



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – I – ALGEBRA-I – SMT1501

Unit-I Introduction

One of the simplest and most basic of all algebraic structures is the *group*. A group is defined to be a set with an operation (let us call it $*$) which is associative, has a neutral element, and for which each element has an inverse. More formally,

By a group we mean a set G with an operation $$ which satisfies the axioms:*

(G1) *$*$ is associative.*

(G2) *There is an element e in G such that $a * e = a$ and $e * a = a$ for every element a in G .*

(G3) *For every element a in G , there is an element a^{-1} in G such that $a * a^{-1} = e$ and $a^{-1} * a = e$.*

The group we have just defined may be represented by the symbol $\langle G, * \rangle$. This notation makes it explicit that the group consists of the *set* G and the *operation* $*$. (Remember that, in general, there are other possible operations on G , so it may not always be clear which is the group's operation unless we indicate it.) If there is no danger of confusion, we shall denote the group simply with the letter G .

The groups which come to mind most readily are found in our familiar number systems. Here are a few examples.

\mathbf{Z} is the symbol customarily used to denote the set

$$\{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$$

of the integers. The set \mathbf{Z} , with the operation of *addition*, is obviously a group. It is called the *additive group of the integers* and is represented by the symbol $\langle \mathbf{Z}, + \rangle$. Mostly, we denote it simply by the symbol \mathbf{Z} .

\mathbf{Q} designates the set of the rational numbers (that is, quotients m/n of integers, where $n \neq 0$). This set, with the operation of addition, is called the *additive group of the rational numbers*, $\langle \mathbf{Q}, + \rangle$. Most often we denote it simply by \mathbf{Q} .

The symbol \mathbf{R} represents the set of the real numbers. \mathbf{R} , with the operation of addition, is called the *additive group of the real numbers*, and is represented by $\langle \mathbf{R}, + \rangle$, or simply \mathbf{R} .

The set of all the *nonzero rational numbers* is represented by \mathbf{Q}^* . This set, with the operation of *multiplication*, is the group $\langle \mathbf{Q}^*, \cdot \rangle$, or simply \mathbf{Q}^* . Similarly, the set of all the *nonzero real numbers* is represented by \mathbf{R}^* . The set \mathbf{R}^* with the operation of multiplication, is the group $\langle \mathbf{R}^*, \cdot \rangle$, or simply \mathbf{R}^* .

Finally, \mathbf{Q}^{pos} denotes the group of all the positive rational numbers, with multiplication. \mathbf{R}^{pos} denotes the group of all the positive real numbers, with multiplication.

Groups occur abundantly in nature. This statement means that a great many of the algebraic structures which can be discerned in natural phenomena turn out to be groups. Typical examples, which we shall examine later, come up in connection with the structure of crystals, patterns of symmetry, and various kinds of geometric transformations. Groups are also important because they happen to be one of the fundamental building blocks out of which more complex algebraic structures are made.

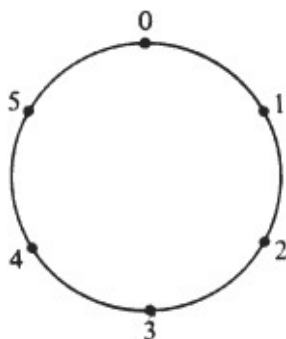
Especially important in scientific applications are the *finite* groups, that is, groups with a finite number of elements. It is not surprising that such groups occur often in applications, for in most situations of the real world we deal with only a finite number of objects.

The easiest finite groups to study are those called the *groups of integers modulo n* (where n is any positive integer greater than 1). These groups will be described in a casual way here, and a rigorous treatment deferred until later.

Let us begin with a specific example, say, the group of integers modulo 6. This group consists of a set of six elements,

$$\{0, 1, 2, 3, 4, 5\}$$

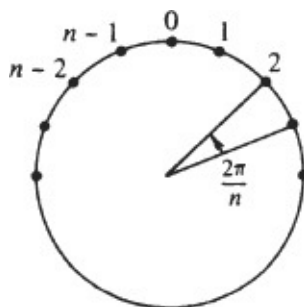
and an operation called *addition modulo 6*, which may be described as follows: Imagine the numbers 0 through 5 as being evenly distributed on the circumference of a circle. To add two numbers h and k , start with h and move clockwise k additional units around the circle: $h + k$ is where you end up. For example, $3 + 3 = 0$, $3 + 5 = 2$, and so on. The set $\{0, 1, 2, 3, 4, 5\}$ with this operation is called the *group of integers modulo 6*, and is represented by the symbol \mathbf{z}_6 .



In general, the group of integers modulo n consists of the set

$$\{0, 1, 2, \dots, n - 1\}$$

with the operation of *addition modulo n* , which can be described exactly as previously. Imagine the numbers 0 through $n - 1$ to be points on the unit circle, each one separated from the next by an arc of length $2\pi/n$.



To add h and k , start with h and go clockwise through an arc of k times $2\pi/n$. The sum $h + k$ will, of

course, be one of the numbers 0 through $n - 1$. From geometrical considerations it is clear that this kind of addition (by successive rotations on the unit circle) is *associative*. Zero is the neutral element of this group, and $n - h$ is obviously the inverse of h [for $h + (n - h) = n$, which coincides with 0]. This group, the *group of integers modulo n* , is represented by the symbol \mathbf{z}_n .

Often when working with finite groups, it is useful to draw up an “operation table.” For example, the operation table of \mathbf{z}_6 is

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

The basic format of this table is as follows:

+	0	1	2	3	4	5
0						
1						
2						
3						
4						
5						

with one *row* for each element of the group and one *column* for each element of the group. Then $3 + 4$, for example, is located in the row of 3 and the column of 4. In general, any finite group $\langle G, * \rangle$ has a table

*	y
⋮			
x		$x * y$	
⋮			

The entry in the row of x and the column of y is $x * y$.

Let us remember that the commutative law is *not* one of the axioms of group theory; hence the identity $a * b = b * a$ is not true in every group. If the commutative law holds in a group G , such a group is called a *commutative group* or, more commonly, an *abelian group*. Abelian groups are named after the mathematician Niels Abel, who was mentioned in Chapter 1 and who was a pioneer in the study of groups. All the examples of groups mentioned up to now are abelian groups, but here is an example which is not.

Let G be the group which consists of the six matrices

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{B} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} -1 & -1 \\ 0 & 1 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} -1 & -1 \\ 1 & 0 \end{pmatrix} \quad \mathbf{K} = \begin{pmatrix} 1 & 0 \\ -1 & -1 \end{pmatrix}$$

with the operation of *matrix multiplication* which was explained on page 8. This group has the following operation table, which should be checked:

	I	A	B	C	D	K
I	I	A	B	C	D	K
A	A	I	C	B	K	D
B	B	K	D	A	I	C
C	C	D	K	I	A	B
D	D	C	I	K	B	A
K	K	B	A	D	C	I

In linear algebra it is shown that the multiplication of matrices is associative. (The details are simple.) It is clear that \mathbf{I} is the identity element of this group, and by looking at the table one can see that each of the six matrices in $\{\mathbf{I}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{K}\}$ has an inverse in $\{\mathbf{I}, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{K}\}$. (For example, \mathbf{B} is the inverse of \mathbf{D} , \mathbf{A} is the inverse of \mathbf{A} , and so on.) Thus, G is a group! Now we observe that $\mathbf{AB} = \mathbf{C}$ and $\mathbf{BA} = \mathbf{K}$, so G is not commutative.

EXERCISES

A. Examples of Abelian Groups

Prove that each of the following sets, with the indicated operation, is an abelian group.

Instructions Proceed as in [Chapter 2, Exercise B](#).

1 $x * y = x + y + k$ (k a fixed constant), on the set \mathbb{R} of the real numbers.

2 $x * y = \frac{xy}{2}$, on the set $\{x \in \mathbb{R} : x \neq 0\}$.

3 $x * y = x + y + xy$, on the set $\{x \in \mathbb{R} : x \neq -1\}$.

4 $x * y = \frac{x + y}{xy + 1}$, the set $\{x \in \mathbb{R} : -1 < x < 1\}$.

B. Groups on the Set $\mathbb{R} \times \mathbb{R}$

The symbol $\mathbb{R} \times \mathbb{R}$ represents the set of all ordered pairs (x, y) of real numbers. $\mathbb{R} \times \mathbb{R}$ may therefore be identified with the set of all the points in the plane. Which of the following subsets of $\mathbb{R} \times \mathbb{R}$, with the indicated operation, is a group? Which is an abelian group?

Instructions Proceed as in the preceding exercise. To find the identity element, which in these problems is an ordered pair (e_1, e_2) of real numbers, solve the equation $(a, b) * (e_1, e_2) = (a, b)$ for e_1 and e_2 . To

find the inverse (a', b') of (a, b) , solve the equation $(a, b) * (a', b') = (e_1, e_2)$ for a' and b' . [Remember that $(x, y) = (x', y')$ if and only if $x = x'$ and $y = y'$.]

1 $(a, b) * (c, d) = (ad + bc, bd)$, on the set $\{(x, y) \in \mathbb{R} \times \mathbb{R} : y \neq 0\}$.

2 $(a, b) * (c, d) = (ac, bc + d)$, on the set $\{(x, y) \in \mathbb{R} \times \mathbb{R} : x \neq 0\}$.

3 Same operation as in part 2, but on the set $\mathbb{R} \times \mathbb{R}$.

4 $(a, b) * (c, d) = (ac - bd, ad + bc)$, on the set $\mathbb{R} \times \mathbb{R}$ with the origin deleted.

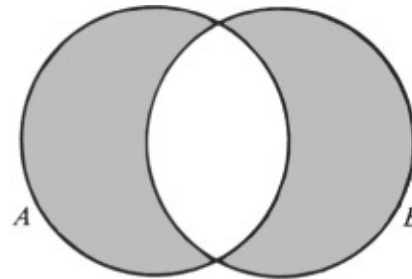
5 Consider the operation of the preceding problem on the set $\mathbb{R} \times \mathbb{R}$. Is this a group? Explain.

C. Groups of Subsets of a Set

If A and B are any two sets, their *symmetric difference* is the set $A + B$ defined as follows:

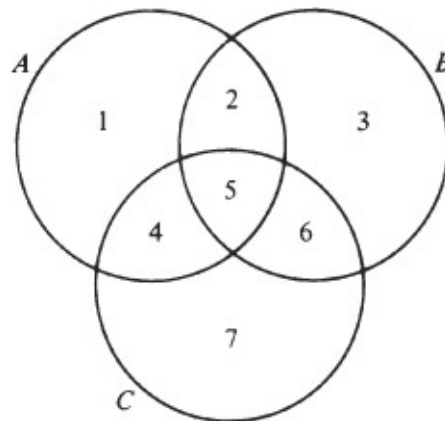
$$A + B = (A - B) \cup (B - A)$$

NOTE: $A - B$ represents the set obtained by removing from A all the elements which are in B .



The shaded area is $A + B$

It is perfectly clear that $A + B = B + A$; hence this operation is commutative. It is also associative, as the accompanying pictorial representation suggests: Let the union of A , B , and C be divided into seven regions as illustrated.



$A + B$ consists of the regions 1, 4, 3, and 6.

$B + C$ consists of the regions 2, 3, 4, and 7.

$A + (B + C)$ consists of the regions 1, 3, 5, and 7.

$(A + B) + C$ consists of the regions 1, 3, 5, and 7.

Thus, $A + (B + C) = (A + B) + C$.

If D is a set, then the *power set* of D is the set P_D of all the subsets of D . That is,

$$P_D = \{A: A \subseteq D\}$$

The operation $+$ is to be regarded as an operation on P_D .

- 1 Prove that there is an identity element with respect to the operation $+$, which is _____.
- 2 Prove every subset A of D has an inverse with respect to $+$, which is _____. Thus, $\langle P_D, + \rangle$ is a group!
- 3 Let D be the three-element set $D = \{a, b, c\}$. List the elements of P_D . (For example, one element is $\{a\}$, another is $\{a, b\}$, and so on. Do not forget the empty set and the whole set D .) Then write the operation table for $\langle P_D, + \rangle$.

D. A Checkerboard Game

1	2
3	4

Our checkerboard has only four squares, numbered 1, 2, 3, and 4. There is a single checker on the board, and it has four possible moves:

- V : Move vertically; that is, move from 1 to 3, or from 3 to 1, or from 2 to 4, or from 4 to 2.
- H : Move horizontally; that is, move from 1 to 2 or vice versa, or from 3 to 4 or vice versa.
- D : Move diagonally; that is, move from 2 to 3 or vice versa, or move from 1 to 4 or vice versa.
- I : Stay put.

We may consider an operation on the set of these four moves, which consists of performing moves successively. For example, if we move horizontally and then vertically, we end up with the same result as if we had moved diagonally:

$$H * V = D$$

If we perform two horizontal moves in succession, we end up where we started: $H * H = I$. And so on. If $G = \{V, H, D, I\}$, and $*$ is the operation we have just described, write the table of G .

*	I	V	H	D
I				
V				
H				
D				

Granting associativity, explain why $\langle G, * \rangle$ is a group.

E. A Coin Game



Imagine two coins on a table, at positions A and B . In this game there are eight possible moves:

- M_1 : Flip over the coin at A .
- M_2 : Flip over the coin at B .
- M_3 : Flip over both coins.
- M_4 : Switch the coins.
- M_5 : Flip coin at A ; then switch.
- M_6 : Flip coin at B ; then switch.
- M_7 : Flip both coins; then switch.
- I : Do not change anything.

We may consider an operation on the set $\{I, M_1, \dots, M_7\}$, which consists of performing any two moves in succession. For example, if we switch coins, then flip over the coin at A , this is the same as first flipping over the coin at B then switching:

$$M_4 * M_1 = M_2 * M_4 = M_6$$

If $G = \{I, M_1, \dots, M_7\}$ and $*$ is the operation we have just described, write the table of $\langle G, * \rangle$.

$*$	I	M_1	M_2	M_3	M_4	M_5	M_6	M_7
I								
M_1								
M_2								
M_3								
M_4								
M_5								
M_6								
M_7								

Granting associativity, explain why $\langle G, * \rangle$ is a group. Is it commutative? If not, show why not.

F. Groups in Binary Codes

The most basic way of transmitting information is to code it into strings of 0s and 1s, such as 0010111, 1010011, etc. Such strings are called *binary words*, and the number of 0s and 1s in any binary word is called its *length*. All information may be coded in this fashion.

When information is transmitted, it is sometimes received incorrectly. One of the most important purposes of coding theory is to find ways of *detecting errors*, and *correcting* errors of transmission.

If a word $\mathbf{a} = a_1a_2 \dots a_n$ is sent, but a word $\mathbf{b} = b_1b_2 \dots b_n$ is received (where the a_i and the b_j are 0s

or 1s), then the *error pattern* is the word $\mathbf{e} = e_1e_2 \dots e_n$ where

$$e_i = \begin{cases} 0 & \text{if } a_i = b_i \\ 1 & \text{if } a_i \neq b_i \end{cases}$$

With this motivation, we define an operation of *adding* words, as follows: If \mathbf{a} and \mathbf{b} are both of length 1, we add them according to the rules

$$0 + 0 = 0 \quad 1 + 1 = 0 \quad 0 + 1 = 1 \quad 1 + 0 = 1$$

If \mathbf{a} and \mathbf{b} are both of length n , we add them by *adding corresponding digits*. That is (let us introduce commas for convenience),

$$(a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (a_1 + b_1, a_2 + b_2, \dots, a_n + b_n)$$

Thus, the sum of \mathbf{a} and \mathbf{b} is the error pattern \mathbf{e} .

For example,

$$\begin{array}{r} 0010110 \\ +0011010 \\ \hline =0001100 \end{array} \qquad \begin{array}{r} 10100111 \\ +11110111 \\ \hline =01010000 \end{array}$$

The symbol \mathbb{B}^n will designate the set of all the binary words of length n . We will prove that the operation of word addition has the following properties on \mathbb{B}^n :

1. It is commutative.
2. It is associative.
3. There is an identity element for word addition.
4. Every word has an inverse under word addition.

First, we verify the commutative law for words of length 1:

$$0 + 1 = 1 = 1 + 0$$

1 Show that $(a_1, a_2, \dots, a_n) + (b_1, b_2, \dots, b_n) = (b_1, b_2, \dots, b_n) + (a_1, a_2, \dots, a_n)$.

2 To verify the associative law, we first verify it for words of length 1:

$$1 + (1 + 1) = 1 + 0 = 1 = 0 + 1 = (1 + 1) + 1$$

$$1 + (1 + 0) = 1 + 1 = 0 = 0 + 0 = (1 + 1) + 0$$

Check the remaining six cases.

3 Show that $(a_1, \dots, a_n) + [(b_1, \dots, b_n) + (c_1, \dots, c_n)] = [(a_1, \dots, a_n) + (b_1, \dots, b_n)] + (c_1, \dots, c_n)$.

4 The identity element of \mathbb{B}^n , that is, the identity element for adding words of length n , is _____.

5 The inverse, with respect to word addition, of any word (a_1, \dots, a_n) is _____.

6 Show that $\mathbf{a} + \mathbf{b} = \mathbf{a} - \mathbf{b}$ [where $\mathbf{a} - \mathbf{b} = \mathbf{a} + (-\mathbf{b}^*)$].

7 If $\mathbf{a} + \mathbf{b} = \mathbf{c}$, show that $\mathbf{a} = \mathbf{b} + \mathbf{c}$.

G. Theory of Coding: Maximum-Likelihood Decoding

We continue the discussion started in [Exercise F](#): Recall that \mathbb{B}^n designates the set of all binary words of length n . By a *code* we mean a subset of \mathbb{B}^n . For example, below is a code in \mathbb{B}^5 . The code, which we shall call C_1 , consists of the following binary words of length 5:

00000
00111
01001
01110
10011
10100
11010
11101

Note that there are 32 possible words of length 5, but only eight of them are in the code C_1 . These eight words are called *codewords*; the remaining words of \mathbb{B}^5 are not codewords. Only codewords are transmitted. If a word is received which is not a codeword, it is clear that there has been an *error of transmission*. In a well-designed code, it is unlikely that an error in transmitting a codeword will produce another codeword (if that were to happen, the error would not be detected). Moreover, in a good code it should be fairly easy to locate errors and correct them. These ideas are made precise in the discussion which follows.

The *weight* of a binary word is the number of 1s in the word: for example, 11011 has weight 4. The *distance* between two binary words is the number of positions in which the words differ. Thus, the distance between 11011 and 01001 is 2 (since these words differ only in their first and fourth positions). The *minimum distance* in a code is the smallest distance among all the distances between pairs of codewords. For the code C_1 , above, pairwise comparison of the words shows that the minimum distance is 2. What this means is that *at least two* errors of transmission are needed in order to transform a codeword into another codeword; single errors will change a codeword into a *noncodeword*, and the error will therefore be detected. In more desirable codes (for example, the so-called Hamming code), the minimum distance is 3, so any one or two errors are *always* detected, and only three errors in a single word (a very unlikely occurrence) might go undetected.

In practice, a code is constructed as follows: in every codeword, certain positions are *information positions*, and the remaining positions are *redundancy positions*. For instance, in our code C_1 , the first three positions of every codeword are the information positions: if you look at the eight codewords (and confine your attention only to the first three digits in each word), you will see that every three-digit sequence of 0s and 1s is there namely,

000, 001, 010, 011, 100, 101, 110, 111

The numbers in the fourth and fifth positions of every codeword satisfy *parity-check equations*.

1 Verify that every codeword $a_1a_2a_3a_4a_5$ in C_1 satisfies the following two parity-check equations: $a_4 = a_1 + a_3$; $a_5 = a_1 + a_2 + a_3$.

2 Let C_2 be the following code in \mathbb{B}^6 . The first three positions are the information positions, and every codeword $a_1a_2a_3a_4a_5a_6$ satisfies the parity-check equations $a_4 = a_2$, $a_5 = a_1 + a_2$, and $a_6 = a_1 + a_2 + a_3$.

(a) List the codewords of C_2 .

(b) Find the minimum distance of the code C_2 .

(c) How many errors in any codeword of C_2 are sure to be detected? Explain.

3 Design a code in \mathbb{B}^4 where the first two positions are information positions. Give the parity-check equations, list the codewords, and find the minimum distance.

If \mathbf{a} and \mathbf{b} are any two words, let $d(\mathbf{a}, \mathbf{b})$ denote the distance between \mathbf{a} and \mathbf{b} . To *decode* a received word \mathbf{x} (which may contain errors of transmission) means to find the codeword closest to \mathbf{x} , that is, the codeword \mathbf{a} such that $d(\mathbf{a}, \mathbf{x})$ is a minimum. This is called *maximum-likelihood decoding*.

4 Decode the following words in C_1 : 11111, 00101, 11000, 10011, 10001, and 10111.

You may have noticed that the last two words in part 4 had ambiguous decodings: for example, 10111 may be decoded as either 10011 or 00111. This situation is clearly unsatisfactory. We shall see next what conditions will ensure that every word can be decoded into only *one* possible codeword.

In the remaining exercises, let C be a code in \mathbb{B}^n , let m denote the minimum distance in C , and let \mathbf{a} and \mathbf{b} denote codewords in C .

5 Prove that it is possible to detect up to $m - 1$ errors. (That is, if there are errors of transmission in $m - 1$ or fewer positions of a codeword, it can always be determined that the received word is incorrect.)

6 By the *sphere of radius k* about a codeword \mathbf{a} we mean the set of all words in \mathbb{B}^n whose distance from \mathbf{a} is no greater than k . This set is denoted by $S_k(\mathbf{a})$; hence

$$S_k(\mathbf{a}) = \{\mathbf{x}: d(\mathbf{a}, \mathbf{x}) \leq k\}$$

If $t = \frac{1}{2}(m - 1)$, prove that any two spheres of radius t , say $S_t(\mathbf{a})$ and $S_t(\mathbf{b})$, have no elements in common. [HINT: Assume there is a word \mathbf{x} such that $\mathbf{x} \in S_t(\mathbf{a})$ and $\mathbf{x} \in S_t(\mathbf{b})$. Using the definitions of t and m , show that this is impossible.]

7 Deduce from part 6 that if there are t or fewer errors of transmission in a codeword, the received word will be decoded correctly.

8 Let C_2 be the code described in part 2. (If you have not yet found the minimum distance in C_2 , do so now.) Using the results of parts 5 and 7, explain why two errors in any codeword can always be detected, and why one error in any codeword can always be corrected.

Is it possible for a group to have *two different* identity elements? Well, suppose e_1 and e_2 are identity elements of some group G . Then

$$e_1 * e_2 = e_2 \quad \text{because } e_1 \text{ is an identity element, and}$$

$$e_1 * e_2 = e_1 \quad \text{because } e_2 \text{ is an identity element}$$

Therefore

$$e_1 = e_2$$

This shows that in every group there is *exactly one* identity element.

Can an element a in a group have *two different inverses*? Well, if a_1 and a_2 are both inverses of a , then

$$a_1 * (a * a_2) = a_1 * e = a_1$$

and

$$(a_1 * a) * a_2 = e * a_2 = a_2$$

By the associative law, $a_1 * (a * a_2) = (a_1 * a) * a_2$; hence $a_1 = a_2$. This shows that in every group, each element has *exactly one* inverse.

Up to now we have used the symbol $*$ to designate the group operation. Other, more commonly used symbols are $+$ and \cdot (“plus” and “multiply”). When $+$ is used to denote the group operation, we say we are using *additive notation*, and we refer to $a + b$ as the *sum* of a and b . (Remember that a and b do not have to be numbers and therefore “sum” does not, in general, refer to adding numbers.) When \cdot is used to denote the group operation, we say we are using *multiplicative notation*, we usually write ab instead of $a \cdot b$, and call ab the *product* of a and b . (Once again, remember that “product” does not, in general, refer

to multiplying numbers.) Multiplicative notation is the most popular because it is simple and saves space. In the remainder of this book multiplicative notation will be used except where otherwise indicated. In particular, when we represent a group by a letter such as G or H , it will be understood that the group's operation is written as multiplication.

