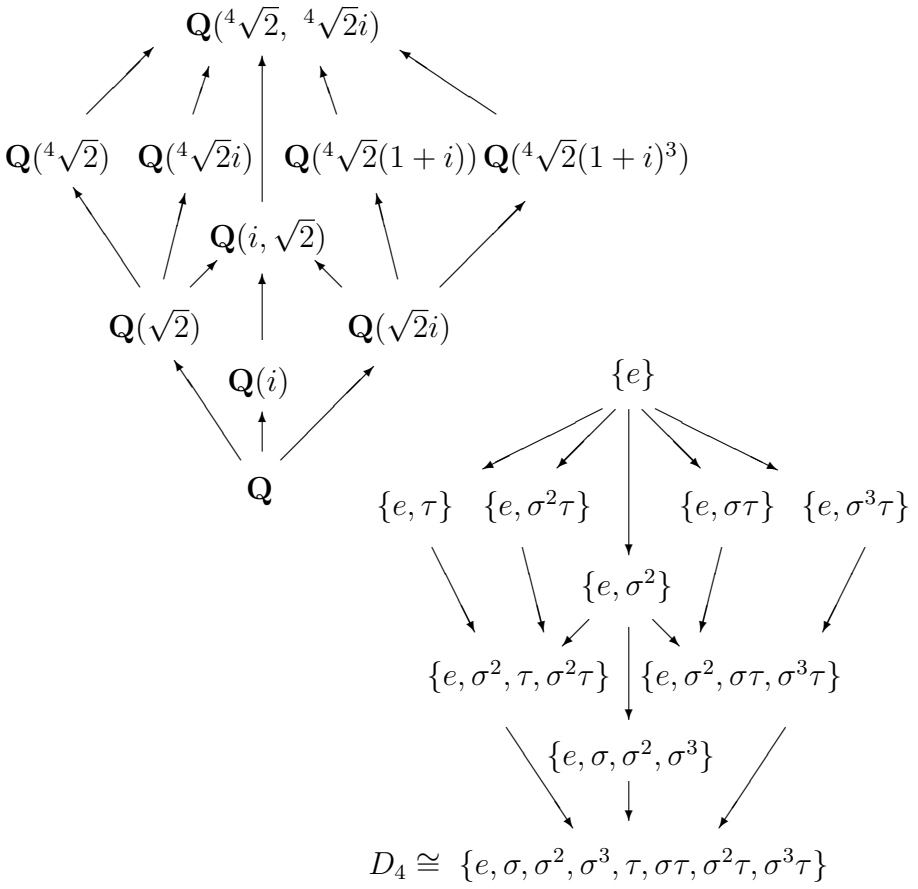
$\underline{x^3 - 2 \text{ over } \mathbf{Q}}$: (Here $\omega = \mathrm{e}^{2\pi i/3}$.)



$\underline{(x^2-2)(x^2-3) \text{ over } \mathbf{Q}}$: Note that the 'lattices' for $x^{12}-1$ are isomorphic to the ones here. We use $\pm$ to denote the automorphism sending $\sqrt{2}$ to $+\sqrt{2}$ and $\sqrt{3}$ to $-\sqrt{3}$; similarly for $\mp$ and $=$.



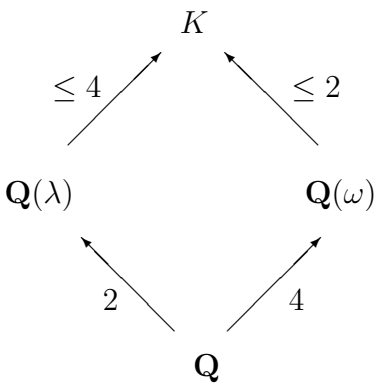
$x^4 - 2$ over $\mathbf{Q}$: Here τ maps $\sqrt[4]{2}$ to itself and maps $\sqrt[4]{2}i$ to $-\sqrt[4]{2}i$; and σ maps $\sqrt[4]{2}$ to $\sqrt[4]{2}i$ and maps $\sqrt[4]{2}i$ to $-\sqrt[4]{2}$.



In the first example, one verifies that $\mathbf{Q}(\omega^2 + \omega^3)$ is the unique strictly intermediate field by calculating that $a_0 + a_1\omega + a_2\omega^2 + a_3\omega^3$ is fixed by θ^2 if and only if $a_1 = 0$ and $a_2 = a_3$ (where each $a_i \in \mathbf{Q}$). The intermediate fields in the other examples are similarly straightforward calculations using the definition of the Galois group.

Large Exercise 38C. Do these calculations.

Suppose one wished to compute $[K : \mathbf{Q}]$, where $K = \mathbf{Q}(e^{\frac{2\pi i}{3}}, e^{\frac{2\pi i}{5}})$, the splitting field of $(x^3-1)(x^5-1)$ over $\mathbf{Q}$. The method of Section **26** narrows the answer down to 4 or 8, using the following diagram, where the arrows are labeled with information about the degree. The two *inequalities* each follow from the label on the parallel arrow by **26A**. Let $\lambda = e^{\frac{2\pi i}{3}}$ and $\omega = e^{\frac{2\pi i}{5}}$.



To verify that 8 is the answer amounts to proving that $\lambda \notin \mathbf{Q}(\omega)$. If it were in $\mathbf{Q}(\omega)$, the first example above shows that $\mathbf{Q}(\lambda) = \mathbf{Q}(\omega^2 + \omega^3)$, so

$$\lambda = a_0 + a_2(\omega^2 + \omega^3) \;=\; a_0 + a_2\mu \;\; \text{(say)} \;\; ,$$

with $a_i \in \mathbf{Q}$. Now

$$\mu^2 \;=\; \omega^4 + 2\omega^5 + \omega^6 \;=\; (-1 - \omega - \omega^2 - \omega^3) + 2 + \omega \;=\; 1 - \mu \; .$$

Thus

$$0 \;=\; \lambda^2 + \lambda + 1 \;=\; (a_0^2 + a_2^2 + a_0 + 1) + (2a_0a_2 - a_2^2 + a_2)\mu \; .$$

It is now easy to see that no rationals a_0, a_2 give zero for both bracketed terms in the last expression.

Exercises 38D. I) Complete the last argument.

(II) Show that, over $\mathbf{Q}$, the polynomial $(x^5-1)(x^7-1)$ has splitting field of degree 24.

(III) See **Goldstein** pp. 250-251 for a proof [using Gauss' theorem **33.1** on the irreducibility of $c_m(x)$] of the special case of Dirichlet's theorem, which

says that there are infinitely many primes in every sequence $1+k$, $1+2k$, $1+3k$, $\cdots$ (for $k > 0$) . Using this, the fundamental theorem, and the fact that, for a primitive n^{th} root, ω, of unity,

$$Aut_{\mathbf{Q}}(\mathbf{Q}(\omega)) \;\cong\; \mathbf{Z}_n^{\times} \;,$$

show that, for any finite abelian group Γ, there exists a 'short' tower of normal extensions

$$\mathbf{Q} \;\subset\; E \;\subset\; \mathbf{Q}(\omega)$$

[for some such ω] such that $Aut_{\mathbf{Q}}(E) \;\cong\; \Gamma$.

Thus every finite *abelian* group is isomorphic to the Galois group over $\mathbf{Q}$ of some polynomial. Some deeper related theorems are as follows.

Šafarevič (1954). *Every finite soluble group is (up to isomorphism) the Galois group over* $\mathbf{Q}$ *of some polynomial.*

Hilbert (1892). *For all* n*, the symmetric group* S_n *is the Galois group over* $\mathbf{Q}$ *of some polynomial (in fact,* $\cdots$ *of some irreducible of degree* n *in* $\mathbf{Q}[x]$*).*
The argument in **37.2** depends on 5 being a *prime*, but general n is harder.

The question of whether *every* finite group is the Galois group over $\mathbf{Q}$ of some polynomial is a notorious unsolved problem (as of 1994).

Kronecker-Weber (1853, 1877). *If* $E \supset \mathbf{Q}$ *is a (finite) normal extension with* $Aut_{\mathbf{Q}}(E)$ *abelian, then for some root,* ω*, of unity, the field* E *is a subfield of* $\mathbf{Q}(\omega)$*.*

39. (Fermat prime)-gons are constructible.

Recall that Gauss' theorem, **33.2**, on the constructibility of regular n-gons was not completely proved. It remained to show that if $p = 2^r + 1$ is a Fermat prime, then the regular p-gon is constructible; equivalently, $\cdots$ then $\mathrm{e}^{2\pi i/p} \in \mathcal{P}$. This is equivalent to finding a tower

$$\mathbf{Q} = F_0 \ \subset \ F_1 \ \subset \ F_2 \ \subset \ \cdots \ \subset \ F_r = \mathbf{Q}(\mathrm{e}^{2\pi i/p}),$$

with each successive extension being of degree 2. Now $\mathbf{Q}(\mathrm{e}^{2\pi i/p}) \supset \mathbf{Q}$ is normal, being the splitting extension for $x^p - 1$. So, by the fundamental theorem, this is equivalent to finding a descending series of subgroups,

$$\mathrm{Aut}_{\mathbf{Q}}(\mathbf{Q}(\mathrm{e}^{2\pi i/p})) = G_0 \ \supset \ G_1 \ \supset \ G_2 \ \supset \ \cdots \ \supset \ G_r = \{e\},$$

such that each group has index 2 in the previous one. But the group $\mathrm{Aut}_{\mathbf{Q}}(\mathbf{Q}(\mathrm{e}^{2\pi i/p}))$ is abelian (by **35.5**) of order 2^r (since the cyclotomic polynomial $c_p(x) = x^{p-1} + x^{p-2} + \cdots + 1$ has degree 2^r), so the existence of this series of subgroups is an easy consequence of **13.2**, which gives the structure of finite abelian groups. In fact, the group at issue is cyclic, making the descending series unique and its existence even more obvious.

Exercise 39A. Give the detailed proof that any abelian group whose order is a power of 2 has such a tower of subgroups.

Large Exercise 39B. For $p = 17$ give explicit details of constructing the tower of fields to show that the regular 17-gon can be produced with a straight-edge and compass. Do it for $p = 5$ first, to warm up.

The other straight-edge & compass leftover is to complete the proof of **27.4**. We should show that any root of $x^4 - 2x - 2$ generates an extension of $\mathbf{Q}$ which has no strictly intermediate fields. Now the polynomial has Galois group S_4, by **39C** below. That group has exactly four subgroups of index four, each being the symmetric group on a three element subset of $\{\ 1,\ 2,\ 3,\ 4\ \}$. None of these is contained in any subgroup of index two, since A_4 is the only subgroup of order 12 in S_4. This completes the proof, using the Galois correspondence.

Exercise 39C. i) Show that the Galois group over F of an irreducible in $F[x]$ acts *transitively* on the set of roots of the polynomial; i.e. for any two roots it has an element mapping the first to the second.

ii) Show that the only proper transitive subgroups of S_4 which contain a transposition are isomorphic to D_4. (To define such groups, place the integers 1 to 4 on the vertices of the square in the definition of D_4.)
iii) **Artin** (14.6.14, p.564) gives a cubic polynomial, the 'resolvent', whose irreducibility would eliminate the D_4 above as a possibility for the Galois group of a given irreducible quartic over $\mathbf{Q}$. Show that this cubic is $x^3 - 4$ when the quartic is $x^4 - 2x - 2$.

Since, as in the proof of **37.2**, complex conjugation gives the needed transposition, this exercise shows that $x^4 - 2x - 2$ has Galois group S_4 over $\mathbf{Q}$, as required. Hilbert's result, getting S_n as a Galois group over $\mathbf{Q}$, is done differently, since the theory of quartics is special.

40. Automorphisms of finite fields.

Since $\mathbf{F}_{p^n}$ is the splitting field of $x^{p^n} - x$ over $\mathbf{F}_p$, the extension $\mathbf{F}_{p^n} \supset \mathbf{F}_p$ satisfies ii), iv) and v) of **35.8**. In fact, it even satisfies i), i.e., it is a *Galois* extension, as discussed before the fundamental theorem. Thus the characteristic p version of that theorem would apply, in fact to any extension $\mathbf{F}_{p^{ab}} \supset \mathbf{F}_{p^a}$ of finite fields.

Without appealing to the fundamental theorem, one can, as follows, directly calculate the group $\mathrm{Aut}_{\mathbf{F}_{p^a}}(\mathbf{F}_{p^{ab}})$, and verify the truth of the fundamental theorem in this instance by inspection. We have

$$|\mathrm{Aut}_{\mathbf{F}_p}(\mathbf{F}_{p^n})| \ \leq \ n \ ,$$

since the extension can be written as a simple extension by an element of degree n; and an automorphism is determined by its effect on that element, for which there are at most "n" choices, the other roots of the relevant minimal polynomial. (As it happens, the proof of **37.1**, that

$$|\mathrm{Aut}_F(K)| \ = \ [K : F] \ ,$$

depended only on $K \supset F$ being simple and being a splitting extension. Therefore it applies to all extensions involving finite fields. So we have equality above, but this will be deduced more concretely below.) The *Frobenius self-map*, $\theta_K : x \mapsto x^p$ of any field K of characteristic p, is evidently bijective of order n when $K = \mathbf{F}_{p^n}$; in fact θ^r maps each x to x^{p^r}, and so has fixed field $\mathbf{F}_{p^s}$ with $s = \mathrm{GCD}\{n \ , \ r\}$. Thus

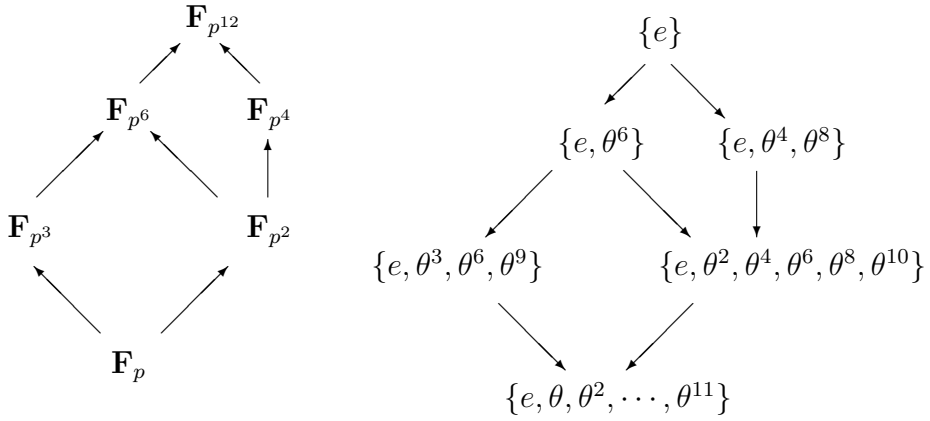
$$\mathrm{Aut}_{\mathbf{F}_p}(\mathbf{F}_{p^n}) \ \cong \ C_n \quad \text{with generator} \ \ \theta_{\mathbf{F}_{p^n}} \ ,$$

since the left hand side is a group of order at most n, and we have found an element of that order. More generally,

$$\mathrm{Aut}_{\mathbf{F}_{p^a}}(\mathbf{F}_{p^{ab}}) \;\cong\; C_b \quad \text{with generator} \quad (\theta_{\mathbf{F}_{p^{ab}}})^a \ .$$

One sees this by noting that the claimed generator generates a cyclic group of order b, which is the same order as that of the automorphism group, since $[\mathbf{F}_{p^{ab}} : \mathbf{F}_{p^a}] = b$.

For example, here is the Galois correspondence for $\mathbf{F}_{p^{12}}$:



41. Galois theoretic proof of the fundamental theorem of (19^{th} century) algebra.

Using the fact that a real polynomial of odd degree has a real root, it follows that

(I) $\mathbf{R} \supset \mathbf{R}$ *is the only extension of* $\mathbf{R}$ *with odd degree.*

For, any element α in such an extension would give $\mathbf{R}(\alpha) \supset \mathbf{R}$ of odd degree. But the minimal polynomial of α over $\mathbf{R}$ is irreducible in $\mathbf{R}[x]$, and so it has degree 1. Thus $\alpha \in \mathbf{R}$, as required.

The fact, that any extension $K \supset F$ of degree 2 has the form $F(\alpha) \supset F$ where $\alpha^2 \in F$ but $\alpha \notin F$, implies that

(II) $\mathbf{C}$ *has no extension of degree 2.*

For, $x^2 - z$ splits as $(x - w)(x + w)$ in $\mathbf{C}[x]$, where, if $z = re^{i\theta}$, then $w = \sqrt{r}e^{i\theta/2}$.

Finally, note that

(III) *if $K_1 \supset \mathbf{R}$ has degree 2, then $K_1 \cong \mathbf{C}$.*

For, with $K_1 = \mathbf{R}(\alpha)$, where $\alpha^2 \in \mathbf{R}$ but $\alpha \notin \mathbf{R}$, then necessarily $\alpha^2 < 0$. Define $\phi : \mathbf{C} \to K_1$ by $\phi(a + ib) = a + (\alpha b/\sqrt{-\alpha^2})$, where $\sqrt{-\alpha^2}$ of course means the *positive real* whose square is $-\alpha^2$. It is easily checked that ϕ is an isomorphism.

Combining Galois theory with some non-trivial group theory (but not using anything dependent on the 'Fundamental Theorem of Algebra') yields:

Lemma 41.1. *If $K \supset \mathbf{R}$ is a normal (finite) extension, then it has degree at most 2.*

It follows from (III) that $K \cong \mathbf{R}$ or $\mathbf{C}$.

Proof. Let $[K : \mathbf{R}] = 2^s q$ where q is odd.

First we'll show that $q = 1$. By the Galois correspondence and the existence of Sylow subgroups (**11.1**), $\mathrm{Aut}_{\mathbf{R}}(K)$ has a subgroup of order 2^s which has the form $\mathrm{Aut}_E(K)$ where $\mathbf{R} \subset E \subset K$. But then the degree $[E : \mathbf{R}]$ is q above, so $q = 1$ by (I).

Now since $|\mathrm{Aut}_{\mathbf{R}}(K)| = 2^s$, there is a series of groups

$$\mathrm{Aut}_{\mathbf{R}}(K) = G_0 \rhd G_1 \rhd \cdots \rhd G_s = \{e\},$$

where each $|G_i/G_{i+1}| = 2$, by Exercise **11D** (solubility of p-groups). Applying the inverse Galois correspondence yields a tower

$$\mathbf{R} = K_0 \subset K_1 \subset \cdots \subset K_s = K ,$$

where each $[K_{i+1} : K_i] = 2$. Assume that $s > 0$. Then $K_1 \cong \mathbf{C}$ by (III). But then K_2 cannot exist by (II), so $s = 1$, as required.

Corollary 41.2. $\mathbf{C}$ *has no finite extensions other than itself; i.e.,* $\mathbf{C}$ *is algebraically closed.*

Proof. Suppose that $L \supset \mathbf{C}$ is a finite extension. Let $\alpha \in L$, let $g(x)$ be the minimal polynomial of α over $\mathbf{R}$, and let $K \supset \mathbf{R}$ be a splitting extension for $(x^2+1)g(x)$. Then $K \cong \mathbf{C}$ by **41.1**, so $g(x)$ splits in $\mathbf{C}[x]$ and $\alpha \in \mathbf{C}$, as required.

Remarks. (1) It is possible to give a 'rabbit out of hat' proof using less field theory than above and no group theory. One avoids the Sylow theorem by a tricky induction and a magic formula.

(2) *A real closed field* R is one which has a linear order—[a transitive relation $<$ with exactly one of $a < b,\ a = b,\ b < a$ holding for each pair $(a,\ b)$] satisfying the axioms for an *ordered field*—[for all $a,\ b$ and c, we have $a < b \Rightarrow a + c < b + c\ ;\ (0 < c\ \&\ a < b) \Rightarrow ac < bc$],
and such that:

i) each positive element has a square root in R—this implies that the order is unique, once the field operations have been given, since then, $a < b \iff b - a$ is a non-zero square; and

ii) each polynomial of odd degree in $R[x]$ has a root in R.
Then $x^2 + 1$ has no root in R. The proof in this section quickly generalizes to show that $R(i)$ is algebraically closed, when $i^2 = -1$. An example of such an R is $\mathbf{R} \cap \mathbf{A}$, the field of real algebraic numbers (as is $\mathbf{R}$ itself, of course).

(3) A surprising fact is that $\mathbf{C}$ has infinitely many automorphisms outside of the group

$$\mathrm{Aut}_{\mathbf{R}}(\mathbf{C}) \ = \ \{\ e\ ,\ \text{complex conjugation}\ \}\ .$$

These cannot map $\mathbf{R}$ to itself, since $\mathrm{Aut}_{\mathbf{Q}}(\mathbf{R}) = \{e\}$.

Exercise 41A. Prove this last statement. First show that an automorphism would preserve positivity and therefore order.

It follows that $\mathbf{C}$ has subfields isomorphic to $\mathbf{R}$ other than $\mathbf{R}$ itself. These extra automorphisms of $\mathbf{C}$ are not continuous. They may be proved to exist by finding an intermediate field E between $\mathbf{Q}$ and $\mathbf{C}$ such that $\mathbf{C} \supset E$ is algebraic and no element of E is algebraic over $\mathbf{Q}$. Take $E = \mathbf{Q}(S)$ (the field generated by $\mathbf{Q} \cup S$), where S is a maximal algebraically independent (over $\mathbf{Q}$) subset of $\mathbf{C}$. (The existence of such an S follows from Zorn's Lemma.) Then E has many automorphisms, for example ones agreeing on S with any

given self-bijection of S. Now extend such automorphisms to automorphisms of $\mathbf{C}$, using the transfinite extension principle alluded to after the proof of **A10** in the appendix following Section **28**.

APPENDIX B. Separability and the Galois correspondence in arbitrary characteristic.

Definition. A polynomial is *separable* if and only if it has distinct roots; that is, $g(x)$ is separable if and only if $\mathrm{GCD}\{g(x),\ Dg(x)\} = 1$.

This is independent of both the field in which we regard $g(x)$ as having its coefficients, and also of the extension in which it splits. If $g(x) \in F[x]$, then clearly $g(x)$ is separable if and only if (I) its irreducible factors in $F[x]$ are separable, and (II) each occurs with exponent 1 in the factorization of $g(x)$. Essentially we're only interested in separability of *irreducibles*, once the base field F has been fixed. Note also that $p_1(x)^{\alpha_1} \cdots p_s(x)^{\alpha_s}$ has the same splitting field over F as $p_1(x) \cdots p_s(x)$, for distinct irreducibles $p_i(x)$ in $F[x]$ and $\alpha_i \geq 1$.

Caution. As an element of $\mathbf{Q}[x]$, the polynomial $x^2 + 1$ is separable; but the element of $\mathbf{Z}_2[x]$ denoted similarly is not separable. Note also that some authors define $p_1(x)^{\alpha_1} \cdots p_s(x)^{\alpha_s}$ as above to be separable as long as each irreducible factor $p_i(x)$ has no repeated roots. This seems to complicate things a bit, since then separability would be a function of both the polynomial and the field you are using.

Definition. Given $K \supset F$, an element $\alpha \in K$ is *separable over* F if and only if its minimal polynomial over F is separable. (This depends on F, but not on K).
We say that $K \supset F$ is a *separable extension* if and only if all α in K are separable over F.

For example, in characteristic zero, all $K \supset F$ are separable. If β is transcendental over $\mathbf{Z}_p$, then $\mathbf{Z}_p(\beta) \supset \mathbf{Z}_p(\beta^p)$ is not separable: β is not separable over $\mathbf{Z}_p(\beta^p)$ since $x^p - \beta^p = (x - \beta)^p$; although β is certainly separable over $\mathbf{Z}_p(\beta)$. The minimal polynomial of β over $\mathbf{Z}_p(\beta^p)$ is in fact

$x^p - \beta^p$, although all that we need to know here is that the former is a non-linear factor of the latter.

Theorem B1. *i) For any finite $K \supset F$, we have*

$$|Aut_F(K)| \ \leq \ [K : F] \ .$$

More precisely, if K is written as $F(\alpha_1, \cdots, \alpha_\ell)$, with $g_i(x)$ being the minimal polynomial of α_i over $F(\alpha_1, \cdots, \alpha_{i-1})$, and if m_i is the number of roots in K of $g_i(x)$, then $|Aut_F(K)| = m_1 m_2 \cdots m_\ell$.

ii) If K is the splitting field over F of a separable polynomial in $F[x]$, then $|Aut_F(K)| = [K : F]$.

Remark. **B1***ii)* is proved below, after the statements of the other three theorems, and is all that we use for **B4** below. The proof of **B1***i)* is similar.

Theorem B2. *For any finite $K \supset F$, the following are equivalent:*
i) for all $\alpha \in K \backslash F$, $\exists \theta \in Aut_F(K)$ with $\theta(\alpha) \neq \alpha$
(i.e. **35.8***i) holds; i.e. $\mathcal{FG}(F) = F$ in* **38.1***, the fundamental theorem);*
ii) the extension $K \supset F$ is both normal and separable;
iii) the field K is the splitting field over F of some separable polynomial in $F[x]$.

Note The implications *i) $\Leftrightarrow$ ii) $\Rightarrow$ iii)* are immediate from the statement and proof of **35.8**.

Remark. **B1***i)* easily implies that $|Aut_F(K)| = [K : F]$ when $K \supset F$ is normal separable. This would give **B1***ii)* if we could use **B2**[*iii) $\Rightarrow$ ii)*], but the proof below of **B2**[*iii) $\Rightarrow$ ii)*] uses **B1***ii)*.

Definition. The extension $K \supset F$ is called a *Galois* extension if and only if the conditions in **B2** hold.

Theorem B3. *i) A finite extension $K \supset F$ is simple if and only if $|\mathcal{INT}(F, \ K)| < \infty$; that is, $\cdots$ iff there are finitely many intermediate fields.*
ii) If $K \supset F$ is separable and finite, then $K \supset F$ is simple.

Remark. B3*i)*$\Leftarrow$ is very easy, and is all we use in proving:

Theorem B4. *The fundamental theorem,* **38.1***, holds for any finite Galois extension in any characteristic.*

The trick is to first prove that

$$\mathcal{INT}(F,\ K)\ \xrightarrow{\mathcal{G}}\ \mathcal{SG}(F,\ K)\ \xrightarrow{\mathcal{F}}\ \mathcal{INT}(F,\ K)$$

is the identity (almost by definition). But then $\mathcal{INT}(F,\ K)$ must be finite, since $\mathcal{SG}(F,\ K)$ is. By **B3***i)*$\Leftarrow$, the extension $K \supset F$ is simple, and now the same proof as in **38.1**, that

$$\mathcal{SG}(F,\ K)\ \longrightarrow\ \mathcal{INT}(F,\ K)\ \longrightarrow\ \mathcal{SG}(F,\ K)$$

is the identity, works. Parts *ii)* and *iv)* of **38.1** are proved as before. Also **38.1***iii)* is just **B1***ii)* above, in view of **B2**, and since $K \supset F$ being normal and separable trivially implies that $K \supset E$ is also normal and separable, for any $E \in \mathcal{INT}(F,\ K)$.

Proof of B1*ii)*. We prove by induction on $[K_1 : F_1]$ that for all

$$\begin{array}{rl} F_1 \hookrightarrow K_1 & \text{where } K_1 \supset F_1 \text{ and } K_2 \supset F_2 \text{ are splitting} \\ \phi \downarrow\cong & \text{field extensions for separable polynomials} \\ F_2 \hookrightarrow K_2 & \text{which correspond under } \phi, \end{array}$$

the number of isomorphisms $\psi : K_1 \to K_2$, completing the diagram to make it commute, is $[K_1 : F_1]$ (which equals $[K_2 : F_2]$, of course). The initial step is trivial. For the inductive step, given two ϕ-corresponding irreducibles, $g_i(x) \in F_i[x]$, which divide the ϕ-corresponding polynomials referred to above, choose a root $\alpha \in K_1$ for $g_1(x)$. Let the polynomials $g_1(x)$ and $g_2(x)$ have degree m. Then there are "m" roots, γ, of $g_2(x)$ in K_2 by separability, so there are "m" choices of commutative diagram as follows:

$$\begin{array}{ccccc} F_1 & \hookrightarrow & F_1(\alpha) & \hookrightarrow & K_1 \\ \phi \downarrow\cong & & \cong\downarrow \theta & & \text{with } \theta(\alpha) = \gamma\ . \\ F_2 & \hookrightarrow & F_2(\gamma) & \hookrightarrow & K_2 \end{array}$$

For each such θ, by induction there are "$[K_1 : F_1(\alpha)]$" isomorphisms $\psi : K_1 \to K_2$ completing the diagram. Since any ψ at the beginning of the proof arises from a unique such θ, and since

$$[K_1 : F_1]\ =\ [K_1 : F_1(\alpha)][F_1(\alpha) : F_1]\ =\ m[K_1 : F_1(\alpha)]\ ,$$

this completes the induction. Statement **B1***ii)* is the special case in which $F_1 = F_2 = F$; $\phi =$ the identity ; and $K_1 = K_2 = K$.

Proof of B2[*iii)* $\Rightarrow$ *ii)*]. Let $E = \mathcal{FG}(F)$, and let K be the splitting field over F of the separable polynomial $g(x) \in F[x]$. Then $F \subset E \subset K$, so K is the splitting field over E of $g(x)$ as well. By **B1***ii)*, we have

$$|Aut_E(K)| \;=\; [K : E] \quad \text{and} \quad |Aut_F(K)| \;=\; [K : F] \; .$$

By the definition of E, we have $Aut_F(K) = Aut_E(K)$. This yields $[K : F] = [K : E]$. Thus $[E : F] = 1$, and so $E = F$, as required.

Additional Galois theory problems—many books have lots of good problems, especially **Hungerford** .

1. If $\Sigma_{i=0}^{m}\ a_i x^i$ is irreducible in $F[x]$, then so is $\Sigma_{i=0}^{m}\ a_{m-i} x^i$.

2. The following are equivalent for a field F of characteristic p :

   i) every irreducible in $F[x]$ is separable;

   ii) the Frobenius map, $F \to F$; $a \mapsto a^p$, is surjective.

   (Such a field is called *perfect*, as are all fields of characteristic 0, where i) holds necessarily.)

3. Every finite field is perfect.

4. Assume that $K \supset F$ and $L \supset F$ are both finite, that K is isomorphic to a subfield of L, and that L is isomorphic to a subfield K, both isomorphisms fixing elements of F. Show that $L \cong K$. Does this hold without the finiteness assumption?

5. For every finite $K \supset F$, there exists a field L such that : i) $L \supset K$; ii) $L \supset F$ is normal; iii) if $M \supset K$ is an extension with $M \supset F$ normal, then L is isomorphic to a subfield of M, fixing elements of F. Prove that L exists and is unique up to isomorphisms fixing elements of F. (The extension $L \supset F$ is called the *normal closure* normal closure of $K \supset F$.)

6. Given $K \supset F$ and B, C both in $\mathcal{INT}(F,\ K)$, let $B \vee C$ (the 'compositum') be the subfield of K generated by $B \cup C$. Prove that the normal closure of $K \supset F$ in problem 5. can be written

$$(K_1 \vee K_2 \vee \cdots \vee K_r) \ \supset \ F \ ,$$

for some $r \geq 1$ and $K_i \in \mathcal{INT}(L,\ F)$ such that each K_i is isomorphic to K over F.

[This $\vee$ together with $\cap$ are the binary operations in the *lattice* lattice $\mathcal{INT}(F, K)$.]

7. If $K \supset F$ is Galois and $E \in \mathcal{INT}(F,\ K)$, then $K \supset E$ is also Galois.

8. For every $m \geq 2$, give an example of $K \supset F$ with $[K : F] = m$ and $K \neq F(\gamma)$ for every γ with $\gamma^m \in F$. To what extent can this be done in characteristic 0?

9. (See problem 6 above.) If $K \supset F$ is Galois and $g(x)$ is irreducible in $F[x]$, then all irreducibles in the factorization of $g(x)$ in $K[x]$ have the same degree. Show also that this can fail for non-Galois extensions.

10. Prove that every finite group is isomorphic to $Aut_F(K)$ for at least one finite $K \supset F$ in characteristic 0, (even with $F \supset \mathbf{Q}$ finite—the solution that will make you famous is doing it for $F = \mathbf{Q}$).

11. Let $g(x) = 0$ be the 'general equation of degree m' over F [as in Section **36**] giving the field F_0.

    i) Is $g(x)$ irreducible in $F_0[x]$?

    ii) Does this hold in arbitrary characteristic?

    iii) Is $g(x)$ separable in arbitrary characteristic?

12. For any positive integer n, find a finite extension $K \supset F$ such that $K \neq F(\gamma_1, \ldots, \gamma_n)$ for any $\gamma_1, \ldots \gamma_n$.

13. Illustrations of subgroup (non-)existence implying subfield (non-)existence:

    i) If $K \supset F$ is Galois (finite) with $Aut_F(K)$ soluble, then there exists $E \in \mathcal{INT}(F,\ K)$ with $[E : F]$ prime.

    ii) Does i) hold without the assumption on $Aut_F(K)$?

    [Hint: A_6 has no subgroups of prime index.]

    iii) If a prime p divides $[K : F]$ where $K \supset F$ is any finite extension in characteristic 0, does there exist $E \in \mathcal{INT}(F,\ K)$ with $[K : E] = p$?

    iv) Does i) hold for *any* prime dividing $[K : F]$?

APPENDIX C. Solubility implies solvability.

Let us again return to the world of *characteristic* 0 (and, as usual, consider only *finite* extensions). We shall prove that if $g(x) \in F[x]$ has soluble Galois group over F, then $g(x)$ is solvable by radicals over F (the converse of **35.11**). This excellent result is due to Galois. It is false in this simple form when the characteristic is nonzero.

What must be done is to show that a normal extension $E \supset F$, with $Aut_F(E)$ soluble, can be embedded in a radical extension; i.e. there is a short tower $F \subset E \subset K$ for some radical extension $K \supset F$. This is Theorem

C7 below. The obvious first try is (using solubility) to write down a tower of groups,

$$Aut_F(E) = G_0 \ \rhd \ G_1 \ \rhd \ G_2 \ \rhd \ \cdots \ \rhd \ G_r = \{1\} \ ,$$

with successive quotients abelian, or stronger, *cyclic of prime order*. Then, by the Galois correspondence, we have $G_i = Aut_{F_i}(E)$ for a tower of fields

$$F = F_0 \ \subset \ F_1 \ \subset \ F_2 \ \subset \ \cdots \ \subset \ F_r = E \ .$$

Letting G_i/G_{i+1} be cyclic of prime order p_i, we have the forlorn hope that $F_{i+1} = F_i(\beta_i)$ with $\beta_i^{p_i} \in F_i$. This at least motivates the following question: When can an extension $K \supset L$ with $[K : L] = p$, a prime, be written as $K = L(\beta)$ with $\beta^p \in L$? There is a nice clean sufficient condition:

Theorem C1. *If $K \supset L$ is a normal extension of prime degree p such that L contains a primitive p^{th} root of unity, then $K = L(\beta)$ for some β with $\beta^p \in L$.*

The need for normality and roots of unity (give examples to show this !) complicates the "first try" above; but below **C7** is proved by an induction on degree which is an amplification of that "first try".

Note that, when $p = 2$, the hypotheses are automatically satisfied; but we already know the conclusion as well.

There is an easy exercise providing a weak partial converse to **C1**: *If p is a prime and $L(\beta) \supset L$ is a normal extension of degree p with $\beta^p \in L$, then $L(\beta)$ (not necessarily L) contains a primitive p^{th} root of unity.*

The proof of **C1** below will use the *norm*

$$N \ : \ K^\times \ \longrightarrow \ L^\times \ ,$$

$$k \ \mapsto \ \prod_{\theta \in G} \theta(k) \ ,$$

defined for any normal extension $K \supset L$, where $G := Aut_L(K)$.

Lemma C2. *The function N is a well defined morphism of (multiplicative) groups. Furthermore, $N(\ell) = \ell^{\,[K:L]}$ for all $\ell \in L^\times$; and $N(\theta(k)) = N(k)$ for all $k \in K^\times$ and $\theta \in Aut_L(K)$.*

Exercise CA. Prove this.

Lemma C3. (Hilbert's Theorem 90.) *If $K \supset L$ is a normal extension with $Aut_L(K)$ being a cyclic group generated by θ, then for all y in $K^\times$, we have $N(y) = 1 \Leftrightarrow$ there exists an $x \in K^\times$ with $y = x^{-1}\theta(x)$.*

Proof. The $\Leftarrow$ half is trivial, since N is a group morphism, and $N(\theta(x)) = N(x)$. For the other half, we need another technical but fundamental result, due to Dedekind:

Theorem C4. *Let G be any group, and let F be any field. Then the set of all group morphisms from G to $F^\times$, regarded as a subset of the vector space, F^G, of all functions from G to F, is linearly independent.*

Remark. Group morphisms from G to $F^\times$ were called *characters*, particularly when $F = \mathbf{C}$. This is still the case; but, since the year 1899, certain other functions from G to F are also called characters, as explained in Section **48**. In **48.7**, we prove a stronger result (orthonormality rather than just linear independence), applying more generally (to these post-1899 characters as well), but also less generally (G is assumed to be finite). Since we need the result for the infinite group $G = F^\times$, an elementary proof of **C4** is provided below.

Corollary C5. *Any set of automorphisms of a field is linearly independent.* (Just take G to be $F^\times$ itself.)

Proof of C4. Suppose, for a contradiction, that we have a non-trivial linear relation

$$f_1\theta_1 \ + \ f_2\theta_2 \ + \ \cdots \ + \ f_n\theta_n \ = \ 0 \qquad \text{(I)}$$

where each f_i is in F, and the θ_i are distinct morphisms from G to $F^\times$. In fact, choose such a relation with n minimal, so that all f_i are non-zero. Multiplying by f_1^{-1}, we may assume that $f_1 = 1$. Clearly $n > 1$. Choose $h \in G$ with $\theta_1(h) \neq \theta_2(h)$. For any $g \in G$, we have

$$\theta_1(g) \ + \ f_2\theta_2(g) \ + \ \cdots \ + \ f_n\theta_n(g) \ = \ 0 \qquad \text{(II)}$$

$$\theta_1(h)\theta_1(g) + f_2\theta_2(h)\theta_2(g) + \cdots + f_n\theta_n(h)\theta_n(g) \ = \ 0 \qquad \text{(III)}$$

by applying (I) to g and to hg. Multiply (III) by $\theta_1(h)^{-1}$, giving

$$\theta_1(g) \ + \ f_2\theta_1(h)^{-1}\theta_2(h)\theta_2(g) \ + \ \cdots \ + \ f_n\theta_1(h)^{-1}\theta_n(h)\theta_n(g) \ = \ 0 \qquad \text{(IV)}$$

Then (II) minus (IV) yields

$$f_2[1-\theta_1(h)^{-1}\theta_2(h)]\theta_2(g) \;+\; \cdots \;+\; f_n[1-\theta_1(h)^{-1}\theta_n(h)]\theta_n(g) \;=\; 0\ .$$

This holds for all g, so by the minimality of n, we have

$$f_i[1-\theta_1(h)^{-1}\theta_i(h)] \;=\; 0$$

for all $i \geq 2$. But $f_i \neq 0$, so $\theta_1(h) = \theta_i(h)$, which for $i = 2$ gives a contradiction.

Continuation of the proof of C3. To prove $\Rightarrow$, suppose that θ has order n. Define $y_0,\ y_1,\ \cdots,\ y_{n-1}$ inductively, by $y_0 := y$ and

$$y_i \;:=\; y\theta(y_{i-1}) \;=\; y\theta(y)\theta^2(y)\cdots\theta^i(y)\ .$$

Note that $y_{n-1} = N(y) = 1$. Using **C5**, the set $\{\ \mathrm{id},\ \theta,\ \theta^2,\ \cdots,\ \theta^{n-1}\ \}$ is linearly independent, so we may choose some z with

$$y_0 z \;+\; y_1\theta(z) \;+\; \cdots \;+\; y_{n-2}\theta^{n-2}(z) \;+\; y_{n-1}\theta^{n-1}(z) \;\neq\; 0\ .$$

Let x^{-1} be the left hand side. A straightforward calculation shows that $y\theta(x^{-1}) = x^{-1}$, as required.

Proof of C1. It suffices to find a $\beta \in K \setminus L$ such that β^p is fixed by all elements of $Aut_L(K)$: This is because $K = L(\beta)$ for *any* $\beta \in K \setminus L$ (since there are no intermediate fields), and $\beta^p \in L$ (since no element of $K \setminus L$ is fixed by the Galois group). If $\omega \in L$ is a primitive p^{th} root of unity, then

$$N(\omega) \;=\; \omega^{[K:L]} \;=\; \omega^p \;=\; 1\ .$$

Now $|Aut_L(K)| = [K : L] = p$, a prime, so $Aut_L(K)$ is a cyclic group. Let θ be a generator. Then, by **C3**$\Rightarrow$, there exists $\beta \in K$ such that $\omega = \beta^{-1}\theta(\beta)$. Rewrite this as $\theta(\beta) = \beta\omega$. Since $\beta\omega \neq \beta$, we get $\beta \in K \setminus L$. Since $(\beta\omega)^p = \beta^p\omega^p = \beta^p$, we get $\theta(\beta^p) = \beta^p$, and so β^p is fixed by the whole Galois group, as required.

Lemma C6. *Suppose given a diagram of field extensions,*

$$F \quad \subset \quad E$$

$$\begin{array}{ccc} \cap & & \cap \\ \hat{F} & \subset & \hat{E} \ , \end{array}$$

where $E \supset F$ and $\hat{E} \supset \hat{F}$ are both splitting extensions for the same *polynomial in $F[x]$. Then the map*

$$Aut_{\hat{F}}(\hat{E}) \longrightarrow Aut_F(E) \ ,$$

$$\theta \mapsto \theta|_E$$

(defined by restricting), is a well defined morphism of groups, and is injective.

Exercise CB. Prove this, using the most basic ideas near the start of Section **35**.

Theorem C7. *Let $E \supset F$ be a normal extension in characteristic 0 for which $Aut_F(E)$ is soluble. Then there exists a radical extension $K \supset F$ such that $E \subset K$.*

Proof. Proceed by induction on $[E : F]$. The initial step is obvious. For the inductive step, by solubility, there is a prime p such that $Aut_F(E)$ has an index p normal subgroup N—'the top of the tower'. Let $\hat{F} = F(\omega)$, where ω is a primitive p^{th} root of unity, and let $\hat{E} = E(\omega)$. If $E \supset F$ is a splitting extension for $g(x) \in F[x]$, then $\hat{E} \supset \hat{F}$ is also one for $g(x)$, and $\hat{E} \supset F$ is one for $(x^p - 1)g(x)$. So the latter two extensions are normal. By **C6**, the group $Aut_{\hat{F}}(\hat{E})$ is isomorphic to a subgroup of $Aut_F(E)$; and so, by **11H**, it is soluble.

Case i). If $[\hat{E} : \hat{F}] < [E : F]$, then, by the inductive hypothesis, there is a radical extension $K \supset \hat{F}$ such that $\hat{E} \subset K$. But $\hat{F} = F(\omega)$, where $\omega^p{=}1 \in F$, so $K \supset F$ is also radical. Since $E \subset \hat{E} \subset K$, we are done.

Case ii). If $[\hat{E} : \hat{F}] = [E : F](= b$, say), then $Aut_{\hat{F}}(\hat{E})$ is isomorphic to $Aut_F(E)$ by the restriction map, since they are finite groups of the same order b. Let $\hat{N} \lhd Aut_{\hat{F}}(\hat{E})$ be the normal subgroup which corresponds to N under this isomorphism. Let $D \in \mathcal{INT}(\hat{E},\ \hat{F})$ be the intermediate field for which $Aut_D(\hat{E}) = \hat{N}$, using the Galois correspondence. Then the extension $\hat{E} \supset D$ is normal with soluble Galois group $\hat{N}$. Its degree is "p" times smaller than $[E : F]$, so by the inductive hypothesis, there is a radical extension $K \supset D$ such that $\hat{E} \subset K$. Now $D \supset \hat{F}$ is normal, because $\hat{N}$ is a normal subgroup of $Aut_{\hat{F}}(\hat{E})$. This extension has degree p. Thus, by **C1**, we can write D as

$\hat{F}(\beta)$ with $\beta^p \in \hat{F}$. As we did in *Case i)*, write $\hat{F}$ as $F(\omega)$ with $\omega^p \in F$. Combining the last two sentences with the radicality of $K \supset D$, it follows that $K \supset F$ is also a radical extension. Since $E \subset \hat{E} \subset K$, that's it.

Exercise CC. (**C**ubi**C**s !)

i) Show how, by 'completing the cube', finding solutions to the general equation, $x^3 + ax^2 + bx + c = 0$, of degree 3 may be reduced to solving $x^3 + px + q = 0$.

Let

$$u \;=\; [\; \frac{1}{2}(-q + \sqrt{q^2 + \frac{4p^3}{27}}) \;]^{1/3}$$

and

$$v \;=\; [\; \frac{1}{2}(-q - \sqrt{q^2 + \frac{4p^3}{27}}) \;]^{1/3}$$

ii) Show that $u^3 + v^3 + q \;=\; 0$.

iii) Show that $3uv + p \;=\; 0$.

iv) Deduce that $u + v$ is a root of $x^3 + px + q \;=\; 0$.

This gives a solution to the general cubic, one which is 'radical', as claimed earlier (and as follows also from what we just proved in this Appendix, since S_3 and all its subgroups are soluble).

What about the general solution? Well, there are lots of combinations of choices for the square and cube roots in our formulae. **OOPS!** It looks as though we have produced far more than three roots for a cubic polynomial !

The square roots are no problem : Clearly ii) only works if we choose the *same* square root in the two formulae; and changing them both to their negatives just interchanges u and v. So the choice of square roots is illusory; it doesn't result in any extra solutions to the cubic.

Let's be a bit more formal about where we're working. Assume that F is a field whose characteristic is not 2 nor 3 (although the previous bracketed comment only applies for characteristic 0). Show that:

i)* If $a,\ b$ and c in i) are in F, then so are p and q.

ii)* The identity in ii) holds when the square root and cube roots in the formulae are interpreted to be any three fixed elements in some extension of F whose square or cubes are as indicated.

iii)* In an extension of F, prove that, for each choice of cube root giving u, there is at most one choice of cube root giving v so that iii) holds; and exactly one choice if the extension is 'big enough'.

This gets us down from nine to (at most) three solutions of the form $u+v$, so we have the 'general solution', as required.

v) Show that if $u + v$ is one solution, then the others are $\omega u + \omega^2 v$ and $\omega^2 u + \omega v$, where ω is a primitive cube root of unity.

vi) To reinforce all of this, multiply out

$$(x - u - v)(x - \omega u - \omega^2 v)(x - \omega^2 u - \omega v) \ ,$$

where u and v are 'compatible', as explained in iii)*.

As readers of this book (but not all students seem to) realize, the command: 'Show that object $\mathcal{O}$ is a solution to equation $\mathcal{E}$' is usually much easier to obey than the commands: 'Find a solution to equation $\mathcal{E}$' or 'Find all solutions to equation $\mathcal{E}$'. The renaissance Italians who first solved cubics and quartics were of course faced with the harder questions. Due to their success, we needed only to consider the easier one. But to see how these formulae for roots of cubics might arise, you can find the analysis of the harder questions in many texts; for example, **Rotman**, pp.25–28.

Exercise CD. (**D**is**C**riminants !) Assuming characteristic 0, let $g(x) \in F[x]$ have splitting field E over F. By indexing its roots—that is, by writing

$$g(x) \ = \ (x - r_1)(x - r_2) \cdots (x - r_n) \ \in E[x] \ ,$$

we get the Galois group as a subgroup of S_n, and we can define

$$\Delta \ = \ \prod_{i<j} (r_i - r_j) \ \in E \ .$$

i) Show that Δ depends, up to sign, only on $g(x)$ and E, and not on the indexing. More precisely, show that, if the roots are re-indexed as $s_1, \ \cdots, \ s_n$, where $s_i = r_{\sigma(i)}$ for some permutation σ, then Δ changes by being multiplied by the sign of σ.

The *discriminant* of $g(x)$ is defined to be

$$D \ = \ \Delta^2 \ ,$$

which is clearly independent of the indexing. Obviously $D = 0$ if and only if $g(x)$ has a repeated root. (This is a minor bit of discrimination practised by

the discriminant.)

ii) Show that $D \in F$.

iii) Let G be the Galois group of $g(x)$ over F, and let $H = G \cap A_n$, regarding $G \subset S_n$ as above. Show that the intermediate field $\mathcal{F}(H)$ (notation from the Galois correspondence) is $F(\Delta)$; in particular,

$$G \subset A_n \quad \text{if and only if} \quad \sqrt{D} \in F \ .$$

(So that's another thing which the discriminant discriminates !)

Exercise CE. Refer to **CC**, and assume the characteristic to be 0.

i) Show that the two corresponding cubics in **CC** i) have the same discriminant, and it is $-4p^3 - 27q^2$. (More precisely, show that

$$\Delta \ = \ 3(1+2\omega)\sqrt{q^2 + \frac{4p^3}{27}} \qquad [\text{ and } \quad (1+2\omega)^2 \ = \ -3 \ ] \ ,$$

where the square root is the one chosen in the formula for the roots of cubics, assuming the roots to be indexed in the following order: $u+v$, $\omega u+\omega^2 v$, $\omega^2 u+\omega v$.) Find another formula for the discriminant of the general (unreduced) cubic, in terms of its coefficients a, b and c. Generalize the first phrase to arbitrary degree : Show how any monic of degree n can be replaced by a 'reduced' one, whose coefficient of x^{n-1} is zero, where the two polynomials have roots differing merely by a 'translation', and have the *same* discriminant.

ii) Check directly from the formula for the discriminant of a cubic, $g(x)$, that $D = 0$ if and only if GCD$\{\ g(x),\ Dg(x)\ \} \neq 1$, where the last D denotes the formal derivative. Now use the explicit formula for the roots to show directly that these conditions are equivalent to the existence of a repeated root.

iii) Prove the following easy facts about Galois groups of cubics:
A product of three linears in $F[x]$ has trivial Galois group.
A linear times an irreducible quadratic in $F[x]$ has Galois group cyclic of order 2.
An irreducible cubic in $F[x]$ has Galois group A_3 or S_3, and we get A_3 if and only if $\Delta \in F$.

iv) Show that when $F = \mathbf{Q}$ and $g(x)$ is irreducible in $\mathbf{Q}[x]$, and if we take the splitting field which is inside $\mathbf{C}$, then $g(x)$ has three real roots when $D > 0$; and only one real root when $D < 0$ [in which case the Galois group is necessarily S_3, by iii)]. (Further discrimination !)

V. Modules over PIDs and Similarity of Matrices.

The next five sections give a straightforward (but very important) generalization of **13.2**, and a major application to matrix theory. Basic facts concerning linear algebra and abelian groups are assumed, as well as unique factorization in the ring (see sections **13, 17** and **25**). The first section, **42**, consists of the minimum needed theory for modules over a commutative ring. The student will likely find that later algebra courses emphasize modules very strongly. Section **43** contains the structure theorem for finitely generated modules over a *Euclideanizable domain*. Thus the above title misleads, motivating:

Project. Reprove everything in sections **43** and **44**, assuming only that R is a PID (or else, consult **Hartley & Hawkes, Jacobson** or **Hungerford**).

In Section **44**, part of the more common approach for proving the structure theorem via Smith normal form is given. All of this specializes, with $R = \mathbf{Z}$, to abelian groups in Section **45**, and to the application, with $R = F[x]$, to similarity of matrices in Section **46**.

We shall label by '**RV**' any statement whose proof is a *routine verification*, left to the reader. These should provide a vehicle for plenty of mathematical recreation. Only a few other exercises are included, except at the end of the final section, **46**.

42. Basics on modules.

In this section, R is any commutative ring (with 1, of course).

Definitions. An *R-module* (if R is a field, also called an *R-vector space*—see **25**) is an abelian group M together with a *scalar multiplication*

$$R \times M \longrightarrow M \quad ; \quad (r, m) \mapsto r \cdot m \ ,$$

such that, for all r, r' in R and m, m' in M :

$$(r + r') \cdot m = r \cdot m + r' \cdot m \qquad ; \qquad r \cdot (r' \cdot m) = (rr') \cdot m \ ;$$

$$r \cdot (m + m') = r \cdot m + r \cdot m' \qquad ; \qquad 1 \cdot m = m \ .$$

Such an M is *finitely generated* (if R is a field, also said to be *spanned by a finite set*, which implies that M is *finite dimensional*) if and only if there is a subset $\{ m_1, \cdots, m_n \}$ of M such that each m can be written in the form $r_1 \cdot m_1 + \cdots + r_n \cdot m_n$ for some r_i in R. The set $\{ m_1, \cdots, m_n \}$ is then a *set of generators* for M. (If R is a field, it is also called a *spanning set,* and, when n is minimal, a *basis*.) A module which can be generated by one element (to be fussy, by a singleton set) is called *cyclic*—when $R = \mathbf{Z}$, this agrees with our previous notation for abelian groups.

A group morphism $\theta : M_1 \to M_2$ between R-modules is a *module morphism* (if R is a field, also called a *linear transformation*) if and only if θ also satisfies $\theta(r \cdot m) = r \cdot \theta(m)$ for all $r \in R$ and $m \in M_1$. Such a θ is a module *isomorphism* if and only if it is bijective.

Having written these definitions down, we'll normally cease using the "$\cdot$", since the ring multiplication and the module scalar multiplication can always be distinguished by the context.

RV i) *If θ is an isomorphism, then θ^{-1} is a module morphism, and so an isomorphism. Then we write $M_1 \cong M_2$, so that $\cong$ is an equivalence relation, using also that identity maps and composites of morphisms are morphisms.*

A subgroup N of an R-module M is a *submodule* (if R is a field, also called a *subspace*) if and only if N also satisfies $rn \in N$ for all $n \in N$ and $r \in R$. Then the quotient group M/N is given the scalar multiplication $r(m + N) := (rm) + N$.

RV ii) *This scalar multiplication is well-defined, and makes M/N into a module, the 'quotient module'.*

Let $\theta : M_1 \to M_2$ be a module morphism.

RV iii) $\mathrm{Ker}\theta$ *and* $\mathrm{Im}\theta$ *are submodules of M_1 and M_2, respectively.*

RV iv) *For submodules $N_1 \subset \mathrm{Ker}\theta \subset M_1$ and $\mathrm{Im}\theta \subset N_2 \subset M_2$, the factorization of θ as a composite,*

$$M_1 \longrightarrow M_1/N_1 \longrightarrow N_2 \longrightarrow M_2 \ ,$$

consists of three module morphisms: the canonical surjection; followed by

the map sending $m_1 + N_1$ to $\theta(m_1)$; followed by the inclusion.

RV v) *Taking $N_1 = \mathrm{Ker}\theta$ and $N_2 = \mathrm{Im}\theta$, the group isomorphism $M_1/\mathrm{Ker}\theta \to \mathrm{Im}\theta$ in the middle above is in fact a module isomorphism.*

If $M_1, \ \cdots, \ M_t$ are R-modules, their *external direct sum* is

$$M_1 \bigoplus \cdots \bigoplus M_t \ := \ \{ \ (m_1, \ \cdots, \ m_t) \ : \ m_i \in M_i \ \} \ ,$$

with its usual abelian group structure using coordinatewise addition, and given the coordinatewise scalar multiplication:

$$r(m_1, \ \cdots, \ m_t) \ := \ (rm_1, \ \cdots, \ rm_t) \ .$$

RV vi) *This makes $M_1 \bigoplus \cdots \bigoplus M_t$ into an R-module.*

RV vii) $M_1 \bigoplus (M_2 \bigoplus \cdots \bigoplus M_t) \ \cong \ M_1 \bigoplus \cdots \bigoplus M_t$.

This is virtually tautological rather than merely routine. The other associativity isomorphisms for $\bigoplus$ hold; **vii)** happens to be the one needed later.

RV viii) *Using its ring operations, R is itself an R-module.*

RV ix) *The submodules of the module R are precisely the ideals I of the ring R.*

Thus, for example, R/I is a quadruply *abused notation*: it denotes a set, an abelian group, a ring, an R-module, and an R/I-module.

RV x) *Let M be an R-module, and S a subring of R. Then restricting*

$$S \times M \ \stackrel{\text{inclusion}}{\longrightarrow} \ R \times M \ \stackrel{\text{scalar mult.}}{\longrightarrow} \ M$$

makes the abelian group M into an S-module.

For example, any $\mathbf{C}$-vector space may be thought of in addition as an $\mathbf{R}$-vector space, of *double* the dimension. Any $F[x]$-module 'is' also an F-vector space.

In **x)** we could have had R and S being any two commutative rings, with a fixed ring morphism $\phi : S \to R$. But only the case when ϕ is inclusion of a subring will be used below.

43. Structure of finitely generated modules over euclidean domains.

Assume now that R is a Euclideanizable domain, although we only need it to be a PID for the proof given of **43.2** to work.

Theorem 43.1. *If M is a finitely generated R-module, then there are integers $k \geq 0$ and $\ell \geq 0$, and non-zero non-invertible elements d_i of R, with $d_1 \mid d_2 \mid \cdots \mid d_\ell$, such that*

$$M \quad \cong \quad R/(d_1) \oplus R/(d_2) \oplus \cdots \oplus R/(d_\ell) \oplus R^k \ .$$

[Here the ideal in R generated by d is denoted (d), and is a submodule of R by **RV ix)**.]

Theorem 43.2. *With notation as in* **43.1**, *if also we have non-negative integers j and m, and elements $e_1 \mid \cdots \mid e_m$, with no e_i zero or invertible, such that*

$$R/(e_1) \oplus \cdots \oplus R/(e_m) \oplus R^j \quad \cong \quad R/(d_1) \oplus \cdots \oplus R/(d_\ell) \oplus R^k \quad ,$$

then $j = k$, $\ell = m$, and, for all i, the elements d_i and e_i are associates in R.

We may paraphrase these theorems as saying that every finitely generated module is 'uniquely' a direct sum of cyclic modules. Combining these two theorems, we have a *classification* for finitely generated modules M over a Euclideanizable domain (which actually holds more generally over any PID): each such M corresponds to an *invariant*

$$(k \ ; \ [d_1] \ , \ \cdots \ , \ [d_\ell]) \ ,$$

where $[d]$ denotes the 'associate class' of d in R. The set of isomorphism classes of such modules is thus in 1-1 correspondence with the set of sequences which can serve as an invariant. Let us re-emphasize that possibly k or ℓ is zero, that no d_i is zero or invertible, and that d_i divides d_{i+1} for each $i < \ell$.

This section will give efficient proofs of these theorems, proofs which don't reveal in any very direct way how to calculate the invariant. An algorithm for doing that is discussed in the next section.

Proof of Theorem 43.1.

If $M = \{0\}$, the theorem holds with $\ell = k = 0$, using the usual conventions. Proceed by induction on $n :=$ the minimum number of generators for M, to prove the slightly stronger statement that

$$\textit{the theorem holds with, in addition, } \; k + \ell \; = \; n \; .$$

Start the induction with $n = 0$ for which $M = \{0\}$ as above. [If this makes you nervous—it shouldn't—, specialize and simplify the following proof to $n = 1$, and it gives a proof for an initial case with $n = 1$.]