There is common agreement that in additive notation the identity element is denoted by 0, and the inverse of a is written as $-a$. (It is called the *negative* of a .) In multiplicative notation the identity element is e and the inverse of a is written as a^{-1} (" a inverse"). It is also a tradition that $+$ is to be used only for commutative operations.

The most basic rule of calculation in groups is the *cancellation law*, which allows us to cancel the factor a in the equations $ab = ac$ and $ab = ca$. This will be our first theorem about groups.

Theorem 1 *If G is a group and a, b, c are elements of G , then*

- (i) $ab = ac$ implies $b = c$ and
- (ii) $ba = ca$ implies $b = c$

It is easy to see why this is true: if we multiply (on the left) both sides of the equation $ab = ac$ by a^{-1} , we get $b = c$. In the case of $ba = ca$, we multiply on the right by a^{-1} . This is the *idea* of the proof; now here is the proof:

Suppose	$ab = ac$
Then	$a^{-1}(ab) = a^{-1}(ac)$
By the associative law,	$(a^{-1}a)b = (a^{-1}a)c$
that is,	$eb = ec$
Thus, finally,	$b = c$

Part (ii) is proved analogously.

In general, we *cannot* cancel a in the equation $ab = ca$. (Why not?)

Theorem 2 *If G is a group and a, b are elements of G , then*

$$ab=e \text{ implies } a=b^{-1} \text{ and } b=a^{-1}$$

The proof is very simple: if $ab = e$, then $ab = aa^{-1}$ so by the cancellation law, $b = a^{-1}$. Analogously, $a = b^{-1}$.

This theorem tells us that if the product of two elements is equal to e , these elements are inverses of each other. In particular, if a is the inverse of b , then b is the inverse of a .

The next theorem gives us important information about computing inverses.

Theorem 3 *If G is a group and a, b are elements of G , then*

- (i) $(ab^{-1} = b^{-1}a^{-1}$ and
- (ii) $(a^{-1})^{-1}=a$

The first formula tells us that the inverse of a product is the product of the inverses in reverse order. The next formula tells us that a is the inverse of the inverse of a . The proof of (i) is as follows:

$$\begin{aligned}
(ab)(b^{-1}a^{-1}) &= a[(bb^{-1})a^{-1}] && \text{by the associative law} \\
&= a[ea^{-1}] && \text{because } bb^{-1} = e \\
&= aa^{-1} \\
&= e
\end{aligned}$$

Since the product of ab and $b^{-1}a^{-1}$ is equal to e , it follows by [Theorem 2](#) that they are each other's inverses. Thus, $(ab)^{-1} = b^{-1}a^{-1}$. The proof of (ii) is analogous but simpler: $aa^{-1} = e$, so by [Theorem 2](#) a is the inverse of a^{-1} , that is, $a = (a^{-1})^{-1}$.

The associative law states that the two products $a(bc)$ and $(ab)c$ are equal; for this reason, no confusion can result if we denote either of these products by writing abc (without any parentheses), and call abc the product of these *three* elements in this order.

We may next define the product of any *four* elements a, b, c , and d in G by

$$abcd = a(bcd)$$

By successive uses of the associative law we find that

$$a(bc)d = ab(cd) = (ab)(cd) = (ab)cd$$

Hence the product $abcd$ (without parentheses, but without changing the order of its factors) is defined without ambiguity.

In general, any two products, each involving the same factors in the same order, are equal. The net effect of the associative law is that *parentheses are redundant*.

Having made this observation, we may feel free to use products of several factors, such as $a_1a_2 \cdots a_n$, without parentheses, whenever it is convenient. Incidentally, by using the identity $(ab)^{-1} = b^{-1}a^{-1}$ repeatedly, we find that

$$(a_1a_2 \cdots a_n)^{-1} = a_n^{-1} \cdots a_2^{-1}a_1^{-1}$$

If G is a finite group, the number of elements in G is called the *order* of G . It is customary to denote the order of G by the symbol

$$|G|$$

EXERCISES

Remark on notation In the exercises below, the exponential notation a^n is used in the following sense: if a is any element of a group G , then a^2 means aa , a^3 means aaa , and, in general, a^n is the product of n factors of a , for any positive integer n .

A. Solving Equations in Groups

Let a, b, c , and x be elements of a group G . In each of the following, solve for x in terms of a, b , and c .

Example Solve simultaneously: $x^2 = b$ and $x^5 = e$

From the first equation, $b = x^2$

Squaring, $b^2 = x^4$

Multiplying on the left by x , $xb^2 = xx^4 = x^5 = e$. (Note: $x^5 = e$ was given.)

Multiplying by $(b^2)^{-1}$, $xb^2(b^2)^{-1}$. Therefore, $x = (b^2)^{-1}$.

Solve:

1 $axb = c$

2 $x^2b = xa^{-1}c$

Solve simultaneously:

3 $x^2a = bxc^{-1}$ and $acx = xac$

4 $ax^2 = b$ and $x^3 = e$

5 $x^2 = a^2$ and $x^5 = e$

6 $(xax)^3 = bx$ and $x^2a = (xa)^{-1}$

B. Rules of Algebra in Groups

For each of the following rules, either prove that it is true in every group G , or give a counterexample to show that it is false in some groups. (All the counterexamples you need may be found in the group of matrices $\{I, A, B, C, D, K\}$ described on page 28.)

1 If $x^2 = e$, then $x = e$.

2 If $x^2 = a^2$, then $x = a$.

3 $(ab)^2 = a^2b^2$

4 If $x^2 = x$, then $x = e$.

5 For every $x \in G$, there is some $y \in G$ such that $x = y^2$. (This is the same as saying that every element of G has a “square root.”)

6 For any two elements x and y in G , there is an element z in G such that $y = xz$.

C. Elements That Commute

If a and b are in G and $ab = ba$, we say that a and b *commute*. Assuming that a and b commute, prove the following:

1 a^{-1} and b^{-1} commute.

2 a and b^{-1} commute. (HINT: First show that $a = b^{-1}ab$.)

3 a commutes with ab .

4 a^2 commutes with b^2 .

5 xax^{-1} commutes with xbx^{-1} , for any $x \in G$.

6 $ab = ba$ iff $aba^{-1} = b$.

(The abbreviation iff stands for “if and only if.” Thus, first prove that if $ab = ba$, then $aba^{-1} = b$. Next, prove that if $aba^{-1} = b$, then $ab = ba$. Proceed roughly as in [Exercise A](#). Thus, assuming $ab = ba$, solve for b . Next, assuming $aba^{-1} = b$, solve for ab .)

7 $ab = ba$ iff $aba^{-1}b^{-1} = e$.

† D. Group Elements and Their Inverses¹

Let G be a group. Let a, b, c denote elements of G , and let e be the neutral element of G .

1 Prove that if $ab = e$, then $ba = e$. (HINT: See [Theorem 2](#).)

2 Prove that if $abc = e$, then $cab = e$ and $bca = e$.

3 State a generalization of parts 1 and 2

Prove the following:

4 If $xay = a^{-1}$, then $yax = a^{-1}$.

5 Let a, b , and c each be equal to its own inverse. If $ab = c$, then $bc = a$ and $ca = b$.

6 If abc is its own inverse, then bca is its own inverse, and cab is its own inverse.

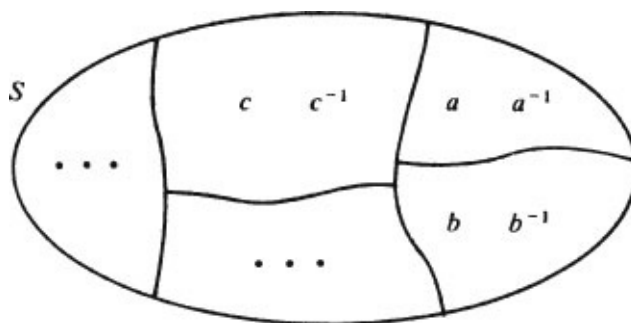
7 Let a and b each be equal to its own inverse. Then ba is the inverse of ab .

8 $a = a^{-1}$ iff $aa = e$. (That is, a is its own inverse iff $a^2 = e$.)

9 Let $c = c^{-1}$. Then $ab = c$ iff $xy^2 abc = e$.

† E. Counting Elements and Their Inverses

Let G be a finite group, and let S be the set of all the elements of G which are *not* equal to their own inverse. That is, $S = \{x \in G : x \neq x^{-1}\}$. The set S can be divided up into pairs so that each element is paired off with its own inverse. (See diagram on the next page.) Prove the following:



1 In any finite group G , the number of elements not equal to their own inverse is an even number.

2 The number of elements of G equal to their own inverse is odd or even, depending on whether the number of elements in G is odd or even.

3 If the order of G is even, there is at least one element x in G such that $x \neq e$ and $x = x^{-1}$.

In parts 4 to 6, let G be a finite *abelian* group, say, $G = \{e, a_1, a_2, \dots, a_n\}$. Prove the following:

4 $(a_1 a_2 \cdots a_n)^2 = e$

5 If there is no element $x \neq e$ in G such that $x = x^{-1}$, then $a_1 a_2 \cdots a_n = e$.

6 If there is exactly one $x \neq e$ in G such that $x = x^{-1}$, then $a_1 a_2 \cdots a_n = x$.

† F. Constructing Small Groups

In each of the following, let G be any group. Let e denote the neutral element of G .

1 If a, b are any elements of G , prove each of the following:

(a) If $a^2 = a$, then $a = e$.

(b) If $ab = a$, then $b = e$.

(c) If $ab = b$, then $a = e$.

2 Explain why every row of a group table must contain each element of the group exactly once. (HINT: Suppose jc appears twice in the row of a :

	\cdots	y_1	\cdots	y_2	\cdots
\vdots		\vdots		\vdots	
a	\cdots	x	\cdots	x	\cdots

Now use the cancellation law for groups.)

3 There is *exactly one group* on any set of three distinct elements, say the set $\{e, a, b\}$. Indeed, keeping in mind parts 1 and 2 above, there is only one way of completing the following table. Do so! *You need not prove associativity.*

	e	a	b
e	e	a	b
a	a		
b	b		

4 There is exactly one group G of four elements, say $G = \{e, a, b, c\}$, satisfying the additional property that $xx = e$ for every $x \in G$. Using only part 1 above, complete the following group table of G :

	e	a	b	c
e	e	a	b	c
a	a			
b	b			
c	c			

5 There is exactly one group G of four elements, say $G = \{e, a, b, c\}$, such that $xx = e$ for some $x \neq e$ in G , and $yy \neq e$ for some $y \in G$ (say, $aa = e$ and $bb \neq e$). Complete the group table of G , as in the preceding exercise.

6 Use [Exercise E3](#) to explain why the groups in parts 4 and 5 are the only possible groups of four elements (except for renaming the elements with different symbols).

G. Direct Products of Groups

If G and H are any two groups, their *direct product* is a new group, denoted by $G \times H$, and defined as follows: $G \times H$ consists of all the ordered pairs (x, y) where x is in G and y is in H . That is,

$$G \times H = \{(x, y) : x \in G \text{ and } y \in H\}$$

The operation of $G \times H$ consists of multiplying corresponding components:

$$(x, y) \cdot (x' y') = (xx', yy')$$

If G and H are denoted additively, it is customary to denote $G \times H$ additively:

$$(x, y) + (x' y') = (x+x', y+y')$$

1 Prove that $G \times H$ is a group by checking the three group axioms, (G1) to (G3):

$$(G1) \quad (x_1, y_1)[(x_2, y_2)(x_3, y_3)] = (\quad , \quad)$$

$$[(x_1, y_1)(x_2, y_2)](x_3, y_3) = (\quad , \quad)$$

$$(G2) \quad \text{Let } e_G \text{ be the identity element of } G, \text{ and } e_H \text{ the identity element of } H.$$

The identity element of $G \times H$ is (\quad, \quad) . Check

$$(G3) \quad \text{For each } (a, b) \in G \times H, \text{ the inverse of } (a, b) \text{ is } (\quad, \quad). \text{ Check.}$$

2 List the elements of $\mathbf{Z}_2 \times \mathbf{Z}_3$, and write its operation table. (NOTE: There are six elements, each of which is an ordered pair. The notation is additive.)

3 If G and H are abelian, prove that $G \times H$ is abelian.

4 Suppose the groups G and H both have the following property:

Every element of the group is its own inverse.

Prove that $G \times H$ also has this property.

H. Powers and Roots of Group Elements

Let G be a group, and $a, b \in G$. For any positive integer n we define a^n by

$$a^n = \underbrace{aaa \cdots a}_{n \text{ factors}}$$

If there is an element $x \in G$ such that $a = x^2$, we say that a has a square root in G . Similarly, if $a = y^3$ for some $y \in G$, we say a has a cube root in G . In general, a has an n th root in G if $a = z^n$ for some $z \in G$. Prove the following:

1 $(bab^{-1})^n = ba^n b^{-1}$, for every positive integer n . Prove by induction. (Remember that to prove a formula such as this one by induction, you first prove it for $n = 1$; next you prove that *if* it is true for $n = k$, *then* it must be true for $n = k + 1$. You may conclude that it is true for every positive integer n . Induction is explained more fully in Appendix C.)

2 If $ab = ba$, then $(ab)^n = a^n b^n$ for every positive integer n . Prove by induction.

3 If $xax = e$, then $(xa)^{2n} = a^n$.

4 If $a^3 = e$, then a has a square root.

5 If $a^2 = e$, then a has a cube root.

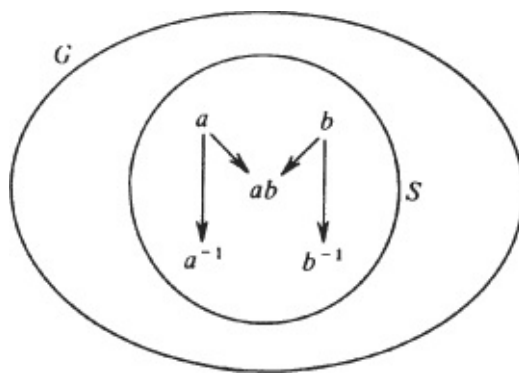
6 If $a \in 1$ has a cube root, so does a .

7 If $x^2 ax = a^{-1}$, then a has a cube root. (HINT: Show that xax is a cube root of a^{-1} .)

8 If $xax = b$, then 06 has a square root.

¹ When the exercises in a set are related, with some exercises building on preceding ones so that they must be done in sequence, this is indicated with a symbol t in the margin to the left of the heading.

Let G be a group, and S a nonempty subset of G . It may happen (though it doesn't have to) that the product of every pair of elements of S is in S . If it happens, we say that S is *closed with respect to multiplication*. Then, it may happen that the inverse of every element of S is in S . In that case, we say that S is *closed with respect to inverses*. If both these things happen, we call S a *subgroup* of G .



When the operation of G is denoted by the symbol $+$, the wording of these definitions must be adjusted: if the *sum* of every pair of elements of S is in S , we say that S is *closed with respect to addition*. If the negative of every element of S is in S , we say that S is *closed with respect to negatives*. If both these things happen, S is a *subgroup* of G .

For example, the *set of all the even integers* is a subgroup of the additive group \mathbf{Z} of the integers. Indeed, the sum of any two even integers is an even integer, and the negative of any even integer is an even integer.

As another example, \mathbf{Q}^* (the group of the nonzero rational numbers, under multiplication) is a subgroup of \mathbf{R}^* (the group of the nonzero real numbers, under multiplication). Indeed, $\mathbf{Q}^* \subseteq \mathbf{R}^*$ because every rational number is a real number. Furthermore, the product of any two rational numbers is rational, and the inverse (that is, the reciprocal) of any rational number is a rational number.

An important point to be noted is this: if S is a subgroup of G , *the operation of S is the same as the operation of G* . In other words, if a and b are elements of S , the product ab computed in S is precisely the product ab computed in G .

For example, it would be meaningless to say that $\langle \mathbb{Q}^*, \cdot \rangle$ is a subgroup of $\langle \mathbb{R}, + \rangle$; for although it is true that \mathbb{Q}^* is a subset of \mathbb{R} , the operations on these two groups are different.

The importance of the notion of subgroup stems from the following fact: *if G is a group and S is a subgroup of G , then S itself is a group.*

It is easy to see why this is true. To begin with, the operation of G , restricted to elements of S , is certainly an operation on S . *It is associative*: for if a , b , and c are in S , they are in G (because $S \subseteq G$); but G is a group, so $a(bc) = (ab)c$. Next, *the identity element e of G is in S* (and continues to be an identity element in S) for S is nonempty, so S contains an element a ; but S is closed with respect to inverses, so S also contains a^{-1} ; thus, S contains $aa^{-1} = e$, because S is closed with respect to multiplication. Finally, *every element of S has an inverse in S* because S is closed with respect to inverses. Thus, S is a group!

One reason why the notion of subgroup is useful is that it provides us with an easy way of showing that certain things are groups. Indeed, if G is already known to be a group, and S is a subgroup of G , we may conclude that S is a group without having to check all the items in the definition of “group.” This conclusion is illustrated by the next example.

Many of the groups we use in mathematics are groups whose elements are functions. In fact, historically, the first groups ever studied as such were groups of functions.

$\mathcal{F}(\mathbb{R})$ represents the *set of all functions from \mathbb{R} to \mathbb{R}* , that is, the set of all real-valued functions of a real variable. In calculus we learned how to add functions: if f and g are functions from \mathbb{R} to \mathbb{R} , their *sum* is the function $f + g$ given by

$$[f + g](x) = f(x) + g(x) \quad \text{for every real number } x$$

Clearly, $f + g$ is again a function from \mathbb{R} to \mathbb{R} , and is uniquely determined by f and g .

$\mathcal{F}(\mathbb{R})$, with the operation $+$ for adding functions, is the group $\langle \mathcal{F}(\mathbb{R}), + \rangle$, or simply $\mathcal{F}(\mathbb{R})$. The details are simple, but first, let us remember what it means for two functions to be equal. If f and g are functions from \mathbb{R} to \mathbb{R} , then f and g are *equal* (that is, $f = g$) if and only if $f(x) = g(x)$ for every real number x . In other words, to be equal, f and g must yield the same value when applied to every real number x .

To check that $+$ is associative, we must show that $f + [g + h] = [f + g] + h$, for every three functions, f , g , and h in $\mathcal{F}(\mathbb{R})$. This means that for any real number x , $\{f + [g + h]\}(x) = \{[f + g] + h\}(x)$. Well,

$$\{f + [g + h]\}(x) = f(x) + [g + h](x) = f(x) + g(x) + h(x)$$

and $\{[f + g] + h\}(x)$ has the same value.

The neutral element of $\mathcal{F}(\mathbb{R})$ is the function $\mathbf{0}$ given by

$$\mathbf{0}(x) = 0 \quad \text{for every real number } x$$

To show that $\mathbf{0} + f = f$, one must show that $[\mathbf{0} + f](x) = f(x)$ for every real number x . This is true because $[\mathbf{0} + f](x) = \mathbf{0}(x) + f(x) = 0 + f(x) = f(x)$.

Finally, the inverse of any function f is the function $-f$ given by

$$[-f](x) = -f(x) \quad \text{for every real number } x$$

One perceives immediately that $f + [-f] = \mathbf{0}$, for every function f .

$\mathcal{C}(\mathbb{R})$ represents the set of all *continuous* functions from \mathbb{R} to \mathbb{R} . Now, $\mathcal{C}(\mathbb{R})$, with the operation $+$, is a subgroup of $\mathcal{F}(\mathbb{R})$, because we know from calculus that the sum of any two continuous functions is a

continuous function, and the negative $-f$ of any continuous function f is a continuous function. Because any subgroup of a group is itself a group, we may conclude that $\mathcal{C}(\mathbb{R})$, with the operation $+$, is a group. It is denoted by $\langle \mathcal{C}(\mathbb{R}), + \rangle$, or simply $\mathcal{C}(\mathbb{R})$.

$\mathcal{D}(\mathbb{R})$ represents the set of all the *differentiable* functions from \mathbb{R} to \mathbb{R} . It is a subgroup of $\mathcal{F}(\mathbb{R})$ because the sum of any two differentiable functions is differentiable, and the negative of any differentiable function is differentiable. Thus, $\mathcal{D}(\mathbb{R})$, with the operation of adding functions, is a group denoted by $\langle \mathcal{D}(\mathbb{R}), + \rangle$, or simply $\mathcal{D}(\mathbb{R})$.

By the way, in any group G the one-element subset $\{e\}$, containing only the neutral element, is a subgroup. It is closed with respect to multiplication because $ee = e$, and closed with respect to inverses because $e^{-1} = e$. At the other extreme, the whole group G is obviously a subgroup of itself. These two examples are, respectively, the smallest and largest possible subgroups of G . They are called the *trivial* subgroups of G . All the other subgroups of G are called *proper* subgroups.

Suppose G is a group and a , b , and c are elements of G . Define S to be the subset of G which contains *all the possible products of a , b , c , and their inverses*, in any order, with repetition of factors permitted. Thus, typical elements of S would be

$$abac^{-1}$$

$$c^{-1}a^{-1}bbc$$

and so on. It is easy to see that S is a subgroup of G : for if two elements of S are multiplied together, they yield an element of S , and the inverse of any element of S is an element of S . For example, the product of aba and $cb^{-1}ac$ is

$$abacb^{-1}ac$$

and the inverse of $ab^{-1}c^{-1}a$ is

$$a^{-1}cba^{-1}$$

S is called the *subgroup of G generated by a , b , and c* .

If a_1, \dots, a_n are any finite number of elements of G , we may define the *subgroup generated by a_1, \dots, a_n* in the same way. In particular, if a is a single element of G , we may consider the subgroup generated by a . This subgroup is designated by the symbol $\langle a \rangle$, and is called a *cyclic subgroup* of G ; a is called its *generator*. Note that $\langle a \rangle$ consists of all the possible products of a and a^{-1} , for example, $a^{-1}aaa^{-1}$ and $aaa^{-1}aa^{-1}$. However, since factors of a^{-1} cancel factors of a , there is no need to consider products involving both a and a^{-1} side by side. Thus, $\langle a \rangle$ contains

$$a, aa, aaa, \dots,$$

$$a^{-1}, a^{-1}a^{-1}, a^{-1}a^{-1}a^{-1}, \dots,$$

as well as $aa^{-1} = e$.

If the operation of G is denoted by $+$, the same definitions can be given with “sums” instead of “products.”

In the group of matrices whose table appears on page 28, the subgroup generated by D is $\langle D \rangle = \{I, B, D\}$ and the subgroup generated by A is $\langle A \rangle = \{I, A\}$. (The student should check the table to verify this.) In fact, the entire group G of that example is generated by the two elements A and B .

If a group G is generated by a single element a , we call G a *cyclic group*, and write $G = \langle a \rangle$. For example, the additive group \mathbf{Z}_6 is cyclic. (What is its generator?)

Every finite group G is generated by one or more of its elements (obviously). A set of equations, involving only the generators and their inverses, is called a set of *defining equations* for G if these equations completely determine the multiplication table of G .

For example, let G be the group $\{e, a, b, b^2, ab, ab^2\}$ whose generators a and b satisfy the equations

$$a^2 = e \quad b^3 = e \quad ba = ab^2 \quad (1)$$

These three equations do indeed determine the multiplication table of G . To see this, note first that the equation $ba = ab^2$ allows us to switch powers of a with powers of b , bringing powers of a to the left, and powers of b to the right. For example, to find the product of ab and ab^2 , we compute as follows:

$$(ab)(ab^2) = \underbrace{abab^2}_{=ab^2} = aab^2b^2 = a^2b^4$$

But by [Equations \(1\)](#), $a^2 = e$ and $b^4 = b^3b = b$; so finally, $(ab)(ab^2) = b$. All the entries in the table of G may be computed in the same fashion.