For the inductive step, take $n > 0$ and fix M such that M can be generated by some n-element set, but not by any $(n-1)$-element set. Define G to be the following subset of R :

$$G \; := \; \{ \; r \in R \; : \; \exists \{m_1, \cdots, m_n\} \text{ generating } M \; , \\ \text{and } \{r_2, \cdots, r_n\} \subset R, \text{ with } rm_1 + \sum_{i>1} r_i m_i = 0 \; \} \; .$$

Note that *no element of G is invertible*, since the minimality of n would be contradicted if we could express m_1 in terms of the other generators: $m_1 \; = \; -\, r^{-1} \sum_{i>1} r_i m_i$. Clearly $0 \in G$.

Case 1. Assume that $G = \{0\}$:
Pick a set $\{ \; m_1, \; \cdots, \; m_n \; \}$ generating M, and define $\theta : R^n \to M$ by

$$\theta(r_1, \; \cdots, \; r_n) \; = \; r_1 m_1 + \cdots + r_n m_n \; .$$

It is straightforward to check that θ is a module morphism. It is surjective by the definition of the statement: '$\{ \; m_1, \; \cdots, \; m_n \; \}$ generates M'. It is injective because $G \; = \{0\}$: [If $(r_1, \cdots, r_n) \in \text{Ker}\theta$, then each $r_i \in G$, since we can re-order the generating set as

$$\{ \; m_i, \; m_1, \; \cdots \; , \; m_{i-1}, \; m_{i+1}, \; \cdots \; , \; m_n \; \} \; .$$

But $r_i \in G$ implies $r_i = 0$, so $\text{Ker}\theta = \{(0, 0, \cdots, 0)\}$, as required.] Thus θ is an isomorphism, so that the theorem holds for M with $\ell = 0$ and $k = n$.

Case 2. Assume that $G \neq \{0\}$:
Let $\delta : R \setminus \{0\} \to \mathbf{N}$ be a fixed function such that (R, δ) is a Euclidean domain. Now choose d_1 so that $d_1 \neq 0$ and $d_1 \in G$, with $\delta(d_1) \leq \delta(r)$ for

all non-zero $r \in G$. [The crucial thing is that $\leq$ is a *discrete* linear order on $\mathbf{N}$. We couldn't do this if we only knew that δ took values in the non-negative rationals or reals, for example.]

Claim i): For all $\{ m_1, \cdots, m_n \}$ generating M, if

$$d_1 m_1 + \sum_{i>1} r_i m_i \;=\; 0 \;,$$

then d_1 divides r_j for all $j > 1$.

Proof. Let $r_j = qd_1 + s$, where either $s = 0$ or $\delta(s) < \delta(d_1)$ (using the division algorithm). Then

$$\begin{aligned} 0 \;&=\; d_1 m_1 + (qd_1 + s)m_j + \sum_{i \neq 1 \text{ or } j} r_i m_i \\ &=\; s m_j + d_1(m_1 + q m_j) + \sum_{i \neq 1 \text{ or } j} r_i m_i \;. \end{aligned}$$

But

$$B \;:=\; \{ m_j,\; m_1 + q m_j,\; m_2,\; \cdots,\; m_{j-1},\; m_{j+1},\; \cdots,\; m_n \}$$

is an n-element set which generates M: [It suffices to express each m_i as a linear combination from B—this is trivial for $i > 1$, and for $i = 1$ we have $m_1 = (1)(m_1 + qm_j) + (-q)(m_j)$.] Thus $s \in G$. But $s \neq 0$ and $\delta(s) < \delta(d_1)$ would contradict the definition of d_1, so $s = 0$, as required.

Claim ii): For all $\{ m_1, \cdots, m_n \}$ generating M, if

$$d_1 m_1 + \sum_{i>1} r_i m_i \;=\; 0 \;=\; \sum_{i \geq 1} r'_i m_i \quad ,$$

then d_1 divides r'_1.

Proof. Write $r'_1 = q' d_1 + s'$ where either $s' = 0$ or $\delta(s') < \delta(d_1)$. Then

$$0 = 0 - q'0 = \sum_{i \geq 1} r'_i m_i - q' \left(d_1 m_1 + \sum_{i>1} r_i m_i \right) = s' m_1 + \sum_{i>1} (r'_i - q' r_i) m_i \;.$$

Thus $s' \in G$, so $\delta(s') < \delta(d_1)$ is not possible, as required.

Now, using **i)**, fix some n-element set $\{ m_1, \cdots, m_n \}$ generating M, and a relation

$$d_1 (m_1 + \sum_{i>1} t_i m_i) \;=\; 0$$

for some $t_2, \cdots, t_n$ in R. Define

$$\bar{m_1} \;=\; m_1 + \sum_{i>1} t_i m_i \ .$$

Claim iii): *The set* $A \;:=\; \{\ \bar{m}_1,\ \ m_2,\ \ m_3,\ \ \cdots,\ m_n\ \}$ *generates* M.

Proof. It suffices to express each m_i as a linear combination from A. This is trivial for $i > 1$, since then $m_i \in A$. For $i = 1$, just note that

$$m_1 \;=\; \bar{m}_1 - t_2 m_2 - t_3 m_3 - \cdots - t_n m_n \ .$$

Let $< \bar{m}_1 >$ be the (cyclic) submodule of M generated by $\bar{m}_1$; that is,

$$< \bar{m}_1 > \;\; := \;\{\ r\bar{m}_1 \ :\ r \in R\ \}\ .$$

Claim iv): *The set*

$$\{\ m_2 + < \bar{m}_1 >,\ m_3 + < \bar{m}_1 >,\ \cdots\ ,\ m_n + < \bar{m}_1 >\ \}$$

generates the quotient module $M/ < \bar{m}_1 >$.

Proof. Given $m \in M$, use **iii)** to choose $r_i \in R$ such that

$$m \;=\; r_1 \bar{m}_1 + \sum_{i>1} r_i m_i \ .$$

Then

$$\begin{aligned} m + < \bar{m}_1 > \;&=\; \left(\sum_{i>1} r_i m_i + r_1 \bar{m}_1 \right) + < \bar{m}_1 > \\ &=\; \left(\sum_{i>1} r_i m_i \right) + < \bar{m}_1 > \\ &=\; \sum_{i>1} r_i \left(m_i + < \bar{m}_1 > \right)\ , \qquad \text{as required.} \end{aligned}$$

Claim v): *We have* $M/ < \bar{m}_1 > \;\; \cong \;\; R/(d_2) \bigoplus \cdots \bigoplus R/(d_\ell) \bigoplus R^k$ *for some* $k \geq 0$, $\ell - 1 \geq 0$, *non-zero non-invertible* $d_2 \mid d_3 \ \cdots \ \mid d_\ell$ *such that* $k + \ell - 1 \leq n - 1$.

We'll soon see that $=$ holds, not $<$, in the last inequality.

Proof. By **iv)**, $M/ < \bar{m}_1 >$ can be generated by "$n-1$" elements, so this is immediate from the inductive hypothesis.

Claim vi): $< \bar{m}_1 > \ \cong \ R/(d_1)$.
Probably (d_1) should be written $< d_1 >$ for consistency, but earlier in field theory we always used (d_1) when thinking of it as an ideal rather than as a submodule.

Proof. Define $\phi : R \to \ < \bar{m}_1 >$ by $\phi(r) = r\bar{m}_1$. Then ϕ is easily seen to be a module morphism. It is surjective by the definition of $< \bar{m}_1 >$. Now $d_1\bar{m}_1 = 0$ by definition of $\bar{m}_1$, so $d_1 \in \text{Ker}\phi$, and thus $(d_1) \subset \text{Ker}\phi$. But

$$r \in \text{Ker}\phi \implies 0 = r\bar{m}_1 = r\bar{m}_1 + 0m_2 + 0m_3 + \cdots + 0m_n$$

$$\implies d_1 \mid r \quad \text{by } \mathbf{ii)} \qquad \implies \ r \in (d_1) \ .$$

Thus $\text{Ker}\phi \subset (d_1)$, giving $\text{Ker}\phi = (d_1)$; and so

$$R/(d_1) \ = \ R/\text{Ker}\phi \ \cong \ \text{Im}\phi \ = \ < \bar{m}_1 > \ ,$$

as required.

Claim vii): $M \ \cong \ < \bar{m}_1 > \ \oplus \ (M/ < \bar{m}_1 >)$.
Proof. Given $m \in M$, write

$$m \ = \ r_1\bar{m}_1 + \sum_{i>1} r_i m_i \ ,$$

using **iii)**. Define $\psi : M \ \longrightarrow \ < \bar{m}_1 > \ \oplus \ (M/ < \bar{m}_1 >)$ by

$$\psi(m) \ := \ (r_1\bar{m}_1 \ , \ m + < \bar{m}_1 >) \ .$$

To show that this is well defined, suppose that

$$r_1\bar{m}_1 + \sum_{i>1} r_i m_i \ = \ r_1'\bar{m}_1 + \sum_{i>1} r_i' m_i \ .$$

Then $(r_1 - r_1')\bar{m}_1 + \sum_{i>1}(r_i - r_i')m_i = 0$. By **ii)** and **iii)**, $d_1 \mid (r_1 - r_1')$. Thus $(r_1 - r_1')\bar{m}_1 = 0$ since $d_1\bar{m}_1 = 0$. So $r_1\bar{m}_1 = r_1'\bar{m}_1$, showing that the first component in the definition of ψ is well-defined, as required. Also ψ is easily seen to be a module morphism. It is surjective, by the definition of $< \bar{m}_1 >$.

Finally, to prove that ψ is injective, suppose that $m = r_1\bar{m}_1 + \sum_{i>1} r_i m_i$ is in Kerψ. Then

$$(0,0) \;=\; \psi(m) \;=\; (r_1\bar{m}_1 \;,\; m + <\bar{m}_1>) \;,$$

so $r_1\bar{m}_1 = 0$ and $\sum_{i>1} r_i m_i \;= m \;\in\; <\bar{m}_1>$. Writing $\sum_{i>1} r_i m_i$ as $r_0\bar{m}_1$, we see that $d_1 \mid (-r_0)$ by **ii).** Therefore $r_0\bar{m}_1 = 0$, that is, $\sum_{i>1} r_i m_i = 0$. Thus $m = 0$.

Claim viii): $M \;\cong\; R/(d_1) \oplus \cdots \oplus R/(d_\ell) \oplus R^k$,
for d_1 as defined at the beginning of **Case 2**, *and for k and $d_2, \cdots, d_\ell$ as in* **v)**.

Proof. This is immediate, combining **vii), vi)** and **v)**.

Claim ix): *d_1 is non-zero and non-invertible, $k + \ell = n$, and $d_1 \mid d_2$ if it happens that $\ell > 1$.*

Proof. The first statement is clear since G contains no invertible elements. Let $\theta : M \to D$ be any isomorphism as in **viii)**, where D is the direct sum in **viii)**. Define elements as follows, where the right-hand sides are non-zero in the i^{th} slot:

$$u_1, \cdots, u_\ell \;\text{ in } M \;, \quad \text{by } \;\theta(u_i) \;:=\; (0, \cdots, 0, 1 + (d_i), 0, \cdots, 0) \;,$$

and

$$v_{\ell+1}, \cdots, v_{\ell+k} \;\text{ in } M \;, \quad \text{by } \;\theta(v_i) \;:=\; (0, \cdots, 0, 1, 0, \cdots, 0) \;.$$

Then $\{\; u_1, \;\cdots,\; u_\ell,\; v_{\ell+1},\; \cdots,\; v_{\ell+k} \;\}$ generates M since it maps under the isomorphism θ into a set of generators for D. Thus $k+\ell \geq n$, since M cannot be generated by fewer than "n" elements. By **v)**, $k + \ell \leq n$, so $k + \ell = n$, as required. Also if $\ell \geq 1$, then

$$d_1(1 + (d_1), 0, \cdots, 0) \;=\; 0_D \;=\; d_2(0, 1 + (d_2), 0, \cdots, 0) \;.$$

Thus, applying θ^{-1} , we see that $d_1 u_1 = 0_M = d_2 u_2$. Thus
$d_1 u_1 + d_2 u_2 + 0 + \cdots + 0 v_{\ell+k} \;=\; 0$. Hence $d_1 \mid d_2$ by **i)**, as required.

Since $d_2 \mid d_3 \;\cdots$ by **v)**, the last two claims complete the induction, and so, the proof of Theorem **43.1**.

Proof of Theorem 43.2.
This will be presented as a sequence of definitions and routine verifications.

Definition. The *torsion submodule*, T_M, of a module M is

$$T_M \;:=\; \{\; m \in M \;:\; rm = 0 \;\text{ for some }\; r \neq 0 \;\}\;.$$

RV i): *T_M is a submodule of M.*

RV ii): *If $\theta : M \to N$ is a module isomorphism, then $\theta(T_M) = T_N$, so that $T_M \cong T_N$. Furthermore,*

$$M/T_M \;\cong\; N/T_N \qquad \text{via} \qquad m + T_M \;\mapsto\; \theta(m) + T_N\;.$$

RV iii): *If D is the direct sum in* **43.1**, *then*

$$T_D \;\;\cong\;\; R/(d_1) \oplus \cdots \oplus R/(d_\ell) \quad \text{and} \quad D/T_D \;\cong\; R^k\;.$$

Combining **iii)** [and its analogue for the other direct sum in **43.2**] with **ii)**, the hypothesis of **43.2** yields

RV iv): $R/(d_1) \;\oplus \cdots \oplus\; R/(d_\ell) \;\;\cong\;\; R/(e_1) \;\oplus \cdots \oplus\; R/(e_m)$

and

RV v): $R^k \;\cong\; R^j\;.$

This has split the proof into two parts: a 'torsion' part, **iv)**; and a 'free' part, **v)**, which we deal with first.

RV vi): *Let Q_R be the field of fractions of R, and let $\theta : R^j \to R^k$ be an isomorphism of R-modules from* **v)**. *Then $\phi : Q_R^j \to Q_R^k$, given by $\phi(a_1/b,\; \cdots,\; a_j/b) := (1/b)\theta(a_1, \cdots, a_j)$, is an isomorphism of vector spaces over Q_R.*

It now follows from **v), vi)**, and elementary linear algebra that $j = k$, as required, since Q_R^j and Q_R^k are isomorphic vector spaces, and so they have the same dimension.

Now let p be any irreducible in R. Thus $R/(p)$ [as a ring] is in fact a field. For any non-negative integer α and R-module M, let

$$p^\alpha M \ := \ \{ \ p^\alpha m \ : \ m \in M \ \} \ .$$

RV vii): *$p^{\alpha+1}M$ is a submodule of $p^\alpha M$, itself a submodule of M.*

RV viii): *The abelian group $p^\alpha M \, / \, p^{\alpha+1}M$ is an $R/(p)$-vector space, using the scalar multiplication*

$$(r + (p)) \left(m + p^{\alpha+1}M \right) \ := \ rm + p^{\alpha+1}M \ .$$

The main point is to check that this is well-defined.

RV ix): *If $M \cong N$ as R-modules, then $p^\alpha M/p^{\alpha+1}M \ \cong \ p^\alpha N/p^{\alpha+1}N$ as $R/(p)$-vector spaces.*

Lemma x): *For $M = R/(d)$, the dimension over $R/(p)$ of the vector space $p^\alpha M/p^{\alpha+1}M$ is*

$$\begin{cases} 1 & \text{if} \quad p^{\alpha+1} \mid d \ \text{ in } R \ ; \\ 0 & \text{if not} \end{cases} \qquad .$$

(For example, when $R = \mathbf{Z}$, we have $2\mathbf{Z}_{12}/4\mathbf{Z}_{12} \ \cong \ \mathbf{Z}_2$ whereas $4\mathbf{Z}_{12}/8\mathbf{Z}_{12} \ \cong \ \{0\}$.)

Proof. Let $\lambda_0 := p^\alpha + (d) = p^\alpha(1 + (d)) \in p^\alpha M$. For any other element $\lambda = p^\alpha b + (d)$ of $p^\alpha M$, clearly $\lambda = b\lambda_0$, and so

$$\lambda + p^{\alpha+1}M \ = \ (b + (p)) \left(\lambda_0 + \ p^{\alpha+1}M \right) \ .$$

Thus $\lambda_0 + p^{\alpha+1}M$ spans $p^\alpha M/p^{\alpha+1}M$ over $R/(p)$. It remains to prove that $\lambda_0 + p^{\alpha+1}M$ is non-zero if and only if $p^{\alpha+1} \mid d$ in R. Now

$$[\lambda_0 + p^{\alpha+1}M \text{ is zero}] \iff \left[p^\alpha + (d) = p^{\alpha+1}c + (d) \text{ for some } c \in R \right]$$

$$\iff \ [p^\alpha(1 - pc) \ \in (d) \text{ for some } c \in R] \ .$$

The last statement implies that $p^{\alpha+1}$ does not divide d, since $p^\alpha(1 - pc) \ = de$ would otherwise contradict unique factorization, the left-hand side not being divisible by $p^{\alpha+1}$.

Conversely, if $p^{\alpha+1}$ does not divide d, we must prove the existence of c and e in R with $p^\alpha(1-pc) = de$. Write this as $d'e + p^{\alpha+1-\mu}c = p^{\alpha-\mu}$, where $d = p^\mu d'$ with GCD $\{ p, d' \} = 1$. It is clear that such e and c can be found, since the equation $d'x + p^{\alpha+1-\mu}y = 1$ has a solution (x, y) in $R \times R$.

RV xi): *As $R/(p)$-vector spaces,*

$$p^\alpha(M_1 \bigoplus M_2)/p^{\alpha+1}(M_1 \bigoplus M_2) \quad \cong \quad (p^\alpha M_1/p^{\alpha+1}M_1) \bigoplus (p^\alpha M_2/p^{\alpha+1}M_2) \ .$$

The proof of Theorem 43.2 is now completed as follows:

Consider $p^\alpha L \,/\, p^{\alpha+1}\ L$, where L is the left-hand side in **iv)**. Using **x)** and the iteration of **xi)** to any finite number of summands, its dimension over $R \,/\, (p)$ is the number of i for which $p^{\alpha+1}$ divides d_i. By **iv)** and **ix)** this must equal the number of i for which $p^{\alpha+1}$ divides e_i. Since this is true for all irreducibles p and non-negative integers α, the divisibility conditions $d_i \mid d_{i+1}$ and $e_i \mid e_{i+1}$ and unique factorization in R show that $\ell = m$ and $d_i \sim e_i$ for all i, since these "numbers of i" referred to above clearly determine the prime power factorizations of all d_i and e_i up to an invertible factor. (See also the definition of *elementary divisors* at the end of Section **44**).

Remark. The proof of **43.2** evidently applies when R is any PID, (any UFD??), although **43.1** fails in general for UFD's.

Exercise 43A. Justify the last remark by an example. Hint: Try taking the ring R to be $\mathbf{Z}[x]$.

(a) Take M as the quotient $(R \bigoplus R)/D$, where D is the cyclic submodule generated by $(2,\ x)$. Show that:

i) M has no torsion (i.e. $rm = 0$ implies $r = 0$ or $m = 0$ for $r \in R$ and $m \in M$);

ii) M is not cyclic;

iii) the module defined in exactly the same way, except that $R = \mathbf{Q}[x]$, *is* cyclic.

Assuming that M were a direct sum of quotients of R, deduce from i) that all the summands are isomorphic to R, from ii) that the number of summands is greater than 1, and from iii) that the number of summands is less than 2.

(b) Perhaps simpler is just to take M to be the ideal generated by $\{2,\ x\}$, as a submodule of $\mathbf{Z}[x]$.

(c) Now show that the modules in parts (a) and (b) are actually isomorphic (so your hard work in one of the previous two parts was unnecessary).

Exercise 43B. Can you generalize **43A**(b) to find an example (at least) for any UFD which has a non-principal ideal generated by two elements?

ADDENDUM. An alternative proof that iv) and the divisibility conditions imply that $\ell = m$ and $d_i \sim e_i$ for all i.

Definition. 'The' *annihilator,* $\mathrm{ann}(M)$, of an R-module M, is any generator for the ideal

$$\{\ r \in R\ :\ rm = 0 \text{ for all } m \in M\ \}$$

(using that R is a PID).

RV xii): *This* is *an ideal*, and so, since R is a PID, *the element ann(M) is well-defined up to associates.*

RV xiii): *If $M \cong N$ as R-modules, then ann(M) and ann(N) are associates.*

RV xiv): *If L is the left-hand side in* **iv)**, *then* $\mathrm{ann}(L) \sim d_\ell$.
(Use the fact that $d_i \mid d_{i+1}$).

It follows from **iv), xiii)** and **xiv)** that $d_\ell \sim e_m$, as required. Now let E be the right-hand side in **iv)**, and let $\theta : L \to E$ be a module isomorphism. Let $m_0 \in L$ be any element for which $rm_0 = 0$ implies that $d_\ell \mid r$ [for example, m_0 may be any element whose last component is $1 + (d_\ell)$]. Let $< m_0 > \ := \ \{\ rm_0\ :\ r \in R\ \}$, the cyclic submodule generated by m_0.

RV xv): *We have* $L/ < m_0 > \ \cong \ E/ < \theta(m_0) >$ *as modules, via*

$$m + < m_0 > \ \mapsto \ \theta(m) + < \theta(m_0) > \ .$$

Lemma xvi): $L/ < m_0 > \ \cong \ R/(d_1) \bigoplus \cdots \bigoplus R/(d_{\ell-1})$.

Given this lemma, the proof is completed by induction on ℓ: combining **xvi)** [and its analogue for $E/ < \theta(m_0) >$] with **xv)** yields

$$R/(d_1)\bigoplus\cdots\bigoplus R/(d_{\ell-1}) \;\cong\; R/(e_1)\bigoplus\cdots\bigoplus R/(e_{m-1}) \ .$$

Thus, by the inductive hypothesis, $\ell - 1 = m - 1$ and $d_i \sim e_i$ for $1 \leq i < \ell$, as required.

To prove the lemma, we use the method of the next section. Letting $m_i = (0, \cdots, 0, 1 + (d_i), 0, \cdots, 0)$, the module L has generators $\{m_1, \cdots, m_\ell\}$ and relations $d_i m_i = 0$ for $1 \leq i \leq \ell$. Thus $L/ < m_0 >$ also has "ℓ" generators with the 'same' relations plus an extra one given by $m_0 + < m_0 >= 0$. Thus the matrix C giving the relations is

$$\begin{pmatrix} d_1 & & & & & \\ & d_2 & & & 0 & \\ & & \cdot & & & \\ & & & \cdot & & \\ & 0 & & & \cdot & \\ & & & & & d_\ell \\ b_1 & b_2 & \cdot & \cdot & \cdot & b_\ell \end{pmatrix} \quad [\text{of shape } (\ell+1) \times \ell \ ],$$

where $m_0 = b_1 m_1 + b_2 m_2 + \cdots + b_\ell \ m_\ell$. But the possible choices of m_0 are exactly those for which

$$\mathrm{GCD}\{d_1 d_2 \ldots d_\ell \ , b_1 d_2 \ldots d_\ell \ , d_1 b_2 \ldots d_\ell \ , \cdots , d_1 d_2 \ldots d_{\ell-1} b_\ell\} \sim d_1 d_2 \ldots d_{\ell-1} \ .$$

(For this proof it is not enough to simply take $m_0 = m_\ell$, that is, $b_\ell = 1$ and all other $b_i = 0$, because we need the analogue for E.) But then the above matrix is easily reduced by row/column operations to

$$\begin{pmatrix} 1 & & & & & \\ & d_1 & & & 0 & \\ & & \cdot & & & \\ & & & \cdot & & \\ & 0 & & & \cdot & \\ & & & & & d_{\ell-1} \\ 0 & 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix} ,$$

as required.

Note that this argument can be re-written so that it does not use anything from Section **44**. On the other hand, nothing from Section **44** which was just used depends at all on the theorem, **43.2**, which was being partially reproved.

44. Generators, relations and elementary operations.

We shall continue to assume that R is a Euclideanizable domain, although the first few paragraphs apply to any commutative ring. Let $C = (c_{ij})$ be a matrix in $R^{p\times n}$, the set of $p \times n$ matrices with entries in R. Define an R-module:

$$M_C \quad := \quad R^n \ / \ < \sum_{j=1}^{n} c_{ij} e_j : 1 \leq i \leq p > \quad ,$$

where $e_j = (0, \cdots, 0, 1, 0, \cdots, 0) \in R^n$, and the denominator is the submodule generated by the given ("p") elements. Thus $M_C = R^n/\mathrm{Im}\theta_C$, where $\theta_C : R^p \to R^n$ is the module morphism uniquely determined by requiring that

$$\theta_C(f_i) \ = \ \sum_j c_{ij} e_j \ , \quad \text{where } f_i := (0, \cdots, 1, \cdots, 0) \in R^p \ .$$

Remark. Any module M which is isomorphic to M_C, via some isomorphism which maps $e_j + \mathrm{Im}\theta_C$ to (say) $m_j \in M$ for each j, will be said to be 'given by generators $m_1, \cdots, m_n$ and relations $\sum_j c_{ij} m_j = 0$'. This is equivalent to the existence of a module surjection $R^n \to M$, mapping e_j to m_j, with kernel equal to $\mathrm{Im}\theta_C$. In M, the m_i generate, and the relations hold; but this also captures the more subtle idea that *all other relations in M follow from the given ones.* For a brief discussion of the analogous matters relating to (non-commutative) groups, see the last paragraphs of Section **11**.

Proposition 44.1. *If also $D \in R^{p\times n}$, and if there exist matrices P, invertible in $R^{p\times p}$, and Q, invertible in $R^{n\times n}$, such that $PCQ = D$, then $M_C \ \cong \ M_D$.*

Proof. First one checks that $\theta_{C'C} = \theta_C \circ \theta_{C'}$ whenever the matrix product $C'C$ is defined, by evaluating both sides on the 'standard basis' generators, and using that both maps are module morphisms.

Thus the diagram

$$\begin{array}{ccc} R^p & \xrightarrow{\theta_C} & R^n \\ \theta_P \uparrow & & \downarrow \theta_Q \\ R^p & \xrightarrow{\theta_D} & R^n \end{array}$$

is commutative. Furthermore, the vertical arrows denote isomorphisms since, for example, $\theta_P \circ \theta_{P^{-1}} = \theta_{P^{-1}P} = \theta_I$, which is the identity map of R^p. It follows that θ_Q maps $\mathrm{Im}\theta_C$ isomorphically onto $\mathrm{Im}\theta_D$, and so determines an isomorphism

$$M_C = R^n/\mathrm{Im}\theta_C \quad \longrightarrow \quad R^n/\mathrm{Im}\theta_D = M_D$$

by

$$v + \mathrm{Im}\theta_C \quad \mapsto \quad \theta_Q(v) + \mathrm{Im}\theta_D \ ,$$

as required.

Exercise 44A. Formulate and prove a converse to **44.1**. (This won't be used below.)

Proposition 44.2. *If, for integers m, ℓ, p and n, the $p \times n$ diagonal matrix D, beginning with "m" 1's, is*

$$\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & 0 & \\ & & 1 & & & & \\ & & & d_1 & & & \\ & & 0 & & \ddots & & \\ & & & & & d_\ell & \\ & & & & & & 0 \end{pmatrix},$$

for $d_i \in R$, then

$$M_D \quad \cong \quad R/(d_1) \oplus \cdots \oplus R/(d_\ell) \oplus R^{n-m-\ell} .$$

Corollary 44.3. *If (as in* **44.1***) $PCQ = D$, for D as in* **44.2** *with non-zero non-invertible d_i such that $d_i \mid d_{i+1}$ for all $i < \ell$, then the invariant of M_C is $(n - m - \ell \ ; \ [d_1], \cdots, [d_\ell])$.*

Proof of 44.2. Again we use the '**RV**-style'; most of the **RV**s below are nearly tautological, particularly **iv)** and **v)**.

Definition. Given module morphisms $\theta_i : M_i' \to M_i$ for $1 \leq i \leq n$, define

$$\theta_1 \oplus \cdots \oplus \theta_n \ : \ M_1' \oplus \cdots \oplus M_n' \ \longrightarrow \ M_1 \oplus \cdots \oplus M_n$$

as usual (coordinatewise) by $(m_1', \cdots, m_n') \ \mapsto \ (\theta_1 m_1', \cdots, \theta_n m_n')$.

RV i): *Im*$(\theta_1 \oplus \cdots \oplus \theta_n) \ = \ (\mathrm{Im}\theta_1) \oplus \cdots \oplus (\mathrm{Im}\theta_n)$ *, so*

$$(M_1 \oplus \cdots \oplus M_n)/\mathrm{Im}(\theta_1 \oplus \cdots \oplus \theta_n) \quad \cong \quad (M_1/\mathrm{Im}\theta_1) \oplus \cdots \oplus (M_n/\mathrm{Im}\theta_n)$$

via the map: coset of $(m_1, \cdots, m_n) \ \mapsto \ (m_1 + \mathrm{Im}\theta_1, \cdots, m_n + \mathrm{Im}\theta_n)$.

RV ii): *If D' is the same as D, except that it is a square matrix of shape $n \times n$, then θ_D and $\theta_{D'}$ have the same image.*

RV iii): *For D' as in* **ii)**, *we have* $\theta_{D'} = \theta_1 \bigoplus \cdots \bigoplus \theta_n$, *where*

$$\theta_i : R \to R \text{ is multiplication by } \begin{cases} 1 & \text{if } 1 \le i \le m ; \\ d_i & \text{if } m+1 \le i \le m+\ell ; \\ 0 & \text{if } i > m+\ell . \end{cases}$$

RV iv): *If* $\theta : R \to R$ *is multiplication by* d, *then* $\mathrm{Im}\theta = (d)$.