When a group is determined by a set of generators and defining equations, its structure can be efficiently represented in a diagram called a *Cayley diagram*. These diagrams are explained in [Exercise G](#).

EXERCISES

A. Recognizing Subgroups

In parts 1–6 below, determine whether or not H is a subgroup of G . (Assume that the operation of H is the same as that of G .)

Instructions If H is a subgroup of G , show that both conditions in the definition of “subgroup” are satisfied. If H is *not* a subgroup of G , explain which condition fails.

Example $G = \mathbf{R}^*$, the multiplicative group of the real numbers.

$$H = \{2^n : n \in \mathbf{Z}\} \quad H \text{ is } \boxed{\text{yes}} \quad \text{is not } \boxed{\text{no}} \quad \text{a subgroup of } G.$$

(i) If $2^n, 2^m \in H$, then $2^n 2^m = 2^{n+m}$. But $n + m \in \mathbf{Z}$, so $2^{n+m} \in H$.

(ii) If $2^n \in H$, then $1/2^n = 2^{-n}$. But $-n \in \mathbf{Z}$, so $2^{-n} \in H$.

(Note that in this example the operation of G and H is multiplication. In the next problem, it is addition.)

1 $G = \langle \mathbf{R}, + \rangle$, $H = \{\log a : a \in \mathbf{Q}, a > 0\}$. H is ☐ is not ☐ a subgroup of G .

2 $G = \langle \mathbf{R}, + \rangle$, $H = \{\log n : n \in \mathbf{Z}, n > 0\}$. H is ☐ is not ☐ a subgroup of G .

3 $G = \langle \mathbf{R}, + \rangle$, $H = \{x \in \mathbf{R} : \tan x \in \mathbf{Q}\}$. H is ☐ is not ☐ a subgroup of G .

HINT: Use the following formula from trigonometry:

$$\tan(x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}$$

- 4 $G = \langle \mathbb{R}^*, \cdot \rangle$, $H = \{2^n 3^m : m, n \in \mathbb{Z}\}$. H is ☐ is not ☐ a subgroup of G .
 5 $G = \langle \mathbb{R} \times \mathbb{R}, + \rangle$, $H = \{(x, y) : y = 2x\}$. H is ☐ is not ☐ a subgroup of G .
 6 $G = \langle \mathbb{R} \times \mathbb{R}, + \rangle$, $H = \{(x, y) : x^2 + y^2 > 0\}$. H is ☐ is not ☐ a subgroup of G .
 7 Let C and D be sets, with $C \subseteq D$. Prove that P_C is a subgroup of P_D .

B. Subgroups of Functions

In each of the following, show that H is a subgroup of G .

Example $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(0) = 0\}$

(i) Suppose $f, g \in H$; then $f(0) = 0$ and $g(0) = 0$, so $[f + g](0) = f(0) + g(0) = 0 + 0 = 0$. Thus, $f + g \in H$.

(ii) If $f \in H$, then $f(0) = 0$. Thus, $[-f](0) = -f(0) = -0 = 0$, so $-f \in H$.

1 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(x) = 0 \text{ for every } x \in [0, 1]\}$

2 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(-x) = -f(x)\}$

3 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f \text{ is periodic of period } \pi\}$

REMARK: A function f is said to be *periodic* of period a if there is a number a , called the period of f , such that $f(x) = f(x + na)$ for every $x \in \mathbb{R}$ and $n \in \mathbb{Z}$.

4 $G = \langle \mathcal{C}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{C}(\mathbb{R}) : \int_0^1 f(x) dx = 0\}$

5 $G = \langle \mathcal{D}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{D}(\mathbb{R}) : df/dx \text{ is constant}\}$

6 $G = \langle \mathcal{F}(\mathbb{R}), + \rangle$, $H = \{f \in \mathcal{F}(\mathbb{R}) : f(x) \in \mathbb{Z} \text{ for every } x \in \mathbb{R}\}$

C. Subgroups of Abelian Groups

In the following exercises, let G be an abelian group.

1 If $H = \{x \in G : x = x^{-1}\}$, that is, H consists of all the elements of G which are their own inverses, prove that H is a subgroup of G .

2 Let n be a fixed integer, and let $H = \{x \in G : x^n = e\}$. Prove that H is a subgroup of G .

3 Let $H = \{x \in G : x = y^2 \text{ for some } y \in G\}$; that is, let H be the set of all the elements of G which have a square root. Prove that H is a subgroup of G .

4 Let H be a subgroup of G , and let $K = \{x \in G : x^2 \in H\}$. Prove that K is a subgroup of G .

5. Let H be a subgroup of G , and let K consist of all the elements x in G such that some power of x is in H . That is, $K = \{x \in G : \text{for some integer } n > 0, x^n \in H\}$. Prove that K is a subgroup of G .

6 Suppose H and K are subgroups of G , and define HK as follows:

$$HK = \{xy : x \in H \text{ and } y \in K\}$$

Prove that HK is a subgroup of G .

7 Explain why parts 4–6 are not true if G is not abelian.

D. Subgroups of an Arbitrary Group

Let G be a group.

1 If H and K are subgroups of a group G , prove that $H \cap K$ is a subgroup of G . (Remember that $x \in H \cap K$ iff $x \in H$ and $x \in K$.)

2 Let H and K be subgroups of G . Prove that if $H \subseteq K$, then H is a subgroup of K .

3 By the *center* of a group G we mean the set of all the elements of G which commute with every element of G , that is,

$$C = \{a \in G : ax = xa \text{ for every } x \in G\}$$

Prove that C is a subgroup of G .

4 Let $C' = \{a \in G : (ax)^2 = (xa)^2 \text{ for every } x \in G\}$. Prove that C' is a subgroup of G .

5 Let G be a *finite* group, and let S be a nonempty subset of G . Suppose S is closed with respect to multiplication. Prove that S is a subgroup of G . (HINT: It remains to prove that S contains e and is closed with respect to inverses. Let $S = \{a_1, \dots, a_n\}$. If $a_i \in S$, consider the *distinct* elements $a_i a_1, a_i a_2, \dots, a_i a_n$.)

6 Let G be a group and $f: G \rightarrow G$ a function. A *period* of f is any element a in G such that $f(x) = f(ax)$ for every $x \in G$. Prove: The set of all the periods of f is a subgroup of G .

7 Let H be a subgroup of G , and let $K = \{x \in G : xax^{-1} \in H \text{ iff } a \in H\}$. Prove:

(a) K is a subgroup of G .

(b) H is a subgroup of K .

8 Let G and H be groups, and $G \times H$ their direct product.

(a) Prove that $\{(x, e) : x \in G\}$ is a subgroup of $G \times H$.

(b) Prove that $\{(x, x) : x \in G\}$ is a subgroup of $G \times G$.

E. Generators of Groups

1 List all the cyclic subgroups of $\langle \mathbf{z}_{10}, + \rangle$.

2 Show that \mathbf{z}_{10} is generated by 2 and 5.

3 Describe the subgroup of \mathbf{z}_{12} generated by 6 and 9.

4 Describe the subgroup of \mathbf{z} generated by 10 and 15.

5 Show that \mathbf{z} is generated by 5 and 7.

6 Show that $\mathbf{z}_2 \times \mathbf{z}_3$ is a cyclic group. Show that $\mathbf{z}_3 \times \mathbf{z}_4$ is a cyclic group.

7 Show that $\mathbf{z}_2 \times \mathbf{z}_4$ is *not* a cyclic group, but is generated by $(1, 1)$ and $(1, 2)$.

8 Suppose a group G is generated by two elements a and b . If $ab = ba$, prove that G is abelian.

F. Groups Determined by Generators and Defining Equations

1 Let G be the group $\{e, a, b, b^2, ab, ab^2\}$ whose generators satisfy $a^2 = e, b^3 = e, ba = ab^2$. Write the table of G .

2 Let G be the group $\{e, a, b, b^2, b^3, ab, ab^2, ab^3\}$ whose generators satisfy $a^2 = e, b^4 = e, ba = ab^3$. Write the table of G . (G is called the *dihedral group* D_4 .)

Just as there are great works in art and music, there are also great creations of mathematics. “Greatness,” in mathematics as in art, is hard to define, but the basic ingredients are clear: a *great* theorem should contribute substantial new information, and it should be *unexpected*!. That is, it should reveal something which common sense would not naturally lead us to expect. The most celebrated theorems of plane geometry, as may be recalled, come as a complete surprise; as the proof unfolds in simple, sure steps and we reach the conclusion—a conclusion we may have been skeptical about, but which is now established beyond a doubt—we feel a certain sense of awe not unlike our reaction to the ironic or tragic twist of a great story.

In this chapter we will consider a result of modern algebra which, by all standards, is a great theorem. It is something we would not likely have foreseen, and which brings new order and simplicity to the relationship between a group and its subgroups.

We begin by adding to our algebraic tool kit a new notion—a conceptual tool of great versatility which will serve us well in all the remaining chapters of this book. It is the concept of a *coset*.

Let G be a group, and H a subgroup of G . For any element a in G , the symbol

$$aH$$

*denotes the set of all products ah , as a remains fixed and h ranges over H . aH is called a **left coset** of H in G .*

In similar fashion

$$Ha$$

*denotes the set of all products ha , as a remains fixed and h ranges over H . Ha is called a **right coset** of H in G .*

In practice, it will make no difference whether we use left cosets or right cosets, just as long as we *remain consistent*. Thus, from here on, whenever we use cosets we will use *right* cosets. To simplify our

sentences, we will say *coset* when we mean “right coset.”

When we deal with cosets in a group G , we must keep in mind that every coset in G is a subset of G . Thus, when we need to prove that two cosets Ha and Hb are equal, we must show that they are *equal sets*. What this means, of course, is that every element $x \in Ha$ is in Hb , and conversely, every element $y \in Hb$ is in Ha . For example, let us prove the following elementary fact:

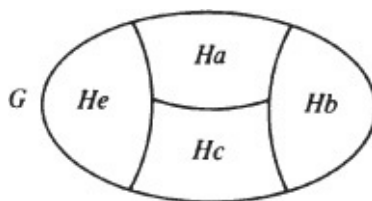
$$\text{If } a \in Hb, \text{ then } Ha = Hb \quad (1)$$

We are given that $a \in Hb$, which means that $a = h_1b$ for some $h_1 \in H$. We need to prove that $Ha = Hb$.

Let $x \in Ha$; this means that $x = h_2a$ for some $h_2 \in H$. But $a = h_1b$, so $x = h_2a = (h_2h_1)b$, and the latter is clearly in Hb . This proves that every $x \in Ha$ is in Hb ; analogously, we may show that every $y \in Hb$ is in Ha , and therefore $Ha = Hb$.

The first major fact about cosets now follows. Let G be a group and let H be a fixed subgroup of G :

Theorem 1 *The family of all the cosets Ha , as a ranges over G , is a partition of G .*



PROOF: First, we must show that any two cosets, say Ha and Hb , are either disjoint or equal. If they are disjoint, we are done. If not, let $x \in Ha \cap Hb$. Because $x \in Ha$, $x = h_1a$ for some $h_1 \in H$. Because $x \in Hb$, $x = h_2b$ for some $h_2 \in H$. Thus, $h_1a = h_2b$, and solving for a , we have

$$a = (h_1^{-1}h_2)b$$

Thus,

$$a \in Hb$$

It follows from Property (1) above that $Ha = Hb$.

Next, we must show that *every* element $c \in G$ is in one of the cosets of H . But this is obvious, because $c = ec$ and $e \in H$; therefore,

$$c = ec \in Hc$$

Thus, the family of all the cosets of H is a partition of G . ■

Before going on, it is worth making a small comment: A given coset, say Hb , may be written in more than one way. By Property (1) *if a is any element in Hb , then Hb is the same as Ha* . Thus, for example, if a coset of H contains n different elements a_1, a_2, \dots, a_n , it may be written in n different ways, namely, Ha_1, Ha_2, \dots, Ha_n .

The next important fact about cosets concerns finite groups. Let G be a finite group, and H a subgroup of G . We will show that *all the cosets of H have the same number of elements!* This fact is a consequence of the next theorem.

Theorem 2 *If Ha is any coset of H , there is a one-to-one correspondence from H to Ha .*

PROOF: The most obvious function from H to Ha is the one which, for each $h \in H$, matches h with ha . Thus, let $f: H \rightarrow Ha$ be defined by

$$f(h) = ha$$

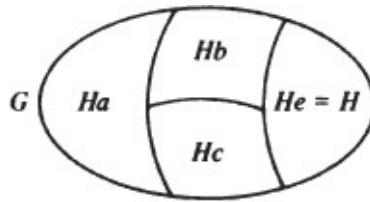
Remember that a remains fixed whereas h varies, and check that f is injective and surjective.

f is injective: Indeed, if $f(h_1) = f(h_2)$, then $h_1a = h_2a$, and therefore $h_1 = h_2$.

f is surjective, because every element of Ha is of the form ha for some $h \in H$, and $ha = f(h)$.

Thus, f is a one-to-one correspondence from H to Ha , as claimed. ■

By [Theorem 2](#), any coset Ha has the same number of elements as H , and therefore all the cosets have the same number of elements!



Let us take a careful look at what we have proved in [Theorems 1](#) and [2](#). Let G be a finite group and H any subgroup of G . G has been partitioned into cosets of H , and all the cosets of H have the same number of elements (which is the same as the number of elements in H). Thus, *the number of elements in G is equal to the number of elements in H , multiplied by the number of distinct cosets of H* . This statement is known as Lagrange's theorem. (Remember that the number of elements in a group is called the group's order.)

Theorem 3: Lagrange's theorem *Let G be a finite group, and H any subgroup of G . The order of G is a multiple of the order of H .*

In other words, the order of any subgroup of a group G is a divisor of the order of G .

For example, if G has 15 elements, its proper subgroups may have either 3 or 5 elements. If G has 7 elements, it has *no* proper subgroups, for 7 has no factors other than 1 and 7. This last example may be generalized:

Let G be a group with a *prime* number p of elements. If $a \in G$ where $a \neq e$, then the order of a is some integer $m \neq 1$. But then the cyclic group $\langle a \rangle$ has m elements. By Lagrange's theorem, m must be a factor of p . But p is a prime number, and therefore $m = p$. It follows that $\langle a \rangle$ has p elements, and is therefore all of G ! Conclusion:

Theorem 4 *If G is a group with a prime number p of elements, then G is a cyclic group. Furthermore, any element $a \neq e$ in G is a generator of G .*

[Theorem 4](#), which is merely a consequence of Lagrange's theorem, is quite remarkable in itself. What it says is that *there is (up to isomorphism) only one group of any given prime order p* . For example, the only group (up to isomorphism) of order 7 is \mathbf{Z}_7 , the only group of order 11 is \mathbf{Z}_{11} , and so on! So we now have complete information about all the groups whose order is a prime number.

By the way, if a is any element of a group G , the order of a is the same as the order of the cyclic

subgroup $\langle a \rangle$, and by Lagrange's theorem this number is a divisor of the order of G . Thus,

Theorem 5 *The order of any element of a finite group divides the order of the group.*

Finally, if G is a group and H is a subgroup of G , the *index of H in G* is the number of cosets of H in G . We denote it by $(G:H)$. Since the number of elements in G is equal to the number of elements in H , multiplied by the number of cosets of H in G ,

$$(G:H) = \frac{\text{order of } G}{\text{order of } H}$$

EXERCISES

A. Examples of Cosets in Finite Groups

In each of the following, H is a subgroup of G . In parts 1–5 list the cosets of H . For each coset, list the elements of the coset.

Example $G = \mathbf{Z}_4$, $H = \{0, 2\}$.

(REMARK: If the operation of G is denoted by $+$, it is customary to write $H + x$ for a coset, rather than Hx .) The cosets of H in this example are

$$H = H + 0 = H + 2 = \{0, 2\} \quad \text{and} \quad H + 1 = H + 3 = \{1, 3\}$$

- 1 $G = S_3$, $H = \{\varepsilon, \beta, \delta\}$.
- 2 $G = S_3$, $H = \{\varepsilon, \alpha\}$.
- 3 $G = \mathbf{Z}_{15}$, $H = \langle 5 \rangle$.
- 4 $G = D_4$, $H = \{R_0, R_4\}$. (For D_4 , see page 73.)
- 5 $G = S_4$, $H = A_4$. (For A_4 , see page 86.)
- 6 Indicate the order and index of each of the subgroups in parts 1 to 5.

B. Examples of Cosets in Infinite Groups

Describe the cosets of the subgroups described in parts 1–5:

- 1 The subgroup $\langle 3 \rangle$ of \mathbf{Z} .
- 2 The subgroup \mathbf{Z} of \mathbb{R} .
- 3 The subgroup $H = \{2^n: n \in \mathbf{Z}\}$ of \mathbb{R}^* .
- 4 The subgroup $\langle \frac{1}{2} \rangle$ of \mathbb{R}^* ; the subgroup $\langle \frac{1}{2} \rangle$ of \mathbb{R} .
- 5 The subgroup $H = \{(x, y): x = y\}$ of $(\mathbb{R} \times \mathbb{R})$.
- 6 For any positive integer m , what is the index of $\langle m \rangle$ in \mathbf{Z} ?
- 7 Find a subgroup of \mathbb{R}^* whose index is equal to 2.

C. Elementary Consequences of Lagrange's Theorem

Let G be a finite group. Prove the following:

- 1 If G has order n , then $x^n = e$ for every x in G .
- 2 Let G have order pq , where p and q are primes. Either G is cyclic, or every element $x \neq e$ in G has

order p or q .

3 Let G have order 4. Either G is cyclic, or every element of G is its own inverse. Conclude that every group of order 4 is abelian.

4 If G has an element of order p and an element of order q , where p and q are distinct primes, then the order of G is a multiple of pq .

5 If G has an element of order k and an element of order m , then $|G|$ is a multiple of $\text{lcm}(k, m)$, where $\text{lcm}(k, m)$ is the least common multiple of k and m .

6 Let p be a prime number. In any finite group, the number of elements of order p is a multiple of $p - 1$.

D. Further Elementary Consequences of Lagrange's Theorem

Let G be a finite group, and let H and K be subgroups of G . Prove the following:

1 Suppose $H \subseteq K$ (therefore H is a subgroup of K). Then $(G: H) = (G: K)(K: H)$.

2 The order of $H \cap K$ is a common divisor of the order of H and the order of K .

3 Let H have order m and K have order n , where m and n are relatively prime. Then $H \cap K = \{e\}$.

4 Suppose H and K are not equal, and both have order the same prime number p . Then $H \cap K = \{e\}$.

5 Suppose H has index p and K has index q , where p and q are distinct primes. Then the index of $H \cap K$ is a multiple of pq .

6 If G is an abelian group of order n , and m is an integer such that m and n are relatively prime, then the function $f(x) = x^m$ is an automorphism of G .

E. Elementary Properties of Cosets

Let G be a group, and H a subgroup of G . Let a and b denote elements of G . Prove the following:

1 $Ha = Hb$ iff $ab^{-1} \in H$.

2 $Ha = H$ iff $a \in H$.

3 If $aH = Ha$ and $bH = Hb$, then $(ab)H = H(ab)$.

4 If $aH = Ha$, then $a^{-1}H = Ha^{-1}$.

5 If $(ab)H = (ac)H$, then $bH = cH$.

6 The number of right cosets of H is equal to the number of left cosets of H .

7 If J is a subgroup of G such that $J = H \cap K$, then for any $a \in G$, $Ja = Ha \cap Ka$. Conclude that if H and K are of finite index in G , then their intersection $H \cap K$ is also of finite index in G .

Theorem 5 of this chapter has a useful converse, which is the following:

Cauchy's theorem *If G is a finite group, and p is a prime divisor of $|G|$, then G has an element of order p .*

For example, a group of order 30 must have elements of orders 2, 3, and 5. Cauchy's theorem has an elementary proof, which may be found on page 340.

In the next few exercise sets, we will survey all possible groups whose order is ≤ 10 . By **Theorem 4** of this chapter, if G is a group with a prime number p of elements, then $G \cong \mathbb{Z}_p$. This takes care of all groups of orders 2, 3, 5, and 7. In Exercise G6 of **Chapter 15**, it will be shown that if G is a group with p^2 elements (where p is a prime), then $G \cong \mathbb{Z}_{p^2}$ or $G \cong \mathbb{Z}_p \times \mathbb{Z}_p$. This will take care of all groups of orders 4 and 9. The remaining cases are examined in the next three exercise sets.

† F. Survey of All Six-Element Groups

Let G be any group of order 6. By Cauchy's theorem, G has an element a of order 2 and an element b of order 3. By [Chapter 10](#), Exercise E3, the elements

$$e, a, b, b^2, ab, ab^2$$

are all distinct; and since G has only six elements, these are all the elements in G . Thus, ba is one of the elements e, a, b, b^2, ab , or ab^2 .

1 Prove that ba cannot be equal to either e, a, b , or b^2 . Thus, $ba = ab$ or $ba = ab^2$.

Either of these two equations completely determines the table of G . (See the discussion at the end of [Chapter 5](#).)

2 If $ba = ab$, prove that $G \cong \mathbf{z}_6$.

3 If $ba = ab^2$, prove that $G \cong S_3$.

It follows that \mathbf{z}_6 and S_3 are (up to isomorphism), the only possible groups of order 6.

† G. Survey of All 10-Element Groups

Let G be any group of order 10.

1 Reason as in Exercise F to show that $G = \{e, a, b, b^2, b^3, b^4, ab, ab^2, ab^3, ab^4\}$, where a has order 2 and b has order 5.

2 Prove that ba cannot be equal to e, a, b, b^2, b^3 , or b^4 .

3 Prove that if $ba = ab$, then $G \cong \mathbf{z}_{10}$.

4 If $ba = ab^2$, prove that $ba^2 = a^2b^4$, and conclude that $b = b^4$. This is impossible because b has order 5; hence $ba \neq ab^2$. (HINT: The equation $ba = ab^2$ tells us that we may move a factor a from the right to the left of a factor b , but in so doing, we must square b . To prove an equation such as the preceding one, move all factors a to the left of all factors b .)

5 If $ba = ab^3$, prove that $ba^2 = a^2b^9 = a^2b^4$, and conclude that $b = b^4$. This is impossible (why?); hence $ba \neq ab^3$.

6 Prove that if $ba = ab^4$, then $G \cong D_5$ (where D_5 is the group of symmetries of the pentagon).

Thus, the only possible groups of order 10 (up to isomorphism), are \mathbf{z}_{10} and D_5 .

† H. Survey of All Eight-Element Groups

Let G be any group of order 8. If G has an element of order 8, then $G \cong \mathbf{z}_8$. Let us assume now that G has no element of order 8; hence all the elements $\neq e$ in G have order 2 or 4.

1 If every $x \neq e$ in G has order 2, let a, b, c be three such elements. Prove that $G = \{e, a, b, c, ab, bc, ac, abc\}$. Conclude that $G \cong \mathbf{z}_2 \times \mathbf{z}_2 \times \mathbf{z}_2$.

In the remainder of this exercise set, assume G has an element a of order 4. Let $H = \langle a \rangle = \{e, a, a^2, a^3\}$. If $b \in G$ is not in H , then the coset $Hb = \{b, ab, a^2b, a^3b\}$. By Lagrange's theorem, G is the union of $He = H$ and Hb ; hence

$$G = \{e, a, a^2, a^3, b, ab, a^2b, a^3b\}$$

2 Assume there is in Hb an element of order 2. (Let b be this element.) If $ba = a^2b$, prove that $b^2a = a^4b^2$, hence $a = a^4$, which is impossible. (Why?) Conclude that either $ba = ab$ or $ba = a^3b$.