RV v): $R/(1) \cong \{0\}$; $R/(0) \cong R$; $\{0\} \bigoplus M \cong M$.

Now the proof is finished as follows:

$$\begin{aligned} M_D \;:=\; R^n/\mathrm{Im}\theta_D \;&=\; R^n/\mathrm{Im}\theta_{D'} && \text{by } \mathbf{ii}) \\ &=\; R^n/\mathrm{Im}(\theta_1 \bigoplus \cdots \bigoplus \theta_n) && \text{by } \mathbf{iii}) \\ &\cong\; (R/\mathrm{Im}\theta_1) \bigoplus \cdots \bigoplus (R/\mathrm{Im}\theta_n) && \text{by } \mathbf{i}) \\ &=\; (R/(1))^m \bigoplus R/(d_1) \bigoplus \cdots \bigoplus R/(d_\ell) \bigoplus (R/(0))^{n-m-\ell} && \text{by } \mathbf{iii}),\ \mathbf{iv}) \\ &\cong\; R/(d_1) \bigoplus \cdots \bigoplus R/(d_\ell) \bigoplus R^{n-m-\ell} && \text{by } \mathbf{v}) \ . \end{aligned}$$

Assertion 44.4. *For each C, there exist P, Q and D as in Corollary* **44.3**. *Rather stronger: there is an 'effective method' (or algorithm) to actually calculate P, Q and D.*

Remarks. The matrix D in **44.3** is sometimes called the *Smith normal form*; and applying an algorithm as below is called *reduction* to Smith normal form.

Before giving an informal description of the algorithm, we relate this to the approach used in several texts on this subject. Given M with generators $\{ m_1, \cdots, m_n \}$, there is obviously a module surjection $R^n \to M$ determined by $e_j \mapsto m_j$. What is not obvious (but is true) is that the kernel of this surjection is finitely generated, say by "p" elements of R^n (stronger: even with $p \le n$). In fact any submodule of a finitely generated module over a PID is itself finitely generated.

Exercise 44B. Prove this, using only the fact that ideals are finitely generated (i.e. R is *Noetherian*)—the case of R^n is the crucial one, since the general case will follow by mapping the free module onto it. (There is a

proof in **Artin**, p.469.)
It follows that $M \cong M_C$ for some matrix C (in fact, many different C); one says that every finitely generated module over a PID is 'finitely presentable'. From this, **44.1, 44.2** and **44.4** evidently give a second proof of Theorem **43.1**. With a little more work, one can also give a 'matrix proof' of Theorem **43.2**; that is, given C, the matrix D in **44.3** is (up to associates) unique (but P and Q are not). Actually, the logic can be reversed to deduce the above from **43.1** and **43.2**.

In many texts, the above approach is used. We chose to separate out the 'effective methods', proofs of whose existence are not necessarily the best proofs of **43.1** and **43.2**. Often the proofs given of the existence and correctness of the algorithm are only part way towards a proof acceptable in computation theory.

Exercise 44C. Show that, when R is a field, the existence of Smith normal form is equivalent to the fact that any linear transformation $V \to W$ between finite dimensional R-vector spaces can be represented by a matrix of the form $\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$. (When $V = W$, one is of course allowed different bases for the domain and codomain.)

Informal description of an algorithm for 44.4. As in linear algebra, there are three types of row operation on matrices in $R^{p\times n}$:
I_r. Interchange two rows.
II_r. Add a multiple of one row to a different row.
III_r. Multiply a row by an invertible in R.

These are reversible, as are the analogous column operations, denoted I_c, II_c, III_c. The following assertion is then proved as in linear algebra over a field: *for all C and D, there exist invertible P, Q with $PCQ = D$ if and only if some sequence of row/column operations leads from C to D.*

The procedure, for converting a given C by r/c operations to a D as in **44.2**, goes by induction on size (as in linear algebra when R is a field). It suffices to show how to get a top left entry which divides all other entries; then one 'kills' the rest of the first row and column, and proceeds (south-east!) inductively.

To do this, pick a Euclidean function δ on R. It suffices show that if

a non-zero matrix entry of minimum δ does not divide some other entry, then certain operations will produce a matrix with smaller such minimum δ. Iteration of this must terminate eventually, at which point a non-zero entry of minimum δ dividing all other entries can be moved the top left, using the operations I_r and I_c. Assume then that T has an entry t_{ij} with $\delta(t_{ij}) \leq \delta(t_{k\ell})$ for all $t_{k\ell} \neq 0$, but t_{ij} fails to divide some entry of T.

i) If it doesn't divide an entry in its own row, say $t_{i\ell}$, write this entry $t_{i\ell} = qt_{ij} + r$ where $r \neq 0$ and $\delta(r) < \delta(t_{ij})$. Then subtracting q times the j^{th} column from the ℓ^{th} has the desired result of lowering the minimum value of δ.

ii) If t_{ij} doesn't divide an entry in its own column, just interchange "row" and "column" in i).

iii) Finally suppose that t_{ij} divides all entries in its own row and column, but does not divide $t_{k\ell}$. Let $t_{i\ell} = rt_{ij}$. Add $(1 - r)$ times the j^{th} column to the ℓ^{th} column. Perhaps this decreases the minimum δ, as required. If not, note that the new $(i, \ell)^{th}$ entry is t_{ij} and the new $(k, \ell)^{th}$ entry is congruent to $t_{k\ell}$ modulo t_{ij}. Thus we are now in case ii) above, with (i, j) replaced by (i, ℓ).

This completes the sketch of the algorithm. Programming and increasing efficiency may be safely left to an underling.

Exercise 44D. As an example with some applicability, the reader can easily show how to convert

$$\begin{pmatrix} e & 0 \\ 0 & f \end{pmatrix} \quad \text{to} \quad \begin{pmatrix} \text{GCD}\{e, f\} & 0 \\ 0 & \text{LCM}\{e, f\} \end{pmatrix}$$

for non-zero e and f in R. (Recall that GCD $\{e, f\}$ can be written as $se+tf$, and that these matrices have the same determinant.)

It follows that

$$R/(e) \oplus R/(f) \quad \cong \quad R/(\text{GCD}\{e, f\}) \oplus R/(\text{LCM}\{e, f\}) \ .$$

This may be iterated to convert any direct sum of modules $R\,/\,(d)$ to one in the *invariant factor form* of Theorem **43.1**. This also shows that if $d = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_s^{\alpha_s}$ is the unique factorization of d in R (with p_i distinct), then

$$R/(d) \quad \cong \quad R/(p_1^{\alpha_1}) \oplus \cdots \oplus R/(p_s^{\alpha_s}) \ .$$

Applying this to each d_j occurring in the invariant of M, we obtain a second canonical decomposition of M, as a direct sum of modules $R/(p^\alpha)$ plus a 'free'

part R^k. This is called the *elementary divisor form* of M. It is unique except that there is no natural way to choose an ordering for the multiset of powers p^α of irreducibles. There is a straightforward way to recover the invariant factors d_j from the elementary divisors p^α: the last one, d_ℓ, is the product, over all p which occur, of the highest power of p occurring; deleting these, $d_{\ell-1}$ is obtained in the same way from the remaining elementary divisors, and so on. This demonstrates the uniqueness of the multiset of elementary divisors as a consequence of the uniqueness of the sequence of invariant factors. It is also perhaps the simplest way to see how to convert an arbitrary direct sum of modules of the form $R/(d)$ to invariant factor form: one first produces the elementary divisor form by factoring each such d, and then passes to the invariant factor form as above.

45. Finitely generated abelian groups revisited.

If $(\ A\ ,\ +\ )$ is an abelian group, then a $\mathbf{Z}$-module, $(\ A\ ,\ +\ ,\ \cdot\ )$, can be defined by specifying

$$0_{\mathbf{Z}} \cdot a \ := \ 0_A \qquad\qquad \text{and, for } n > 0 \ ,$$

$$n \cdot a \ := \ a + (n-1) \cdot a \qquad \text{[inductively]}$$

and

$$(-n) \cdot a \ := \ -\,(n \cdot a) \ .$$

It is elementary to check that this gives a well defined $\mathbf{Z}$-module, and that a group morphism is also a $\mathbf{Z}$-module morphism. Thus the notions of 'abelian group' and '$\mathbf{Z}$-module' are interchangeable. Furthermore, being finitely generated as a group and as a $\mathbf{Z}$-module are equivalent. The previous sections therefore give the classification of finitely generated abelian groups from Section **13**, including the effective method to reduce a finitely presented abelian group to invariant factor form

$$\mathbf{Z}_{d_1} \oplus \cdots \oplus \mathbf{Z}_{d_\ell} \oplus \mathbf{Z}^k \qquad \text{for } \ 1 < d_1 \mid d_2 \mid \cdots \mid d_\ell \ ,$$

and to elementary divisor form

$$\mathbf{Z}_{p_1^{\alpha_1}} \oplus \cdots \oplus \mathbf{Z}_{p_s^{\alpha_s}} \oplus \mathbf{Z}^k \ .$$

FROM HERE ONWARDS, THE READER WILL NEED A MORE THOROUGH BACKGROUND IN LINEAR ALGEBRA—Chapters 1, 3 and 4 of **Artin** would suffice.

46. Similarity of matrices.

Let F be a field and let $A \in F^{n\times n}$. Recall from linear algebra that another $n \times n$ matrix B is *similar* to A if and only if there is an invertible S in $F^{n\times n}$ such that $B = S^{-1}AS$. It is easy to see that similarity is an equivalence relation. By regarding A as a linear transformation, and then thinking of that transformation as the scalar multiplication by x in a module over $F[x]$, and finally applying the basic theorems **43.1** and **43.2**, we shall get a 1–1 correspondence between similarity classes of $n \times n$ matrices and sequences of non-constant monic polynomials, $d_1(x) \mid d_2(x) \mid d_3(x) \mid \cdots \mid d_\ell(x)$, whose degrees sum to n. By knowing some of the details of how this works, one can then find a set of representatives (for similarity classes) which are quite simple matrices—the *rational canonical form* and (when F is algebraically closed) the *Jordan form*, both given near the end of this section. One needs only to write down exactly one matrix which produces each given sequence of polynomials. In contrast with what appears to be the common approach in texts for undergraduates, we shall go almost immediately to the heart of the matter in **46.2** below to show how, given the matrix A, to actually *calculate* the $d_i(x)$, which are called the *invariant factors of* A (since they are the invariant factors of the module associated to A). More precisely, we find a matrix, with polynomial entries, which gives a presentation by generators and relations of the module associated to the scalar-entried matrix A—the former matrix turns out to be simply $xI - A$. Then r/c operations complete the job of getting the invariant factors. From this, all the theory related to: the *minimal polynomial*, $d_\ell(x)$; the *characteristic polynomial*, which is the product of all the $d_i(x)$; etc... drops out without any fuss.

First we show how to get a 1–1 correspondence between similarity classes (of matrices) and isomorphism classes (of $F[x]$-modules which are finite dimensional as F-vector spaces). The 'information' contained in the matrix A above may be 'coded' as a pair consisting of $F^{n\times 1}$ together with the linear operator on $F^{n\times 1}$ which is defined to be left multiplication by A. Up to similarity, this information is coded by:

i) *any* n-dimensional F-vector space M, together with
ii) any F-linear operator on M which happens to be representable by the matrix A in some F-basis for M.

Now suppose instead that M denotes an $F[x]$-module which, when regarded as an F-vector space, is n-dimensional. The map $m \mapsto x \cdot m$ gives an F-linear operator on M. In a sense, there is no additional information to be had from M, besides its F-vector space structure together with this operator: for each $g(x) \in F[x]$, the map $m \mapsto g(x) \cdot m$ is determined by this information using the module axioms:

$$(a_0 + a_1 x + a_2 x^2 + \cdots) \cdot m \;=\; a_0 \cdot\; m \;+\; a_1 \cdot (x \cdot m) + a_2 \cdot (x \cdot (x \cdot m)) + \cdots \quad .$$

Below we shall make precise this 1-1 correspondence, between the set of similarity classes of square F-matrices and the set of isomorphism classes of $F[x]$-modules M with $\dim_F M < \infty$, and then use the results from sections **43** and **44** to completely 'solve' the 'problem' of similarity.

Definition. For $A \in F^{n\times n}$, let $N^{(A)}$ be the $F[x]$-module whose underlying abelian group is $F^{n\times 1}$, with scalar multiplication

$$g(x) \cdot \mathbf{v} \;:=\; g(A)\mathbf{v} \qquad \text{(matrix multiplication)}$$

$$\text{for all} \quad \mathbf{v} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in F^{n\times 1} \; .$$

Exercise 46A. Check that $N^{(A)}$ is a module.

Clearly $\dim_F N^{(A)} = n$.

Theorem 46.1. *i) The matrix A is similar to B if and only if we have $N^{(A)} \cong N^{(B)}$ as $F[x]$-modules.*
ii) Furthermore, every $F[x]$-module M with $\dim_F(M) < \infty$ is isomorphic to some module $N^{(A)}$.

Proof. Assuming that A is similar to B, choose an invertible S with $S^{-1}AS = B$. Then, by induction on j , $A^j S = SB^j$ for all $j \geq 0$. Thus

$$(\sum_j c_j A^j)S \;=\; S \sum_j c_j B^j \; ;$$

that is, $g(A)S = Sg(B)$ for any $g(x) \in F[x]$. Using $\cdot$ and $\star$ for scalar multiplications in $N^{(A)}$ and $N^{(B)}$ respectively, this says that

$$g(x) \cdot (S\mathbf{v}) \;=\; S(g(x) \star \mathbf{v})$$

for all $\mathbf{v} \in F^{n\times 1}$. Thus left multiplication by S defines a module morphism from $N^{(B)}$ to $N^{(A)}$, which is an isomorphism since S^{-1} exists.

Conversely, suppose that $\theta : N^{(B)} \to N^{(A)}$ is an $F[x]$-module isomorphism. It is, in particular, an invertible F-linear operator on $F^{n\times 1}$, and therefore$^{(*)}$ it is given as left multiplication by some invertible $S \in F^{n\times n}$; i.e. $\theta(\mathbf{v}) = S\mathbf{v}$. But now, since $\theta(x \star \mathbf{v}) = x \cdot (\theta(\mathbf{v}))$, we have $SB\mathbf{v} = AS\mathbf{v}$ for all $\mathbf{v}$. Thus$^{(*)}$ $SB = AS$ or $S^{-1}AS = B$, as required. (At $^{(*)}$, some very basic linear algebra has been used.)

As for the last claim in the proposition, take $n = \dim_F(M)$, choose any basis $\{\ b_1,\ \cdots,\ b_n\ \}$ for M as a vector space over F, and determine an F-linear map

$$\phi \;:\; F^{n\times 1} \;\longrightarrow\; M$$

by mapping the i^{th} standard basis vector to b_i. It is mechanical to check that ϕ is a module isomorphism from $N^{(A)}$ to M, where A is the matrix which represents $(x\cdot)$ with respect to the given basis for M.

We now have the desired 1–1 correspondence between similarity classes of matrices and isomorphism classes of modules. It takes the similarity class of A to the isomorphism class of $N^{(A)}$. This map is well defined by half of **46.1**i); injective by the other half; and surjective by **46.1**ii).

Theorem 46.2. *As $F[x]$-modules,*

$$N^{(A)} \;\;\cong\;\; M_{xI-A} \;\;.$$

That is, taking $R = F[x]$ and $C = xI - A \in (F[x])^{n\times n}$ in Section **44** *(with $p = n$), the module M_C is isomorphic to $N^{(A)}$. Equivalently, a presentation of $N^{(A)}$ by "n" generators and "n" relations is defined using the matrix $xI - A$.*

Proof. Let $\psi : (F[x])^n \to N^{(A)}$ be the unique $F[x]$-module morphism which maps $e_i = (0, \cdots, 1, \cdots, 0) \in (F[x])^n$ to

$g_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \in N^{(A)}$. (Recall that the latter module equals $F^{n\times 1}$ as a set. It is helpful here to distinguish e_i (whose components are constant polynomials) from g_i (which is almost the same and would often be identified with e_i). There are lots of other surjective module morphisms besides ψ which will produce various other choices for generators and relations; but ψ seems to be particularly natural, and the resulting presentation matrix, $xI - A$, is as simple as one could hope for.)

Continuing with the proof, we shall be looking at the following morphisms of modules:

$$(F[x])^n \ \stackrel{\theta_{xI-A}}{\longrightarrow}\ (F[x])^n \ \stackrel{\psi}{\longrightarrow}\ N^{(A)} \ .$$

Clearly ψ is surjective. We must show that its kernel agrees with $\mathrm{Im}(\theta_{xI-A})$ (in the notation of Section **44**). This image is the submodule generated by $\{\ c_1,\ \cdots,\ c_n\ \}$, where

$$c_i \ = \ (-a_{1i},\ -a_{2i},\ \cdots,\ x-a_{ii},\ \cdots,\ -a_{ni}) \ ,$$

since $c_i = \theta_{xI-A}(e_i)$ [and the e_i's generate the domain of θ_{xI-A}]. Each c_i is in the kernel of ψ, since

$$\psi(c_i) \ = \ \psi\left(xe_i - \sum_j a_{ji}e_j\right) \ = \ xg_i \ - \ \sum_j a_{ji}g_j \ = \ 0 \ ,$$

the last equality being given by

$$xg_i \ = \ Ag_i \ = \ \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix} \ = \ \sum_j a_{ji}g_j \ ,$$

from the definition of scalar multiplication in $N^{(A)}$. Now let J be an abbreviated name for $\mathrm{Im}(\theta_{xI-A})$, the $F[x]$-submodule generated by $\{\ c_1,\ \cdots,\ c_n\ \}$. We have just proved that ψ factors through a surjective morphism $(F[x])^n/J \to N^{(A)}$. It remains to show that the latter is an isomorphism. For this it suffices to check that

$$\dim_F\left((F[x])^n/J\right) \ \leq \ n \ ,$$

which follows by showing that

$$X \ := \ \{ \ e_1 + J, \ \cdots, \ e_n + J \ \}$$

spans $(F[x])^n/J$ over F. Since the elements $h_{N,i} := (0, \cdots, x^N, \cdots, 0)$ [with the power of x in the i^{th} position] span $(F[x])^n$ over F, we shall prove by induction on N that

$$h_{N,i} + J \ \in \ Span_F(X) \ .$$

This is clear for $N = 0$ since $h_{0,i} = e_i$. For the inductive step, note that a straightforward calculation with the definitions of $h_{N,i}$ and c_i yields

$$h_{N,i} \ = \ x^{N-1}c_i + a_{1i}h_{N-1,1} + a_{2i}h_{N-1,2} + \cdots + a_{ni}h_{N-1,n} \ .$$

But $x^{N-1}c_i \in J$ by the definition of J, and $a_{ki}h_{N-1,k} + J \in Span_F(X)$ for each k by the inductive hypothesis, completing the proof.

Proposition 46.3. *If, for invertible $F[x]$-matrices P and Q of size $n \times n$, we have*

$$P(xI - A)Q \ = \ D \ := \ \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & 0 & & \\ & & 1 & & & & \\ & & & d_1(x) & & & \\ & 0 & & & \ddots & & \\ & & & & & d_\ell(x) & \\ & & & & & & 0 \end{pmatrix} ,$$

with "m" 1's on the diagonal together with monics $d_j(x)$, then

i) $m + \ell = n = \sum_1^\ell \deg d_j(x)$, so that there are, in fact, no zeros on the diagonal of D, and

$$\det(xI - A) \ = \ d_1(x)d_2(x)\cdots d_\ell(x) \ .$$

ii) Assume also that $1 \neq d_1(x) \mid d_2(x) \mid \cdots \mid d_\ell(x)$. Then D is uniquely determined by A, and the invariant of the module $N^{(A)}$ (or of M_{xI-A} by **46.2***) is $(0 \ ; \ [d_1(x)], \cdots, [d_\ell(x)])$.*

Proof. i) Since P is invertible, we have

$$1 \ = \ \det(P^{-1}P) \ = \ \det(P^{-1})\det(P) \ ,$$

and so det$P$ is invertible in $F[x]$, and therefore lies in F. (In fact, *for any R and any $P \in R^{n\times n}$, the inverse, P^{-1}, exists in $R^{n\times n}$ if and only if* det$P$ *is invertible in R.*

Exercise 45B. Use the 'adjoint formula' for the inverse to prove the "if" half.)

Similarly, det$Q$ is invertible in $F[x]$, and so is also in F. Thus, applying *det* to the display in the proposition, det$D$ is a non-zero constant multiple of $\det(xI - A)$. But the latter is clearly monic of degree n. Thus D has no zeros on the diagonal, so its determinant is $d_1(x)\cdots d_\ell(x)$, which is monic, and therefore agrees with $\det(xI - A)$, and has degree n.

ii) By **44.1**, M_{xI-A} has the same invariant as M_D. By **44.2** and i) above, this invariant is $(0\ ;\ [d_1(x)], \cdots [d_\ell(x)])$. By **43.2**, there can be no other such diagonal matrix D_1 satisfying $P_1(xI - A)Q_1 = D_1$ for invertible P_1 and Q_1.

Assertion 46.4. *For each A in $F^{n\times n}$, there exist P, Q and D as in* **46.3***ii). They may be obtained by applying r/c operations over $F[x]$:*

$$\begin{array}{c|c} I & xI - A \\ \hline & I \end{array} \quad \rightsquigarrow \quad \rightsquigarrow \quad \cdots \quad \rightsquigarrow \quad \begin{array}{c|c} P & D \\ \hline & Q \end{array} \qquad (\textbf{BACH}?)$$

This is the specialization of **44.4**, taking $R = F[x]$.

Definition. The invariant factors, $d_1(x), \cdots, d_\ell(x)$, of $N^{(A)}$ will also be referred to as the *invariant factors of A*—similarly for the *elementary divisors of A*, the multiset consisting of the powers of monic irreducibles occurring in the unique factorizations of all the $d_j(x)$.

Corollary 46.5. *Two square F-matrices are similar iff they have the same sequence of invariant factors (or equivalently, $\cdots$ the same multiset of elementary divisors).*

This is immediate from **43.1, 43.2, 46.1, 46.2** and **46.3**, since we have:

$$A \text{ is similar to } B \iff N^{(A)} \cong N^{(B)} \iff M_{xI-A} \cong M_{xI-B} \iff$$

$xI - A$ and $xI - B$ reduce by r/c ops. to the *same* Smith normal form .

The latter refers to the matrix D in **46.3**ii), requiring all the diagonal entries to be monic.

Much of the above is summarized by four bijections:

Set of similarity classes of square F-matrices of all sizes .

$\updownarrow$ basic linear algebra

Set of $\cong$ classes of linear operators $V \to V$,

where V varies over finite dimensional F-vector spaces,

and $(V \xrightarrow{T} V) \cong (W \xrightarrow{S} W) \iff_{\text{defn.}}$ there is a linear isomorphism $P : W \to V$ such that $P^{-1}TP = S$.

$\updownarrow$ start of this section

Set of isomorphism classes of $F[x]$ -modules M for which $\dim_F M < \infty$.

$\updownarrow$ major theorem(**43.1**, **43.2**)

Set of all sequences $1 \neq d_1(x) \mid d_2(x) \mid \cdots \mid d_\ell(x)$ of monics in $F[x]$.

$\updownarrow$ rinky – dink process

Set of finite multisets of positive powers of monic irreducibles in $F[x]$.

A *canonical form* for similarity is a set of square F-matrices (of hopefully rather simple description) such that every square F-matrix is similar to exactly one in the set. In other words, the set is formed by picking out just one (hopefully rather handsome) matrix from each similarity class.

With the theory above, it suffices to find one matrix, for each multiset consisting of powers of monic irreducibles in $F[x]$, whose elementary divisors

form that multiset. Note that r/c operations over $F[x]$ exist to convert $xI-A$ into

$$\begin{pmatrix} 1 & & & & & \\ & \ddots & & & 0 & \\ & & 1 & & & \\ & & & d_1(x) & & \\ & 0 & & & \ddots & \\ & & & & & d_\ell(x) \end{pmatrix},$$

where $d_1(x), \cdots, d_\ell(x)$ are the invariant factors of A, and others to convert the latter to

$$\begin{pmatrix} q_1(x) & & 0 \\ & \ddots & \\ 0 & & q_n(x) \end{pmatrix},$$

where the $q_i(x)$ are the elementary divisors of A, supplemented by 1's to make their number up to n, and planted in any desired order on the diagonal. Now if

$$A \;=\; \begin{pmatrix} A' & 0 \\ 0 & A'' \end{pmatrix},$$

with A' and A'' square, then

$$xI - A \;=\; \begin{pmatrix} xI' - A' & 0 \\ 0 & xI'' - A'' \end{pmatrix},$$

for suitable identity matrices I' and I''. It follows that the multiset of elementary divisors for such an A is the union of those for A' and A''. (The same is not true of the invariant factors—one must first 'fracture' the monics in the union into elementary divisors and then re-assemble. These last points are probably more easily seen in terms of modules than of matrices—after shuffling the free part, it is obvious that a direct sum of modules in elementary divisor form is still in that form; but a direct sum of two in invariant factor form can seldom be put into that form just by permuting summands.) Thus it suffices to produce one canonical form matrix for each power, $p(x)^\alpha$, of a monic irreducible $p(x)$. Then, for a given elementary divisor multiset, to produce the corresponding canonical form matrix, one assembles these as

blocks down the diagonal, the number of blocks being the size of the multiset, with the blocks in some arbitrarily chosen order.

Rational Canonical Form. For each monic in $F[x]$,

$$q(x) \;=\; x^s + a_{s-1}x^{s-1} + \cdots + a_0 \;,$$

it is easy to find r/c operations over $F[x]$ which convert

$$xI - B_{q(x)} \quad \text{to} \quad \begin{pmatrix} 1 & & & \\ & \ddots & & 0 \\ & & 1 & \\ 0 & & & q(x) \end{pmatrix} ,$$

where the *companion matrix*

$$B_{q(x)} \;:=\; \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & 0 \\ & & \ddots & \ddots & \\ 0 & & & 0 & 1 \\ -a_0 & -a_1 & \cdots & \cdots & -a_{s-1} \end{pmatrix} .$$

(Do the cases $s = 2$ and 3 to convince yourself of this—the formal proof of an assertion like this deserves to be left in the closet!)

Thus, given $A \in F^{n\times n}$, taking $q_i(x)$ to be $p(x)^\alpha$ once for each occurrence of $p(x)^\alpha$ among the elementary divisors of A, 'the' *rational canonical form* of A will be a matrix

$$\text{BLOCK DIAGONAL} \; (\; B_{q_1(x)} \; , \; B_{q_2(x)} \; , \; \cdots \;)$$

Jordan Canonical Form. Here assume that F is *algebraically closed* (which is equivalent to saying that all monic irreducibles in $F[x]$ have the form $x - \lambda$, as λ varies over F; for example, $F = \mathbf{C}$.) It is easy to find r/c

operations which convert

$$xI - J_{\alpha,\lambda} \quad \text{to} \quad \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ 0 & & & (x-\lambda)^\alpha \end{pmatrix},$$

where the *Jordan block*

$$J_{\alpha,\lambda} := \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \lambda & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix} \in F^{\alpha\times\alpha}.$$

(Do a couple of examples to convince yourself of this.)

'The' *Jordan canonical form* of A will then be block diagonal, with one Jordan block as above for each occurrence of $(x-\lambda)^\alpha$ among the elementary divisors of A.

By using **46.2**, we have quickly derived both of these canonical forms, as well as algorithms to produce them. One can derive, by a lengthier argument, just the existence of these canonical forms, without first proving **46.2**. One need only analyze the general cyclic $F[x]$-module. This may make the rational and Jordan blocks appear to be more natural. Here is a quick summary—see **Artin**, pp.478–482, for a more detailed treatment. Let $\mathbf{v}_1$ be a module generator for the cyclic $F[x]$-module $F[x]/ < d(x) >$, where $d(x)$ is monic of degree s. Then $g(x)\mathbf{v}_1 = 0$ only for multiples $g(x)$ of $d(x)$. Thus $\{\mathbf{v}_1,\ \mathbf{v}_2,\ \cdots,\ \mathbf{v}_s\}$ is linearly independent over F, where $\mathbf{v}_i := x^{i-1}\mathbf{v}_1$. With respect to that basis, it is easily calculated that the matrix of the linear operator $(x\cdot)$ is the rational block whose bottom row consists of the negatives of the coefficients of $d(x)$ [i.e. the companion matrix of $d(x)$]. To get the Jordan block from before, assume that $d(x) = (x-\lambda)^\alpha$, so that $\alpha = s$. This time use instead the basis whose i^{th} member is $(x-\lambda)^{i-1}\mathbf{v}_1$.

A number of otherwise rather difficult results in matrix theory are easy consequences of this theory. In linear algebra, $\det(xI - A)$ is called the

characteristic polynomial of A. By **46.3**, this is the product, $d_1(x) \cdots d_\ell(x)$, of the invariant factors of A. The *minimal polynomial* of A is the monic which generates the ideal $\{\ g(x) \in F[x]\ :\ g(A) = 0\ \}$. But this, by definition, is the annihilator of the module $N^{(A)}$. Now, for *any* PID, say R, when $d_i \mid d_{i+1}$ for all i, we have

$$\text{ann}\,(\ R/(d_1) \oplus \cdots \oplus R/(d_\ell)\ ) \quad = \quad d_\ell\ .$$

(See the alternative proof of Theorem **43.2**). Thus the minimal polynomial of A is $d_\ell(x)$, the 'largest' invariant factor. We immediately deduce two results:

i) The minimal polynomial divides the characteristic polynomial; or equivalently, $f(A) = 0$, where $f(x) = \det(xI - A)$. This is known as the **Cayley-Hamilton theorem**.

ii) Every irreducible factor of the characteristic polynomial occurs as a factor of the minimal polynomial.