3 Let b be as in part 2. Prove that if $ba = ab$, then $G \cong \mathbf{Z}_4 \times \mathbf{Z}_2$.

4 Let b be as in part 2. Prove that if $ba = a^3b$, then $G \cong D_4$.

5 Now assume the hypothesis in part 2 is false. Then b, ab, a^2b , and a^3b all have order 4. Prove that $b^2 = a^2$. (HINT: What is the order of b^2 ? What element in G has the same order?)

6 Prove: If $ba = ab$, then $(a^3b)^2 = e$, contrary to the assumption that $\text{ord}(a^3b) = 4$. If $ba = a^2b$, then $a = b^4a = e$, which is impossible. Thus, $ba = a^3b$.

7 The equations $a^4 = b^4 = e$, $a^2 = b^2$, and $ba = a^3b$ completely determine the table of G . Write this table. (G is known as the *quaternion group* Q .)

Thus, the only groups of order 8 (up to isomorphism) are \mathbf{Z}_8 , $\mathbf{Z}_2 \times \mathbf{Z}_2 \times \mathbf{Z}_2$, $\mathbf{Z}_4 \times \mathbf{Z}_2$, D_4 , and Q .

† I. Conjugate Elements

If $a \in G$, a *conjugate* of a is any element of the form xax^{-1} , where $x \in G$. (Roughly speaking, a conjugate of a is any product consisting of a sandwiched between any element and its inverse.) Prove each of the following:

1 The relation “ a is equal to a conjugate of b ” is an equivalence relation in G . (Write $a \sim b$ for “ a is equal to a conjugate of b .”)

This relation \sim partitions any group G into classes called *conjugacy classes*. (The conjugacy class of a is $[a] = \{xax^{-1} : x \in G\}$.)

For any element $a \in G$, the *centralizer* of a , denoted by C_a , is the set of all the elements in G which commute with a . That is,

$$C_a = \{x \in G : xa = ax\} = \{x \in G : xax^{-1} = a\}$$

Prove the following:

2 For any $a \in G$, C_a is a subgroup of G .

3 $x^{-1}ax = y^{-1}ay$ iff xy^{-1} commutes with a iff $xy^{-1} \in C_a$.

4 $x^{-1}ax = y^{-1}ay$ iff $C_ax = C_ay$. (HINT: Use Exercise El.)

5 There is a one-to-one correspondence between the set of all the conjugates of a and the set of all the cosets of C_a . (HINT: Use part 4.)

6 The number of distinct conjugates of a is equal to $(G : C_a)$, the index of C_a in G . Thus, *the size of every conjugacy class is a factor of $|G|$* .

† J. Group Acting on a Set

Let A be a set, and let G be any subgroup of S_A . G is a group of permutations of A ; we say it is a *group acting on the set A* . Assume here that G is a finite group. If $u \in A$, the *orbit of u* (with respect to G) is the set

$$O(u) = \{g(u) : g \in G\}$$



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – II – ALGEBRA-I – SMT1501

Unit-II Homomorphism

In this chapter we continue our discussion of functions, but we confine our discussions to functions *from a set to itself*. In other words, we consider only functions $f: A \rightarrow A$ whose domain is a set A and whose range is in the same set A .

To begin with, we note that any two functions f and g (from A to A) are *equal* if and only if $f(x) = g(x)$ for every element x in A .

If f and g are functions from A to A , their composite $f \circ g$ is also a function from A to A . We recall that it is the function defined by

$$[f \circ g](x) = f(g(x)) \quad \text{for every } x \text{ in } A \quad (1)$$

It is a very important fact that the *composition of functions is associative*. Thus, if f , g , and h are three functions from A to A , then

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

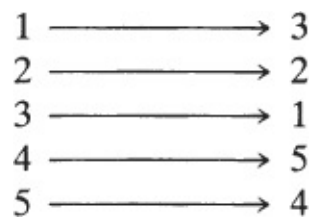
To prove that the functions $f \circ (g \circ h)$ and $(f \circ g) \circ h$ are equal, one must show that for every element x in A ,

$$\{f \circ [g \circ h]\}(x) = \{[f \circ g] \circ h\}(x)$$

We get this by repeated use of [Equation \(1\)](#):

$$\begin{aligned} \{f \circ [g \circ h]\}(x) &= f([g \circ h](x)) = f(g(h(x))) \\ &= [f \circ g](h(x)) = \{[f \circ g] \circ h\}(x) \end{aligned}$$

By a *permutation* of a set A we mean a *bijective function* from A to A , that is, a one-to-one correspondence between A and itself. In elementary algebra we learned to think of a permutation as a *rearrangement* of the elements of a set. Thus, for the set $\{1,2,3,4,5\}$, we may consider the rearrangement which changes $(1,2,3,4,5)$ to $(3,2,1,5,4)$; this rearrangement may be identified with the function



which is obviously a one-to-one correspondence between the set $\{1,2,3,4,5\}$ and itself. It is clear, therefore, that there is no real difference between the new definition of permutation and the old. The new definition, however, is more general in a very useful way since it allows us to speak of permutations of sets A even when A has infinitely many elements.

In [Chapter 6](#) we saw that the composite of any two bijective functions is a bijective function. Thus, *the composite of any two permutations of A is a permutation of A* . It follows that we may regard the operation \circ of composition as *an operation on the set of all the permutations of A* . We have just seen that composition is an associative operation. Is there a neutral element for composition?

For any set A , the *identity function on A* , symbolized by ε_A or simply ε , is the function $x \rightarrow x$ which carries every element of A to itself. That is, it is defined by

$$\varepsilon(x) = x \quad \text{for every element} \quad x \in A$$

It is easy to see that ε is a permutation of A (it is a one-to-one correspondence between A and itself); and if f is any other permutation of A , then

$$f \circ \varepsilon = f \quad \text{and} \quad \varepsilon \circ f = f$$

The first of these equations asserts that $[f \circ \varepsilon](x) = f(x)$ for every element x in A , which is quite obvious, since $[f \circ \varepsilon](x) = f(\varepsilon(x)) = f(x)$. The second equation is proved analogously.

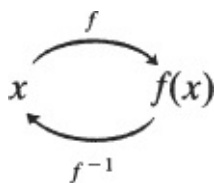
We saw in [Chapter 6](#) that the inverse of any bijective function exists and is a bijective function. Thus, *the inverse of any permutation of A is a permutation of A* . Furthermore, if f is any permutation of A and f^{-1} is its inverse, then

$$f^{-1} \circ f = \varepsilon \quad \text{and} \quad f \circ f^{-1} = \varepsilon$$

The first of these equations asserts that for any element x in A ,

$$[f^{-1} \circ f](x) = \varepsilon(x)$$

that is, $f^{-1}(f(x)) = x$:



This is obviously true, by the definition of the inverse of a function. The second equation is proved analogously.

Let us recapitulate: The operation \circ of composition of functions qualifies as an operation on the set of all the permutations of A . This operation is associative. There is a permutation ε such that $\varepsilon \circ f = f$ and f

$\circ \varepsilon = f$ for any permutation f of A . Finally, for every permutation f of A there is another permutation f^{-1} of A such that $f \circ f^{-1} = \varepsilon$ and $f^{-1} \circ f = \varepsilon$. Thus, *the set of all the permutations of A , with the operation \circ of composition, is a group.*

For any set A , the group of all the permutations of A is called the *symmetric group on A* , and it is represented by the symbol S_A . For any positive integer n , the symmetric group on the set $\{1, 2, 3, \dots, n\}$ is called the *symmetric group on n elements*, and is denoted by S_n .

Let us take a look at S_3 . First, we list all the permutations of the set $\{1, 2, 3\}$:

$$\begin{array}{lll} \varepsilon = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} & \alpha = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix} & \beta = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \\ \gamma = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} & \delta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} & \kappa = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \end{array}$$

This notation for functions was explained on page 57; for example,

$$\beta = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$$

is the function such that $\beta(1) = 3$, $\beta(2) = 1$, and $\beta(3) = 2$. A more graphic way of representing the same function would be

$$\beta = \begin{pmatrix} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 3 & 1 & 2 \end{pmatrix}$$

The operation on elements of S_3 is composition. To find $\alpha \circ \beta$, we note that

$$[\alpha \circ \beta](1) = \alpha(\beta(1)) = \alpha(3) = 2$$

$$[\alpha \circ \beta](2) = \alpha(\beta(2)) = \alpha(1) = 1$$

$$[\alpha \circ \beta](3) = \alpha(\beta(3)) = \alpha(2) = 3$$

Thus

$$\alpha \circ \beta = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix} = \gamma$$

Note that in $\alpha \circ \beta$, β is applied first and α next. A graphic way of representing this is

$$\begin{array}{c} \beta = \begin{pmatrix} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 3 & 1 & 2 \end{pmatrix} \\ \swarrow \quad \searrow \quad \swarrow \quad \searrow \\ \alpha = \begin{pmatrix} 1 & 2 & 3 \\ \downarrow & \downarrow & \downarrow \\ 1 & 3 & 2 \end{pmatrix} \end{array}$$

The other combinations of elements of S_3 may be computed in the same fashion. The student should

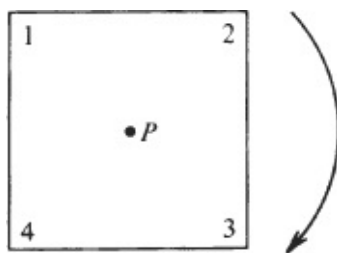
check the following table, which is the table of the group S_3 :

\circ	ε	α	β	γ	δ	κ
ε	ε	α	β	γ	δ	κ
α	α	ε	γ	β	κ	δ
β	β	κ	δ	α	ε	γ
γ	γ	δ	κ	ε	α	β
δ	δ	γ	ε	κ	β	α
κ	κ	β	α	δ	γ	ε

By a *group of permutations* we mean any group S_A or S_n , or any subgroup of one of these groups. Among the most interesting groups of permutations are the groups of symmetries of geometric figures. We will see how such groups arise by considering the *group of symmetries of the square*.

We may think of a symmetry of the square as any way of moving a square to make it coincide with its former position. Every time we do this, vertices will coincide with vertices, so a symmetry is completely described by its effect on the vertices.

Let us number the vertices as in the following diagram:



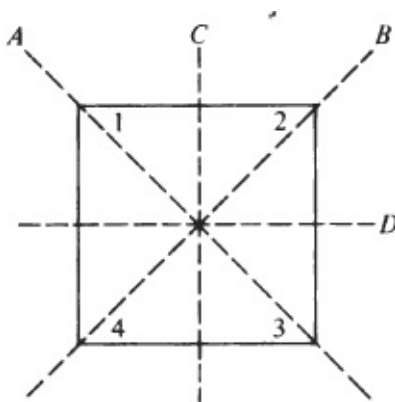
The most obvious symmetries are obtained by rotating the square clockwise about its center P , through angles of 90° , 180° , and 270° , respectively. We indicate each symmetry as a permutation of the vertices; thus a clockwise rotation of 90° yields the symmetry

$$R_1 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$$

for this rotation carries vertex 1 to 2, 2 to 3, 3 to 4, and 4 to 1. Rotations of 180° and 270° yield the following symmetries, respectively:

$$R_2 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \quad \text{and} \quad R_3 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix}$$

The remaining symmetries are flips of the square about its axes A , B , C , and D :



For example, when we flip the square about the axis A , vertices 1 and 3 stay put, but 2 and 4 change places; so we get the symmetry

$$R_4 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix}$$

In the same way, the other flips are

$$R_5 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} \quad R_6 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

and

$$R_7 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

One last symmetry is the *identity*

$$R_0 = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

which leaves the square as it was.

The operation on symmetries is composition: $R_i \circ R_j$ is the result of first performing R_j , and then R_i . For example, $R_1 \circ R_4$ is the result of first flipping the square about its axis A , then rotating it clockwise 90° :

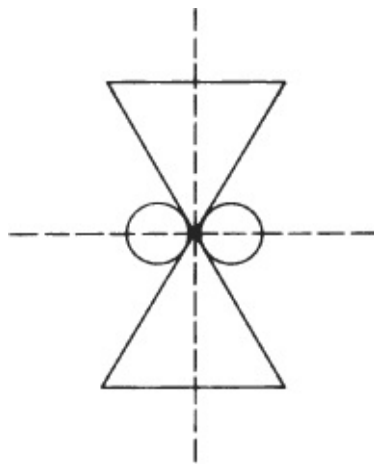
$$\begin{aligned} R_1 \circ R_4 &= \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} = R_6 \end{aligned}$$

Thus, the net effect is the same as if the square had been flipped about its axis C .

The eight symmetries of the square form a group under the operation \circ of composition, called the *group of symmetries of the square*.

For every positive integer $n \geq 3$, the regular polygon with n sides has a group of symmetries, symbolized by D_n , which may be found as we did here. These groups are called the *dihedral groups*. For example, the group of the square is D_4 , the group of the pentagon is D_5 , and so on.

Every plane figure which exhibits regularities has a group of symmetries. For example, the following figure, has a group of symmetries consisting of two rotations (180° and 360°) and two flips about the indicated axes. Artificial as well as natural objects often have a surprising number of symmetries.



Far more complicated than the plane symmetries are the symmetries of objects in space. Modern-day crystallography and crystal physics, for example, rely very heavily on knowledge about groups of symmetries of three-dimensional shapes.

Groups of symmetry are widely employed also in the theory of electron structure and of molecular vibrations. In elementary particle physics, such groups have been used to predict the existence of certain elementary particles before they were found experimentally!

Symmetries and their groups arise everywhere in nature: in quantum physics, flower petals, cell division, the work habits of bees in the hive, snowflakes, music, and Romanesque cathedrals.

EXERCISES

A. Computing Elements of S_6

1 Consider the following permutations f , g , and h in S_6 :

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 1 & 3 & 5 & 4 & 2 \end{pmatrix} \quad g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 1 & 6 & 5 & 4 \end{pmatrix}$$

$$h = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 6 & 4 & 5 & 2 \end{pmatrix}$$

Compute the following:

$$f^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} \quad g^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix}$$

$$h^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix}$$

$$f \circ g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} \quad g \circ f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix}$$

$$2 f \circ (g \circ h) =$$

$$3 g \circ h^{-1} =$$

$$4 \ h \circ g^{-1} \circ f^{-1} =$$

$$5 \ g \circ g \circ g =$$

B. Examples of Groups of Permutations

1 Let G be the subset of S_4 consisting of the permutations

$$\varepsilon = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} \quad f = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} \quad h = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

Show that G is a group of permutations, and write its table:

\circ	ε	f	g	h
ε				
f				
g				
h				

2 List the elements of the cyclic subgroup of S_6 generated by

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 1 & 6 & 5 \end{pmatrix}$$

3 Find a four-element abelian subgroup of S_5 . Write its table.

4 The subgroup of S_5 generated by

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \end{pmatrix} \quad g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix}$$

has six elements. List them, then write the table of this group:

$$\varepsilon = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 4 & 5 & 3 \end{pmatrix}$$

$$h = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

$$k = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

$$l = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

\circ	ε	f	g	h	k	l
ε						
f						
g						
h						
k						
l						

C. Groups of Permutations of \mathbb{R}

In each of the following, A is a subset of \mathbb{R} and G is a set of permutations of A . Show that G is a subgroup of S_A , and write the table of G .

- 1 A is the set of all $x \in \mathbb{R}$ such that $x \neq 0, 1$. $G = \{\varepsilon, f, g\}$, where $f(x) = 1/(1-x)$ and $g(x) = (x-1)/x$.
- 2 A is the set of all the nonzero real numbers. $G = \{\varepsilon, f, g, h\}$, where $f(x) = 1/x$, $g(x) = -x$, and $h(x) = -1/x$.
- 3 A is the set of all the real numbers $x \neq 0, 1$. $G = \{\varepsilon, f, g, h, k\}$, where $f(x) = 1-x$, $g(x) = 1/x$, $h(x) = 1/(1-x)$, $j(x) = (x-1)/x$, and $k(x) = x/(x-1)$.
- 4 A is the set of all the real numbers $x \neq 0, 1, 2$. G is the subgroup of S_A generated by $f(x) = 2-x$ and $g(x) = 2/x$. (G has eight elements. List them, and write the table of G .)

† D. A Cyclic Group of Permutations

For each integer n , define f_n by $f_n(x) = x + n$.

- 1 Prove: For each integer n , f_n is a permutation of \mathbb{R} , that is, $f_n \in S_{\mathbb{R}}$.
- 2 Prove that $f_n \circ f_m = f_{n+m}$ and $f_n^{-1} = f_{-n}$.
- 3 Let $G = \{f_n : n \in \mathbb{Z}\}$. Prove that G is a subgroup of $S_{\mathbb{R}}$.
- 4 Prove that G is cyclic. (Indicate a generator of G .)

† E. A Subgroup of $S_{\mathbb{R}}$

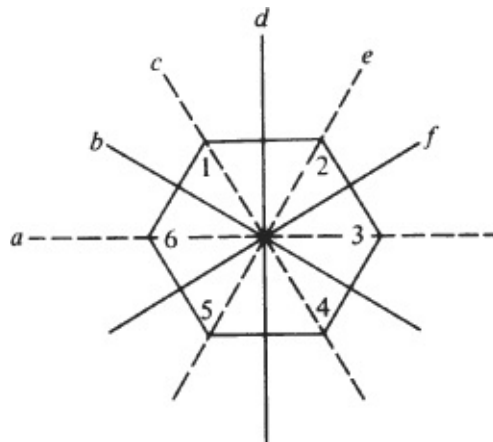
For any pair of real numbers $a \neq 0$ and b , define a function $f_{a,b}$ as follows:

$$f_{a,b}(x) = ax + b$$

- 1 Prove that $f_{a,b}$ is a permutation of \mathbb{R} , that is, $f_{a,b} \in S_{\mathbb{R}}$.
- 2 Prove that $f_{a,b} \circ f_{c,d} = f_{ac, ad+b}$.
- 3 Prove that $f_{a,b}^{-1} = f_{1/a, -b/a}$.
- 4 Let $G = \{f_{a,b} : a \in \mathbb{R}, b \in \mathbb{R}, a \neq 0\}$. Show that G is a subgroup of $S_{\mathbb{R}}$.

F. Symmetries of Geometric Figures

- 1 Let G be the group of symmetries of the regular hexagon. List the elements of G (there are 12 of them), then write the table of G .



$$R_0 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} \quad R_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 6 & 1 \end{pmatrix}$$

$$R_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 1 & 2 \end{pmatrix} \quad R_3 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 1 & 2 & 3 \end{pmatrix}$$

etc. . . .

2 Let G be the group of symmetries of the rectangle. List the elements of G (there are four of them), and write the table of G .

3 List the symmetries of the letter **Z** and give the table of this group of symmetries. Do the same for the letters **V** and **H**.

4 List the symmetries of the following shape, and give the table of their group.



(Assume that the three arms are of equal length, and the three central angles are equal.)

G. Symmetries of Polynomials

Consider the polynomial $p = (x_1 - x_2)^2 + (x_3 - x_4)^2$. It is unaltered when the subscripts undergo any of the following permutations:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

For example, the first of these permutations replaces p by

$$(x_2 - x_1)^2 + (x_3 - x_4)^2$$

the second permutation replaces p by $(x_1 - x_2)^2 + (x_4 - x_3)^2$; and so on. The *symmetries of a polynomial* p are all the permutations of the subscripts which leave p unchanged. They form a group of permutations.

List the symmetries of each of the following polynomials, and write their group table.

1 $p = x_1x_2 + x_2x_3$

2 $p = (x_1 - x_2)(x_2 - x_3)(x_1 - x_3)$

3 $p = x_1x_2 + x_2x_3 + x_1x_3$

4 $p = (x_1 - x_2)(x_3 - x_4)$

H. Properties of Permutations of a Set A

1 Let A be a set and $a \in A$. Let G be the subset of S_A consisting of all the permutations f of A such that $f(a) = a$. Prove that G is a subgroup of S_A .

2 If f is a permutation of A and $a \in A$, we say that f *moves* a if $f(a) \neq a$. Let A be an infinite set, and let

G be the subset of S_A which consists of all the permutations f of A which move *only a finite number of elements* of A . Prove that G is a subgroup of S_A .

3 Let A be a finite set, and B a subset of A . Let G be the subset of S_A consisting of all the permutations f of A such that $f(x) \in B$ for every $x \in B$. Prove that G is a subgroup of S_A .

4 Give an example to show that the conclusion of part 3 is not necessarily true if A is an infinite set.

I. Algebra of Kinship Structures (Anthropology)

Anthropologists have used groups of permutations to describe kinship systems in primitive societies. The algebraic model for kinship structures described here is adapted from *An Anatomy of Kinship* by H. C. White. The model is based on the following assumptions, which are widely supported by anthropological research:

- (i) The entire population of the society is divided into *clans*. Every person belongs to one, and only one, clan. Let us call the clans k_1, k_2, \dots, k_n .
- (ii) In every clan k_i , all the men must choose their wives from among the women of a specified clan k_j . We symbolize this by writing $w(k_i) = k_j$.
- (iii) Men from two different clans cannot marry women from the same clan. That is, if $k_i \neq k_j$, then $w(k_i) \neq w(k_j)$.
- (iv) All the children of a couple are assigned to some fixed clan. So if a man belongs to clan k_i , all his children belong to a clan which we symbolize by $c(k_i)$.
- (v) Children whose fathers belong to different clans must themselves be in different clans. That is, if $k_i \neq k_j$, then $c(k_i) \neq c(k_j)$.
- (vi) A man cannot marry a woman of his own clan. That is, $w(k_i) \neq k_i$.

Now let $K = \{k_1, k_2, \dots, k_n\}$ be the set of all the distinct clans. By (ii), w is a function from K to K , and by (iv), c is a function from K to K . By (iii), w is an injective function; hence (see [Exercise F2 of Chapter 6](#)) w is a permutation of K . Likewise, by (v), c is a permutation of K .

Let G be the group of permutations generated by c and w ; that is, G consists of c, w, c^{-1}, w^{-1} , and all possible composites which can be formed from these—for example, $c \circ w \circ w \circ c^{-1} \circ w^{-1}$. Clearly the identity function ε is in G since, for example, $\varepsilon = c \circ c^{-1}$. Here are two final assumptions:

- (vii) Every person, in any clan, has a relation in every other clan. This means that for any k_i and k_j in K , there is a permutation α in G such that $\alpha(k_i) = k_j$.
- (viii) Rules of kinship apply uniformly to all clans. Thus, for any α and β in G , if $\alpha(k_j) = \beta(k_j)$ for some specific clan k_j , it necessarily follows that $\alpha(k_i) = \beta(k_i)$ for every clan k_i .

Prove parts 1–3:

1 Let $\alpha \in G$. If $\alpha(k_i) = k_i$ for any given k_i , then $\alpha = \varepsilon$.

2 Let $\alpha \in G$. There is a positive integer $m \leq n$ such that $\alpha^m = \varepsilon$.

[$\alpha^m = \alpha \circ \alpha \circ \dots \circ \alpha$ (m factors of α). HINT: Consider $\alpha(k_1), \alpha^2(k_1)$, etc.]

3 The group G consists of exactly n permutations.

Explain parts 4–9.

4 If a person belongs to clan k_i , that person's father belongs to clan $c^{-1}(k_i)$. If a woman belongs to clan k_j ,

her husband belongs to clan $w^{-1}(k_j)$.

5 If any man is in the same clan as his son, then $c = \varepsilon$. If any woman is in the same clan as her son, then $c = w$.

6 If a person belongs to clan k_i , the son of his mother's sister belongs to clan $c \circ w^{-1} \circ w \circ c^{-1}(k_i)$. Conclude that marriage between matrilineal parallel cousins (marriage between a woman and the son of her mother's sister) is prohibited.