In particular, every root of the characteristic polynomial occurs as a root of the minimal polynomial (which is equivalent to ii) when F is algebraically closed). These roots are known also as *eigenvalues.* When F is algebraically closed, they are the diagonal entries in the Jordan form; and the dimension of the eigenspace corresponding to λ is the number of Jordan blocks in which λ is the diagonal element [i.e. the number of elementary divisors of the form $(x-\lambda)^\alpha$]. The diagonalizability theorems from linear algebra all follow immediately. Note that for *any* F, a matrix has a Jordan form as long as its minimal polynomial is a product of linear factors (and conversely). It is diagonalizable as long as these factors are distinct.

Many readers will have first encountered $xI - A$ in dealing with eigenvalues. It is perhaps ironic that deciding whether two matrices are similar (that is, carrying out reduction of $xI - A$ to Smith normal form) is in general easier than finding exact eigenvalues (which requires one to solve polynomial equations).

We have used in this section the ubiquitous passage back and forth between matrices and linear transformations which is a central topic in elementary linear algebra. But all the main results have been stated in terms of square matrices, since this is simpler and more direct. The student should translate each such statement to the equivalent statement concerning linear operators. For example, 'block diagonal matrix' will translate to 'direct sum decomposition into invariant subspaces'.

We have emphasized finding the invariant factors of a matrix, and thereby finding its 'canonical form(s)'. The question of actually finding an invertible matrix S which conjugates the given matrix to its canonical form is now a straightforward application of basic linear algebra—the methods here will produce F-bases for the relevant vector spaces, and S is then a 'change-of-basis' matrix. Multiplying will produce a matrix which conjugates between any two given similar matrices, once both conjugating matrices have been found which convert the givens to their common canonical form.

Exercise 46C. Try to analyze your example from **43A** for the non-PID, $\mathbf{Z}[x]$, by the methods of this chapter (U.S.–style?), and analyse what goes 'wrong' (British–style?).

Exercise 46D. Look for matrices in the similarity section of your linear algebra text(s) and calculate their rational and (when existing) Jordan forms.

Exercise 46E. i) For $n = 1,\ 2$, and 3, find formulae for the numbers of similarity classes in $(\mathbf{F}_p)^{n\times n}$.
ii) Also find formulae for the numbers of conjugacy classes in their groups, $GL(n,\ \mathbf{F}_p)$, of invertibles.
Can you generalize these to arbitrary n? (See **31C**.)

Exercise 46F. Find all matrices in $F^{2\times 2}$ which are similar only to themselves. Now generalize this to $F^{n\times n}$. Explain this in terms of linear operators.

Exercise 46G. Is the following 'proof' of the Cayley-Hamilton theorem convincing? If not, why not?
"Since $f(x) = \det(xI - A)$, we have

$$f(A) \ = \ \det(AI - A) \ = \ \det(0) \ = \ 0 \ \text{!!"}$$

Exercise 46H. Show that A and B are similar (over F) if and only if $xI - A$ and $xI - B$ are 'row/column equivalent' (over $F[x]$).

Exercise 46I. i) Write down the short proof, based on the results of this section, that, over an algebraically closed field, A is similar to a diagonal matrix if and only if its minimal polynomial has no repeated roots.
ii) Give an example to show that the "if" above can fail over 'general' fields.
iii) Show that, if $A^2 = A$ (i.e. A is *idempotent*), then A is diagonalizable (over any field).

iv) Prove that two idempotent matrices are similar if and only if they are 'row/column equivalent'.

Explain these in terms of linear operators.

Exercise 46J. Prove that, if A is a matrix which is a Jordan or rational block, then A is not similar to any block diagonal matrix with more than one block down the diagonal. Explain this in terms of linear operators.

Exercise 46K. i) Prove that every square matrix over an algebraically closed field is similar to a matrix of the form $D + N$, where D is a diagonal matrix, $N^i = 0$ for some i (i.e. N is *nilpotent*), and $DN = ND$.
ii) Prove that a matrix over such a field is nilpotent if and only if it has only 0 as an eigenvalue.
iii) Prove that a matrix over *any* field is invertible if and only if it *doesn't* have 0 as an eigenvalue.

Explain these in terms of linear operators.

Exercise 46L. Prove that an element of finite order in $GL(n,\ \mathbf{C})$ is similar to a diagonal matrix whose diagonal entries are roots of unity; and conversely. Explain this in terms of linear operators.

Exercise 46M. Is every square matrix similar to its transpose?

Exercise 46N. 'Classify' finitely generated modules over
i) $\mathbf{Z}[i]$, where $i^2 = -1$;
ii) $\mathbf{C}[\mu]$, where $\mu^2 = 0$.

Exercise 46O. Find all the similarity classes in $\mathbf{C}^{5\times 5}$ whose characteristic polynomial is $(x+5)^2(x-7)^3$. In each such class, write down both a sample matrix and the dimensions of the eigenspaces for all eigenvalues.

Exercise 46P. Show that, if A and B commute and both are diagonalizable, then there is an invertible S such that both $S^{-1}AS$ and $S^{-1}BS$ are diagonal matrices (over any field). Explain this in terms of linear operators. Show that the commuting condition is needed.

Exercise 46Q. Prove that the minimal polynomial of a block diagonal matrix is the LCM of the minimal polynomials of its blocks.

Exercise 46R. Let J and R be the Jordan and rational blocks, respectively, corresponding to $(x-\lambda)^n$. For small (all?) n, find S such that $S^{-1}JS = R$.

Exercise 46S. Prove that, if A and B are in $F^{n\times n}$, and $K \supset F$ is a field extension, then A and B are similar over F if and only if they are similar over K.

VI. Group Representations

Sections **47** to **50** study the ways in which a finite group can be related to subgroups of the group, $GL(n, \mathbf{C})$, of all invertible $n \times n$ matrices, by morphisms from the former to the latter. These are called *representations* of the group, and called *faithful* representations when the morphism is injective. When faithful, the representation gives an isomorphism, 'realizing' the group as a group of matrices. In general, it realizes a quotient of the group as a matrix group. We shall only make a small start on this subject, which is probably larger than the rest of group theory put together. There are some surprisingly nice results, even at the very beginning of the subject. We'll finish with the famous Burnside (p, q)-theorem which says that a group whose order is divisible by fewer than three primes is necessarily soluble. This is an excellent example of an application of representations to the structure theory of groups, one which had no proof outside representation theory for many decades. The applications of group representations in geometry, topology, harmonic analysis, differential equations, physics, chemistry, number theory (for example, Wiles' proof, in 1994, of Fermat's last theorem), and probability theory are undoubtedly more important even than its applications purely within algebra. Interesting applications to combinatorics also occur.

47. G-modules & representations.

There are two equivalent ways of thinking about representations of a finite group—as G-modules and as G-matreps, both defined below. We'll pass back and forth freely between the two concepts, once we see this equivalence in **47.1**.

Definition. Let G be a group and let S be a set. An *action of G on S* is a function

$$G \times S \longrightarrow S \quad ; \quad (g,\ s) \mapsto g \cdot s\ ,$$

which satisfies $1 \cdot s = s$ and $a \cdot (b \cdot s) = (ab) \cdot s$ for all a and b in G and all $s \in S$.

We shall say that two actions of G, on S by $\cdot$, and on T by $\star$, are *isomorphic* if and only if there is a bijective function $\Gamma : S \to T$ such that $\Gamma(g \cdot s) = g \star \Gamma(s)$ for all $g \in G$ and $s \in S$.

See the remarks after **11.1** giving a list of examples of G-actions. We used them to prove theorems **11.1** (Sylow) and **11.4**, and they occurred on every page in the sections on Galois theory. A set together with a given G-action on the set is sometimes called a *G-set*.

Definition. Let G be a finite group. A *G-module* is a finite dimensional $\mathbf{C}$-vector space V together with an action of G on the set V such that the action is linear; i.e. for each g, the map $v \mapsto g \cdot v$ is a linear operator on V; i.e.

$$g \cdot (\alpha v + \beta w) \;=\; \alpha(g \cdot v) + \beta(g \cdot w)$$

for all complex numbers α and β, all $g \in G$, and all v and w in V.
Two G-modules are said to be *isomorphic* if and only if they are isomorphic (as in the previous definition) by some isomorphism, Γ, which is also $\mathbf{C}$-linear. The *dimension* of V (some call it *degree*) is just its vector space dimension.

Remark. There is a ring called the *complex group algebra of G* (see the beginning of Section **50**), which is non-commutative if G is, and such that a G-module is virtually the same thing as a module over the group algebra (in the sense of the previous chapter except for the non-commutativity). A more sophisticated treatment of the subject dealt with here sees it as a special case of module theory over non-commutative rings with a certain property, *semisimplicity*, which the group algebra has.

Definition. Let G be a finite group. A *G-matrep* is a group morphism $\rho : G \to GL(n, \mathbf{C})$ for some $n \geq 0$. (By convention, when $n = 0$, the general linear group is the trivial group—after all, the group of linear operators on the zero vector space is a trivial group.) The jargon 'matrep' is not used elsewhere— it saves us from saying 'matrix representation' quite a few times (and permits the joke before **47N**)—I request the reader's forgiveness for resorting to bad literary taste in the interests of economy.
The above G-matrep is said to be *equivalent* to some other given G-matrep $\lambda : G \to GL(m, \mathbf{C})$ if and only if both $m = n$ and there exists an invertible matrix P such that $\lambda(g) = P^{-1}\rho(g)P$ for all $g \in G$. (This last definition may seem a bit unnatural—it is justified by the next proposition.)

Exercise 47A. Show that *isomorphism*, on the class of G-modules, and *equivalence*, on the class of G-matreps, are both equivalence relations.

The ideas of G-module and G-matrep are almost identical, in the following

sense. Given a G-matrep $\rho : G \to GL(n, \mathbf{C})$, let $V = \mathbf{C}^{n\times 1}$ and let the action of every g on V be left multiplication by the matrix $\rho(g)$. Given a G-module V, choose any basis for V and map g to $\rho(g) :=$ the matrix representing the operator $(g\ \cdot)$ with respect to the chosen basis.

Proposition 47.1. *i) These are well defined maps back and forth, from the class of G-matreps to the class of G-modules, and from the class of pairs, (G-module, basis), to the class of G-matreps.*
ii) Equivalent G-matreps go to isomorphic G-modules.
iii) Starting with a G-module, the equivalence class of the G-matrep produced is independent of which basis is used.
iv) Isomorphic G-modules produce equivalent G-matreps.
v) Starting with a G-module and doing the two maps consecutively gives a G-module isomorphic to the one you started with.
vi) The same holds for G-matreps and equivalence.

Thus we obtain a bijection between the set of isomorphism classes of G-modules and the set of equivalence classes of G-matreps.

Proof. It is a routine verification to check that starting from a G-matrep produces a G-module, and also the other way round, giving *i)*. As for *ii)*, if the matrix P demonstrates that ρ is equivalent to λ, then left multiplication by P gives a linear operator on $\mathbf{C}^{n\times 1}$ which may easily be seen to be a G-module isomorphism mapping the G-module obtained from ρ to that obtained from λ. For *iii)* and *iv)*, if Γ is an isomorphism between two G-modules, and they have been given bases in order to produce corresponding G-matreps, then the matrix P which represents Γ using these two bases is easily checked to provide an equivalence between the G-matreps corresponding to the given G-modules. When the two G-modules are equal, this proves *iii)*, and when they're not necessarily equal it also does *iv)*. Starting from a G-matrep, the corresponding G-module is $\mathbf{C}^{n\times 1}$, which has the standard basis. Using this basis, the G-matrep you get is actually the one you started with (not just equivalent to it). This proves *vi)*. Starting with a G-module and a basis for it, let Γ be the unique linear isomorphism from it to $\mathbf{C}^{n\times 1}$ which takes the given basis to the standard basis. Then Γ is readily checked to be an isomorphism of G-modules. Its codomain is the G-module associated to that G-matrep which is associated to the starter with its given basis. This proves *v)*.

Exercise 47B. Do all the routine verifications in detail to convert this into a complete proof.

Definition. Define the *direct sum*, of two G-modules V and W, to be the vector space $V \bigoplus W$ of ordered pairs, with the G-action

$$g \cdot (v,\ w) \ := \ (g \cdot v,\ g \cdot w) \ .$$

Define the direct sum $\rho \oplus \lambda$ of two G-matreps to be the map which sends g to the block diagonal matrix

$$\begin{pmatrix} \rho(g) & 0 \\ 0 & \lambda(g) \end{pmatrix} \ .$$

Exercise 47C. i) Show that both of these external direct sum operations are well defined.
ii) Prove that if $V \cong V'$ and $W \cong W'$, then $V \bigoplus W \cong V' \bigoplus W'$.
iii) Prove the analogue of ii) for G-matreps and equivalence.
iv) Show that if V and ρ correspond under the bijection in **47.1**, and W and λ do as well, then so do $V \bigoplus W$ and $\rho \oplus \lambda$.

Definition. A *G-invariant subspace* (also called a *G-submodule*), of a G-module V, is any vector subspace W such that $g \cdot w \in W$ for all $w \in W$ and $g \in G$.
A G-module V is called *irreducible* if and only if it has precisely two invariant subspaces (which are therefore $\{0\}$ and V itself—in particular, $V \neq \{0\}$).

Exercise 47D. Determine what the corresponding concepts are in the language of G-matreps.

Definition. A G-module V is the *internal direct sum* of a finite sequence $V_1,\ V_2, \cdots$ of G-submodules if and only if each element of V can be written uniquely as a sum of vectors, one from each V_i . This is equivalent to stating that the map

$$V_1 \bigoplus V_2 \bigoplus \cdots \quad \longrightarrow \quad V \ ,$$

which sends $(v_1,\ v_2, \cdots)$ to $v_1 + v_2 + \cdots$, is an isomorphism of G-modules.

Examples. The representations of the trivial group are obvious, so let's try to determine all the representations of a cyclic group, $\{1,\ g\}$, of order 2.

For such a matrep ρ, any linear operator T with matrix $\rho(g)$ satisfies $T^2 = I$. Therefore the space on which T acts splits uniquely as the internal direct sum of the $(+1)$-eigenspace of T and the (-1)-eigenspace of T. Picking any bases in these two eigenspaces, each of them splits, but usually non-uniquely, into internal direct sums of one dimensional invariant subspaces. Putting this together we see that $G = \{1,\ g\}$ has only two irreducible modules, each of dimension 1 over $\mathbf{C}$, on which $(g\ \cdot)$ acts as the identity in one case, and as $-I$ on the other 'irrep'. Furthermore an arbitrary G-module splits uniquely as the internal direct sum of two invariant subspaces, each of which splits non-uniquely as an internal direct sum of irreducible invariant submodules. All of the irreducibles within either one of the two earlier invariant direct summands are isomorphic to each other.

We'll see that the general features just noted carry over to an arbitrary finite group, the number of irreducibles being finite but usually larger than 2, and the dimensions of some of these irreducibles being larger than 1 when the group is non-abelian.

Definition. A G-module is *trivial* when $(g\ \cdot)$ is the identity map for all group elements g. A G-matrep is *trivial* when it maps all group elements to the identity matrix. These are corresponding concepts. The one dimensional trivial representation is the only one of these which is irreducible. The *zero representation* (i.e. $n = 0$ for matreps and dimension is zero for G-modules) is trivial, not irreducible, and even more uninteresting than the other trivial representations; but it is needed to avoid extra phrases in definitions and results.

Examples continued. Let's look at the cyclic group $G = \{1,\ g,\ g^2\}$ of order three, the cyclic group $H = \{1,\ h,\ h^2,\ h^3\}$ of order four, and the Klein group $K = \{1,\ a,\ b,\ ab\}$ of order four. A G-matrep, ρ, is determined by knowing $\rho(g)$ since $\rho(g^2) = \rho(g)^2$ and $\rho(1) = I$. Also $\rho(g)^3 = I$. Similarly , an H-matrep, λ is determined by knowing $\lambda(h)$; and $\lambda(h)^4 = I$. Finally, a K-matrep μ is determined by knowing both $\mu(a) = A$ and $\mu(b) = B$; and they are matrices satisfying $A^2 = I = B^2$ and $AB = BA$.

By elementary considerations of eigenspaces, it follows that a G-module is uniquely the internal direct sum of three invariant subspaces, corresponding to the three cube roots of unity, occurring as eigenvalues for $\rho(g)$. Each of these three is itself a direct sum of 1-dimensional irreducibles, all isomorphic.

The same is true of an H-module, except that there are four distinct irreducibles. The fact that A and B commute yields a similar result for K, with four irreducibles possible, all of dimension one and corresponding to the four combinations of ± 1, occurring as eigenvalues for A and B. (In each of these cases, the number of copies of a given irreducible within a given G-module might be zero!)

Exercise 47E. Write out the complete details for these claims above.

Examples cont'd. Let's now look at the smallest non-commutative group, S_3. It has at least two 1-dimensional (and therefore irreducible) representations: the trivial one, and the action on $\mathbf{C}$ given by $(\sigma \cdot) :=$ multiplication by $\text{sign}(\sigma)$. It also has a linear action on $\mathbf{C}^3$ given by permuting the coordinates, i.e.

$$\sigma \cdot (z_1,\ z_2,\ z_3) \ :=\ (z_{\sigma^{-1}(1)},\ z_{\sigma^{-1}(2)},\ z_{\sigma^{-1}(3)}) \ .$$

Exercise 47F. Check that this is a representation, and that we must use the inverse of σ on the right-hand side—we've been working with *left* actions; with right actions the inverse wouldn't occur.

This last representation is not irreducible; the subspace $\{(z,\ z,\ z)\}$ is a 1-dimensional trivial invariant submodule. A complementary submodule is

$$\{\ (z_1,\ z_2,\ z_3)\ :\ z_1 + z_2 + z_3 = 0\ \} \ .$$

Exercise 47G. i) Prove that this last set is an invariant S_3-submodule, and is irreducible.
ii) Try to prove that S_3 has only these three irreducible representations, up to isomorphism of course.

Here is a table summarizing these facts and a few others concerning the five examples.

group	*#conj.classes*	*#irreps.*	*order*	$\sum \dim^2(\textit{irreps.})$
C_2	2	2	2	$2 = 1^2 + 1^2$
C_3	3	3	3	$3 = 1^2 + 1^2 + 1^2$
C_4	4	4	4	$4 = 1^2 + 1^2 + 1^2 + 1^2$
D_2	4	4	4	$4 = 1^2 + 1^2 + 1^2 + 1^2$
S_3	3	3	6	$6 = 1^2 + 1^2 + 2^2$

The numerical coincidences in the table are not coincidental: in order of relative difficulty, here are three facts, which we'll prove in sections **47, 48** and **50** respectively.
The squared dimensions of the irreps add up to the order of the group!
The number of irreps is the number of conjugacy classes in the group!
The dimension of each irrep divides the order of the group!
Our 'empirical' evidence for these is admittedly not especially strong just yet. In fact it doesn't seem obvious that a finite group cannot have infinitely many non-isomorphic irreducibles, possibly even ones whose dimensions get arbitrarily large.

First let's establish, for any finite group G, the decomposition of every G-module into irreducibles.

Maschke's Theorem 47.2. *If W is an invariant subspace of a G-module, then there is another invariant subspace U such that V is the internal direct sum of W and U.*

Corollary 47.3. *Any G-module is the direct sum of (finitely many) irreducible submodules.*
The corollary is immediate by induction on the dimension of the G-module. By 'convention' (actually by logic), the zero module is the direct sum of the empty set of irreducibles.

Definition. A *G-map*, $\theta : V \to V'$, between two G-modules is a linear map such that $\theta(g \cdot v) = g \cdot \theta(v)$ for all $g \in G$ and $v \in V$.
Thus an isomorphism is a bijective G-map.

Lemma 47.4. *The kernel and image of a G-map both are invariant subspaces.*

Exercise 47H. Prove this.

Proof of 47.2. It suffices to find a G-map $\theta : V \to W$ such that $\theta(w) = w$ for all $w \in W$—for linear algebra shows that V then splits as the direct sum of W and the kernel of θ, so the result follows from half of **47.4**. By linear algebra, we know that there is a *linear* map $\phi : V \to W$ which has the splitting property $\phi(w) = w$ for all $w \in W$. Use the following 'averaging

trick' to define θ so that it will be a G-map with the splitting property. Let

$$\theta(v) \ := \ |G|^{-1} \sum_{g \in G} g^{-1} \cdot \phi(g \cdot v) \ .$$

It is a routine verification that θ has the required properties. Let's just check that it is a G-map:

$$\theta(g_0 \cdot v) = |G|^{-1} \sum_{g \in G} g^{-1} \cdot \phi(g g_0 \cdot v) = |G|^{-1} \sum_{h \in G} g_0 h^{-1} \cdot \phi(h \cdot v) = g_0 \cdot \theta(v) \ ,$$

letting $h = g g_0$.

Exercise 47I. Do the other verifications to finish this proof.

Definition. Given two G-modules V and W, define $Hom_G(V,\ W)$ to be the set of all G-maps from V to W.

Proposition 47.5. *The set $Hom_G(V,\ W)$ is a subspace of the vector space, $Hom_{\mathbf{C}}(V,\ W)$, of* all *linear maps from V to W.*

Exercise 47J. Prove this.

Exercise 47K. Prove that if $V \cong_G V'$ and $W \cong_G W'$, then $Hom_G(V,\ W) \cong_{\mathbf{C}} Hom_G(V',\ W')$. Here we are using $\cong_G$ to denote isomorphisms of G-modules, and $\cong_{\mathbf{C}}$ to denote isomorphisms of complex vector spaces.

Schur's Lemma 47.6. *Let V and W be* **irreducible** *G-modules.*
i) If $V \ncong W$, then $Hom_G(V,\ W)$ is the zero subspace.
ii) If $V \cong W$, then $Hom_G(V,\ W)$ is a 1-dimensional space, all of whose non-zero elements are isomorphisms.
iii) Taking $V = W$, the space $Hom_G(V,\ V)$ consists of all the scalar multiples of the identity.

Proof. *i)* Suppose that $\theta \in Hom_G(V,\ W)$ is non-zero. By **47.4**, its image is a non-zero invariant subspace of W, and so is all of W, by irreducibility. By **47.4**, its kernel is a proper invariant subspace of V, and so is zero, by irreducibility. Thus θ is an isomorphism, contradicting hypothesis.

ii) follows immediately from *iii)*, by taking $V' = W' = V$ in **47K**. Its second half also follows from the argument given for *i)*.

iii) Let $\theta \in Hom_G(V,\ V)$. Let α be an eigenvalue of θ. Then $\theta - \alpha I$ is a G-map with a non-zero kernel, so its kernel is V, by irreducibility. Thus it is the zero map, as required.

Lemma 47.7. *The 'Hom_G functor' distributes over direct sums; that is, for G-modules $V_\alpha,\ W_\beta,\ V$ and W,*

$$Hom_G(\oplus_\alpha V_\alpha,\ W) \ \cong_{\mathbf{C}} \ \bigoplus_\alpha Hom_G(V_\alpha,\ W)\ ,$$

and

$$Hom_G(V,\ \oplus_\beta W_\beta) \ \cong_{\mathbf{C}} \ \bigoplus_\beta Hom_G(V,\ W_\beta)\ .$$

Proof. Consider the canonical maps from linear algebra, which prove the same isomorphisms, except that Hom_G is replaced by $Hom_{\mathbf{C}}$. A routine calculation shows that these maps restrict to isomorphisms between the subspaces in the lemma.

Now we can prove the most important aspect of the uniqueness of the decomposition of a G-module into a direct sum of irreducibles.

Theorem 47.8. *Let V be a G-module. Suppose that*

$$\bigoplus_\alpha V_\alpha \ \cong_G \ V \ \cong_G \ \bigoplus_\gamma V_\gamma\ ,$$

where all the V_α and V_γ are irreducible G-modules. Let W be any irreducible G-module. Then the number of α for which V_α is isomorphic to W equals the number of γ for which V_γ is isomorphic to W.

This says that the number of times that a given irreducible appears in a decomposition of a G-module depends only on the G-module and not on the particular decomposition. We have seen that there is no uniqueness of the actual irreducible direct summands, when a given isomorphism class of irreducible occurs more than once in a decomposition of some G-module into an internal direct sum.

Proof. We show below that the number at issue in the case of subscripts α is the dimension of $Hom_G(V,\ W)$. By symmetry this will also be true

for subscripts γ, as required. (*The main point is to find a formula for the number which is manifestly independent of the details of any decomposition!*)

$$\dim_{\mathbf{C}} Hom_G(V,\ W) \ =_{by\ \mathbf{47K}}\ \dim_{\mathbf{C}} Hom_G(\textstyle\bigoplus_\alpha V_\alpha,\ W)$$

$$=_{by\ \mathbf{47.7}}\ \dim_{\mathbf{C}} \bigoplus_\alpha Hom_G(V_\alpha,\ W) \ = \ \sum_\alpha \dim_{\mathbf{C}} Hom_G(V_\alpha,\ W)$$

$$= \sum_{V_\alpha \cong W} \dim_{\mathbf{C}} Hom_G(V_\alpha,\ W) \ + \ \sum_{V_\alpha \ncong W} \dim_{\mathbf{C}} Hom_G(V_\alpha,\ W)$$

$$=_{by\ \mathbf{47.6}}\ \sum_{V_\alpha \cong W} 1 \ + \ \sum_{V_\alpha \ncong W} 0 \ = \ \#\{\alpha : V_\alpha \cong W\}\ .$$

This completes the proof.

Corollary to the proof 47.9. *The number of times that an irreducible G-module W occurs as a direct summand in a G-module V is* $\dim_{\mathbf{C}} Hom_G(V,\ W)$.

Exercise 47L. Prove that it's also $\dim_{\mathbf{C}} Hom_G(W,\ V)$.

We shall have quite a few references to Schur's lemma. Here's another one.

Theorem 47.10. *If G is abelian, then every irreducible G-module is 1-dimensional.*

Proof. For every $h \in G$, the map $(h\ \cdot)$ is a G-map on any G-module. For,

$$h\ \cdot (g\ \cdot v) \ = \ (hg)\ \cdot v \ = \ (gh)\ \cdot v \ = \ g\ \cdot (h\ \cdot v)\ .$$

If the G-module is irreducible, it is immediate from *iii)* of Schur's lemma that $(h\ \cdot)$ is multiplication by a scalar. This being true for all h, *every* vector subspace is invariant. But any vector space of dimension greater than 1 has plenty of non-trivial proper vector subspaces.

So far we haven't produced even one construction for a non-trivial G-module for an arbitrary finite group G. Here is one which will have plenty of use.

Definition. For each G, define a G-module R_G, called the *regular representation of* G, as follows. As a vector space, let R_G be the space with basis $\{r_g : g \in G\}$, so that the dimension is $|G|$. Define the action by specifying $h \cdot r_g := r_{hg}$ and extending linearly. This gives the formula

$$h \cdot \sum_{g\in G} \alpha_g r_g \;=\; \sum_{g\in G} \alpha_{h^{-1}g} r_g \ ,$$

for complex numbers α_g.

Exercise 47M. Verify that R_G *is* a G-module.

Lemma 47.11. *For any G-module W, we have*

$$Hom_G(R_G,\ W) \ \cong_{\mathbf{C}}\ W \ .$$

On the left, W is regarded as a G-module; but on the right it is regarded as merely a vector space.

Using **47.9**, and applying **47.11** with W irreducible, we obtain

Corollary 47.12. *Every irreducible G-module occurs as a direct summand in R_G a number of times equal to its dimension.*
This could be written

$$R_G \ \cong_G \ \bigoplus_W W^{\oplus(\dim W)} \ ,$$

where the direct sum is over irreducible G-modules, one for each isomorphism class, and $W^{\oplus n}$ denotes a direct sum of "n" copies of W.
Taking dimensions, we get

$$|G| \ = \ \sum_W (\dim W)^2 \ ,$$

proving the first claim after the table following **47G**. *In particular, the number of isomorphism classes of irreducibles is finite.*

Corollary to the Corollary 47.13. *An abelian group G has a total of "$|G|$" irreducible representations (all 1-dimensional).*

Proof of 47.11. Define a map

$$\Gamma :\ Hom_G(R_G,\ W) \ \longrightarrow\ W \ ,$$

by

$$\theta \;\mapsto\; \theta(r_1)\ .$$

It is evident that Γ is a linear map. Let $\theta \in \mathrm{Ker}(\Gamma)$. Then

$$\theta(r_g) \;=\; \theta(g\cdot r_1) \;=\; g\cdot\theta(r_1) \;=\; g\cdot\Gamma(\theta) \;=\; g\cdot 0 = 0\ .$$

Since θ is linear, and $\{r_g\}$ is a basis, we get $\theta = 0$. Thus Γ is injective. It remains to show that Γ is surjective. Given $w \in W$, define a linear map θ_w from R_G to W by specifying it on the basis we're using:

$$\theta_w(r_g) \;:=\; g\cdot w\ .$$

As long as we can show that θ_w is a G-map, we're finished, since

$$\Gamma(\theta_w) \;=\; \theta_w(r_1) \;=\; 1\cdot w \;=\; w\ ,$$

as required. Let $v = \sum_{g\in G}\alpha_g r_g$. The needed calculation is:

$$\begin{aligned}
\theta_w(h\cdot v) &= \theta_w[\textstyle\sum_{g\in G}\alpha_g(h\cdot r_g)]\\
&= \theta_w(\textstyle\sum_{g\in G}\alpha_g r_{hg})\\
&= \theta_w(\textstyle\sum_{k\in G}\alpha_{h^{-1}k} r_k)\\
&= \textstyle\sum_{k\in G}\alpha_{h^{-1}k}(k\cdot w)\\
&= \textstyle\sum_{g\in G}\alpha_g(hg\cdot w)\\
&= h\cdot\textstyle\sum_{g\in G}\alpha_g(g\cdot w)\\
&= h\cdot\textstyle\sum_{g\in G}\alpha_g\theta_w(r_g)\\
&= h\cdot\theta_w(\textstyle\sum_{g\in G}\alpha_g r_g) \quad = \quad h\cdot\theta_w(v)\ .
\end{aligned}$$

Remarks. i) The representation R_G may appear to you to have been 'pulled out of a hat'. Here is a general construction which converts a finite G-set, X, into a G-module R_X. When the G-set is G acting on itself by left multiplication, this construction produces a G-module isomorphic to R_G.