7 Marriage between a man and the daughter of his father's sister is prohibited.

8 If matrilineal cross-cousins may marry (that is, a woman may marry the son of her mother's brother), then $c \circ w = w^{-1} \circ c$.

9 If patrilineal cross-cousins may marry (a woman may marry the son of her father's sister), then c and w^{-1} commute.

Permutations of finite sets are used in every branch of mathematics—for example, in geometry, in statistics, in elementary algebra—and they have a myriad of applications in science and technology. Because of their practical importance, this chapter will be devoted to the study of a few special properties of permutations of finite sets.

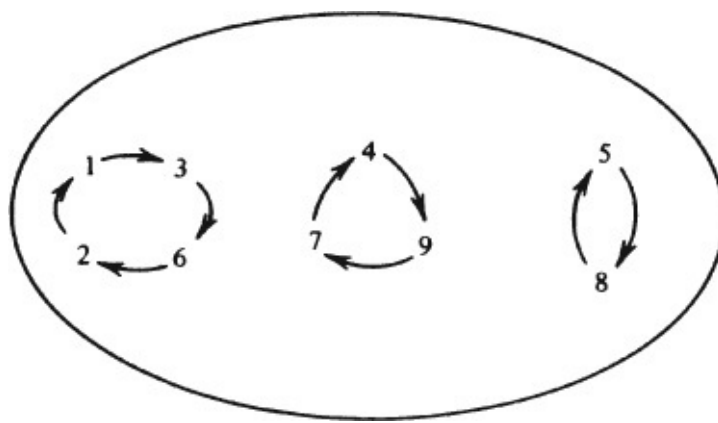
If n is a positive integer, consider a set of n elements. It makes no difference which specific set we consider, just as long as it has n elements; so let us take the set $\{1, 2, \dots, n\}$. We have already seen that the group of all the permutations of this set is called the *symmetric group on n elements* and is denoted by S_n . In the remainder of this chapter, when we say “permutation” we will invariably mean a permutation of the set $\{1, 2, \dots, n\}$ for an arbitrary positive integer n .

One of the most characteristic activities of science (*any* kind of science) is to try to separate complex things into their simplest component parts. This intellectual “divide and conquer” helps us to understand complicated processes and solve difficult problems. The savvy mathematician never misses the chance of doing this whenever the opportunity presents itself. We will see now that every permutation can be decomposed into simple parts called “cycles,” and these cycles are, in a sense, the most basic kind of permutations.

We begin with an example: take, for instance, the permutation

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 1 & 6 & 9 & 8 & 2 & 4 & 5 & 7 \end{pmatrix}$$

and look at how f moves the elements in its domain:



Notice how f decomposes its domain into three separate subsets, so that, in each subset, the elements are permuted cyclically so as to form a closed chain. These closed chains may be considered to be the component parts of the permutation; they are called “cycles.” (This word will be carefully defined in a moment.) Every permutation breaks down, just as this one did, into separate cycles.

Let a_1, a_2, \dots, a_s be distinct elements of the set $\{1, 2, \dots, n\}$. By the *cycle* $(a_1 a_2 \dots a_s)$ we mean the permutation

$$\underbrace{a_1 \rightarrow a_2 \rightarrow a_3 \rightarrow \dots \rightarrow a_{s-1} \rightarrow a_s}_{\text{cycle}}$$

of $\{1, 2, \dots, n\}$ which carries a_1 to a_2 , a_2 to a_3 , ..., a_{s-1} to a_s , and a_s to a_1 , while leaving all the remaining elements of $\{1, 2, \dots, n\}$ fixed.

For instance, in S_6 , the cycle (1426) is the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 6 & 3 & 2 & 5 & 1 \end{pmatrix}$$

In S_5 , the cycle (254) is the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 5 & 3 & 2 & 4 \end{pmatrix}$$

Because cycles are permutations, we may form the *composite* of two cycles in the usual manner. The composite of cycles is generally called their *product* and it is customary to omit the symbol \circ . For example, in S_5 ,

$$\begin{aligned} (245)(124) &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 3 & 5 & 2 \end{pmatrix} \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 1 & 5 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 5 & 3 & 1 & 2 \end{pmatrix} \end{aligned}$$

Actually, it is very easy to compute the product of two cycles by reasoning in the following manner: Let us continue with the same example,

$$\underbrace{(2 \ 4 \ 5)}_{\alpha} \underbrace{(1 \ 2 \ 4)}_{\beta}$$

Remember that the permutation on the right is applied first, and the permutation on the left is applied next.

Now,

β carries 1 to 2, and α carries 2 to 4; hence $\alpha\beta$ carries 1 to 4.

β carries 2 to 4, and α carries 4 to 5; hence $\alpha\beta$ carries 2 to 5.

β leaves 3 fixed and so does α ; hence $\alpha\beta$ leaves 3 fixed.

β carries 4 to 1 and α leaves 1 fixed, so $\alpha\beta$ carries 4 to 1.

β leaves 5 fixed and α carries 5 to 2; hence $\alpha\beta$ carries 5 to 2.

If $(a_1 a_2 \dots a_s)$ is a cycle, the integer s is called its *length*; thus, $(a_1 a_2 \dots a_s)$ is a *cycle of length s* . For example, (1532) is a cycle of length 4.

If two cycles have no elements in common they are said to be *disjoint*. For example, (132) and (465) are disjoint cycles, but (132) and (453) are not disjoint. *Disjoint cycles commute*: that is, if $(a_1 \dots a_r)$ and $(b_1 \dots b_s)$ are disjoint, then

$$\underbrace{(a_1 \dots a_r)}_{\alpha} \underbrace{(b_1 \dots b_s)}_{\beta} = \underbrace{(b_1 \dots b_s)}_{\beta} \underbrace{(a_1 \dots a_r)}_{\alpha}$$

It is easy to see why this is true: α moves the a 's but not the b 's, while β moves the b 's but not the a 's. Thus, if β carries b_i to b_j , then $\alpha\beta$ does the same, and so does $\beta\alpha$. Similarly, if α carries a_h to a_k then $\beta\alpha$ does the same, and so does $\alpha\beta$.

We are now ready to prove what was asserted at the beginning of this chapter: Every permutation can be decomposed into cycles—in fact, into *disjoint* cycles. More precisely, we can state the following:

Theorem 1 *Every permutation is either the identity, a single cycle, or a product of disjoint cycles.*

We begin with an example, because the proof uses the same technique as the example. Consider the permutation

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 2 & 1 & 6 \end{pmatrix}$$

and let us write f as a product of disjoint cycles. We begin with 1 and note that

$$1 \xrightarrow{f} 3 \xrightarrow{f} 5 \xrightarrow{f} 1$$

We have come a complete circle and found our first cycle, which is (135) . Next, we take the first number which hasn't yet been used, namely, 2. We see that

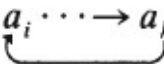
$$2 \xrightarrow{f} 4 \xrightarrow{f} 2$$

Again we have come a complete circle and found another cycle, which is (24) . The only remaining number is 6, which/leaves fixed. We are done:

$$f = (135)(24)$$

The proof for *any* permutation f follows the same pattern as the example. Let a_1 be the first number in $\{1, \dots, n\}$ such that $f(a_1) \neq a_1$. Let $a_2 = f(a_1)$, $a_3 = f(a_2)$, and so on in succession until we come to our first repetition, that is, until $f(a_k)$ is equal to one of the numbers a_1, a_2, \dots, a_{k-1} . Say $f(a_k) = a_i$. If a_i is not

a_1 , we have

$$a_1 \rightarrow a_2 \rightarrow \cdots \rightarrow a_{i-1} \rightarrow a_i \cdots \rightarrow a_k$$


so a_i is the image of two elements, a_k and a_{i-1} , which is impossible because f is bijective. Thus, $a_i = a_1$, and therefore $f(a_k) = a_1$. We have come a complete circle and found our first cycle, namely, $(a_1 a_2 \cdots a_k)$.

Next, let b_1 be the first number which has not yet been examined and such that $f(b_1) \neq b_1$. We let $b_2 = f(b_1)$, $b_3 = f(b_2)$, and proceed as before to obtain the next cycle, say $(b_1 \cdots b_t)$. Obviously $(b_1 \cdots b_t)$ is disjoint from $(a_1 \dots, a_k)$. We continue this process until all the numbers in $\{1, \dots, n\}$ have been exhausted. This concludes the proof.

Incidentally, it is easy to see that this product of cycles is unique, except for the order of the factors.

Now our curiosity may prod us to ask: once a permutation has been written as a product of disjoint cycles, *has it been simplified as much as possible?* Or is there some way of simplifying it further?

A cycle of length 2 is called a *transposition*. In other words, a transposition is a cycle (a_i, a_j) which interchanges the two numbers a_i and a_j . It is a fact both remarkable and trivial that every cycle can be expressed as a product of one or more transpositions. In fact,

$$(a_1 a_2 \dots a_r) = (a_r a_{r-1})(a_r a_{r-2}) \dots (a_r a_3)(a_r a_2 a_r a_1)$$

which may be verified by direct computation. For example,

$$(12345) = (54)(53)(52)(51)$$

However, there is more than one way to write a given permutation as a product of transpositions. For example, (12345) may also be expressed as a product of transpositions in the following ways:

$$(12345) = (15)(14)(13)(12)$$

$$(12345) = (54)(52)(51)(14)(32)(41)$$

as well as in many other ways.

Thus, every permutation, after it has been decomposed into disjoint cycles, may be broken down further and expressed as a product of transpositions. However, the expression as a product of transpositions is not unique, and even the *number of transpositions* involved is not unique.

Nevertheless, when a permutation π is written as a product of transpositions, *one* property of this expression is unique: the number of transpositions involved is either *always even* or *always odd*. (This fact will be proved in a moment.) For example, we have just seen that (12345) can be written as a product of four transpositions and also as a product of six transpositions; it can be written in many other ways, but always as a product of an *even* number of transpositions. Likewise, (1234) can be decomposed in many ways into transpositions, but always an *odd* number of transpositions.

A permutation is called *even* if it is a product of an even number of transpositions, and *odd* if it is a product of an odd number of transpositions. What we are asserting, therefore, is that every permutation is unambiguously either odd or even.

This may seem like a pretty useless fact—but actually the very opposite is true. A number of great theorems of mathematics depend for their proof (at that crucial step when the razor of logic makes its

decisive cut) on none other but the distinction between even and odd permutations.

We begin by showing that the identity permutation, ε , is an even permutation.

Theorem 2 *No matter how ε is written as a product of transpositions, the number of transpositions is even.*

PROOF: Let t_1, t_2, \dots, t_m be m transpositions, and suppose that

$$\varepsilon = t_1 t_2 \dots t_m \quad (1)$$

We aim to prove that ε can be rewritten as a product of $m - 2$ transpositions. We will then be done: for if ε were equal to a product of an odd number of transpositions, and we were able to rewrite this product repeatedly, each time with two fewer transpositions, then eventually we would get ε equal to a single transposition (ab) , and this is impossible.

Let x be any numeral appearing in one of the transpositions t_2, \dots, t_m . Let $t_k = (xa)$, and suppose t_k is the last transposition in Equation (1) (reading from left to right) in which x appears:

$$\varepsilon = t_1 t_2 \dots t_{k-1} \underbrace{t_k}_{=(xa)} \underbrace{t_{k+1} \dots t_m}_{x \text{ does not appear here}}$$

Now, t_{k-1} is a transposition which is either equal to (xa) , or else one or both of its components are different from x and a . This gives four possibilities, which we now treat as four separate cases.

Case I $t_{k-1} = (xa)$.

Then $t_{k-1} t_k = (xa)(xa)$, which is equal to the identity permutation. Thus, $t_{k-1} t_k$ may be removed without changing Equation (1). As a result, ε is a product of $m - 2$ transpositions, as required.

Case II $t_{k-1} = (xb)$ where $b \neq x, a$.

Then $t_{k-1} t_k = (xb)(xa)$

But $(xb)(xa) = (xa)(ab)$

We replace $t_{k-1} t_k$ by $(xa)(ab)$ in Equation (1). As a result, the last occurrence of x is one position further left than it was at the start.

Case III $t_{k-1} = (ca)$, where $c \neq x, a$.

Then $t_{k-1} t_k = (ca)(xa)$

But $(ca)(xa) = (xc)(ca)$

We replace $t_{k-1} t_k$ by $(xc)(ca)$ in Equation (1), as in Case II.

Case IV $t_{k-1} = (bc)$, where $b \neq x, a$ and $c \neq x, a$

Then $t_{k-1} t_k = (bc)(xa)$

But $(bc)(xa) = (xa)(bc)$

We replace $t_{k-1} t_k$ by $(xa)(bc)$ in Equation (1), as in Cases II and III.

In Case I, we are done. In Cases II, III, and IV, we repeat the argument one or more times. Each time, the last appearance of x is one position further left than the time before. This must eventually lead to Case I. For otherwise, we end up with the last (hence the only) appearance of x being in t_1 . This cannot be: for

if $t_1 = (xa)$ and x does not appear in t_2, \dots, t_m , then $\varepsilon(x) = a$, which is impossible! ■

(The box ■ is used to mark the ending of a proof.)

Our conclusion is contained in the next theorem.

Theorem 3 *If $\pi \in S_n$, then π cannot be both an odd permutation and an even permutation.*

Suppose π can be written as the product of an even number of transpositions, and differently as the product of an odd number of transpositions. Then the same would be true for π^{-1} . But $\varepsilon = \pi^\circ \pi^{-1}$: thus, writing π^{-1} as a product of an even number of transpositions and π as a product of an odd number of transpositions, we get an expression for ε as a product of an odd number of transpositions. This is impossible by [Theorem 2](#).

The set of all the even permutations in S_n is a subgroup of S_n . It is denoted by A_n , and is called the *alternating group* on the set $\{1, 2, \dots, n\}$.

EXERCISES

A. Practice in Multiplying and Factoring Permutations

1 Compute each of the following products in S_9 . (Write your answer as a single permutation.)

(a) $(145)(37)(682)$

(b) $(17)(628)(9354)$

(c) $(71825)(36)(49)$

(d) $(12)(347)$

(e) $(147)(1678)(74132)$

(f) $(6148)(2345)(12493)$

2 Write each of the following permutations in S_9 as a product of disjoint cycles:

(a) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 4 & 9 & 2 & 5 & 1 & 7 & 6 & 8 & 3 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 7 & 4 & 9 & 2 & 3 & 8 & 1 & 6 & 5 \end{pmatrix}$

(c) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 7 & 9 & 5 & 3 & 1 & 2 & 4 & 8 & 6 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 9 & 8 & 7 & 4 & 3 & 6 & 5 & 1 & 2 \end{pmatrix}$

3 Express each of the following as a product of transpositions in S_8 :

(a) (137428)

(b) $(416)(8235)$

(c) $(123)(456)(1574)$

(d) $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 1 & 4 & 2 & 8 & 7 & 6 & 5 \end{pmatrix}$

4 If $\alpha = (3714)$, $\beta = (123)$, and $\gamma = (24135)$ in S_7 , express each of the following as a product of disjoint cycles:

(a) $\alpha^{-1} \beta$

(b) $\gamma^{-1} \alpha$

(c) $\alpha^2 \beta$

(d) $\beta^2 \alpha \gamma$

- (e) γ^4
- # (f) $\gamma^3\alpha^{-1}$
- (g) $\beta^{-1}\gamma$
- (h) $\alpha^{-1}\gamma^2\alpha$

(NOTE: $\alpha^2 = \alpha \circ \alpha$, $\gamma^3 = \gamma \circ \gamma \circ \gamma$, etc.)

5 In S_5 , write (12345) in five different ways as a cycle, and in five different ways as a product of transpositions.

6 In S_5 , express each of the following as the square of a cycle (that is, express as α^2 where α is a cycle):

- (a) (132)
- (b) (12345)
- (c) (13)(24)

B. Powers of

If π is any permutation, we write $\pi \circ \pi = \pi^2$, $\pi \circ \pi \circ \pi = \pi^3$, etc. The convenience of this notation is evident.

1 Compute α^{-1} , α^2 , α^3 , α^4 , α^5 where

- (a) $\alpha = (123)$
- (b) $\alpha = (1234)$
- (c) $\alpha = (123456)$.

In the following problems, let α be a cycle of length s , say $\alpha = (\alpha_1\alpha_2 \dots \alpha_s)$.

2 Describe all the *distinct* powers of α . How many are there? Note carefully the connection with addition of integers modulo s (page 27).

3 Find the inverse of α , and show that $\alpha^{-1} = \alpha^s$

Prove each of the following:

4 α^2 is a cycle iff s is odd.

5 If s is odd, α is the square of some cycle of length α . (Find it. HINT: Show $\alpha = \alpha^{s+1}$.)

6 If s is even, say $s = 2t$, then α^2 is the product of two cycles of length t . (Find them.)

7 If s is a multiple of k , say $s = kt$, then α^k is the product of k cycles of length t .

8 If s is a prime number, every power of α is a cycle.

C. Even and Odd Permutations

1 Determine which of the following permutations is even, and which is odd.

- (a) $\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 7 & 4 & 1 & 5 & 6 & 2 & 3 & 8 \end{pmatrix}$
- (b) (71864)
- (c) (12)(76)(345)
- (d) (1276)(3241)(7812)
- (e) (123)(2345)(1357)

Prove each of the following:

2 (a) The product of two even permutations is even.

(b) The product of two odd permutations is even.

(c) The product of an even permutation and an odd permutation is odd.

3 (a) A cycle of length l is even if l is odd.

(b) A cycle of length l is odd if l is even.

4 (a) If α and β are cycles of length l and m , respectively, then $\alpha\beta$ is even or odd depending on whether $l + m - 2$ is even or odd.

(b) If $\pi = \beta_1 \dots \beta_r$ where each β_i is a cycle of length l_i , then π is even or odd depending on whether $l_1 + l_2 + \dots + l_r - r$ is even or odd.

D. Disjoint Cycles

In each of the following, let α and β be disjoint cycles, say

$$\alpha = (a_1 a_2 \dots a_s) \quad \text{and} \quad \beta = (b_1 b_2 \dots b_r)$$

Prove parts 1–3:

1 For every positive integer n , $(\alpha\beta)^n = \alpha^n \beta^n$.

2 If $\alpha\beta = \varepsilon$, then $\alpha = \varepsilon$ and $\beta = \varepsilon$.

3 If $(\alpha\beta)^t = \varepsilon$, then $\alpha^t = \varepsilon$ and $\beta^t = \varepsilon$ (where t is any positive integer). (Use part 2 in your proof.)

4 Find a transposition γ such that $\alpha\beta\gamma$ is a cycle.

5 Let γ be the same transposition as in the preceding exercise. Show that $\gamma\alpha\beta$ and $\gamma\alpha\beta$ are cycles.

6 Let α and β be cycles of odd length (not disjoint). Prove that if $\alpha^2 = \beta^2$, then $\alpha = \beta$.

† E. Conjugate Cycles

Prove each of the following in S_n :

1 Let $\alpha = (a_1, \dots, a_s)$ be a cycle and let π be a permutation in S_n . Then $\pi\alpha\pi^{-1}$ is the cycle $(\pi(a_1), \dots, \pi(a_s))$.

If α is any cycle and π any permutation, $\pi\alpha\pi^{-1}$ is called a *conjugate* of α . In the following parts, let π denote any permutation in S_n .

2 Conclude from part 1: Any two cycles of the same length are conjugates of each other.

3 If α and β are disjoint cycles, then $\pi\alpha\pi^{-1}$ and $\pi\beta\pi^{-1}$ are disjoint cycles.

4 Let σ be a product $\alpha_1 \dots \alpha_t$ of t disjoint cycles of lengths l_1, \dots, l_t , respectively. Then $\pi\sigma\pi^{-1}$ is also a product of t disjoint cycles of lengths l_1, \dots, l_t .

5 Let α_1 and α_2 be cycles of the same length. Let β_1 and β_2 be cycles of the same length. Let α_1 and β_1 be disjoint, and let α_2 and β_2 be disjoint. There is a permutation $\pi \in S_n$ such that $\alpha_1\beta_1 = \pi\alpha_2\beta_2\pi^{-1}$.

† F. Order of Cycles

1 Prove in S_n : If $\alpha = (a_1 \dots a_s)$ is a cycle of length s , then $\alpha^s = \varepsilon$, $\alpha^{2s} = \varepsilon$, and $\alpha^{3s} = \varepsilon$. Is $\alpha^k = \varepsilon$ for any positive integer $k < s$? (Explain.)

If α is any permutation, the least positive integer n such that $\alpha^n = \varepsilon$ is called the *order* of α .

- 2** Prove in S_n : If $\alpha = (\alpha_1 \dots \alpha_s)$ is any cycle of length s , the order of α is s .
- 3** Find the order of each of the following permutations:
- (a) $(12)(345)$
 - (b) $(12)(3456)$
 - (c) $(1234)(56789)$
- 4** What is the order of $\alpha\beta$, if α and β are disjoint cycles of lengths 4 and 6, respectively? (Explain why. Use the fact that disjoint cycles commute.)
- 5** What is the order of $\alpha\beta$ if α and β are disjoint cycles of lengths r and s , respectively? (Venture a guess, explain, but do not attempt a rigorous proof.)

† G. Even/Odd Permutations in Subgroups of S_n

Prove each of the following in S_n :

- 1** Let $\alpha_1, \dots, \alpha_r$ be distinct even permutations, and β an odd permutation. Then $\alpha_1\beta, \dots, \alpha_r\beta$ are r *distinct* odd permutations. (See [Exercise C2](#).)
- 2** If β_1, \dots, β_r are distinct odd permutations, then $\beta_1\beta_1, \beta_1\beta_2, \dots, \beta_1\beta_r$ are r *distinct* even permutations.
- 3** In S_n , there are the same number of odd permutations as even permutations. (HINT: Use part 1 to prove that the number of even permutations \leq is the number of odd permutations. Use part 2 to prove the reverse of that inequality.)
- 4** The set of all the even permutations is a subgroup of S_n . (It is denoted by A_n and is called the *alternating group* on n symbols.)
- 5** Let H be any subgroup of S_n . H either contains only even permutations, or H contains the same number of odd as even permutations. (Use parts 1 and 2.)

† H. Generators of A_n and S_n

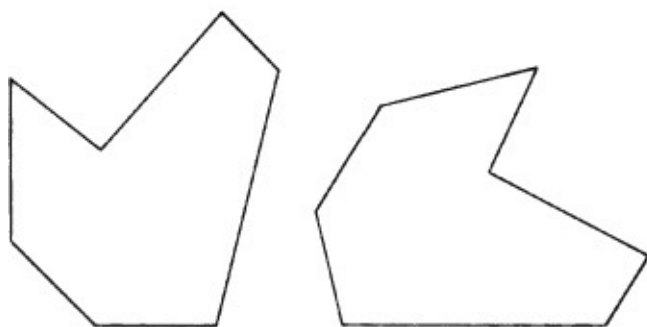
Remember that in any group G , a set S of elements of G is said to *generate* G if every element of G can be expressed as a product of elements in S and inverses of elements in S . (See page 47.)

- 1** Prove that the set T of all the transpositions in S_n generates S_n .
- # **2** Prove that the set $T_1 = \{(12), (13), \dots, (1n)\}$ generates S_n .
- 3** Prove that every even permutation is a product of one or more cycles of length 3. [HINT: $(13)(12) = (123)$; $(12)(34) = (321)(134)$.] Conclude that the set U of all cycles of length 3 generates A_n .
- 4** Prove that the set $U_1 = \{(123), (124), \dots, (12n)\}$ generates A_n . [HINT: $(abc) = (1ca)(1ab)$, $(1ab) = (1b2)(12a)(12b)$, and $(1b2) = (12b)^2$.]
- 5** The pair of cycles (12) and $(12 \dots n)$ generates S_n . [HINT: $(1 \dots n)(12)(1 \dots n)^{-1} = (23)$; $(12)(23)(12) = (13)$.]

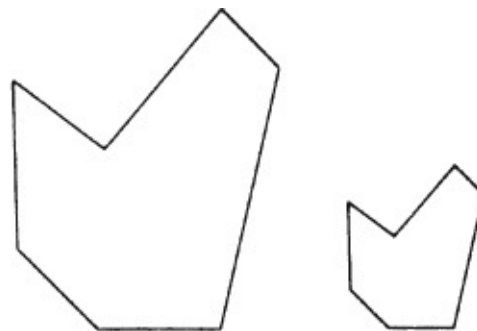
Human perception, as well as the “perception” of so-called intelligent machines, is based on the ability to recognize the same structure in different guises. It is the faculty for discerning, in *different* objects, the same relationships between their parts.

The dictionary tells us that two things are “isomorphic” if they *have the same structure*. The notion of isomorphism—of having the same structure—is central to every branch of mathematics and permeates all of abstract reasoning. It is an expression of the simple fact that objects may be different in substance but identical in form.

In geometry there are several kinds of isomorphism, the simplest being congruence and similarity. Two geometric figures are congruent if there exists a plane motion which makes one figure coincide with the other; they are similar if there exists a transformation of the plane, magnifying or shrinking lengths in a fixed ratio, which (again) makes one figure coincide with the other.



These two figures are congruent



These two figures are similar

We do not even need to venture into mathematics to meet some simple examples of isomorphism. For instance, the two palindromes

M A D A M
A
M A D

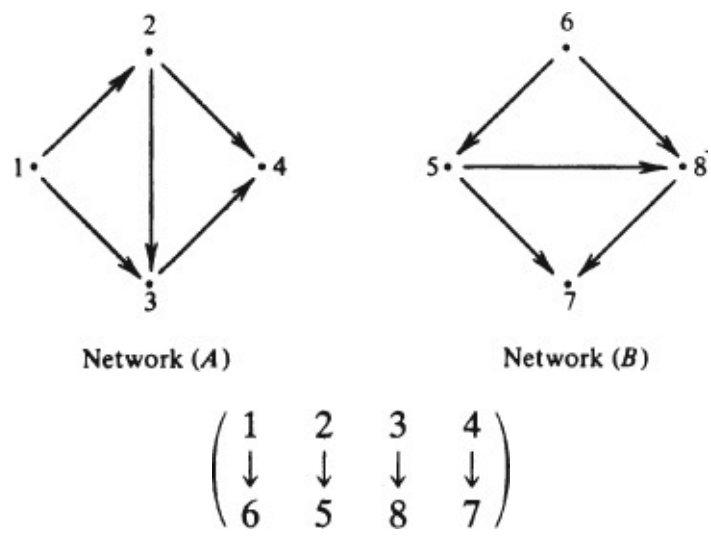
and

R O T O R
O
R O T

are different, but obviously isomorphic; indeed, the first one coincides with the second if we replace M

by R, A by O, and D by T.

Here is an example from applied mathematics: A flow network is a set of points, with arrows joining some of the points. Such networks are used to represent flows of cash or goods, channels of communication, electric circuits, and so on. The flow networks (A) and (B), below, are different, but can be shown to be isomorphic. Indeed, (A) can be made to coincide with (B) if we superimpose point 1 on point 6, point 2 on point 5, point 3 on point 8, and point 4 on point 7. (A) and (B) then coincide in the sense of having the same points joined by arrows in the same direction. Thus, *network (A) is transformed into network (B)* if we replace points 1 by 6, 2 by 5, 3 by 8, and 4 by 7. The one-to-one correspondence which carries out this transformation, namely,



is called an *isomorphism* from network (A) to network (B), for it transforms (A) into (B).
Incidentally, the one-to-one correspondence

$$\left(\begin{array}{ccc} \mathbf{M} & \mathbf{A} & \mathbf{D} \\ \downarrow & \downarrow & \downarrow \\ \mathbf{R} & \mathbf{O} & \mathbf{T} \end{array} \right)$$

is an *isomorphism* between the two palindromes of the preceding example, for it transforms the first palindrome into the second.

Our next and final example is from algebra. Consider the two groups G_1 and G_2 described below:

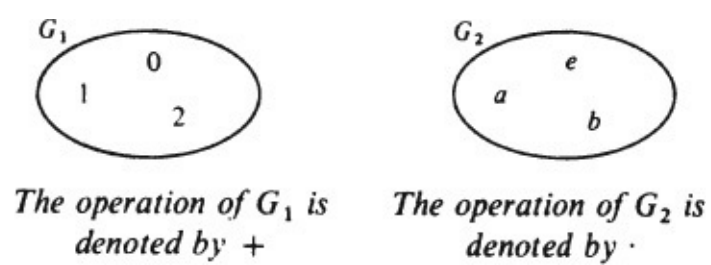


Table of G_1

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

Table of G_2

\cdot	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

G_1 and G_2 are different, but isomorphic. Indeed, if in G_1 we replace 0 by e , 1 by a , and 2 by b , then G_1 coincides with G_2 , the table of G_1 being transformed into the table of G_2 . In other words, the one-to-one correspondence

$$\begin{pmatrix} 0 & 1 & 2 \\ \downarrow & \downarrow & \downarrow \\ e & a & b \end{pmatrix}$$

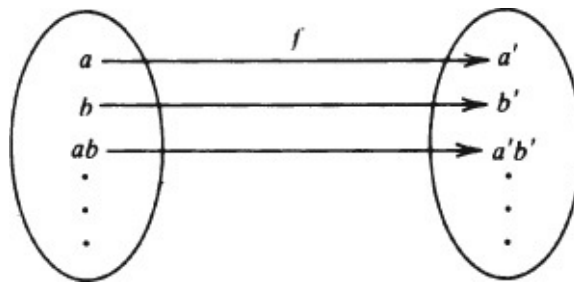
transforms G_1 to G_2 . It is called an *isomorphism* from G_1 to G_2 . Finally, because there exists an isomorphism from G_1 to G_2 , G_1 and G_2 are isomorphic to each other.

In general, by an isomorphism between two groups we mean a one-to-one correspondence between them which transforms one of the groups into the other. If there *exists* an isomorphism from one of the groups to the other, we say they are isomorphic. Let us be more specific:

If G_1 and G_2 are any groups, an *isomorphism* from G_1 to G_2 is a one-to-one correspondence f from G_1 to G_2 with the following property: For every pair of elements a and b in G_1 ,

$$\text{If } f(a) = a' \text{ and } f(b) = b' \text{ then } f(ab) = a'b' \quad (1)$$

In other words, if f matches a with a' and b with b' it *must* match ab with $a'b'$.



It is easy to see that if f has this property it transforms the table of G_1 into the table of G_2 :

G_1	b		G_2	b'
	\vdots			\vdots
a	$\dots ab$	$\xrightarrow[\text{replace } x \text{ by } f(x)]{\text{For every } x}$	a'	$\dots a'b'$

There is another, equivalent way of looking at this situation: If two groups G_1 and G_2 are isomorphic, we can say the two groups are actually the same, except that the elements of G_1 have *different names* from the elements of G_2 . G_1 becomes exactly G_2 if we *rename* its elements. The function which does the renaming is an isomorphism from G_1 to G_2 . Thus, in our last example, if 0 is renamed e , 1 is renamed a ,

and 2 is renamed 6, G_1 becomes exactly G_2 , with the same table. (Note that we have also renamed the operation: it was called $+$ in G_1 and \cdot in G_2 .)

By the way, property (1) may be written more concisely as follows:

$$f(ab) = f(a)f(b) \quad (2)$$

So we may sum up our definition of isomorphism in the following way:

Definition Let G_1 and G_2 be groups. A bijective function $f : G_1 \rightarrow G_2$ with the property that for any two elements a and b in G_1 ,

$$f(ab) = f(a)f(b) \quad (2)$$

is called an **isomorphism** from G_1 to G_2 .

If there exists an isomorphism from G_1 to G_2 , we say that G_1 is **isomorphic** to G_2 .

If there exists an isomorphism f from G_1 to G_2 , in other words, if G_1 is isomorphic to G_2 , we symbolize this fact by writing

$$G_1 \cong G_2$$

to be read, “ G_1 is isomorphic to G_2 .”

How does one recognize if two groups are isomorphic? This is an important question, and not quite so easy to answer as it may appear. There is no way of spontaneously recognizing whether two groups G_1 and G_2 are isomorphic. Rather, the groups must be carefully tested according to the above definition.

G_1 and G_2 are isomorphic if *there exists* an isomorphism from G_1 to G_2 . Therefore, the burden of proof is upon us to *find* an isomorphism from G_1 to G_2 , and show that it *is* an isomorphism. In other words, we must go through the following steps:

1. Make an educated guess, and come up with a function $f : G_1 \rightarrow G_2$ which looks as though it might be an isomorphism.
2. Check that f is *injective* and *surjective* (hence bijective).
3. Check that f satisfies the identity

$$f(ab) = f(a)f(b)$$

Here's an example: \mathbb{R} is the group of the real numbers with the operation of *addition*. \mathbb{R}^{pos} is the group of the *positive* real numbers with the operation of *multiplication*. It is an interesting fact that \mathbb{R} and \mathbb{R}^{pos} are isomorphic. To see this, let us go through the steps outlined above:

1. The educated guess: The exponential function $f(x) = e^x$ is a function from \mathbb{R} to \mathbb{R}^{pos} which, if we recall its properties, might do the trick.
2. f is injective: Indeed, if $f(a) = f(b)$, that is, $e^a = e^b$, then, taking the natural log on both sides, we get $a = b$.

f is surjective: Indeed, if $y \in \mathbb{R}^{\text{pos}}$, that is, if y is any positive real number, then $y = e^{\ln y} = f(\ln y)$; thus, $y = f(x)$ for $x = \ln y$.

3. It is well known that $e^{a+b} = e^a \cdot e^b$, that is,

$$f(a + b) = f(a) \cdot f(b)$$

Incidentally, note carefully that the operation of \mathbb{R} is $+$, whereas the operation of \mathbb{R}^{pos} is \cdot . That is the reason we have to use $+$ on the left side of the preceding equation, and \cdot on the right side of the equation.

How does one recognize when two groups are not isomorphic? In practice it is usually easier to show that two groups are *not* isomorphic than to show they *are*. Remember that if two groups are isomorphic they are replicas of each other; their elements (and their operation) may be named differently, but in all other respects they are the same and share the same properties. Thus, if a group G_1 has a property which group G_2 does not have (or vice versa), they are not isomorphic! Here are some examples of properties to look out for:

1. Perhaps G_1 is commutative, and G_2 is not.
2. Perhaps G_1 has an element which is its own inverse, and G_2 does not.
3. Perhaps G_1 is generated by two elements, whereas G_2 is not generated by any choice of two of its elements.
4. Perhaps every element of G_1 is the square of an element of G_1 , whereas G_2 does not have this property.

This list is by no means exhaustive; it merely illustrates the kind of things to be on the lookout for. Incidentally, the kind of properties to watch for are properties which do not depend merely on the *names* assigned to individual elements; for instance, in our last example, $0 \in G_1$ and $0 \notin G_2$, but nevertheless G_1 and G_2 are isomorphic.

Finally, let us state the obvious: if G_1 and G_2 cannot be put in one-to-one correspondence (say, G_1 has more elements than G_2), clearly they cannot be isomorphic.

In the early days of modern algebra the word “group” had a different meaning from the meaning it has today. In those days a group always meant a *group of permutations*. The only groups mathematicians used were groups whose elements were permutations of some fixed set and whose operation was composition.

There is something comforting about working with tangible, concrete things, such as groups of permutations of a set. At all times we have a clear picture of what it is we are working with. Later, as the axiomatic method reshaped algebra, a group came to mean *any* set with *any* associative operation having a neutral element and allowing each element an inverse. The new notion of group pleases mathematicians because it is simpler and more lean and sparing than the old notion of groups of permutations; it is also more general because it allows many new things to be groups which are not groups of permutations. However, it is harder to visualize, precisely because so many different things can be groups.

It was therefore a great revelation when, about 100 years ago, Arthur Cayley discovered that *every group is isomorphic to a group of permutations*. Roughly, this means that the groups of permutations are actually all the groups there are! Every group is (or is a carbon copy of) a group of permutations. This great result is a classic theorem of modern algebra. As a bonanza, its proof is not very difficult.

Cayley’s Theorem *Every group is isomorphic to a group of permutations.*

PROOF: Let G be a group; we wish to show that G is isomorphic to a group of permutations. The first question to ask is, “*What* group of permutations? Permutations of *what* set?” (After all, every permutation

must be a permutation of some fixed set.) Well, the one set we have at hand is the set G , so we had better fix our attention on *permutations of G* . The way we match up elements of G with permutations of G is quite interesting:

With each element a in G we associate a function $\pi_a : G \rightarrow G$ defined by

$$\pi_a(x) = ax$$

In other words, π_a is the function whose rule may be described by the words “multiply on the left by a ,” We will now show that π_a is a permutation of G :

1. π_a is *injective*: Indeed, if $\pi_a(x_1) = \pi_a(x_2)$, then $ax_1 = ax_2$, so by the cancellation law, $x_1 = x_2$.
2. π_a is *surjective*: For if $y \in G$, then $y = a(a^{-1}y) = \pi_a(a^{-1}y)$. Thus, each y in G is equal to $\pi_a(x)$ for $x = a^{-1}y$.
3. Since π_a is an injective and surjective function from G to G , π_a is a *permutation of G* .

Let us remember that we have a permutation π_a for *each* element a in G ; for example, if b and c are other elements in G , π_b is the permutation “multiply on the left by b ,” π_c is the permutation “multiply on the left by c ,” and so on. In general, let G^* denote the set of *all* the permutations π_a as a ranges over all the elements of G :

$$G^* = \{\pi_a : a \in G\}$$

Observe now that G^* is a set consisting of permutations of G —but not necessarily *all* the permutations of G . In [Chapter 7](#) we used the symbol S_G to designate the group of *all* the permutations of G . We must show now that G^* is a *subgroup* of S_G , for that will prove that G^* is a *group of permutations*.

To prove that G^* is a subgroup of S_G , we must show that G^* is closed with respect to composition, and closed with respect to inverses. That is, we must show that if π_a and π_b are any elements of G^* , their composite $\pi_a \circ \pi_b$ is also in G^* ; and if π_a is any element of G^* , its inverse is in G^* .

First, we claim that if a and b are any elements of G , then

$$\pi_a \circ \pi_b = \pi_{ab} \quad (3)$$

To show that $\pi_a \circ \pi_b$ and π_{ab} are the same, we must show that they have the same effect on every element x : that is, we must prove the identity $[\pi_a \circ \pi_b](x) = \pi_{ab}(x)$. Well, $[\pi_a \circ \pi_b](x) = \pi_a(\pi_b(x)) = \pi_a(bx) = a(bx) = (ab)x = \pi_{ab}(x)$. Thus, $\pi_a \circ \pi_b = \pi_{ab}$; this proves that the composite of any two members π_a and π_b of G^* is another member π_{ab} of G^* . Thus, G^* is *closed with respect to composition*.

It is each to see that π_e is the identity function: indeed,

$$\pi_e(x) = ex = x$$

In other words, π_e is the identity element of S_G .

Finally, by [Equation \(3\)](#),

$$\pi_a \circ \pi_a - 1 = \pi_{aa} - 1 = \pi_e$$

So by [Theorem 2](#) of [Chapter 4](#), the inverse of π_a is $\pi_{a^{-1}}$. This proves that the inverse of any member π_a of G^* is another member $\pi_{a^{-1}}$ of G^* . Thus, G^* is closed with respect to inverses.

Since G^* is closed with respect to composition and inverses, G^* is a subgroup of S_G .

We are now in the final lap of our proof. We have a group of permutations G^* , and it remains only to show that G is isomorphic to G^* . To do this, we must find an isomorphism $f: G \rightarrow G^*$. Let f be the function

$$f(a) = \pi_a$$

In other words, f matches each element a in G with the permutation π_a in G^* . We can quickly show that f is an isomorphism:

1. f is injective: Indeed, if $f(a) = f(b)$ then $\pi_a = \pi_b$. Thus, $\pi_a(e) = \pi_b(e)$, that is, $ae = be$, so, finally, $a = b$.
2. f is surjective: Indeed, every element of G^* is some π_a , and $\pi_a = f(a)$.
3. Lastly, $f(ab) = \pi_{ab} = \pi_a \circ \pi_b = f(a) \circ f(b)$.

Thus, f is an isomorphism, and so $G \cong G^*$. ■

EXERCISES

A. Isomorphism Is an Equivalence Relation among Groups

The following three facts about isomorphism are true for all groups:

- (i) Every group is isomorphic to itself.
- (ii) If $G_1 \cong G_2$, then $G_2 \cong G_1$.
- (iii) If $G_1 \cong G_2$ and $G_2 \cong G_3$, then $G_1 \cong G_3$.

Fact (i) asserts that for any group G , there exists an isomorphism from G to G .

Fact (ii) asserts that, if there is an isomorphism f from G_1 to G_2 , there must be some isomorphism from G_2 to G_1 . Well, the inverse of f is such an isomorphism.

Fact (iii) asserts that, if there are isomorphisms $f: G_1 \rightarrow G_2$ and $g: G_2 \rightarrow G_3$, there must be an isomorphism from G_1 to G_3 . One can easily guess that $g \circ f$ is such an isomorphism. The details of facts (i), (ii), and (iii) are left as exercises.

1 Let G be any group. If $\varepsilon: G \rightarrow G$ is the identity function, $\varepsilon(x) = x$, show that ε is an isomorphism.

2 Let G_1 and G_2 be groups, and $f: G_1 \rightarrow G_2$ an isomorphism. Show that $f^{-1}: G_2 \rightarrow G_1$ is an isomorphism. [HINT: Review the discussion of inverse functions at the end of [Chapter 6](#). Then, for arbitrary elements $c, d \in G_2$, there exist $a, b \in G_1$, such that $c = f(a)$ and $d = f(b)$. Note that $a = f^{-1}(c)$ and $b = f^{-1}(d)$. Show that $f^{-1}(cd) = f^{-1}(c)f^{-1}(d)$.]

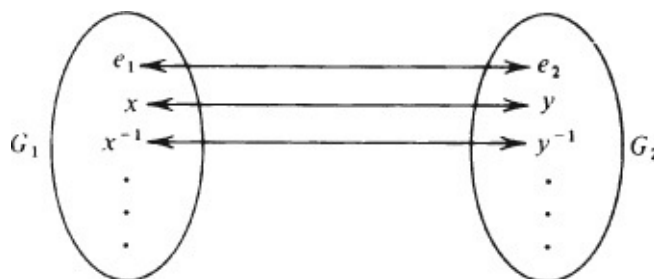
3 Let G_1, G_2 , and G_3 be groups, and let $f: G_1 \rightarrow G_2$ and $g: G_2 \rightarrow G_3$ be isomorphisms. Prove that $g \circ f$:

$G_1 \rightarrow G_3$ is an isomorphism.

B. Elements Which Correspond under an Isomorphism

Recall that an isomorphism f from G_1 to G_2 is a one-to-one correspondence between G_1 and G_2 satisfying $f(ab) = f(a)f(b)$. f matches every element of G_1 with a corresponding element of G_2 . It is important to note that:

- (i) f matches the neutral element of G_1 with the neutral element of G_2 .
- (ii) If f matches an element x in G_1 with y in G_2 , then, necessarily, f matches x^{-1} with y^{-1} . That is, if $x \leftrightarrow y$, then $x^{-1} \leftrightarrow y^{-1}$.
- (iii) f matches a generator of G_1 with a generator of G_2 .



The details of these statements are now left as an exercise. Let G_1 and G_2 be groups, and let $f: G_1 \rightarrow G_2$ be an isomorphism.

- 1 If e_1 denotes the neutral element of G_1 and e_2 denotes the neutral element of G_2 , prove that $f(e_1) = e_2$. [HINT: In any group, there is exactly one neutral element; show that $f(e_1)$ is the neutral element of G_2 .]
- 2 Prove that for each element a in G_1 $f(a^{-1}) = [f(a)]^{-1}$. (HINT: You may use [Theorem 2](#) of [Chapter 4](#).)
- 3 If G_1 is a cyclic group with generator a , prove that G_2 is also a cyclic group, with generator $f(a)$.

C. Isomorphism of Some Finite Groups

In each of the following, G and H are finite groups. Determine whether or not $G \cong H$. Prove your answer in either case.

To find an isomorphism from G to H will require a little ingenuity. For example, if G and H are cyclic groups, it is clear that we must match a generator a of G with a generator b of H ; that is, $f(a) = b$. Then $f(aa) = bb$, $f(aaa) = bbb$, and so on. If G and H are not cyclic, we have other ways: for example, if G has an element which is its own inverse, it must be matched with an element of H having the same property. Often, the specifics of a problem will suggest an isomorphism, if we keep our eyes open.

To prove that a specific one-to-one correspondence $f: G \rightarrow H$ is an isomorphism, we may check that it transforms the table of G into the table of H .

- # 1 G is the checkerboard game group of [Chapter 3](#), [Exercise D](#). H is the group of the complex numbers $\{i, -i, 1, -1\}$ under multiplication.
- 2 G is the same as in part 1. $H = \mathbf{Z}_4$.
- 3 G is the group P_2 of subsets of a two-element set. (See [Chapter 3](#), [Exercise C](#).) H is as in part 1.
- # 4 G is S_3 , H is the group of matrices described on page 28 of the text.

5 G is the coin game group of Chapter 3, Exercise E. H is D_4 , the group of symmetries of the square.

6 G is the group of symmetries of the rectangle. H is as in part 1.

D. Separating Groups into Isomorphism Classes

Each of the following is a set of four groups. In each set, determine which groups are isomorphic to which others. Prove your answers, and use Exercise A3 where convenient.

1 \mathbb{Z}_4 $\mathbb{Z}_2 \times \mathbb{Z}_2$ P_2 V

[P_2 denotes the group of subsets of a two-element set. (See Chapter 3, Exercise C.) V denotes the group of the four complex numbers $\{i, -i, 1, -1\}$ with respect to multiplication.]

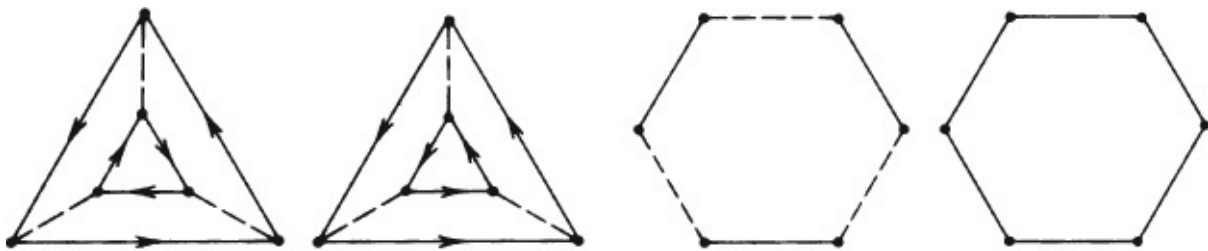
2 S_3 \mathbb{Z}_6 $\mathbb{Z}_3 \times \mathbb{Z}_2$ \mathbb{Z}_7^*

(\mathbb{Z}_7^* denotes the group $\{1, 2, 3, 4, 5, 6\}$ with multiplication modulo 7. The product modulo 7 of a and b is the remainder of ab after division by 7.)

3 \mathbb{Z}_8 P_3 $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ D_4

(D_4 is the group of symmetries of the square.)