Define R_X to be the complex vector space with basis X, and let the linear action on this vector space be the unique linear extension of the given action on the basis X. For example, the 3-dimensional representation of S_3 earlier was of this form, with X being the S_3-set which *defines* S_3. It is unusual for R_X to be irreducible. It never is for *decomposable* G-sets—those which can be written as the disjoint union of two non-empty G-subsets—and seldom is even when X is *in*decomposable.

ii) One can think of the proof of **47.11** as saying that R_G is the 'free G-module on one generator r_1' : for any G-module W and any $w \in W$, there is a unique G-map $\theta_w : R_G \to W$ mapping r_1 to w. This is analogous to $\mathbf{Z}[x]$ being *the free commutative ring on one generator* x: it maps uniquely by a ring morphism into any given ring S, sending x to any preassigned $s \in S$. There are *free groups, free abelian groups, free modules,* $\cdots$, in each case on one generator or, more generally, on any set of generators. See the mapping properties in Section **21**, called 'extension principles' there. It is often best to emphasize the mapping property rather than any particular construction of the free object. The mapping property guarantees that the free object is unique up to a unique isomorphism. Of course a particular construction is needed to be sure that the free object exists.

Exercise 47N. Prove in detail the statements above about mapping properties.

The results we have just proved are more often proved using the character theory of the next section. The methods just used are among Schur's many basic contributions to representation theory. About ten years earlier in 1899, Frobenius had invented the subject of character theory of finite groups (see Section **48**), and had proved versions of most of what we're doing in this chapter. There had been a long history of studying characters for *abelian* groups before that, with applications to physics, number theory, and probability; see **Mackey**.

Representations of finite abelian groups. A major difficulty often occurs in trying to give all the irreducible representations of some group explicitly (—**inventing a better matrep?)**. Abelian groups are not typical of the subject in most ways, including this. We know that such a G has exactly "$|G|$" irreducible representations, all 1-dimensional. Let's write them down explicitly. We can let G be the product $C_a \times C_b \times \cdots$, by the structure

theorem for finite abelian groups in multiplicative notation. We don't need the divisibility conditions on the orders of the factors for what follows. Let $\alpha,\ \beta, \cdots$ be fixed complex numbers which are primitive roots of unity of orders $a,\ b,\ \cdots$. Let $g,\ h,\ \cdots$ be generators for the factors in the previous direct product. For each sequence of integers $i,\ j,\ \cdots$ with $0 \leq i < a,\ 0 \leq j < b,\ \cdots$, let $\rho_{i,j,\ldots}$ be the representation in which $g \mapsto \alpha^i,\ h \mapsto \beta^j,\ \cdots$. It is easy to see that this determines a unique 1-dimensional representation. To see that this is the complete list as we vary the sequence $i,\ j,\ \cdots$, use the following easy exercise to deduce that no two of these representations are equivalent.

Exercise 47O. For *any* group G, suppose that ρ and μ are equivalent G-matreps. Show that if $\rho(g) = \alpha I$, then $\mu(g) = \alpha I$.

Exercise 47P. Find four distinct 1-dimensional representations of the dihedral group D_4 ; also find a 2-dimensional irreducible—recall the definition of that group! (But give *complex*, not real, representations.) Deduce that you've found all of its representations. Do the same for the quaternion group of order 8.

Addendum. We haven't made precise the claim after the first example, concerning the canonical decomposition of a G-module V into a direct sum of invariant subspaces, indexed by the irreducibles W, and each containing all copies of W which occur as invariant subspaces of V. This summand is called the *W-isotypical component*. A character theory version accessible to the masses is given in **48M**. For the elite who have learned about tensor products (see Appendix $\bigotimes$), here is a more elegant version.

Given also a finite dimensional vector space U, we can make $U \bigotimes_{\mathbf{C}} W$ into a G-module with action $\star$, by requiring that $(g\star)$ should be the unique linear operator on $U \bigotimes_{\mathbf{C}} W$ which maps $u \otimes w$ to $u \otimes (g \cdot w)$. This doesn't depend at all on the G-module W being irreducible. We'll revert now to "$\cdot$" rather than "$\star$".

Exercise 47Q. a) Prove that $U \bigotimes_{\mathbf{C}} W$ *is* a G-module.
b) Prove that this G-module is the direct sum of "dim$U$" copies of W.

Now revert to the assumption that W is irreducible, so that b) gives the decomposition of $U \bigotimes_{\mathbf{C}} W$ into irreducibles.
c) Show that there is a 1-1 correspondence between decompositions of $U \bigotimes_{\mathbf{C}} W$

into an internal direct sum of irreducibles and decompositions of U into an internal direct sum of 1-dimensional subspaces.

For V above, take $U = Hom_G(W,\ V)$, and define a map

$$\Gamma\ :\ \bigoplus_W Hom_G(W,\ V)\otimes_{\mathbf{C}} W\ \longrightarrow\ V\ ,$$

(where the direct sum is over all isomorphism classes of irreducible G-modules) by requiring that, on the W^{th} component of the domain of Γ, we set $\Gamma(\theta \otimes w) = \theta(w)$. So the map Γ is basically *evaluation of G-maps*. Denote the image of that W^{th} component under Γ as V_W, and call it *the W^{th} isotypical component of V*.

Theorem 47.14. *i) The map Γ is an isomorphism of G-modules.*
ii) Thus V is the internal direct sum of G-invariant subspaces V_W.
iii) Every invariant subspace of V is uniquely of the form

$$\sum_W \Gamma(U_W \otimes W)$$

for a collection of vector subspaces $U_W \subset Hom_G(W,\ V)$, and all such internal direct sums are invariant.

iv) An invariant subspace is a direct sum of copies of an irreducible W_0 if and only if all the corresponding U_W for $W \neq W_0$ are zero. It is isomorphic to W_0 itself if, in addition, U_{W_0} is 1-dimensional. Thus there is a 1-1 correspondence between invariant subspaces of V which are isomorphic to an irreducible W and 1-dimensional vector subspaces of $Hom_G(W,\ V)$.

Notice how this last statement includes Schur's lemma. The proof is straightforward, and will be left as **Exercise 47R.**

Exercise 47S. If X is a G-set and $x \in X$, the *orbit of x* is the set $\{\ gx\ :\ g \in G\ \}$. Show that there is a natural 1-1 correspondence between it and the cosets of G modulo the *stabilizer subgroup of x*, which is defined to be $\{\ g \in G\ :\ gx = x\ \}$.

This principle is used in the proofs of **11.1, 11.4, 50.10,** and **51.2**.

48. Characters of representations.

For any finite group G, let $\mathbf{C}^{G/\sim}$ be the vector space of all those functions $\Theta : G \to \mathbf{C}$ such that $\Theta(a^{-1}ba) = \Theta(b)$ holds for all a and b. This we'll call the space of *class functions on G*. The dimension of $\mathbf{C}^{G/\sim}$ is clearly equal to the number of conjugacy classes in G. In this section, we shall associate, to each representation of G, an element of $\mathbf{C}^{G/\sim}$, called the *character* of the representation. The *irreducible characters* will be shown to be a basis for $\mathbf{C}^{G/\sim}$ (but not the obvious basis). Thus we'll have proved the second claim after the table following **47G**—*the number of irreps equals the number of conjugacy classes.*

Aside. (Aside to the aside—as Shakespeare could have told you (but for having better taste in humour), **an aside is a digression which**

gets twice the gas kilometreage.) The genre of proof above—establishing an enumerative result with an algebraic proof—occurs often. (**Ex.** By considering the homogeneous components of the ring $F[s_1,\ s_2,\ \cdots]/\text{Ideal}\{s_i^2 - s_{2i} : i > 0\}$, where s_i has degree i, show that the number of partitions of an integer into odd parts is equal to its number of partitions into distinct parts.) Combinatorialists spend much worthwhile effort on making such proofs more direct. (**Ex.** Do it for this last example.) However, for the theorem above, there is no known explicit bijection, defined 'simultaneously' for all groups, which maps the set of irreps onto the set of conjugacy classes.

There are many other applications of characters besides the above theorem, which also has a '(group algebra)-style' proof. Indeed, as indicated earlier, Frobenius' original formulation of the subject was expressed entirely in terms of characters; representations didn't occur at all. We begin by defining the character of a representation.

Definition. For $A = (A_{ij}) \in \mathbf{C}^{n\times n}$, define the *trace of* A to be the sum of its diagonal entries:

$$\text{tr}(A) \ := \ \sum_1^n A_{ii} \ .$$

Proposition 48.1. *The trace of a matrix is also the sum of its eigenvalues, counted with their multiplicities as roots of its characteristic polynomial. For all square matrices A and invertible P, we have*

$$\text{tr}(P^{-1}AP) \ = \ \text{tr}(A) \ .$$

Proof. The first statement follows by examining the coefficient of the second highest power of x in

$$\det(xI - A) \ = \ \prod_j (x - \lambda_j) \ .$$

The second follows from the first, since $P^{-1}AP$ and A have the same characteristic polynomial. Here's a direct proof. Let $Q = (Q_{ij}) = P^{-1}$. Then

$$\begin{aligned}
\mathrm{tr}(P^{-1}AP) &= \textstyle\sum_i (QAP)_{ii} \\
&= \textstyle\sum_i \sum_{j,k} Q_{ij} A_{jk} P_{ki} \\
&= \textstyle\sum_{j,k} (\sum_i P_{ki} Q_{ij}) A_{jk} \\
&= \textstyle\sum_{j,k} (PQ)_{kj} A_{jk} \\
&= \textstyle\sum_{j,k} \delta_{kj} A_{jk} \qquad \text{(Kronecker delta)} \\
&= \textstyle\sum_j A_{jj} \quad = \quad \mathrm{tr}(A)\ .
\end{aligned}$$

Corollary 48.2. *For any linear operator $T : V \to V$, we have a well defined scalar $\mathrm{tr}(T) := \mathrm{tr}(A)$, where A represents T with respect to any basis of the finite dimensional vector space V.*

Remark. For readers who have learned about tensor products, here is the direct definition of the trace function on operators:

$$\text{trace} \ : \ Hom_{\mathbf{C}}(V,\ V) \ \longrightarrow \ V^* \bigotimes\nolimits_{\mathbf{C}} V \ \longrightarrow \ \mathbf{C}\ ,$$

where V^* is the dual of V, the right-hand map is *evaluation*, the linear map which sends $f \otimes v$ to $f(v)$, and the left-hand map is the inverse of the linear isomorphism which sends $f \otimes v$ to $[w \mapsto f(w)v]$. This will not be used below.

Definition. Let V be a G-module. The *character* of V is the function $\chi_V : G \to \mathbf{C}$ defined by $\chi_V(g) := \mathrm{tr}[(g\ \cdot) : V \to V\ ]$.
The corresponding object χ_ρ for a G-matrep ρ is evidently given by defining $\chi_\rho(g) := \mathrm{tr}[\rho(g)]$.

Corollary 48.3. *The function χ_ρ (resp. χ_V) is a class function, and depends only on the equivalence class of ρ (resp., on the isomorphism class of V).*
In the case of a matrep ρ, this is clear, since $\rho(a^{-1}ba) = \rho(a)^{-1}\rho(b)\rho(a)$, and since an equivalent matrep takes values of the form $P^{-1}\rho(a)P$. For G-modules, the proof is similar, or may be done by passing to the corresponding matrep.

Remarks. i) Since the eigenvalues of a finite order matrix are roots of unity, character values are sums of roots of unity, and therefore algebraic numbers. They are not arbitrary algebraic numbers, by the results of Abel/Galois/Ruffini. As well, in the next section, **49**, we'll see that they are so-called *algebraic integers*.

ii) For 1-dimensional matreps, there is no distinction between ρ and its character χ_ρ. In that case (**but only in that case**), we have

$$\chi_\rho(ab) \ = \ \chi_\rho(a)\chi_\rho(b) \ ,$$

since ρ is a morphism of groups.

Exercise 48A.

iii) Prove that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for *all* matrices A and B.
(For invertible matrices, this is immediate from **48.1**). It follows that $\chi_\rho(ab) = \chi_\rho(ba)$ (or, use that ab and ba are conjugate).

iv) The dimension of ρ is $\chi_\rho(1)$.

Proposition 48.4. *The character of the regular representation is given by*

$$\chi_{R_G}(g) \ = \ \begin{cases} |G| & \text{if } \ g = 1 \ ; \\ 0 & \text{if } \ g \neq 1 \ . \end{cases}$$

Proof. Recall that $g \cdot r_h = r_{gh}$. Thus, for fixed g, the coefficient of r_h in $g \cdot r_h$ is 0 if $g \neq 1$, and is 1 if $g = 1$. Summing over h gives the result.

Exercise 48B. More generally, given a finite G-set X, find a formula for the character of the associated linear representation, R_X, in terms of fixed points; and show that it reduces to **48.4** when X is G under left multiplication.

The following easy fact shows that the characters of all representations are known once we have the irreducible characters.

Proposition 48.5. *We have* $\chi_{V\oplus W}(g) \ = \ \chi_V(g) + \chi_W(g)$.

Exercise 48C. Prove this.

Exercise 48D. Calculate the characters of all the irreducible representations in the previous section, without peeking below until you've finished.

Each finite group has a *character table*. In it, one lists the names of the irreducible representations down the left-hand side and the conjugacy classes across the top, and enters the character values into the body of the table;

and then decorates it to taste with various bells and whistles. A handy one to include with each conjugacy class is the *size* of the class. Here are the examples from the previous section, where I've added an extra bottom line (not officially part of the character table) giving the characters of the regular representation (which is not irreducible for non-trivial groups). The names $\rho^2,\ \mu\nu,\ \cdots$ will be explained later in **48H**.

C_2	[1] 1	[1] g
χ_{triv}	1	1
χ_ρ	1	-1
χ_{reg}	2	0

C_3	[1] 1	[1] g	[1] g^2
χ_{triv}	1	1	1
χ_ρ	1	ω	ω^2
χ_{ρ^2}	1	ω^2	ω
χ_{reg}	3	0	0

$(\omega = \mathrm{e}^{2\pi i/3})$

D_2	[1] 1	[1] a	[1] b	[1] ab
χ_{triv}	1	1	1	1
χ_μ	1	-1	1	-1
χ_ν	1	1	-1	-1
$\chi_{\mu\nu}$	1	-1	-1	1
χ_{reg}	4	0	0	0

C_4	[1] 1	[1] h	[1] h^2	[1] h^3
χ_{triv}	1	1	1	1
χ_ρ	1	ω	ω^2	ω^3
χ_{ρ^2}	1	ω^2	ω^4	ω^6
χ_{ρ^3}	1	$-i$	-1	i
χ_{reg}	4	0	0	0

$(\omega = \mathrm{e}^{2\pi i/4})$

S_3	[1] $1-$cycle	[3] $2-$cycles	[2] $3-$cycles
χ_{triv}	1	1	1
χ_{sign}	1	-1	1
χ_{other}	2	0	-1
χ_{reg}	6	0	0
χ_{standard}	3	1	0

For S_3 we've added *two* illegitimate lines at the bottom. The standard representation is the 3-dimensional linearization, after **47E** and from Remark i) before **47N**, of the defining action for S_3 on $\{\ 1,\ 2,\ 3\ \}$.

Looking at these we see that the extra row(s) on the bottom are in agreement with **48.5**. If you add up each column in the character table itself, weighting each entry by multiplying by the entry to its left in the first column, you get the entry for the regular representation, i.e.

$$\chi_{\text{reg}} \;=\; \sum_{\text{irreps } \rho} \dim\rho \; \chi_\rho \; .$$

This is equivalent to **47.12**, and is also a special case of orthogonality below. Similarly, the equation

$$\chi_{\text{standard}} \;=\; \chi_{\text{other}} \;+\; \chi_{\text{triv}}$$

checks out. Note also that it is normal practice to put the identity element first, so that the leftmost column lists the dimensions of the irreducibles. The overworked word 'degree' is often used to mean dimension.

Exercise 48E. Expanding on **47P**, calculate the character tables for both non-abelian groups of order 8.

Now look closely at the seven character tables you've got [not at the illegitimate lower line(s)].

Firstly they are *square*, with (say) "m" rows and "m" columns; that's the theorem (yet to be proved) from the introductory paragraph of this section.

Next, *the columns*, regarded as vectors in $\mathbf{C}^m$, *form an orthogonal basis* with respect to the standard (hermitian) inner product, with each vector of length $\sqrt{|G|/(\text{class size})}$.

Finally, *the rows*, regarded as vectors in $\mathbf{C}^m$, *form an orthogonal basis*, with each vector of length $\sqrt{|G|}$, with respect to a weighting of the standard (Hermitian) inner product, where we "weight" by multiplying each term in the calculation of the standard inner product by the number of elements in that conjugacy class.

These last three statements hold for any finite group. (If they don't hold for your answer to **48E**, then start again. If they do, then you're almost certainly correct.) The last two observations are called the *character orthogonality relations*. They are far easier to remember in the above form, as handwaving facts with a schematic picture of a table in your head, than to remember as the formulae given in **48.12** and in the proof of **48.7** below.

Another way to remember the row orthogonality is that the irreducible characters form an ortho*normal* set with respect to a standard Hermitian inner product on the space of *all* complex valued functions on G. We discuss this further after the proof of **48.6**.

Perhaps one can prove orthogonality without using the equality between the numbers of conjugacy classes and of irreps. Suppose that we've done so, and that these two numbers differ for some group. By taking the entries of its character table and dividing each by the square root of the number of elements in the conjugacy class label for that entry's column, form a non-square matrix, M. Orthogonality tells us that we have a matrix for which both of the square matrices $M^{tr}M$ and MM^{tr} are diagonal matrices with non-zero diagonal entries. But this is ridiculous, since there's certainly no pair of non-square matrices M and N of complementary shape such that the square matrices MN and NM both have maximal rank—the rank cannot exceed the smaller of the two matrix dimensions involved. We shall redo this proof of the squareness of the character table directly in **48.11** below, without having to quote these rank facts from linear algebra.

The proofs of orthogonality below look a bit formidable. However, they could be reproduced by a well trained drone who was provided with the following information—so remember this and the rest is essentially calculation.

In both cases you apply Schur's Lemma.

For row orthogonality, apply it to the following map, depending on an arbitrary linear map $\alpha : V_1 \to V_2$ between G-modules with actions $\cdot$ and $\star$ (this is the 'Maschke averaging trick') :

$$\phi_\alpha \ : \ V_1 \longrightarrow V_2 \ ,$$

$$v \longmapsto \sum_{g \in G} g^{-1} \star \alpha(g \cdot v) \ .$$

For column orthogonality, apply it to the following map, defined for any G-module U and any conjugacy class C in G:

$$\begin{aligned} \psi_C \ : \ U \ &\longrightarrow \ U \ , \\ u \ &\mapsto \ \sum_{h \in C} h \cdot u \ . \end{aligned}$$

Lemma 48.6 *These two maps are in fact G-maps.*

Proof. The proof for ϕ_α is exactly the same as that within Maschke's theorem, **47.2**. As for the other, using the fact that C is a conjugacy class to justify the fourth equality,

$$\begin{aligned} \psi_C(g \cdot u) \ &= \ \sum_{h \in C} hg \cdot u \ = \ \sum_{h \in C} gg^{-1}hg \cdot u \\ &= \ g \cdot \sum_{h \in C} g^{-1}hg \cdot u \ = \ g \cdot \sum_{k \in C} k \cdot u \ = \ g \cdot \psi_C(u) \ . \end{aligned}$$

as required.

To discuss row orthogonality, here's some useful notation. Let $\mathbf{C}^G$ denote the vector space of *all* functions from G to $\mathbf{C}$. Give it the following Hermitian inner product:

$$< \alpha \mid \beta > \ := \ |G|^{-1} \sum_{g \in G} \alpha(g)\overline{\beta(g)} \ .$$

The standard basis for $\mathbf{C}^G$ consists of the functions which take one value equal to 1, and the rest equal to 0. Then the above is the standard inner product which would make this into an orthonormal basis, except that we've divided by $|G|$. Now row orthogonality may be stated as follows.

Theorem 48.7. *The set of irreducible characters is an orthonormal set in $\mathbf{C}^G$, with respect to the inner product $<\ \mid\ >$. In particular, that set is linearly independent.*

Remark. Note that characters have the property

$$\chi_\mu(g^{-1}) \ = \ \overline{\chi_\mu(g)} \ ,$$

since the eigenvalues for A^{-1} are the reciprocals of those for A, and since $z^{-1} = \overline{z}$ for a root of unity z. Thus we have

$$< \chi_\rho \mid \chi_\mu > \ = \ |G|^{-1} \sum_{g \in G} \chi_\mu(g^{-1})\chi_\rho(g) \ .$$

Exercise 48F. Assume that $\{\ z(g)\ :\ g \in G\ \}$ is a collection of numbers for which $z(a^{-1}ba) = z(b)$ for all a and b; i.e. a class function. Let the right-hand sum below be over all the conjugacy classes C in G, with g_C being some chosen element in C. Prove that

$$\sum_{g\in G} z(g) \ = \ \sum_{C} |C| z(g_C)\ .$$

Deduce that **48.7** is really the same as row orthogonality of the character table.

Proof of 48.7. As we said earlier, this will follow fairly mechanically by applying Schur's lemma to the G-map, ϕ_α, between two irreducible G-modules V_1 and V_2. For any choice of $\mathbf{C}$-linear map α, that G-map must be zero if V_1 and V_2 are not isomorphic, and is a scalar multiple of the identity map when $V_1 = V_2$.

First suppose that ρ and μ are inequivalent irreducible matreps, obtained by choosing bases in the non-isomorphic irreducible G-modules V_1 and V_2 respectively. Let X be the matrix representing α with respect to these bases. We therefore have, from the definition of ϕ_α,

$$\sum_{g\in G} \mu(g^{-1}) X \rho(g) \ = \ 0$$

for all matrices X of a suitable size, where the right-hand side is a zero matrix. Looking at the $(i,\ell)^{\text{th}}$ entry in this identity, we get

$$\sum_{j,k;\ g\in G} \mu(g^{-1})_{ij}\ X_{jk}\ \rho(g)_{k\ell} \ = \ 0$$

for all (i,ℓ) and all choices of complex numbers X_{jk}. Making one of these choices to be 1, and the rest to be 0, it follows that for all i, j, k, and ℓ, we have

$$\sum_{g\in G} \mu(g^{-1})_{ij}\ \rho(g)_{k\ell} \ = \ 0\ .$$

Specialize this by taking $i = j$ and $k = \ell$, sum over all (i,k), and divide by $|G|$:

$$|G|^{-1} \sum_{g\in G}\ (\sum_i \mu(g^{-1})_{ii})\ (\sum_k \rho(g)_{kk}) \ = \ 0\ .$$

This is exactly

$$|G|^{-1} \sum_{g \in G} \chi_\mu(g^{-1}) \chi_\rho(g) \;=\; 0 \;,$$

that is, $< \chi_\rho \mid \chi_\mu > \;=\; 0$, as required.

Now instead take $\mu = \rho$, an irreducible matrep coming from the irreducible G-module $V_1 (= V_2)$ with some basis. Again let X represent α using that basis. By the first paragraph of the proof, we get

$$\sum_{g \in G} \rho(g^{-1}) X \rho(g) \;=\; zI \;,$$

for some complex number z. Letting $n = \dim V_1$ be the size of the matrices, and taking the trace of both sides, we obtain $z = |G|\ \mathrm{tr}(X)/n$. The $(i,\ell)^{\text{th}}$ entry gives

$$\sum_{j,k;\ g \in G} \rho(g^{-1})_{ij}\ X_{jk}\ \rho(g)_{k\ell} \;=\; \begin{cases} n^{-1}|G| \sum_k X_{kk} & \text{if } \; i = \ell \;; \\ \\ 0 & \text{if } \; i \neq \ell \;. \end{cases}$$

This holds for all (i, ℓ) and all choices of complex numbers X_{jk}. As in the paragraph above, 'equating coefficients' of the X's yields

$$\sum_{g \in G} \rho(g^{-1})_{ij}\ \rho(g)_{k\ell} \;=\; \delta_{jk}\delta_{i\ell}|G|/n \;, \quad \text{(Kronecker deltas)}.$$

This leads quickly to

$$\sum_{i,k;\ g \in G} \rho(g^{-1})_{ii}\ \rho(g)_{kk} \;=\; |G| \;,$$

which is the same as $< \chi_\rho \mid \chi_\rho > \;=\; 1$, as required.

Corollary 48.8. *i) Two G-modules are isomorphic if (and only if) their characters are equal.*

ii) If $V \cong \bigoplus U^{\oplus n_U}$ and $W \cong \bigoplus U^{\oplus m_U}$ are the decompositions of two G-modules into direct sums of irreducibles U, then

$$< \chi_V \mid \chi_W > \ = \ \sum_U n_U m_U \ = \ \dim Hom_G(V,\ W) \ .$$

iii) In particular, the number of copies of an irreducible U in the direct sum decomposition of V is $< \chi_V \mid \chi_U >$.

iv) A G-module V is irreducible if and only if $< \chi_V \mid \chi_V > \ = \ 1$.

Proof. If V and W are two G-modules, decomposed into direct sums of irreducibles as in *ii)*, then $\chi_V = \sum_U n_U \chi_U$ and similarly for W, using the m_U. Then *i)* is immediate, since the linear independence of the irreducible characters yields $n_U = m_U$ for all U. The first equality in *ii)* is clear from the usual calculation of an inner product of two vectors expressed in terms of an orthonormal set. The second equality in *ii)* is immediate from **47.9** and **47L**. Part *iii)* is a special case of *ii)*. Part *iv)* follows because a sum, $\sum_U (n_U)^2$, of squares of non-negative integers can only equal 1 when one of them is 1 and the rest are 0.

Remark. Note how the two significant, but initially obscure, technical objects, $<\ \mid\ >$ and $\dim Hom_G$, from this and the previous sections, have turned out to be the same object in different clothing !

Now let's return to the other two observations concerning character tables, namely squareness and column orthogonality.

Lemma 48.9. *Let ρ be an irreducible G-matrep of dimension n, and let $g_C \in C$, a conjugacy class in G. Then*

$$\sum_{g \in C} \rho(g) \ = \ n^{-1}|C|\ \chi_\rho(g_C) I \ .$$

Proof. If ρ is the matrep associated to an irreducible G-module U with some basis, then the left-hand side of the identity to be proved is the matrix of the G-map ψ_C (defined before **48.6**). By Schur's lemma, it must be zI for some z. Taking traces immediately yields the fact that $z = n^{-1}|C|\ \chi_\rho(g_C)$, as required.

Corollary 48.10. *For any class function $f \in \mathbf{C}^{G/\sim}$, and any irreducible G-matrep ρ of dimension n, we have*

$$|G|^{-1} \sum_{g \in G} f(g)\rho(g) \;=\; n^{-1} < \chi_\rho \mid \overline{f} > I \ .$$

Proof. With the outside sum over all conjugacy classes C, and with notation as in **48.9**, the left-hand side above is

$$|G|^{-1} \sum_C \sum_{g \in C} f(g)\rho(g) \;=\; |G|^{-1} \sum_C f(g_C) \sum_{g \in C} \rho(g)$$

$$=\; n^{-1}|G|^{-1} \sum_C f(g_C)\ |C|\ \chi_\rho(g_C)\ I \;=\; n^{-1}|G|^{-1} \sum_{g \in G} f(g)\ \chi_\rho(g)\ I \ ,$$

which is the right-hand side. (Alternatively, there is a G-module with a self-G-map for which the left-hand side is the matrix.)

Theorem 48.11. *The set of irreducible characters of a finite group is an orthonormal basis for the subspace of class functions on that group (with respect to the restriction of our standard Hermitian inner product on the space of* all *functions). In particular, the number of irreducible representations of a group equals its number of conjugacy classes.*

Proof. Since the irreducible characters form an orthonormal set, they will be a basis for $\mathbf{C}^{G/\sim}$, as required, as long as no non-zero class function, f, is orthogonal to all of them. But if $< \chi_\rho \mid f >= 0$ for all irreducible characters ρ, then from **48.10** we obtain

$$\sum_{g \in G} \overline{f(g)}\rho(g) \;=\; 0 \ ,$$

for all ρ which are irreducible, and therefore also for all ρ whatsoever. Taking the latter to be the regular representation, and applying this to the element r_1, we get

$$\sum_{g \in G} \overline{f(g)} r_g \;=\; 0 \ .$$

Thus $f(g) = 0$ for all g, as required, since $\{\ r_g \ :\ g \in G\ \}$ is linearly independent.

Exercise 48G. Show that if all the irreducible representations of G are 1-dimensional, then G is abelian (the converse of **47.13**).

Here are the column orthogonality relations for the character table.