4 The groups having the following Cayley diagrams:



E. Isomorphism of Infinite Groups

1 Let E designate the group of all the even integers, with respect to addition. Prove that $\mathbb{Z} \cong E$.

2 Let G be the group $\{10^n : n \in \mathbb{Z}\}$ with respect to multiplication. Prove that $G \cong \mathbb{Z}$. (Remember that the operation of \mathbb{Z} is addition.)

3 Prove that $\mathbb{C} = \mathbb{R} \times \mathbb{R}$.

4 We have seen in the text that \mathbb{R} is isomorphic to \mathbb{R}^{pos} . Prove that \mathbb{R} is *not* isomorphic to \mathbb{R}^* (the multiplicative group of the nonzero real numbers). (HINT: Consider the properties of the number -1 in \mathbb{R}^* . Does \mathbb{R} have *any* element with those properties?)

5 Prove that \mathbb{Z} is not isomorphic to \mathbb{Q} .

6 We have seen that $\mathbb{R} \cong \mathbb{R}^{\text{pos}}$. However, prove that \mathbb{Q} is *not* isomorphic to \mathbb{Q}^{pos} . (\mathbb{Q}^{pos} is the multiplicative group of positive rational numbers.)

F. Isomorphism of Groups Given by Generators and Defining Equations

If a group G is generated, say, by a , b , and c , then a set of equations involving a , b , and c is called a set of *defining equations* for G if these equations completely determine the table of G . (See end of Chapter 5.) If G' is another group, generated by elements a' , b' , and c' satisfying the same defining equations as a , b , and c , then G' has the same table as G (because the tables of G and G' are completely determined by the defining equations, which are the same for G as for G').

Consequently, if we know generators and defining equations for two groups G and G' , and if we are

able to match the generators of G with those of G' so that the defining equations are the same, we may conclude that $G \cong G'$.

Prove that the following pairs of groups G, G' are isomorphic.

- # 1 G = the subgroup of S_4 generated by (24) and (1234); $G' = \{e, a, b, b^2, b^3, ab, ab^2, ab^3\}$ where $a^2 = e$, $b^4 = e$, and $ba = ab^3$.
- 2 $G = S_3$; $G' = \{e, a, b, ab, aba, abab\}$ where $a^2 = e$, $b^2 = e$, and $bab = aba$.
- 3 $G = D_4$; $G' = \{e, a, b, ab, aba, (ab)^2, ba, bab\}$ where $a^2 = b^2 = e$ and $(ab)^4 = e$.
- 4 $G = \mathbf{Z}_2 \times \mathbf{Z}_2 \times \mathbf{Z}_2$; $G' = \{e, a, b, c, ab, ac, bc, abc\}$ where $a^2 = b^2 = c^2 = e$ and $(ab)^2 = (bc)^2 = (ac)^2 = e$.

G. Isomorphic Groups on the Set \mathbb{R}

- 1 G is the set $\{x \in \mathbb{R} : x \neq -1\}$ with the operation $x * y = x + y + xy$. Show that $f(x) = x - 1$ is an isomorphism from \mathbb{R}^* to G . Thus, $\mathbb{R}^* \cong G$.
- 2 G is the set of the real numbers with the operation $x * y = x + y + 1$. Find an isomorphism $f: \mathbb{R} \rightarrow G$ and show that it is an isomorphism.
- 3 G is the set of the nonzero real numbers with the operation $x * y = xy/2$. Find an isomorphism from \mathbb{R}^* to G .
- 4 Show that $f(x, y) = (-1)^y x$ is an isomorphism from $\mathbb{R}^{\text{pos}} \times \mathbf{Z}_2$ to \mathbb{R}^* . (REMARK: To combine elements of $\mathbb{R}^{\text{pos}} \times \mathbf{Z}_2$, one multiplies first components, adds second components.) Conclude that $\mathbb{R}^* \cong \mathbb{R}^{\text{pos}} \times \mathbf{Z}_2$.

H. Some General Properties of Isomorphism

- 1 Let G and H be groups. Prove that $G \times H \cong H \times G$.
- # 2 If $G_1 \cong G_2$ and $H_1 \cong H_2$, then $G_1 \times H_1 \cong G_2 \times H_2$.
- 3 Let G be any group. Prove that G is abelian iff the function $f(x) = x^{-1}$ is an isomorphism from G to G .
- 4 Let G be any group, with its operation denoted multiplicatively. Let H be a group with the same set as G and let its operation be defined by $x * y = y \cdot x$ (where \cdot is the operation of G). Prove that $G \cong H$.
- 5 Let c be a fixed element of G . Let H be a group with the same set as G , and with the operation $x * y = xcy$. Prove that the function $f(x) = c^{-1}x$ is an isomorphism from G to H .

I. Group Automorphisms

If G is a group, an *automorphism* of G is an isomorphism from G to G . We have seen (Exercise A1) that the identity function $\varepsilon(x) = x$ is an automorphism of G . However, many groups have *other* automorphisms besides this obvious one.

- 1 Verify that

$$f = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$$

is an automorphism of \mathbf{Z}_6 .

- 2 Verify that

$$f_1 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 2 & 4 & 1 & 3 \end{pmatrix} \quad f_2 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 3 & 1 & 4 & 2 \end{pmatrix}$$

and

$$f_3 = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 0 & 4 & 3 & 2 & 1 \end{pmatrix}$$

are all automorphisms of \mathbf{z}_5 .

3 If G is any group, and a is any element of G , prove that $f(x) = axa^{-1}$ is an automorphism of G .

4 Since each automorphism of G is a bijective function from G to G , it is a *permutation* of G . Prove the set

$$\text{Aut}(G)$$

of all the automorphisms of G is a subgroup of S_G . (Remember that the operation is composition.)

J. Regular Representation of Groups

By Cayley's theorem, every group G is isomorphic to a group G^* of permutations of G . Recall that we match each element a in G with the permutation π_a defined by $\pi_a = ax$, that is, the rule “multiply on the left by a .” We let $G^* = \{\pi_a : a \in G\}$; with the operation \circ of composition it is a group of permutations, called the *left regular representation* of G . (It is called a “representation” of G because it is isomorphic to G .)

Instead of using the permutations π_a , we could just as well have used the permutations ρ_a defined by $\rho_a(x) = xa$, that is, “multiply on the right by a .” The group $G^\rho = \{\rho_a : a \in G\}$ is called the *right regular representation* of G .

If G is commutative, there is no difference between right and left multiplication, so G^* and G^ρ are the same, and are simply called the *regular representation* of G . Also, if the operation of G is denoted by $+$, the permutation corresponding to a is “add a ” instead of “multiply by a .”

Example The regular representation of \mathbf{z}_3 consists of the following permutations:

$$\begin{aligned} \pi_0 &= \begin{pmatrix} 0 & 1 & 2 \\ 0 & 1 & 2 \end{pmatrix} && \text{that is, the identity permutation} \\ \pi_1 &= \begin{pmatrix} 0 & 1 & 2 \\ 1 & 2 & 0 \end{pmatrix} && \text{that is, the rule “add 1”} \\ \pi_2 &= \begin{pmatrix} 0 & 1 & 2 \\ 2 & 0 & 1 \end{pmatrix} && \text{that is, the rule “add 2”} \end{aligned}$$

The regular representation of \mathbf{z}_3 has the following table:

\circ	π_0	π_1	π_2
π_0	π_0	π_1	π_2
π_1	π_1	π_2	π_0
π_2	π_2	π_0	π_1

The function

$$f = \begin{pmatrix} 0 & 1 & 2 \\ \pi_0 & \pi_1 & \pi_2 \end{pmatrix}$$

is easily seen to be an isomorphism from \mathbf{z}_3 to its regular representation.

Find the right and left regular representation of each of the following groups, and compute their tables. (If the group is abelian, find its regular representation.)

1 P_2 , the group of subsets of a two-element set. (See [Chapter 3](#), [Exercise C](#).)

2 \mathbf{z}_4 .

3 The group G of matrices described on page 28 of the text.

HOMOMORPHISMS

We have seen that if two groups are isomorphic, this means there is a one-to-one correspondence between them which transforms one of the groups into the other. Now if G and H are any groups, it may happen that there is a function which transforms G into H , although this function is *not* a one-to-one correspondence. For example, \mathbb{Z}_6 is transformed into \mathbb{Z}_3 by

$$f = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 2 & 0 & 1 & 2 \end{pmatrix}$$

as we may verify by comparing their tables:

+	0	1	2	3	4	5		+	0	1	2	0	1	2
0	0	1	2	3	4	5	Replace x by $f(x)$ →	0	0	1	2	0	1	2
1	1	2	3	4	5	0		1	1	2	0	1	2	0
2	2	3	4	5	0	1		2	2	0	1	2	0	1
3	3	4	5	0	1	2		0	0	1	2	0	1	2
4	4	5	0	1	2	3		1	1	2	0	1	2	0
5	5	0	1	2	3	4		2	2	0	1	2	0	1

Eliminate duplicate
information
→
(For example, $2 + 2 = 1$
appears four separate
times in table above.)

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

If G and H are any groups, and there is a function f which transforms G into H , we say that if is a *homomorphic image* of G . The function f is called a *homomorphism* from G to f . This notion of homomorphism is one of the skeleton keys of algebra, and this chapter is devoted to explaining it and

defining it precisely.

First, let us examine carefully what we mean by saying that “ f transforms G into H .” To begin with, f must be a function from G onto H ; but that is not all, because f must also transform the table of G into the table of H . To accomplish this, f must have the following property: for any two elements a and b in G ,

$$\text{if } f(a) = a' \quad \text{and} \quad f(b) = b', \quad \text{then} \quad f(ab) = a' b' \quad (1)$$

Graphically,

$$\begin{array}{lcl} \text{if} & a & \xrightarrow{f} a' \\ \text{and} & b & \xrightarrow{f} b' \\ \text{then} & ab & \xrightarrow{f} a' b' \end{array}$$

Condition (1) may be written more succinctly as follows:

$$f(ab) = f(a)f(b) \quad (2)$$

Thus,

Definition if G and H are groups, a **homomorphism** from G to H is a function $f: G \rightarrow H$ such that for any two elements a and b in G ,

$$f(ab) = f(a)f(b)$$

If there exists a homomorphism from G **onto** H , we say that H is a **homomorphic image** of G .

Groups have a very important and privileged relationship with their homomorphic images, as the next few examples will show.

Let P denote the group consisting of two elements, e and o , with the table

$+$	e	o
e	e	o
o	o	e

We call this group the *parity group* of even and odd numbers. We should think of e as “even” and o as “odd,” and the table as describing the rule for adding even and odd numbers. For example, even + odd = odd, odd + odd = even, and so on.

The function $f: \mathbf{Z} \rightarrow P$ which carries every even integer to e and every odd integer to o is clearly a homomorphism from \mathbf{Z} to P . This is easy to check because there are only four different cases: for arbitrary integers r and s , r and s are either both even, both odd, or mixed. For example, if r and s are both odd, their sum is even, so $f(r) = o$, $f(s) = o$, and $f(r + s) = e$. Since $e = o + o$,

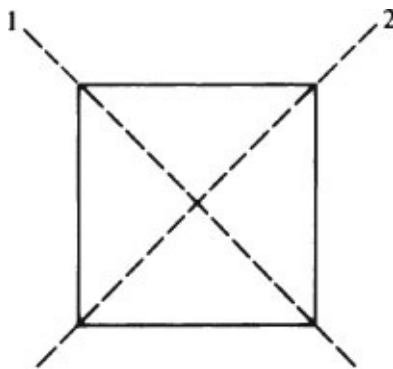
$$f(r + s) = f(r) + f(s)$$

This equation holds analogously in the remaining three cases; hence f is a homomorphism. (Note that the symbol $+$ is used on both sides of the above equation because the operation, in \mathbf{Z} as well as in P , is denoted by $+$.)

It follows that P is a homomorphic image of \mathbf{z} !

Now, what do P and \mathbf{z} have in common? P is a much smaller group than \mathbf{z} , therefore it is not surprising that very few properties of the integers are to be found in P . Nevertheless, *one* aspect of the structure of \mathbf{z} is retained absolutely intact in P , namely, the structure of the odd and even numbers. (The fact of being odd or even is called the *parity* of integers.) In other words, *as we pass from \mathbf{z} to P* we deliberately lose every aspect of the integers except their parity; their parity alone (with its arithmetic) is retained, and faithfully preserved.

Another example will make this point clearer. Remember that D_4 is the group of the symmetries of the square. Now, every symmetry of the



square either interchanges the two diagonals here labeled 1 and 2, or leaves them as they were. In other words, every symmetry of the square brings about one of the permutations

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}$$

of the diagonals.

For each $R_i \in D_4$, let $f(R_i)$ be the permutation of the diagonals produced by R_i . Then f is clearly a homomorphism from D_4 onto S_2 . Indeed, it is clear on geometrical grounds that when we perform the motion R_i followed by the motion R_j on the square, we are, at the same time, carrying out the motions $f(R_i)$ followed by $f(R_j)$ on the diagonals. Thus,

$$f(R_j \circ R_i) = f(R_j) \circ f(R_i)$$

It follows that S_2 is a homomorphic image of D_4 . Now S_2 is a smaller group than D_4 , and therefore very few of the features of D_4 are to be found in S_2 . Nevertheless, *one* aspect of the structure of D_4 is retained absolutely intact in S_2 , namely, the diagonal motions. Thus, as we pass from D_4 to S_2 , we deliberately lose every aspect of plane motions except the motions of the diagonals; these alone are retained and faithfully preserved.

A final example may be of some help; it relates to the group \mathbb{B}^n described in [Chapter 3, Exercise E](#). Here, briefly, is the context in which this group arises: The most basic way of transmitting information is to code it into strings of 0s and 1s, such as 0010111, 1010011, etc. Such strings are called *binary words*, and the number of 0s and 1s in any binary word is called its *length*. The symbol \mathbb{B}^n designates the group consisting of all binary words of length n , with an operation of addition described in [Chapter 3, Exercise E](#).

Consider the function $f: \mathbb{B}^7 \rightarrow \mathbb{B}^5$ which consists of *dropping the last two digits* of every seven-digit

word. This kind of function arises in many practical situations: for example, it frequently happens that the first five digits of a word carry the message while the last two digits are an error check. Thus, f separates the message from the error check.

It is easy to verify that f is a homomorphism; hence \mathbb{B}^5 is a homomorphic image of \mathbb{B}^7 . As we pass from \mathbb{B}^7 to \mathbb{B}^5 , the message component of words in \mathbb{B}^7 is exactly preserved while the error check is deliberately lost.

These examples illustrate the basic idea inherent in the concept of a homomorphic image. The cases which arise in practice are not always so clear-cut as these, but the underlying idea is still the same: In a homomorphic image of G , some aspect of G is isolated and faithfully preserved while all else is deliberately lost.

The next theorem presents two elementary properties of homomorphisms.

Theorem 1 *Let G and H be groups, and $f: G \rightarrow H$ a homomorphism. Then*

- (i) $f(e) = e$, and
- (ii) $f(a^{-1}) = [f(a)]^{-1}$ for every element $a \in G$.

In the equation $f(e) = e$, the letter e on the left refers to the neutral element in G , whereas the letter e on the right refers to the neutral element in H .

To prove (i), we note that in any group,

$$\text{if } yy = y \quad \text{then} \quad y = e$$

(Use the cancellation law on the equation $yy = ye$.) Now, $f(e)f(e) = f(ee) = f(e)$; hence $f(e) = e$.

To prove (ii), note that $f(a)f(a^{-1}) = f(aa^{-1}) = f(e)$. But $f(e) = e$, so $f(a)f(a^{-1}) = e$. It follows by [Theorem 2](#) of [Chapter 4](#) that $f(a^{-1})$ is the inverse of $f(a)$, that is, $f(a^{-1}) = [f(a)]^{-1}$.

Before going on with our study of homomorphisms, we must be introduced to an important new concept. If a is an element of a group G , a *conjugate* of a is any element of the form xax^{-1} , where $x \in G$. For example, the conjugates of α in S_3 are

$$\beta \circ \alpha \circ \beta^{-1} = \gamma$$

$$\gamma \circ \alpha \circ \gamma^{-1} = \kappa$$

$$\delta \circ \alpha \circ \delta^{-1} = \kappa$$

$$\kappa \circ \alpha \circ \kappa^{-1} = \gamma$$

as well as a itself, which may be written in two ways, as $\varepsilon \circ \alpha \circ \varepsilon^{-1}$ or as $\alpha \circ \alpha \circ \alpha^{-1}$. If H is any subset of a group G , we say that H is *closed with respect to conjugates* if every conjugate of every element of H is in H . Finally,

Definition *Let H be a subgroup of a group G . H is called a **normal** subgroup of G if it is closed with respect to conjugates, that is, if*

$$\text{for any } a \in H \quad \text{and} \quad x \in G \quad xax^{-1} \in H$$

(Note that according to this definition, a normal subgroup of G is any nonempty subset of G which is closed with respect to products, with respect to inverses, and with respect to conjugates.)

We now return to our discussion of homomorphisms.

Definition Let $f: G \rightarrow H$ be a homomorphism. The **kernel** of f is the set K of all the elements of G which are carried by f onto the neutral element of H . That is,

$$K = \{x \in G: f(x) = e\}$$

Theorem 2 Let $f: G \rightarrow H$ be a homomorphism.

- (i) The kernel of f is a normal subgroup of G , and
- (ii) The range of f is a subgroup of H .

PROOF: Let K denote the kernel of f . If $a, b \in K$, this means that $f(a) = e$ and $f(b) = e$. Thus, $f(ab) = f(a)f(b) = ee = e$; hence $ab \in K$.

If $a \in K$, then $f(a) = e$. Thus, $f(a^{-1}) = [f(a)]^{-1} = e^{-1} = e$, so $a^{-1} \in K$.

Finally, if $a \in K$ and $x \in G$, then $f(xax^{-1}) = f(x)f(a)f(x^{-1}) = f(x)f(a)[f(x)]^{-1} = e$, which shows that $xax^{-1} \in K$. Thus, K is a normal subgroup of G .

Now we must prove part (ii). If $f(a)$ and $f(b)$ are in the range of f , then their product $f(a)f(b) = f(ab)$ is also in the range of f .

If $f(a)$ is in the range of f , its inverse is $[f(a)]^{-1} = f(a^{-1})$, which is also in the range of f . Thus, the range of f is a subgroup of H . ■

If f is a homomorphism, we represent the kernel of f and the range of f with the symbols

$$\ker(f) \quad \text{and} \quad \text{ran}(f)$$

EXERCISES

A. Examples of Homomorphisms of Finite Groups

1 Consider the function $f: \mathbf{z}_8 \rightarrow \mathbf{z}_4$ given by

$$f = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \end{pmatrix}$$

Verify that f is a homomorphism, find its kernel K , and list the cosets of K . [REMARK: To verify that f is a homomorphism, you must show that $f(a + b) = f(a) + f(b)$ for all choices of a and b in \mathbf{z}_8 ; there are 64 choices. This may be accomplished by checking that f transforms the table of \mathbf{z}_8 to the table of \mathbf{z}_4 , as on page 136.]

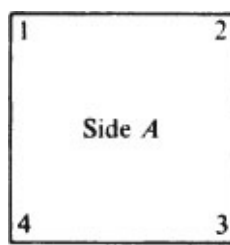
2 Consider the function $f: S_3 \rightarrow \mathbf{z}_2$ given by

$$f = \begin{pmatrix} \varepsilon & \alpha & \beta & \gamma & \delta & \kappa \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Verify that f is a homomorphism, find its kernel K , and list the cosets of K .

3 Find a homomorphism $f: \mathbf{z}_{15} \rightarrow \mathbf{z}_5$, and indicate its kernel. (Do not actually verify that f is a homomorphism.)

4 Imagine a square as a piece of paper lying on a table. The side facing you is side

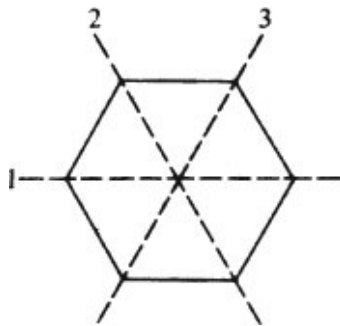


A. The side hidden from view is side B . Every motion of the square either interchanges the two sides (that is, side B becomes visible and side A hidden) or leaves the sides as they were. In other words, every motion R_i of the square brings about one of the permutations

$$\varepsilon = \begin{pmatrix} A & B \\ A & B \end{pmatrix} \quad \text{or} \quad \alpha = \begin{pmatrix} A & B \\ B & A \end{pmatrix}$$

of the sides; call it $g(R_i)$. Verify that $g: D_4 \rightarrow S_2$ is a homomorphism, and give its kernel.

5 Every motion of the regular hexagon brings about a permutation of its diagonals, labeled 1, 2, and 3. For each $R_i \in D_6$, let $f(R_i)$ be the permutation of



the diagonals produced by R_i . Argue informally (appealing to geometric intuition) to explain why $f: D_6 \rightarrow S_3$ is a homomorphism. Then complete the following:

$$f\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} = \varepsilon \quad f\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 6 & 1 \end{pmatrix} = \delta \quad \dots$$

(That is, find the value of f on all 12 elements of D_6 .)

6 Let $B \subset A$. Let $h: P_A \rightarrow P_B$ be defined by $h(C) = C \cap B$. For $A = \{1, 2, 3\}$ and $B = \{1, 2\}$, complete the following:

$$h = \begin{pmatrix} \emptyset & \{1\} & \{2\} & \{3\} & \{1, 2\} & \{1, 3\} & \{2, 3\} & A \end{pmatrix}$$

For any A and $B \subset A$, show that h is a homomorphism.

B. Examples of Homomorphisms of Infinite Groups

Prove that each of the following is a homomorphism, and describe its kernel.

- 1 The function $\phi: \mathcal{F}(\mathbb{R}) \rightarrow \mathbb{R}$ given by $\phi(f) = f(0)$.
- 2 The function $\phi: \mathcal{D}(\mathbb{R}) \rightarrow \mathcal{F}(\mathbb{R})$ given by $\phi(f) = f'$. $\mathcal{D}(\mathbb{R})$ is the group of differentiable functions from \mathbb{R} to \mathbb{R} . f' is the derivative of f .
- 3 The function $f: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x, y) = x + y$.

4 The function $f: \mathbb{R}^* \rightarrow \mathbb{R}^{\text{pos}}$ defined by $f(x) = |x|$.

5 The function $f: \mathbb{C}^* \rightarrow \mathbb{R}^{\text{pos}}$ defined by $f(a + bi) = \sqrt{a^2 + b^2}$.

6 Let G be the multiplicative group of all 2×2 matrices

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

satisfying $ad - bc \neq 0$. Let $f: G \rightarrow \mathbb{R}^*$ be given by $f(A) = \text{determinant of } A = ad - bc$.

C. Elementary Properties of Homomorphisms

Let G, H , and K be groups. Prove the following:

1 If $f: G \rightarrow G$ and $g: H \rightarrow K$ are homomorphisms, then their composite $g \circ f: G \rightarrow K$ is a homomorphism.

2 If $f: G \rightarrow H$ is a homomorphism with kernel K , then f is injective iff $K = \{e\}$.

3 If $f: G \rightarrow H$ is a homomorphism and J is any subgroup of G , then $f(J) = \{f(x) : x \in J\}$ is a subgroup of H .

4 If $f: G \rightarrow H$ is a homomorphism and J is any subgroup of H , then

$$f^{-1}(J) = \{x \in G : f(x) \in J\}$$

is a subgroup of G . Furthermore, $\ker f \subseteq f^{-1}(J)$.

5 If $f: G \rightarrow H$ is a homomorphism with kernel K , and J is a subgroup of G , let f_J designate the restriction of f to J . (In other words f_J is the same function as f , except that its domain is restricted to J .) Then $\ker f_J = J \cap K$.