Theorem 48.12. *Let g and h be elements of G. Summing over all equivalence classes of G-matreps ρ, we have*

$$\sum_\rho \overline{\chi_\rho(g)}\chi_\rho(h) = \begin{cases} |G|/|C| & \text{if } g, h \text{ are in the } \textit{same} \text{ conjugacy class } C; \\ \\ 0 & \text{if } g \text{ and } h \text{ are not conjugate .} \end{cases}$$

Proof. Let $\Delta_C : G \to \mathbf{C}$ be the characteristic function of the conjugacy class C ; i.e. $\Delta_C(g)$ is 1 or 0 according as $g \in C$ or $g \notin C$. Directly from the definition of the inner product we see that

$$< \chi_\rho \mid \Delta_C > \quad = \quad |G|^{-1}|C|\ \chi_\rho(g_C)\ ,$$

where, as usual, g_C is a chosen element in C. Therefore, using the ancient formula for writing a vector in terms of an orthonormal basis,

$$\Delta_C \quad = \quad \sum_\rho < \Delta_C \mid \chi_\rho > \chi_\rho \quad = \quad |G|^{-1}|C| \sum_\rho \overline{\chi_\rho(g_C)}\ \chi_\rho\ .$$

Let D also denote a conjugacy class. Using a daring form of the Kronecker delta, we get

$$\delta_{C,D} \quad = \quad \Delta_C(g_D) \quad = \quad |G|^{-1}|C| \sum_\rho \overline{\chi_\rho(g_C)}\ \chi_\rho(g_D)\ .$$

This is just what the theorem is asserting.

Orthogonality can be used to help complete a partly cooked character table. For example, if you know all but the last row, then, by row orthogonality, the last row is determined up to multiplying its elements by a complex number of modulus 1. But its first entry is a positive integer, so this determines the row completely. This may easily result in a situation where you know all the characters of a group, but don't have explicit formulae for all of its representations. You'll be in good company—Frobenius determined

the characters of all the symmetric groups (these characters actually happen to be integers), but it was several decades later when Young and Specht wrote down the actual representations. More dramatically: in 1911, Schur determined the characters of a sequence of groups, of orders $n!2$, which map surjectively onto the symmetric groups ('the *projective* characters of the symmetric group'); but it was 1988 before the representations themselves were finally found by Maxim Nazarov.

The vector space $\mathbf{C}^G$ is a ring, and $\mathbf{C}^{G/\sim}$ a subring, using multiplication of values (so-called 'pointwise multiplication'). The subset consisting of the characters of all (not necessarily irreducible) representations is obviously closed under addition. But it is also closed under multiplication—it is a 'semi-ring',—as we see from **48K** and **48L** below, which depend on the tensor product.

Exercise 48H. (This explains the names of some of the representations in the character tables exhibited earlier.) Show that if ρ and μ are G-matreps, with ρ of dimension 1, then $g \mapsto \rho(g)\mu(g)$ is also a matrep, which is irreducible if and only if μ is. Furthermore, $g \mapsto \rho(g)^{-1}$ is a matrep.

Exercise 48I. Find two 1-dimensional irreps of S_n for any $n > 1$. Show that the standard representation splits as the direct sum of irreps of dimensions $n-1$ and 1. Find a fourth irrep, also of dimension $n-1$, when $n > 3$. Determine the character table of S_4—note how only integers somehow magically occur.

Exercise 48J. Show that if each element of a group is conjugate to its inverse, then all of the character values of the group are real.

Exercise 48K. Let V be a G-module and W an H-module, for any two finite groups, G and H. Determine an action of $G \times H$ on $V \otimes_{\mathbf{C}} W$ by requiring

$$(g,\ h) \cdot (v \otimes w) \ := \ (g \cdot v) \otimes (h \cdot w) \ .$$

Explain how this, for fixed $(g,\ h)$, determines a well defined linear map. Prove that you get an action of $G \times H$. Show that

$$\chi_{V \otimes W}(g,\ h) \ = \ \chi_V(g)\chi_W(h) \ .$$

How does this relate to **48H**? Prove that

$$< \chi_{V\otimes W} \mid \chi_{V'\otimes W'} > \quad = \quad < \chi_V \mid \chi_{V'} > \ < \chi_W \mid \chi_{W'} > \ .$$

Deduce that if V and W are irreducible, then so is $V \bigotimes W$. Deduce further that, if also V' and W' are irreducible, then $V \bigotimes W$ and $V' \bigotimes W'$ are isomorphic as $G \times H$-modules if and only if both $V \cong_G W$ and $V' \cong_H W'$. Conclude that the irreducible $(G \times H)$-modules are exactly the $V \bigotimes W$, as V ranges over all irreducible G-modules, and W ranges over all irreducible H-modules.

Exercise 48L. Given a group morphism $\beta : G \to K$, and a K-module V, define $\beta^* V$, the *restriction of* V *along* β, to be V with the G-action $g \cdot v \ := \ \beta(g) \cdot v$. Prove that this is a G-module. (Sometimes this is given less generally in two cases: when β is inclusion of a subgroup, it is called restriction; and when β is projection onto a quotient group, it's called 'lifting'.) When $K = G \times G$ and β is the diagonal map $g \mapsto (g,\ g)$, we can apply **48K** and form $V \bullet W := \beta^*(V \bigotimes W)$ for a pair of G-modules. This is also called their tensor product, and is seldom irreducible. It generalizes **48H** : prove that

$$\chi_{V\bullet W} \ = \ \chi_V \chi_W \ .$$

Deduce that the set of characters is closed under multiplication. By multiplying known characters, one can often produce new representations of interest—starting with the trivial, sign, and standard representations, generate the complete character table of S_5 using various tensor products and restrictions.

Exercise 48M. Show that the projection of V to its U-isotypical component is given by

$$v \ \mapsto \ |G|^{-1} \mathrm{dim} U \sum_{g\in G} \overline{\chi_U(g)}\ g \cdot v \ .$$

That is, this map is the identity on the U-isotypical component, and is zero on all the other isotypical components. (See **47.14**.)

49. Algebraic integers.

The following concept has several applications in mathematics besides the applications in the next section to representation theory. Readers who have not yet studied the definitions beginning Section **43** can specialize R to $\mathbf{Z}$ below, and change the term *R-module* to *abelian group*. All of our applications will be in that case.

Definition. Let $S \supset R$ be an extension of commutative rings. An element $s \in S$ is said to be *integral over* R if and only if the conditions in the following theorem hold. The usual definition is condition *i)*. The term *'algebraic integer'* is synonymous with *'integral over* $\mathbf{Z}$*'*.

Theorem 49.1. *With notation as above, the following conditions are equivalent.*

i) There is a monic polynomial in $R[x]$ *having* s *as a root.*

ii) The ring $R[s]$ *is finitely generated as an* R*-module.*

iii) There exists an intermediate ring T *(i.e.* $S \supset T \supset R$*) with* $s \in T$ *and such that* T *is finitely generated as an* R*-module.*

Proof. To prove that *i)* implies *ii)*, note that if some monic polynomial as in *i)* has degree n, then all higher powers of s can be expressed as R-linear combinations of the set $\{ 1, s, \cdots, s^{n-1} \}$, using that polynomial. Therefore that finite set of powers generates $R[s]$ as an R-module. That *ii)* implies *iii)* is trivial, taking T to be $R[s]$. To see that *iii)* implies *i)*, see **Hungerford** VIII 5.3 for the general case. Here is a shorter argument when $R = \mathbf{Z}$, the only case which we'll use. By **13F**, any subset of a finitely generated abelian group generates a subgroup which can be generated by a finite subset of the given set. Taking the group to be T and the set to consist of the powers of s, this provides a guarantee that some power of s is a $\mathbf{Z}$-linear combination of lower powers of s, as required.

Corollary 49.2. *The subset of elements in* S *which are integral over* R *is a subring of* S *and an extension of* R*.*

Proof. The set of products of pairs of elements, the first from a set of R-module generators for $R[s]$ and the second from such a set for $R[t]$, is a set of R-module generators for $R[s, t]$. Taking T in **49.1***iii)* to be $R[s, t]$, we see that if s and t are both integral over R, then so are $s \pm t$ and st, as required. (This proves generally that the ring generated by the union of any two 'integral extensions' is also integral.)

Corollary 49.3. *If S is finitely generated as an R-module, then all of its elements are integral over R.*

Proof. Take $T = S$ in *iii)* of **49.1**.

Proposition 49.4. *The intersection of $\mathbf{Q}$ with the ring of algebraic integers in $\mathbf{C}$ is $\mathbf{Z}$. That is, a rational which is an algebraic integer is actually an integer.*

Proof. If the rational b/c in lowest form is a root of $\sum_0^n a_i x^i$ for integers a_i with $a_n = 1$, then

$$b^n \ + \ a_{n-1}b^{n-1}c \ + \ a_{n-2}b^{n-2}c^2 \ + \ \cdots \ + \ a_0c^n \ = \ 0 \ .$$

It follows easily that any prime dividing c would divide b also. Therefore $c = \pm 1$, as required.

Any root of unity is an algebraic integer, and therefore any sum of roots of unity is also. Thus, since character values for finite groups are always sums of roots of unity, we immediately get the following result, taking us back to the central topic of this chapter.

Proposition 49.5. *If G is a finite group, then its character values, $\chi_\rho(g)$, are algebraic integers for all $g \in G$ and all G-matreps ρ.*

Proposition 49.6. *Let S and S' both be extensions of R. A morphism of rings, $\phi : S \to S'$, which fixes all elements of R, necessarily maps elements integral over R in S to elements integral over R in S'. In particular, any ring morphism between commutative rings maps algebraic integers to algebraic integers.*

Proof. If, for a polynomial f, we have $f(s) = 0$, then

$$f(\phi(s)) \ = \ \phi(f(s)) \ = \ \phi(0) \ = \ 0 \ .$$

We finish this section with a few unmotivated special facts, which will be crucial in the next section.

Proposition 49.7. *Suppose that $\alpha \in \mathbf{C}$ is an algebraic integer with $|\alpha| < 1$ and such that $|\beta| \leq 1$ for all (Galois) conjugates β of α. Then $\alpha = 0$.*

Proof. Let γ be the product of all the conjugates of α. Then γ is fixed by the Galois group over $\mathbf{Q}$ of the minimal polynomial of α. Therefore it is in $\mathbf{Q}$. But γ is an algebraic integer, so it's in $\mathbf{Z}$ by **49.4**. Being less than 1 in absolute value, it must be 0. Hence $\alpha = 0$ as well.

Lemma 49.8. *If each α_i is a complex root of unity, then*

$$| \ \alpha_1 \ + \ \cdots \ + \ \alpha_k \ | \ \ \leq \ \ k \ .$$

Equality holds if and only if all the α_i are equal.

Proof. This is a standard application of the triangle inequality in $\mathbf{R}^2$, since each α_i has modulus 1.

Proposition 49.9. *If $\alpha = (\alpha_1 + \cdots + \alpha_k)/k$ is an algebraic integer in $\mathbf{C}$, and if each α_i is a root of unity, then either $\alpha = 0$ or else $\alpha_i = \alpha$ for all i.*

Proof. If the α_i are not all equal, then **49.8** implies that the hypotheses of **49.7** hold, and then its conclusion gives us exactly what we want. (This uses that 'conjugation' commutes with addition, and 'maps' roots of unity to roots of unity.)

Proposition 49.10. *Suppose that a and b are relatively prime integers, and z is a complex number such that both z and az/b are algebraic integers. Then z/b is also an algebraic integer.*

Proof. We have $z/b = (u)(az/b) + (v)(z)$ for integers u and v such that $ua + vb = 1$. Now apply **49.2**.

50. Dimension divides order & the Burnside $(p,\ q)$-theorem.

First let's define the group algebra of a finite group. This will not be used in any serious way; only the fact that the set in **50.3** below is a commutative ring satisfying **50.3** is needed later.

Definition. The *group algebra* of G is denoted $\mathbf{C}[G]$. As a set, it is the set of all formal linear combinations, $\sum_{g \in G} z_g g$, where the coefficients are

complex numbers. These are added and multiplied in the obvious way, where the multiplication uses the operation in the group:

$$(\sum_{g\in G} z_g g)\ (\sum_{g\in G} w_g g)\ :=\ \sum_{g\in G}\ (\sum_{ab=g} z_a w_b)\ g\ .$$

Proposition 50.1. *i) This produces a ring*—in fact, a **C**-algebra in the sense of Section **52**—*which is in general non-commutative.*

ii) It is commutative if and only if G is an abelian group.

Exercise 50A. Prove this.

We'll only be using the following commutative subring.

Proposition 50.2. *The* centre *of* $\mathbf{C}[G]$ (i.e. the set of elements which commute with every element) *is the vector subspace with basis over* **C** *equal to* $\{\ e_C : C$ is a conjugacy class in $G\ \}$, *where*

$$e_C\ :=\ \sum_{g\in C} g\ .$$

Proof. The linear independence of $\{\ e_C\ :\ \cdots\ \}$ is obvious. That e_C commutes with everything is a routine verification, so we get one inclusion. For the opposite one, let $s = \sum_{g\in G} z_g g$ be any element in the centre of the group algebra. To show that $z_{h^{-1}gh} = z_g$, just calculate with the equality $sh = hs$.

Exercise 50B. Give all the details of this proof.

Proposition 50.3. *The set of* integer *linear combinations of the basis above, $Span_{\mathbf{Z}}\{\ e_C : C$ is a conjugacy class in $G\ \}$, consists entirely of algebraic integers.*

Proof. This is immediate from **49.3**, observing that the set is a subring.

Recall **48.9**, which said that if ρ is an irreducible G-matrep and C is a conjugacy class in G with $g_C \in C$, then

$$\sum_{g\in C} \rho(g)\ =\ m(C,\ \rho)I\ ,$$

where

$$m(C,\ \rho)\ :=\ |C|\ \chi_{\rho}(g_{C})/\dim\rho\ .$$

Proposition 50.4. *The map $\sum_C z_C e_C \mapsto \sum_C z_C m(C,\ \rho)$ is a ring morphism from the centre of the group algebra to* $\mathbf{C}$.

Exercise 50C. Prove this.

Combining **50.4, 50.3** and **49.6**, we get

Corollary 50.5. *For all conjugacy classes C in G and all irreducible G-matreps ρ, the complex number $m(C,\ \rho)$ is an algebraic integer.*

Now we can prove the last of the three claims made after the table in Section **47**.

Theorem 50.6. *The dimension of any irreducible G-module divides the order of the finite group G.*

Proof. Denoting the irreducible matrep as ρ, we have

$$1 \quad = \quad <\chi_{\rho}\mid\chi_{\rho}> \quad = \quad |G|^{-1}\sum_{g\in G}\chi_{\rho}(g)\overline{\chi_{\rho}(g)}$$

$$= \quad |G|^{-1}\dim\rho\sum_{C}\{|C|\ \chi_{\rho}(g_{C})/\dim\rho\}\ \{\overline{\chi_{\rho}(g_{C})}\}\ .$$

Both factors {between the brace-brackets} in the last sum are algebraic integers, by **50.5** and **49.5**, respectively. Since the set of algebraic integers is closed under multiplication and addition, we see that $|G|/\dim\rho$ is an algebraic integer. But it's a rational number. Hence it is an integer, as required, by **49.4**.

Combining this divisibility condition with the squareness of the character table and the 'sum of squares' condition puts strong constraints on what the dimensions of the irreducibles can be. Every group has at least one 1-dimensional representation, the trivial one.

Thus, for example, a group of order p^2 can only have 1-dimensional representations, since all other divisors are already too big. Therefore the group is necessarily abelian by **48G**, an immediate consequence of the squareness

of the character table. This is a new proof of **11.5**, giving a low-level illustration of the use of representation theory to get results about the structure of finite groups.

Similarly, a group of order p^3 will have, say, "a" irreducibles of dimension p, and "$p^3 - ap^2$" of dimension 1. Thus its number of conjugacy classes has the form $p^3 - ap^2 + a$ for some integer a between 0 and $p-1$. The number of conjugacy classes determines the dimensions of the irreducibles completely. In fact, only the values $a = 0$ and 1 occur.

Lemma 50.7. *Suppose that $\chi_\rho(g)/\dim(\rho)$ is an algebraic integer, where $g \in G$ and ρ is an irreducible G-matrep. Then either $\chi_\rho(g) = 0$ or $\rho(g)$ is a scalar multiple of the identity matrix.*

Proof. In **49.9**, take $k = \dim\rho$, and let $\alpha_1, \cdots,\ \alpha_k$ be the list of eigenvalues of $\rho(g)$, listed once for each time they occur as roots of the characteristic polynomial.

Lemma 50.8. *If C is a conjugacy class in G with $g_{_C} \in C$, and ρ is an irreducible G-matrep such that $|C|$ and $\dim\rho$ are relatively prime, then $\chi_\rho(g_{_C})/\dim\rho$ is an algebraic integer.*

Proof. This is immediate from **49.10** with $a = |C|$, $b = \dim\rho$ and $z = \chi_\rho(g_{_C})$, using **49.5** and **50.5**.

Theorem 50.9. *If some conjugacy class in a finite group has prime power order larger than 1, then the group is not simple.*

Proof. Let C be that conjugacy class in G, with $|C| = p^s$ and $g_{_C} \in C$. By column orthogonality, **48.12**, we have

$$\sum_\rho \chi_\rho(1)\chi_\rho(g_{_C}) \ = \ 0\ ,$$

so that

$$1 \ + \sum_{\rho\ \text{non}-\text{trivial}} \dim\rho\ \chi_\rho(g_{_C}) \ \ = \ \ 0\ .$$

Now $\chi_\rho(g_{_C})$ is an algebraic integer by **49.5**, and for no algebraic integer z is $1 + pz = 0$, by **49.4**. Thus we can find an irreducible ρ such that

i) the dimension of ρ is not divisible by p;

ii) $\chi_\rho(g_{_C}) \ \neq \ 0$;

iii) $\chi_\rho(g_C)/\dim\rho$ is an algebraic integer (by **50.8**).

From **50.7**, we conclude that $\rho(g_C)$ is a scalar multiple of the identity. But the scalar matrices form a normal subgroup, N, of $GL(n,\ \mathbf{C})$, so $\rho^{-1}N$ is now a non-trivial normal subgroup of G. If it's proper, we're finished. If not, then use instead the normal subgroup $\mathrm{Ker}\rho$, which is

a) proper, because ρ is not the trivial representation, and

b) non-trivial, because ρ maps the non-abelian group G into the abelian group of scalar matrices.

Theorem 50.10. (Burnside) *If a finite group has order divisible by fewer than three primes, then it is soluble.*

Proof. Suppose, for a contradiction, that G is a counterexample of smallest order. A non-trivial proper normal subgroup, N, of G would then produce soluble groups N and G/N. But then G would be soluble, by **11G**. Therefore N can't exist, and so G is a simple non-abelian group. Let its order be p^aq^b for distinct primes p and q, with $b > 0$. Let H be a Sylow subgroup of order q^b. By **11.4**, the group H has a non-trivial centre, since it has prime power order. Let h be a non-identity element in the centre of H. Let $K = \{\ g \in G\ :\ gh = hg\ \}$, the centralizer of h in G. Since K is a subgroup containing H, we have $|K| = p^cq^b$, where $0 \leq c \leq a$. If $c = a$, then h is in the centre of G, and the latter would be a proper non-trivial normal subgroup of G. Therefore $c < a$. But the conjugacy class of h has order $|G|/|K| = p^{a-c}$ (see **47S**). By **50.9**, we have our contradiction.

See the remarks after **12.2** about finite simple groups—a re-phrasing of **50.10** is that a finite non-abelian simple group must have order divisible by at least three primes. The simplicity of A_5 shows that this is a best possible result in the simplest sense. The renowned Feit-Thompson theorem says that one of the prime divisors of the order of a finite non-abelian simple group is necessarily 2. Examples show that there aren't any odd primes with that privileged status.

We have pushed ahead into deeper waters in this section, and omitted, for example, some quite important methods for the construction of representations—induced representations, exterior and symmetric powers, further development of the products in the exercises at the end of **48**, $\cdots$. Another topic of importance (which is less difficult than some of what we've

done) is the analysis of representations over general fields of characteristic zero by means of comparison with representations over the algebraic closure—see **50E** below.

The student should now read **Serre**, a beautiful, economical exposition. After this, much of the research literature on the representation theory of finite groups, and its applications to number theory, will be accessible. To learn about infinite groups, read **Adams** and **Fulton-Harris**. Also **Artin** has a chapter on group representations, with some more complicated examples, some material on infinite groups, and a good set of exercises.

Exercise 50D. i) Invent an averaging trick analogous to Maschke's to do the following: Let V be a G-module with an inner product $< \ | \ >$. Construct another inner product $(\ | \ )$ on V which is *G-invariant*, i.e. for all $g \in G$ and all $v,\ w \in V$, we have

$$(g \cdot v \mid g \cdot w) \ = \ (v \mid w) \ .$$

ii) Deduce that any G-module *has* a G-invariant inner product.

iii) Deduce that any G-matrep is equivalent to a G-matrep which takes all of its values in the unitary group,

$$U(n) \ := \ \{ \ A \in GL(n) \ : \ A \text{ is unitary } \} \ .$$

In particular, any finite subgroup of $GL(n)$ is conjugate to a subgroup of $U(n)$.

iv) Use the existence of an invariant inner product and the idea of 'orthogonal complement' to give a new proof of the fact that every G-module is a direct sum of irreducibles.

Exercise 50E. Here is something about G-modules, where we modify the definition in assuming that the ground field has characteristic zero, but might not be algebraically closed.

i) Show that i) of Schur's Lemma, **47.6**, continues to hold.

ii) Find an example where the group is abelian but has an irreducible representation of dimension larger than 1—*think rotation!*

iii) Show that parts ii) and iii) of Schur's Lemma can fail when the field is not algebraically closed.

iv) Prove the following more general version of Schur's Lemma: If F is any

field and V is an irreducible G-module, then $Hom_G(V,\ V)$ is a finite dimensional *division algebra over* F—see the definition after **52D** ahead—using composition as multiplication operation. (We don't need characteristic zero here, nor that G is finite.)

This leads nicely into the last two sections of the book, which study division algebras. A generalization of the ideas in Section **52** can be applied (among other places) to representation theory, because of iv) above.

APPENDIX $\otimes$. Tensor Products.

Let R be a commutative ring (with 1, of course), which will be fixed until the last few paragraphs. Let M and N be R-modules. We'll define another R-module, $M \bigotimes N$, called the *tensor product* of M and N.

If you have not studied the basics on modules yet (given here in Section **42**), you can take R to be a field, and replace the word "module" by "vector space" (and below, also replace "module morphism" by "linear transformation"). The few references in earlier sections to tensor products have involved only that case, indeed with only finite dimensional vector spaces being needed. For vector spaces, the dimension of $M \bigotimes N$ will turn out to be the product of the dimensions of M and N.

In the general case, the module $M \bigotimes N$ will contain elements denoted $m \otimes n$, where $m \in M$ and $n \in N$. (Note that this last symbol $\otimes$ is smaller than the earlier $\bigotimes$.) But *not all elements of* $M \bigotimes N$ *will take the form* $m \otimes n$; a general element will be writeable as a finite sum of such elements; and this is a bit messy, since there is no uniqueness in writing elements that way. It turns out that, for most purposes, it is better to think of $M \bigotimes N$ in terms of a *mapping property* (also called a *universal property*) which implicitly defines it, as we do just below, rather than in terms of notations for its elements. See Remark ii) after **47.11** and paragraphs after **21B** for other information concerning mapping properties.

For fixed M and N, consider all R-modules P , and all maps

$$\beta\ :\ M\ \times\ N\ \longrightarrow\ P$$

which are *bilinear over* R; i.e. for fixed $n \in N$, the map sending m to $\beta(m,\ n)$ is a module morphism from M to P; and, for fixed $m \in M$, the map $N \to P$ defined by $n \mapsto \beta(m,\ n)$ is also a morphism of R-modules.

Suggestion. The set $M \times N$ is the underlying set of the module $M \bigoplus N$. *But don't think of it that way in the present context.* This is as you avoid doing in linear algebra, when studying bilinear maps defined on $V \times V$ (for example, an inner product). In that context, you don't usually picture $V \times V$ as the vector space which is the direct sum of V with itself.

Definition. The tensor product of M and N is defined to be any R-module T, together with any bilinear map

$$\iota : \ M \ \times \ N \ \longrightarrow \ T \ ,$$

which has the following 'universal property' : *for every pair* $(P,\ \beta)$ *as above, there is a unique module morphism* $\hat{\beta} : T \to P$ *such that* $\beta = \hat{\beta} \circ \iota$.

Any such T is denoted $M \bigotimes N$.

This definition probably bugs you for at least three reasons : first of all, we haven't defined the module T in a unique way; secondly, it is not at all obvious that any such module T should exist; and finally, the tensor product has just been defined not only to be a module, but also to include the function ι.

The last objection can only be smoothed over by saying that we preferred to not complicate things at the beginning; certainly the module is the more important thing, and often the term 'tensor product' in a discussion means just the module, not ι. Furthermore, the function ι serves to fix the notation for the elements referred to earlier: *the element* $m \otimes n$ *is defined to be* $\iota(m,\ n)$. Thus the set of elements which can be written in the form $m \otimes n$ is exactly the image of ι.

The first two objections are dealt with by proving a couple of propositions.

Proposition $\otimes$1. *Suppose that* $(T',\ \iota')$ *is another choice for a tensor product of* M *and* N. *Then there is a unique module isomorphism,* $\gamma : T \to T'$, *such that* $\iota' = \gamma \circ \iota$.

Thus (if it exists) the tensor product module is unique up to isomorphism; and that isomorphism is itself unique, subject to the last equality. This is what justifies the use of the notation $M \bigotimes N$ for T.

Proof. Treating $(T,\ \iota)$ as the tensor product, and taking $(P,\ \beta)$ in the definition to be $(T',\ \iota')$, we get a module map γ such that $\iota' = \gamma \circ \iota$. Now just reverse the roles : taking $(T',\ \iota')$ as the tensor product, and $(T,\ \iota)$ as

any old pair '(module, bilinear map)', we get a module map $\alpha : T' \to T$ such that $\iota = \alpha \circ \iota'$.

To show that γ is an isomorphism, we'll show that α is its inverse. Now the module morphism $\alpha \circ \gamma : T \to T$ has the following property : $\iota = (\alpha \circ \gamma) \circ \iota$. This says that $\alpha \circ \gamma$ is the unique module morphism which arises in the definition when we take $(T,\ \iota)$ to play both the roles: as the tensor product and as the general pair. But clearly the identity map from T to itself is the unique module morphism in question. Thus $\alpha \circ \gamma$ is the identity. Symmetrically, $\gamma \circ \alpha$ is the identity map from T' to itself.

It remains only to prove that γ is unique. But, by the definition of the tensor product, the uniqueness of γ is immediate from the equality $\iota' = \gamma \circ \iota$ and the fact that γ is a module map; bijectivity isn't needed to prove its uniqueness.

Proposition ⊗2. *For any commutative ring R, and any pair M and N of R-modules, there* is *a tensor product of M and N.*

Proof. For the moment, think of $M \times N$ as merely a *set*. Let F be the free R-module with this set as basis. Thus F consists of all finite linear combinations of pairs $(m,\ n)$ with coefficients in R. Such a pair itself could be identified with that element of F, which we'll denote as $[m,\ n]$, for which the coefficient of $(m,\ n)$ is 1, and all other coefficients are 0.

Let S be the submodule of F generated by the union of the following four sets :

$$\{\ [m+m',\ n]-[m,\ n]-[m',\ n]\ :\ m,m' \in M,\ n \in N\ \}\ ,$$

$$\{\ [m,\ n+n']-[m,\ n]-[m,\ n']\ :\ m \in M,\ n,n' \in N\ \}\ ,$$

$$\{\ [rm,\ n]-r[m,\ n]\ :\ r \in R,\ m \in M,\ n \in N\ \}\ ,$$

and

$$\{\ [m,\ rn]-r[m,\ n]\ :\ r \in R,\ m \in M,\ n \in N\ \}\ .$$

Define T to be the quotient module F/S. Define

$$\iota\ :\ M\ \times\ N\ \longrightarrow\ T$$

by

$$(m,\ n) \longmapsto [m,\ n]+S\ .$$

To prove that ι is bilinear is a straightforward calculation, which will be left as an exercise. (Note however that bilinearity would fail for the map $M \times N \ \rightarrow \ F$ sending $(m,\ n)$ to $[m,\ n]$; it is essential that we factor out the 'relations' given by the four sets.)

To prove the universal property, suppose given $(P,\ \beta)$ as in the definition. Let $\mu : F \rightarrow P$ be the unique module morphism which sends each $[m,\ n]$ to $\beta(m,\ n)$; that is,

$$\mu(\sum_j r_j[m_j,\ n_j]) \ := \ \sum_j r_j \beta(m_j,\ n_j) \ .$$

It is now a straightforward calculation to see that each of the elements in the four sets defining S is in the kernel of μ. Thus there is a module morphism $\hat{\beta} \ : \ T \ = \ F/S \ \longrightarrow \ P$ for which $\hat{\beta}(x + S) \ = \ \mu(x)$ for all $x \ \in \ F$. [This comes from the basic material on the first isomorphism theorem—Section **42; RVii), iii), iv), v)**.] Now the equality $\beta \ = \ \hat{\beta} \circ \iota$ is immediate by calculating both mappings on $(m,\ n)$.

To prove the uniqueness of $\hat{\beta}$, note that the set of elements $[m,\ n] + S$ generates T, since the elements $[m,\ n]$ generate F. A module morphism with domain T is completely determined once we know its values on some set of generators. But any module morphism $\nu : T \rightarrow P$ for which

$\beta = \nu \circ \iota$ satisfies

$$\nu([m,\ n] + S) \ = \ \nu(\iota(m,\ n)) \ = \ \beta(m,\ n) \ = \ \hat{\beta}([m,\ n] + S) \ .$$

Thus $\nu(y) = \hat{\beta}(y)$ for all y, as required.

Proposition $\bigotimes$3. *Suppose that R is a field, and that the vector spaces M and N have bases $\{m_i\}$ and $\{n_j\}$, respectively. Then $M \bigotimes N$ has basis $\{m_i \otimes n_j\}$. In particular, the dimension of the tensor product of two vector spaces is the product of the dimensions of its two ingredients.*

Sketch Proof. This will illustrate the following point : In dealing with the tensor product, it is almost always a bad idea to use a particular construction, such as the one in the last proof. It's usually better to go to the universal property. Let T be the vector space with basis consisting of 'symbols' $m_i \otimes n_j$. Let $\iota : M \times N \to T$ be the function sending $(\sum_i a_i m_i,\ \sum_j b_j n_j)$ to $\sum_{(i,j)} a_i b_j m_i \otimes n_j$, where the elements a_i and b_j are in the field R (and are almost all 0). We leave it as an exercise for the reader to show that $(T,\ \iota)$ satisfies the definition of the tensor product.

Remark. In the case of general R, the following partial generalization can be proved: *If M and N have generating sets $\{m_i\}$ and $\{n_j\}$ respectively, then $M \bigotimes N$ has generating set $\{m_i \otimes n_j\}$. In particular, a tensor product of a pair of finitely generated modules is also finitely generated.* The first assertion is equivalent to the earlier statement that every element in $M \bigotimes N$ can be written as a sum of elements of the form $m \otimes n$.

The tensor product has properties resembling the associative, commutative and distributive laws, and the existence of the identity ele- ment :

$$\begin{aligned}
(M \bigotimes N) \bigotimes P \ &\cong \ M \bigotimes (N \bigotimes P) \quad ; \\
M \bigotimes N \ &\cong \ N \bigotimes M \quad ; \\
(M \bigoplus N) \bigotimes P \ &\cong \ (M \bigotimes P) \bigoplus (N \bigotimes P) \ ; \\
M \bigotimes R \ &\cong \ M \ .
\end{aligned}$$

It now follows that we could write down the tensor product of any two finitely generated modules over a PID in elementary divisor form (see the

remarks after **44D**), once the following formulae are known, where p and q are non-associated irreducibles in R :

$$(R/ < p^i >) \bigotimes (R/ < p^j >) \quad \cong \quad R/ < p^{\min\{i,j\}} > \ ;$$

$$(R/ < p^i >) \bigotimes (R/ < q^j >) \quad \cong \quad \{0\} \ .$$

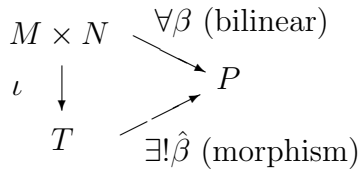
We shall leave all the above as exercises. Remember that the mapping property is the easy way to carry out most of the proofs. For example, if $\iota : M \times N \to M \bigotimes N$ gives the tensor product of M and N (in that order), then it's easy to prove that $\lambda : N \times M \to M \bigotimes N$, given by $\lambda(n,\ m) = \iota(m,\ n)$, gives the tensor product of N and M (in the reverse order to before). This will prove the 'commutativity' above.

The information in our three propositions suffices for completing the proofs in previous sections where the tensor product occurs. These are : the alternative proof of the uniqueness of splitting fields in the remarks after **28.4**, including **28A**; the remark after **48.2**; and Exercises **47Q, 48K, 48L**, as well as Theorem **47.14**.

Aside. In differential geometry and physics, the word "tensor" is used in a way which relates to the material here, with one important extra-algebraic wrinkle. The concept of a *manifold, M, of dimension* n is a generalization of the idea of a surface (the case of dimension 2), such as a sphere, cylinder or torus (surface of a doughnut). At each point $x \in M$, there is the *tangent space*—a vector space, V_x, of dimension n. A *vector field on* M is a choice of one vector, $v_x \in V_x$, for each x in M. This is always assumed to vary continuously with x, and usually to vary smoothly (i.e. as a function, it should have derivatives of all orders). A *tensor field on* M is a choice of one vector in $V_x \bigotimes V_x \bigotimes \cdots \bigotimes V_x$ for each $x \in M$, where the number of factors in the tensor product is fixed. This is what the word "tensor" is often used to mean; the word "field" is omitted, and smoothness tends to go without saying. Often, some or all of the factors in the last tensor product are replaced by $(V_x)^*$, the dual space of V_x. If you see a quite ugly proliferation of sub- and superscripts when reading material on this, what you are seeing are the component scalars of a tensor with respect to a basis for the iterated tensor product. This basis will be the one obtained, starting from some fixed basis for the tangent space, by iterating the basis recipe in the last proposition from

two factors to the number of factors in the tensor product $V_x \bigotimes V_x \bigotimes \cdots \bigotimes V_x$ above. In the case where some factors are the cotangent space $(V_x)^*$, one uses the basis dual to that for V_x.

You will want to get used to 'diagram chasing', also called 'abstract nonsense', before considering yourself to be a fully fledged mathematician. The definition and many of the proofs concerning tensor products can be carried out efficiently using this cultural pursuit. For example, the definition of the tensor product may be expressed as follows :



There are occasions when several 'ground rings' enter a discussion. To be clear about which one is being 'tensored over', the notation $M \bigotimes N$ is expanded to $M \bigotimes_R N$. For example, if R is any ring whose underlying abelian group is isomorphic to $\mathbf{Z}^2$, then $R \bigotimes_{\mathbf{Z}} R \cong \mathbf{Z}^4$, where R is just thought of as an abelian group. However, $R \bigotimes_R R \cong R \cong \mathbf{Z}^2$, where the first isomorphism is of R-modules, but the second only of abelian groups. ($R = \mathbf{Z}[i]$ would do here.) This shows that the result can depend on which ring is 'tensored over'.

An example using vector spaces is as follows : $\mathbf{C} \bigotimes_{\mathbf{C}} \mathbf{C} \cong \mathbf{C} \cong \mathbf{R}^2$, whereas $\mathbf{C} \bigotimes_{\mathbf{R}} \mathbf{C} \cong \mathbf{R}^2 \bigotimes_{\mathbf{R}} \mathbf{R}^2 \cong \mathbf{R}^4$. All of these are (at least) isomorphisms of real vector spaces.

Here is a question which gives some play with individual elements in a tensor product : Is $\mathbf{Q} \bigotimes_{\mathbf{Z}} \mathbf{Q} \cong \mathbf{Q} \bigotimes_{\mathbf{Q}} \mathbf{Q}$ (as abelian groups)? Working on this will probably produce instances where $a \otimes b = c \otimes d$ with $a \neq c$ and $b \neq d$. For example, in $\mathbf{Q} \bigotimes_{\mathbf{Z}} \mathbf{Q}$, is

$$(1/2) \otimes (1/2) \; - \; 1 \otimes (1/4)$$

an element of order 2?—or is it zero? The element denoted in the same way in $\mathbf{Q} \bigotimes_{\mathbf{Q}} \mathbf{Q}$ is certainly 0 , since it is

$$(1/2)[1 \otimes (1/2)] \; - \; (1/2)[1 \otimes (1/2)] \; .$$

Note also that $a \otimes b$ could be zero in $A \bigotimes B$, and yet non-zero in $A' \bigotimes B'$, for submodules A' of A, and B' of B. Give an example ! So the notation $a \otimes b$ on its own is inherently ambiguous.

VI. A Dearth of Division Rings

Sections **51** and **52** each discuss a classic fact about division rings: 1) there are no finite non-commutative ones (Wedderburn) ; and 2) a finite dimensional non-commutative division algebra over $\mathbf{R}$ is necessarily isomorphic to the quaternions (Frobenius). We assume that $1 \neq 0$ in a division ring, although that wasn't done in the original definition.

51. Finite division rings are fields.

To begin, here is a number theoretic lemma whose proof uses the simplest facts about the cyclotomic polynomials c_n from Section **33**.

Lemma 51.1. *Given integers $k > 1$ and $n \geq 1$, suppose that M is a multiset of positive integers $m < n$ with $m \mid n$, such that*

$$k^n \;=\; k \;+\; \sum_M (k^n - 1)/(k^m - 1) \ .$$

Then $n = 1$ (and so M is empty).

Proof. When $m \mid n$ and $m < n$, the cyclotomic polynomial $c_n(x)$ divides $(x^n - 1)/(x^m - 1)$ in $\mathbf{Z}[x]$. Therefore the integer $c_n(k)$ divides

$$k^n - 1 - \sum_M (k^n - 1)/(k^m - 1) \ ,$$

which equals $k - 1$. So write $k - 1 = r c_n(k)$ for some integer r. Thus

$$1 \;=\; (k-1)/(k-1) \;=\; r c_n(k)/(k-1) \;=\; (k-1)^{-1} r \prod_\xi (k - \xi) \ ,$$

with the product being over all the complex primitive n^{th} roots, ξ, of unity. But for $n > 1$, distances in $\mathbf{R}^2$ give $|k - \xi| \;>\; (k-1)$, so

$$|(k-1)^{-1} r \prod_\xi (k - \xi)| \;>\; (k-1)^{-1} |r| \prod_\xi (k-1) \;=\; |r| (k-1)^{\Phi(n)-1} \;\geq\; 1 \ ,$$

contradicting the equality in the previous display. Therefore $n = 1$.

Exercise 51A. Prove that for positive integers $k > 1,\ m$, and n,

$$(k^m - 1) \mid (k^n - 1) \quad \Longleftrightarrow \quad m \mid n \ .$$

Recall that the group of invertibles in a (not necessarily commutative) ring R is denoted $R^\times$. The centre of the ring is

$$Z(R) \ := \ \{ \ r \in R \ : \ rs = sr \text{ for all } s \in R \ \} \ .$$

The centralizer of an element r is

$$C_R(r) \ := \ \{ \ s \in R \ : \ sr = rs \ \} \ .$$

These are subrings. We use the same notation for the centralizer subgroup of an element in a group.

Exercise 51B. Suppose that R is a division ring. Prove that $C_R(r)$ also is. Deduce that, for non-zero r,

$$C_R(r)^\times \ = \ C_{R^\times}(r) \ .$$

Exercise 51C. Recall, from **47S**, that if $h \in C$, a conjugacy class in a group G, then the following map is bijective :

$$G/C_G(h) \longrightarrow C$$

$$gC_G(h) \longmapsto ghg^{-1} \ .$$

Theorem 51.2. (Wedderburn) *Any finite division ring is a field.*

Proof. Let D be a finite division ring. Then its centre, F, is a field, and D is a vector space over F, of dimension n, say. Let $|F| = k$. Thus $|D| = k^n$. Then, summing over *non-singleton* conjugacy classes, C, of the group $D^\times$, with $d_{_C} \in C$ and with $C_D(d_{_C})$ of dimension $m_{_C}$ over F, we get

$$k^n - 1 \ = \ |D^\times| \ = \ |Z(D^\times)| + \sum_C |C| \ = \ |F^\times| + \sum_C |D^\times|/|C_{D^\times}(d_{_C})|$$

$$= \ k - 1 + \sum_C (k^n - 1)/(k^{m_C} - 1) \ .$$

Each term in the summations is an integer greater than 1, so all the $m_{_C}$ are proper divisors of n, by **51A**. But then **51.1** gives $n = 1$. Therefore $D = F$, as required.

Remarks. i) The finite fields were classified in Section **31**, so the finite division rings are completely known.

ii) Desarguean projective geometries are *coordinatized* by division rings (see **Samuel**). It was Hilbert who first observed that Pappus' Theorem holds if and only if the coordinate ring is a field. In particular, there is no 'deduction' of Pappus' from Desargues'. However, by Wedderburn's theorem, every *finite* Desarguean projective geometry is Pappian.

iii) There is a very substantial generalization of Wedderburn's theorem due to Jacobson: *If R is a ring such that, for all its elements r there exists an integer $n > 1$ (possibly depending on r) such that $r^n = r$, then R is commutative.* Wedderburn's theorem follows by taking n to be the order of the division ring.

Jacobson's theorem does not require the assumption that rings have a 1. Could this also be deduced 'after the fact' from the following elementary exercise or an analogue?

Exercise 51D. Show that every (associative) $\mathbf{Z}$-algebra A can be embedded in a ring R, as follows. As an abelian group, take $R = \mathbf{Z} \times A$. Define multiplication by

$$(n,\ x)(m,\ y)\ :=\ (nm,\ ny + mx + xy)\ .$$

Show that R is commutative if and only if A is commutative.

Exercise 51E. Assume that R is a finite ring with no divisors of zero, i.e. the integral domain 'property' holds, except that commutativity is not assumed—($ab = 0 \Rightarrow a = 0$ or $b = 0$). Show that R is a finite field. (See **4A**.)

52. Uniqueness of the quaternions.

We begin with more definitions than are really needed here, but it is helpful to see the larger context. We'll talk about *associative algebras*, and drop the word "associative". But you should be aware that there are non-associative algebras, such as *Lie algebras* (see **Fulton-Harris**), of tremendous importance in mathematics and physics.

Let R be a commutative ring (with 1, as usual). Recall that a $\mathbf{Z}$-algebra satisfies the same axioms as a (not necessarily commutative) ring, except that no identity element is assumed to be in it.

Definition. An *(associative) R-algebra* is an object A which is simultaneously a $\mathbf{Z}$-algebra and an R-module, such that the two multiplications are related by the 'law', (!!), which occurred in the discussion of the quaternions $\mathbf{H}$ in Section **14** :
for all $r \in R$, and $a,\ b \in A$,

$$r \cdot (ab) \ = \ (r \cdot a)(b) \ = \ (a)(r \cdot b) \ .$$

A morphism between R-algebras is a map which preserves all three operations (so it is a morphism of modules, and would be a morphism of rings, except that no reference to an element 1_A occurs).

A canonical example of a non-commutative R-algebra is the algebra, $R^{n \times n}$, of all $n \times n$ matrices over R, for $n > 1$. Another example is the group algebra from Section **50**.

To justify the a priori notation when $R = \mathbf{Z}$, recall from Section **45** that an abelian group is automatically a $\mathbf{Z}$-module.

Exercise 52A. Show that the 'law' in the last definition holds, so that there is a $\mathbf{Z}$-algebra structure in the new sense on any $\mathbf{Z}$-algebra in the old sense.

Exercise 52B. Assume that F is a field, and A is an F-algebra which is finite dimensional as an F-vector space. Assume that A has no zero divisors. Show that A, under its two binary operations, is a division ring. Relate this to Section **19**, especially **19.3**, and to **51E**.

Exercise 52C. Define the *centre*, $Z(A)$, of an R-algebra A, as done in the last section. Show that $Z(A)$ is a commutative R-algebra. Given an R-algebra A with 1_A, define a map $\phi : R \to A$ by sending r to $r \cdot 1_A$. Prove that ϕ is a morphism of rings whose image is contained in the centre of A. Conversely, given a ring A, a commutative ring R, and a ring morphism from R into the centre of A, show how A may be made into an R-algebra with 1. (In particular, commutative ring extensions $S \supset R$ give R-algebras S.)

Exercise 52D. Show how any ring may be 'canonically' converted into an R-algebra, where R is its centre.

Definitions. Let F be a field. A *division algebra over F* is an F-algebra, D, whose underlying $\mathbf{Z}$-algebra is actually a division ring. In particular D

has a 1, so the scalar multiplication by F may alternatively be regarded as an embedding of F into D, by **52C**. Then we may regard F as a subfield of D. Thus the idea of an element $a \in D$ being *algebraic over* F may be defined just as in field theory : there exist $c_i \in F$, not all zero, such that $c_0 + c_1 a + c_2 a^2 + \cdots = 0$. Because $F \subset Z(D)$, it is immaterial whether the c_i are positioned to the left of the a^i or not. As one would expect, D is said to be *algebraic over* F if all of its elements are. Note also that a morphism of division algebras over a field F can now be defined to be a ring morphism which fixes all elements of F. Such a map is necessarily injective, since a division ring has no interesting ideals, two-sided or otherwise.

Exercise 52E. Since an F-module is just a vector space, the meaning of D being a *finite dimensional algebra* is clear. Show that the same argument as in field theory demonstrates that such an algebra is algebraic over F.

Note that any division ring is a division algebra over its centre, by **52D**. This is really the way we were viewing finite division rings in the last section, but we didn't need all these definitions to do so.

We shall deal with Frobenius' theorem (stated below) by chopping it into pieces labeled (I) to (IX). In Exercise **52H** ahead, a different, better (and more 'exercisable'!) proof is suggested. You may want to go directly there and do it.

(I). First note that there are no proper (not necessarily commutative) ring extensions $R \supset \mathbf{C}$ such that

i) $\mathbf{C}$ is inside the centre of R;

ii) R has no zero divisors; and

iii) R is algebraic over $\mathbf{C}$.

For if $r \in R$ is a root of $\prod_j (x - z_j) \in \mathbf{C}[x]$, then $\prod_j (r - z_j) = 0$, so, for some j, we have $r = z_j \in \mathbf{C}$.

In particular, there are no finite dimensional division algebras over $\mathbf{C}$, other than $\mathbf{C}$ itself.

Exercise 52F. Where is i) used above?

(II). Another way of saying that $\mathbf{R}$ has only two algebraic field extensions is that there are no algebraic *commutative* division algebras over $\mathbf{R}$, up to isomorphism, other than $\mathbf{C}$ and $\mathbf{R}$ itself. In particular, this is true as well

with "algebraic" replaced by "finite dimensional".

Theorem 52.1. (Frobenius) *If D is an algebraic division algebra over $\mathbf{R}$ which is not commutative, then D is isomorphic to the quaternions, $\mathbf{H}$, as a real division algebra.*

This applies in particular to finite dimensional non-commutative division algebras over $\mathbf{R}$. Combined with (II), we see that there are only three algebraic division algebras over $\mathbf{R}$, namely $\mathbf{H}$, $\mathbf{C}$ and $\mathbf{R}$. In particular, an algebraic division algebra over $\mathbf{R}$ is necessarily finite dimensional.

Exercise 52G. Produce from your memory of previous sections several examples of fields F and algebraic F-division algebras which are not finite dimensional over F. Now produce one which isn't commutative.

Proof of 52.1. Assume below in (III) to (VIII) that D denotes a *central* algebraic $\mathbf{R}$-division algebra; that is $Z(D) = \mathbf{R}$, where we regard $\mathbf{R}$ as a subfield of D via the embedding of $\mathbf{R}$ which gives D its module structure.

(III). If $\ell \in D \setminus \mathbf{R}$, then

$$\mathbf{R}(\ell) \ := \ \{ \ r_0 + r_1\ell \ : \ r_i \in \mathbf{R} \ \}$$

is a subfield of D isomorphic to $\mathbf{C}$ as an $\mathbf{R}$-algebra.

For ℓ is the root of a real irreducible of degree 2, by algebraicity, so the set is easily seen to be an algebraic field extension of $\mathbf{R}$. Now apply paragraph (II) above.

(IV). If $m \in D \setminus \mathbf{R}$ and $m^2 \in \mathbf{R}$, then $m^2 < 0$.

For $m^2 \neq 0$ since a division ring has no zero divisors; and $x^2 - m^2$ cannot have three roots, m and $\pm\sqrt{m^2}$, in the field $\mathbf{R}(m) \cong \mathbf{C}$. Better: D can be replaced by $\mathbf{R}(m)$, which, by (III), is replaceable by $\mathbf{C}$, for which the result is obvious.

(V). If $\mathbf{R}(\ell) \subset D$ with $\ell^2 = -1$, then there exists $m \in D \setminus \mathbf{R}(\ell)$ with $m\ell = -\ell m$.

Because D is central over $\mathbf{R}$, we can choose an element $s \in D$ with $m := s\ell - \ell s \neq 0$. A trivial calculation yields the identity (i.e. that ℓ and m *anti-commute*). Since $\mathbf{R}(\ell) \cong \mathbf{C}$ by (III), and m doesn't commute with ℓ (WHY?), we see that $m \notin \mathbf{R}(\ell)$.

(VI). If $\mathbf{R}(\ell) \subset D$ with $\ell^2 = -1$, and $m \in D \setminus \mathbf{R}(\ell)$ with $m\ell = -\ell m$, then $m^2 \in \mathbf{R}$.

By (III), we have $m^2 = c_0 + c_1 m$ for $c_i \in \mathbf{R}$. The fact that ℓ and m anti-commute gives that ℓ and m^2 commute. Thus

$$c_0\ell + c_1\ell m \;=\; \ell m^2 \;=\; m^2\ell \;=\; c_0\ell - c_1\ell m \;.$$

Thus $2c_1\ell m = 0$ and so $c_1 = 0$ as required, since we can cancel in a division ring.

(VII). Given ℓ and v in D with $v \notin \mathbf{R}(\ell)$ [and hence $\ell \notin \mathbf{R}(v)$], no element of D outside of $\mathbf{R}$ commutes with both v and ℓ.

For, by (I) and (III), such an element would need to lie in both $\mathbf{R}(\ell)$ and $\mathbf{R}(v)$. But we cannot have $c_0 + c_1\ell = c_0' + c_1' v$ with the c's in $\mathbf{R}$ unless $c_1 = 0 = c_1'$.

(VIII). Given ℓ and v in D with $v^2 = \ell^2 = -1$, the element $\ell v + v\ell$ commutes with both v and ℓ.

This is a trivial calculation.

(IX). Now assume that D is any $\mathbf{R}$-division algebra which contains $\mathbf{H}$ as a subring. Then no non-zero element of D anti-commutes with all three of $i,\ j$ and k.

For such an element y, we have

$$0 \;=\; yk + ky \;=\; yij + ijy \;=\; (yi + iy)j - i(yj - jy) \;=\; -\,i(yj - jy) \;.$$

By cancellation, $yj = jy$. But y and j also anti-commute, yielding $2yj = 0$. Thus $y = 0$, as required.

Now let's polish off the proof. Let D be a non-commutative algebraic division algebra over $\mathbf{R}$. By (II), its centre is isomorphic to $\mathbf{R}$ or $\mathbf{C}$. But the latter is impossible by (I). Thus D is a central $\mathbf{R}$-algebra. By (III), we may choose $\ell \in D$ with $\ell^2 = -1$. By (V), we may choose $m \in D$ which anti-commutes with ℓ. By (IV) and (VI), the element m^2 is a negative real. Multiplying by a suitable real, we may rechoose m so that $m^2 = -1$, and m still anti-commutes with ℓ. Let $n = \ell m$. It is now routine to calculate that $\ell,\ m$ and n satisfy all the relations given for $i,\ j$, and k in the definition of $\mathbf{H}$, the quaternions. From this, it is routine to demonstrate that the map

$$\phi : \mathbf{H} \longrightarrow D \qquad ,$$

$$c_0 + c_1 i + c_2 j + c_3 k \longmapsto c_0 + c_1 \ell + c_2 m + c_3 n \ ,$$

is a morphism of rings fixing each real. Its kernel is zero since $\mathbf{H}$ is a division ring. It remains only to show that ϕ is surjective. Let $u \in D \setminus \mathbf{R}$. By (III), there is an element v of the form $c_0 + c_1 u$ with $v^2 = -1$, where the c's are reals (and so $c_1 \neq 0$). It suffices to exhibit an element $w \in \phi(\mathbf{H})$ with $v = -w$, since then $v \in \phi(\mathbf{H})$, and so $u \in \phi(\mathbf{H})$. We may assume that $v \notin \mathbf{R}(\ell) \cup \mathbf{R}(m) \cup \mathbf{R}(n)$, since that union is contained in $\phi(\mathbf{H})$. Let

$$w \ := \ \frac{1}{2}(v\ell + \ell v)\ell \ + \ \frac{1}{2}(vm + mv)m \ + \ \frac{1}{2}(vn + nv)n \ .$$

Now $v\ell + \ell v \in \mathbf{R}$ by (VIII) and (VII). The same applies to the other two 'coefficients' in the definition of w. Thus w is certainly in the image of ϕ. It is routine to check that $v + w$ anti-commutes with each of ℓ, m and n. Now (IX) evidently holds with $\mathbf{H}$ replaced by any isomorphic $\mathbf{R}$-algebra, and i, j, and k replaced by their images under the isomorphism. Therefore, from (IX), we get $v + w = 0$, completing the proof.

This proof is something of a bag of tricks. There is a more systematic theory of central division algebras, with applications to number theory and quadratic forms. Frobenius' theorem above drops out as a byproduct of this theory. See **Hungerford** and **Lam**(1973).

Exercise 52H. This gives another, and a very slick, proof of Frobenius' theorem. See **Lam** (1990), pp. 219-220, for the details, if necessary. Let D be a non-commutative algebraic division algebra over $\mathbf{R}$.

(i) Argue as in (III) above that D contains a subalgebra isomorphic to $\mathbf{C}$. So, without loss of generality, we have $\mathbf{C} \subset D$.

(ii) Prove that, as a $\mathbf{C}$-vector space, D decomposes as the direct sum of D^+ and D^-, where

$$D^{\pm} := \{\ d \in D\ :\ di = \pm id\ \}\ .$$

(iii) Show that $D^+ = \mathbf{C}$.

(iv) Show that $\mu : D^- \to D^+$, sending x to xy, for any fixed non-zero $y \in D^-$, is well defined and an isomorphism of complex vector spaces.

(v) Prove that any such y satisfies $y^2 \in \mathbf{R}$ and $y^2 < 0$.

(vi) Deduce that $D \cong \mathbf{H}$ as real algebras.

You may be wondering about the following. We gave a definition of $\mathbf{H}$ as $\mathbf{C} \times \mathbf{C}$, similar to the construction of $\mathbf{C}$ from $\mathbf{R}$. This used complex conjugation; but there is also a quaternionic conjugation. Why can't we produce an algebra structure on $\mathbf{H}^2$ using analogous formulae?—and wouldn't this contradict Theorem **52.1**? The answer is that the proof of associativity for $\mathbf{H}$ uses commutativity of $\mathbf{C}$ in an essential way. There *is* a construction of a *non-associative* $\mathbf{R}$-algebra along the lines above. It is called the *Cayley numbers* and, of course, has dimension over $\mathbf{R}$ equal to $8 = 2\dim_{\mathbf{R}}\mathbf{H}$. With a suitable definition of *not-necessarily-associative division algebra*, it is a very difficult theorem, proved first in 1957 by Bott, Kervaire, and Milnor using algebraic topology, that only in dimensions 1, 2, 4 and 8 can there exist such algebras (when the field is $\mathbf{R}$, and we restrict to finite dimensions). Perhaps the neatest proof so far would use the 'postcard-size' argument in the first section of **Adams-Atiyah**. This argument appears to be just an elementary algebraic manipulation; but it is based on Bott periodicity, which is a topological result of great depth. The argument shows that certain topological

spaces (which not-necessarily-associative division algebra structures on $\mathbf{R}^{2n}$ would allow you to construct) can only exist when n is 1, 2 or 4. The spaces which do exist are projective planes based on the algebras. The 'postcard' argument ends when the condition $2^n \mid 3^n - 1$ has been deduced.

Exercise 52I. Show that, indeed,

$$2^n \mid 3^n - 1 \iff n = 1,\ 2 \text{ or } 4\ .$$

References

Adams, J. F. *Lectures on Lie Groups.* W. A. Benjamin, New York, 1969.

Adams, J. F. and Atiyah, M. F. *K-theory and the Hopf invariant.* Quart. J. Math. (Oxford)(2) **17** (1966) 31–38.

Artin, M. *Algebra.* Prentice Hall, Englewood Cliffs, N.J., 1991.

Atiyah, M. F. and Macdonald, I. G. *Introduction to Commutative Algebra.* Addison-Wesley, Reading, Mass., 1969.

Fulton, W. and Harris, J. *Representation Theory—A first course.* Springer-Verlag, New York, 1991.

Greub, W. *Multilinear Algebra.* Springer-Verlag, New York, 1978.

Goldstein, L. J. *Abstract Algebra.* Addison-Wesley, Reading, Mass., 1973.

Hartley, B. and Hawkes, T. O. *Rings, Modules and Linear Algebra.* Chapman & Hall, London, 1970.

Herstein, I. N. *Topics in Algebra.* 2nd ed., Wiley, New York, 1975.

Hungerford, T. W. *Algebra.* Springer-Verlag, New York, 1974.

Jacobson, N. *Basic Algebra I, II.* Freeman, San Francisco, 1974, 1980.

Lam, T. Y. *The Algebraic Theory of Quadratic Forms.* W. A. Benjamin, Reading, Mass., 1973.

Lam, T. Y. *A first course in noncommutative ring theory.* Springer-Verlag, New York, 1991.

Lang, S. *Undergraduate Algebra.* 2nd ed., Springer-Verlag, New York, 1990.

Macdonald, I. G. *Symmetric Functions and Hall Polynomials.* Oxford Univ. Press, Oxford, 1979 (enlarged 2nd ed. to appear).

Mackey, G. W. *Harmonic analysis as the exploitation of symmetry—a historical survey.* Bull. A.M.S. **3** (1980) 543–699.

Rotman, J. *Galois Theory.* Springer-Verlag, New York, 1990.

Samuel, P. *Projective Geometry.* Springer-Verlag, New York, 1988.

Serre, J. P. *Linear Representations of Finite Groups.* Springer-Verlag, New York, 1977.

Stewart, I. *Galois Theory.* Chapman & Hall, London, 1989.

Index
of Notation

REFERENCE MATERIAL

Adams, 249
Adams-Atiyah, 266
Artin, 38, 80, 86, 116, 142, 159, 195, 198, 207, 249
Atiyah-Macdonald, 106
Fulton-Harris, 249, 261
Goldstein, 142, 143, 156
Greub, 106
Hartley-Hawkes, 177
Herstein, 76, 143
Hungerford, 166, 177, 242, 266
Jacobson, 177
Lam(1973), 265
Lam(1991), 61
Lang, 143
Macdonald, 90
Mackey, 225
Rotman, 174
Samuel, 260
Serre, 249
Stewart, 142

Index