6 For any group G , the function $f: G \rightarrow G$ defined by $f(x) = e$ is a homomorphism.

7 For any group G , $\{e\}$ and G are homomorphic images of G .

8 The function $f: G \rightarrow G$ defined by $f(x) = x^2$ is a homomorphism iff G is abelian.

9 The functions $f_1(x, y) = x$ and $f_2(x, y) = y$, from $G \times H$ to G and H , respectively, are homomorphisms.

D. Basic Properties of Normal Subgroups

In the following, let G denote an arbitrary group.

1 Find all the normal subgroups (a) of S_3 and (b) of D_4 .

Prove the following:

2 Every subgroup of an abelian group is normal.

3 The center of any group G is a normal subgroup of G .

4 Let H be a subgroup of G . H is normal iff it has the following property: For all a and b in G , $ab \in H$ iff $ba \in H$.

5 Let H be a subgroup of G . H is normal iff $aH = Ha$ for every $a \in G$.

6 Any intersection of normal subgroups of G is a normal subgroup of G .

E. Further Properties of Normal Subgroups

Let G denote a group, and H a subgroup of G . Prove the following:

- # **1** If H has index 2 in G , then H is normal. (HINT: Use [Exercise D5](#).)
- 2** Suppose an element $a \in G$ has order 2. Then $\langle a \rangle$ is a normal subgroup of G iff a is in the center of G .
- 3** If a is any element of G , $\langle a \rangle$ is a normal subgroup of G iff a has the following property: For any $x \in G$, there is a positive integer k such that $xa = a^k x$.
- 4** In a group G , a *commutator* is any product of the form $aba^{-1}b^{-1}$, where a and b are any elements of G . If a subgroup H of G contains *all* the commutators of G , then H is normal.
- 5** If H and K are subgroups of G , and K is normal, then HK is a subgroup of G . (HK denotes the set of all products hk as h ranges over H and k ranges over K .)
- # **6** Let S be the union of all the cosets Ha such that $Ha = aH$. Then S is a subgroup of G , and H is a normal subgroup of S .

F. Homomorphism and the Order of Elements

If $f: G \rightarrow H$ is a homomorphism, prove each of the following:

- 1** For each element $a \in G$, the order of $f(a)$ is a divisor of the order of a .
- 2** The order of any element $b \neq e$ in the range of f is a common divisor of $|G|$ and $|H|$. (Use part 1.)
- 3** If the range of f has n elements, then $x^n \in \ker f$ for every $x \in G$.
- 4** Let m be an integer such that m and $|H|$ are relatively prime. For any $x \in G$, if $x^m \in \ker f$, then $x \in \ker f$.
- 5** Let the range of f have m elements. If $a \in G$ has order n , where m and n are relatively prime, then a is in the kernel of f . (Use part 1.)
- 6** Let p be a prime. If $\text{ran } f$ has an element of order p , then G has an element of order p .

G. Properties Preserved under Homomorphism

A property of groups is said to be “preserved under homomorphism” if, whenever a group G has that property, every homomorphic image of G does also. In this exercise set, we will survey a few typical properties preserved under homomorphism. If $f: G \rightarrow H$ is a homomorphism of G onto H , prove each of the following:

- 1** If G is abelian, then H is abelian.
- 2** If G is cyclic, then H is cyclic.
- 3** If every element of G has finite order, then every element of H has finite order.
- 4** If every element of G is its own inverse, every element of H is its own inverse.
- 5** If every element of G has a square root, then every element of H has a square root.
- 6** If G is finitely generated, then H is finitely generated. (A group is said to be “finitely generated” if it is generated by finitely many of its elements.)

† H. Inner Direct Products

If G is any group, let H and K be normal subgroups of G such that $H \cap K = \{e\}$. Prove the following:

- 1** Let h_1 and h_2 be any two elements of H , and k_1 and k_2 any two elements of K .

$$h_1 k_1 = h_2 k_2 \quad \text{implies} \quad h_1 = h_2 \quad \text{and} \quad k_1 = k_2$$

(HINT: If $h_1 k_1 = h_2 k_2$, then $h_2^{-1} h_1 \in H \cap K$ and $k_2 k_1^{-1} \in H \cap K$. Explain why.)

2 For any $h \in H$ and $k \in K$, $hk = kh$. (HINT: $hk = kh$ iff $hkh^{-1}k^{-1} = e$. Use the fact that H and K are normal.)

3 Now, make the additional assumption that $G = HK$ that is, every x in G can be written as $x = hk$ for some $h \in H$ and $k \in K$. Then the function $\phi(h,k) = hk$ is an isomorphism from $H \times K$ onto G .

We have thus proved the following: *If H and K are normal subgroups of G , such that $H \cap K = \{e\}$ and $G = HK$, then $G \cong H \times K$. G is sometimes called the inner direct product of H and K .*

† I. Conjugate Subgroups

Let H be a subgroup of G . For any $a \in G$, let $aHa^{-1} = \{axa^{-1} : x \in H\}$; aHa^{-1} is called a conjugate of H . Prove the following:

1 For each $a \in G$, aHa^{-1} is a subgroup of G .

2 For each $a \in G$, $H \cong aHa^{-1}$.

3 H is a normal subgroup of G iff $H = aHa^{-1}$ for every $a \in G$.

In the remaining exercises of this set, let G be a finite group. By the normalizer of H we mean the set $N(H) = \{a \in G : axa^{-1} \in H \text{ for every } x \in H\}$.

4 If $a \in N(H)$, then $aHa^{-1} = H$. (Remember that G is now a finite group.)

5 $N(H)$ is a subgroup of G .

6 $H \subseteq N(H)$. Furthermore, H is a normal subgroup of $N(H)$.

In parts 7–10, let $N = N(H)$.

7 For any $a, b \in G$, $aHa^{-1} = bHb^{-1}$ iff $b^{-1}a \in N(H)$.

8 There is a one-to-one correspondence between the set of conjugates of H and the set of cosets of N . (Thus, there are as many conjugates of H as cosets of N .)

9 H has exactly $(G : N)$ conjugates. In particular, the number of distinct conjugates of H is a divisor of $|G|$.

10 Let K be any subgroup of G , let $K^* = \{Na : a \in K\}$, and let

$$X_K = \{aHa^{-1} : a \in K\}$$

Argue as in part 8 to prove that X_K is in one-to-one correspondence with K^* . Conclude that the number of elements in X_K is a divisor of $|K|$.

QUOTIENT GROUPS

In [Chapter 14](#) we learned to recognize when a group H is a homomorphic image of a group G . Now we will make a great leap forward by learning a method for actually *constructing all the homomorphic images of any group*. This is a remarkable procedure, of great importance in algebra. In many cases this construction will allow us to deliberately select *which* properties of a group G we wish to preserve in a homomorphic image, and which other properties we wish to discard.

The most important instrument to be used in this construction is the notion of a normal subgroup. Remember that a *normal* subgroup of G is any subgroup of G which is closed with respect to conjugates. We begin by giving an elementary property of normal subgroups.

Theorem 1 *If H is a normal subgroup of G , then $aH = Ha$ for every $a \in G$.*

(In other words, there is no distinction between left and right cosets for a normal subgroup.)

PROOF: Indeed, if x is any element of aH , then $x = ah$ for some $h \in H$. But H is closed with respect to conjugates; hence $aha^{-1} \in H$. Thus, $x = ah = (aha^{-1})a$ is an element of Ha . This shows that every element of aH is in Ha ; analogously, every element of Ha is in aH . Thus, $aH = Ha$. ■

Let G be a group and let H be a subgroup of G . There is a way of combining cosets, called *coset multiplication*, which works as follows: *the coset of a , multiplied by the coset of b , is defined to be the coset of ab* . In symbols,

$$Ha \cdot Hb = H(ab)$$

This definition is deceptively simple, for it conceals a fundamental difficulty. Indeed, it is not at all clear that the product of two cosets Ha and Hb , multiplied together in this fashion, is *uniquely defined*. Remember that Ha may be the same coset as Hc (this happens iff c is in Ha), and, similarly, Hb may be the same coset as Hd . Therefore, the product $Ha \cdot Hb$ is the same as the product $He \cdot Hd$. Yet it may easily happen that $H(ab)$ is *not* the same coset as $H(cd)$. Graphically,

$$\begin{array}{ccc} Ha \cdot Hb = H(ab) \\ \parallel \quad \parallel \quad \nparallel \\ Hc \cdot Hd = H(cd) \end{array}$$

For example, if $G = S_3$ and $H = \{\varepsilon, \alpha\}$, then

$$H\beta = \{\beta, \gamma\} = H\gamma$$

$$H\delta = \{\delta, \kappa\} = H\kappa$$

and yet

$$H(\beta \circ \delta) = H\varepsilon \neq H\beta = H(\gamma \circ \kappa)$$

Thus, coset multiplication *does not work* as an operation on the cosets of $H = \{\varepsilon, \alpha\}$ in S_3 . The reason is that, although H is a subgroup of S_3 , H is *not a normal subgroup* of S_3 . If H were a normal subgroup, coset multiplication would work. The next theorem states exactly that!

Theorem 2 *Let H be a normal subgroup of G . If $Ha = He$ and $Hb = Hd$, then $H(ab) = H(cd)$.*

PROOF: If $Ha = He$, then $a \in Hc$; hence $a = h_1c$ for some $h_1 \in H$. If $Hb = Hd$, then $b \in Hd$; hence $b = h_2d$ from some $h_2 \in H$. Thus,

$$ab = h_1ch_2d = h_1(ch_2)d$$

But $ch_2 \in cH = Hc$ (the last equality is true by [Theorem 1](#)). Thus, $ch_2 = h_3c$ for some $h_3 \in H$. Returning to ab ,

$$ab = h_1(ch_2)d = h_1(h_3c)d = (h_1h_3)(cd)$$

and this last element is clearly in $H(cd)$.

We have shown that $ab \in H(cd)$. Thus, by Property (1) in [Chapter 13](#), $H(ab) = H(cd)$. ■

We are now ready to proceed with the construction promised at the beginning of the chapter. Let G be a group and let H be a normal subgroup of G . Think of the set which consists of *all the cosets of H* . This set is conventionally denoted by the symbol G/H . Thus, if Ha, Hb, Hc, \dots are cosets of H , then

$$G/H = \{Ha, Hb, Hc, \dots\}$$

We have just seen that *coset multiplication* is a valid operation on this set. In fact,

Theorem 3 *G/H with coset multiplication is a group.*

PROOF: Coset multiplication is associative, because

$$\begin{aligned} Ha \cdot (Hb \cdot Hc) &= Ha \cdot H(bc) = Ha(bc) = H(ab)c \\ &= H(ab) \cdot Hc = (Ha \cdot Hb) \cdot Hc \end{aligned}$$

The identity element of G/H is $H = He$, for $Ha \cdot He = Ha$ and $He \cdot Ha = Ha$ for every coset Ha .

Finally, the inverse of any coset Ha is the coset Ha^{-1} , because $Ha \cdot Ha^{-1} = Haa^{-1} = He$ and $Ha^{-1} \cdot Ha = Ha^{-1}aHe$.

The group G/H is called the *factor group*, or *quotient group* of G by H .

And now, the pièce de résistance:

Theorem 4 G/H is a homomorphic image of G .

PROOF: The most obvious function from G to G/H is the function f which carries every element to its own coset, that is, the function given by

$$f(x) = Hx$$

This function is a homomorphism, because

$$f(xy) = Hxy = Hx \cdot Hy = f(x)f(y)$$

f is called the *natural homomorphism* from G onto G/H . Since there is a homomorphism from G onto G/H , G/H is a homomorphic image of G . ■

Thus, when we construct quotient groups of G , we are, in fact, constructing homomorphic images of G . The quotient group construction is useful because it is a way of actually manufacturing homomorphic images of any group G . In fact, as we will soon see, it is a way of manufacturing *all* the homomorphic images of G .

Our first example is intended to clarify the details of quotient group construction. Let \mathbf{z} be the group of the integers, and let $\langle 6 \rangle$ be the cyclic subgroup of \mathbf{z} which consists of all the multiples of 6. Since \mathbf{z} is abelian, and every subgroup of an abelian group is normal, $\langle 6 \rangle$ is a normal subgroup of \mathbf{z} . Therefore, we may form the quotient group $\mathbf{z}/\langle 6 \rangle$. The elements of this quotient group are all the cosets of the subgroup $\langle 6 \rangle$, namely:

$$\langle 6 \rangle + 0 = \{ \dots, -18, -12, -6, 0, 6, 12, 18, \dots \}$$

$$\langle 6 \rangle + 1 = \{ \dots, -17, -11, -5, 1, 7, 13, 19, \dots \}$$

$$\langle 6 \rangle + 2 = \{ \dots, -16, -10, -4, 2, 8, 14, 20, \dots \}$$

$$\langle 6 \rangle + 3 = \{ \dots, -15, -9, -3, 3, 9, 15, 21, \dots \}$$

$$\langle 6 \rangle + 4 = \{ \dots, -14, -8, -2, 4, 10, 16, 22, \dots \}$$

$$\langle 6 \rangle + 5 = \{ \dots, -13, -7, -1, 5, 11, 17, 23, \dots \}$$

These are *all* the different cosets of $\langle 6 \rangle$, for it is easy to see that $\langle 6 \rangle + 6 = \langle 6 \rangle + 0$, $\langle 6 \rangle + 7 = \langle 6 \rangle + 1$, $\langle 6 \rangle + 8 = \langle 6 \rangle + 2$, and so on.

Now, the operation on \mathbf{z} is denoted by $+$, and therefore we will call the operation on the cosets *coset addition* rather than coset multiplication. But nothing is changed except the name; for example, the coset $\langle 6 \rangle + 1$ added to the coset $\langle 6 \rangle + 2$ is the coset $\langle 6 \rangle + 3$. The coset $\langle 6 \rangle + 3$ added to the coset $\langle 6 \rangle + 4$ is the coset $\langle 6 \rangle + 7$, which is the same as $\langle 6 \rangle + 1$. To simplify our notation, let us agree to write the cosets in the following shorter form:

$$\bar{0} = \langle 6 \rangle + 0 \quad \bar{1} = \langle 6 \rangle + 1 \quad \bar{2} = \langle 6 \rangle + 2$$

$$\bar{3} = \langle 6 \rangle + 3 \quad \bar{4} = \langle 6 \rangle + 4 \quad \bar{5} = \langle 6 \rangle + 5$$

Then $\mathbf{z}/\langle 6 \rangle$ consists of the six elements $\bar{0}$, $\bar{1}$, $\bar{2}$, $\bar{3}$, $\bar{4}$, and $\bar{5}$, and its operation is summarized in the following table:

+	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$
$\bar{3}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{4}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$
$\bar{5}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$

The reader will perceive immediately the similarity between this group and \mathbf{z}_6 . As a matter of fact, the quotient group construction of $\mathbf{z}/\langle 6 \rangle$ is considered to be the rigorous way of constructing \mathbf{z}_6 . So from now on, we will consider \mathbf{z}_6 to be the same as $\mathbf{z}/\langle 6 \rangle$; and, in general, we will consider \mathbf{z}_n to be the same as $\mathbf{z}/\langle n \rangle$. In particular, we can see that for any n , \mathbf{z}_n is a homomorphic image of \mathbf{z} .

Let us repeat: The motive for the quotient group construction is that it gives us a way of actually *producing* all the homomorphic images of any group G . However, what is even more fascinating about the quotient group construction is that, in practical instances, we can often choose H so as to “factor out” unwanted properties of G , and preserve in G/H only “desirable” traits. (By “desirable” we mean desirable within the context of some specific application or use.) Let us look at a few examples.

First, we will need two simple properties of cosets, which are given in the next theorem.

Theorem 5 *Let G be a group and H a subgroup of G . Then*

- (i) $Ha = Hb$ iff $ab^{-1} \in H$ and
- (ii) $Ha = H$ iff $a \in H$

PROOF: If $Ha = Hb$, then $a \in Hb$, so $a = hb$ for some $h \in H$. Thus,

$$ab^{-1} = h \in H$$

If $ab^{-1} \in H$, then $ab^{-1} = h$ for $h \in H$, and therefore $a = hb \in Hb$. It follows by Property (1) of [Chapter 13](#) that $Ha = Hb$.

This proves (i). It follows that $Ha = He$ iff $ae^{-1} = a \in H$, which proves (ii). ■

For our first example, let G be an abelian group and let H consist of *all the elements of G which have finite order*. It is easy to show that H is a subgroup of G . (The details may be supplied by the reader.) Remember that in an abelian group every subgroup is normal; hence H is a normal subgroup of G , and therefore we may form the quotient group G/H . We will show next that in G/H , *no* element except the neutral element has finite order.

For suppose G/H has an element Hx of finite order. Since the neutral element of G/H is H , this means there is an integer $m \neq 0$ such that $(Hx)^m = H$, that is, $Hx^m = H$. Therefore, by [Theorem 5\(ii\)](#), $x^m \in H$, so x^m has finite order, say t :

$$(x^m)^t = x^{mt} = e$$

But then x has finite order, so $x \in H$. Thus, by [Theorem 5\(ii\)](#), $Hx = H$. This proves that in G/H , the only element Hx of finite order is the neutral element H .

Let us recapitulate: If H is the subgroup of G which consists of all the elements of G which have

finite order, then in G/H , *no* element (except the neutral element) has finite order. Thus, in a sense, we have “factored out” *all the elements of finite order (they are all in H) and produced a quotient group G/H whose elements all have infinite order* (except for the neutral element, which necessarily has order 1).

Our next example may bring out this idea even more clearly. Let G be an arbitrary group; by a *commutator* of G we mean any element of the form $aba^{-1}b^{-1}$ where a and b are in G . The reason such a product is called a commutator is that

$$aba^{-1}b^{-1} = e \quad \text{iff} \quad ab = ba$$

In other words, $aba^{-1}b^{-1}$ reduces to the neutral element whenever a and b commute—and *only* in that case! Thus, in an abelian group all the commutators are equal to e . In a group which is not abelian, the number of distinct commutators may be regarded as a measure of the extent to which G departs from being commutative. (The fewer the commutators, the closer the group is to being an abelian group.)

We will see in a moment that if H is a subgroup of G which *contains all the commutators of G* , then G/H is abelian! What this means, in a fairly accurate sense, is that *when we factor out the commutators of G we get a quotient group which has no commutators* (except, trivially, the neutral element) *and which is therefore abelian*.

To say that G/H is abelian is to say that for any two elements Hx and Hy in G/H , $HxHy = HyHx$; that is, $Hxy = Hyx$. But by [Theorem 5\(ii\)](#),

$$Hxy = Hyx \quad \text{iff} \quad xy(yx)^{-1} \in H$$

Now $xy(yx)^{-1}$ is the commutator $xyx^{-1}y^{-1}$; so if all commutators are in H , then G/H is abelian.

EXERCISES

A. Examples of Finite Quotient Groups

In each of the following, G is a group and H is a normal subgroup of G . List the elements of G/H and then write the table of G/H .

Example $G = \mathbf{Z}_6$ and $H = \{0, 3\}$

The elements of G/H are the three cosets $H = H + 0 = \{0, 3\}$, $H + 1 = \{1, 4\}$, and $H + 2 = \{2, 5\}$. (Note that $H + 3$ is the same as $H + 0$, $H + 4$ is the same as $H + 1$, and $H + 5$ is the same as $H + 2$.) The table of G/H is

+	H	$H + 1$	$H + 2$
H	H	$H + 1$	$H + 2$
$H + 1$	$H + 1$	$H + 2$	H
$H + 2$	$H + 2$	H	$H + 1$

1 $G = \mathbf{Z}_{10}$, $H = \{0, 5\}$. (Explain why $G/H \cong \mathbf{Z}_5$.)

2 $G = S_3$, $H = \{\varepsilon, \beta, \delta\}$.

3 $G = D_4$, $H = \{R_0, R_2\}$. (See page 73.)

4 $G = D_4$, $H = \{R_0, R_2, R_4, R_5\}$.

5 $G = \mathbf{Z}_4 \times \mathbf{Z}_2$, $H = \langle (0,1) \rangle$ = the subgroup of $\mathbf{Z}_4 \times \mathbf{Z}_2$ generated by $(0,1)$.

6 $G = P_3$, $H = \{\emptyset, \{1\}\}$. (P_3 is the group of subsets of $\{1, 2, 3\}$.)

B. Examples of Quotient Groups of $\mathbb{R} \times \mathbb{R}$

In each of the following, H is a subset of $\mathbb{R} \times \mathbb{R}$.

(a) Prove that H is a normal subgroup of $\mathbb{R} \times \mathbb{R}$. (Remember that every subgroup of an abelian group is normal.)

(b) In geometrical terms, describe the elements of the quotient group G/H .

(c) In geometrical terms or otherwise, describe the operation of G/H .

1 $H = \{(x,0): x \in \mathbb{R}\}$

2 $H = \{(x,y): y = -x\}$

3 $H = \{(x,y): y = 2x\}$

C. Relating Properties of H to Properties of G/H

In parts 1-5 below, G is a group and H is a normal subgroup of G . Prove the following ([Theorem 5](#) will play a crucial role):

1 If $x^2 \in H$ for every $x \in G$, then every element of G/H is its own inverse. Conversely, if every element of G/H is its own inverse, then $x^2 \in H$ for all $x \in G$.

2 Let m be a fixed integer. If $x^m \in H$ for every $x \in G$, then the order of every element in G/H is a divisor of m . Conversely, if the order of every element in G/H is a divisor of m , then $x^m \in H$ for every $x \in G$.

3 Suppose that for every $x \in G$, there is an integer n such that $x^n \in H$; then every element of G/H has finite order. Conversely, if every element of G/H has finite order, then for every $x \in G$ there is an integer n such that $x^n \in H$.

4 Every element of G/H has a square root iff for every $x \in G$, there is some $y \in G$ such that $xy^2 \in H$.

5 G/H is cyclic iff there is an element $a \in G$ with the following property: for every $x \in G$, there is some integer n such that $xa^n \in H$.

6 If G is an abelian group, let H_p be the set of all $x \in H$ whose order is a power of p . Prove that H_p is a subgroup of G . Prove that G/H_p has no elements whose order is a nonzero power of p .

7 (a) If G/H is abelian, prove that H contains all the commutators of G .

(b) Let K be a normal subgroup of G , and H a normal subgroup of K . If G/H is abelian, prove that G/K and K/H are both abelian.

D. Properties of G Determined by Properties of G/H and H

There are some group properties which, if they are true in G/H and in H , must be true in G . Here is a sampling. Let G be a group, and H a normal subgroup of G . Prove the following:

1 If every element of G/H has finite order, and every element of H has finite order, then every element of G has finite order.

- 2 If every element of G/H has a square root, and every element of H has a square root, then every element of G has a square root. (Assume G is abelian.)
- 3 Let p be a prime number. If G/H and H are p -groups, then G is a p -group. A group G is called a p -group if the order of every element x in G is a power of p .
- # 4 If G/H and H are finitely generated, then G is finitely generated. (A group is said to be finitely generated if it is generated by a finite subset of its elements.)

E. Order of Elements in Quotient Groups

Let G be a group, and H a normal subgroup of G . Prove the following:

- 1 For each element $a \in G$, the order of the element Ha in G/H is a divisor of the order of a in G . (HINT: Use [Chapter 14, Exercise F1](#).)
- 2 If $(G:H) = m$, the order of every element of G/H is a divisor of m .
- 3 If $(G:H) = p$, where p is a prime, then the order of every element $a \notin H$ in G is a multiple of p . (Use part 1.)
- 4 If G has a normal subgroup of index p , where p is a prime, then G has at least one element of order p .
- 5 If $(G:H) = m$, then $a^m \in H$ for every $a \in G$.
- # 6 In \mathbb{Q}/\mathbb{Z} , every element has finite order.

† F. Quotient of a Group by Its Center

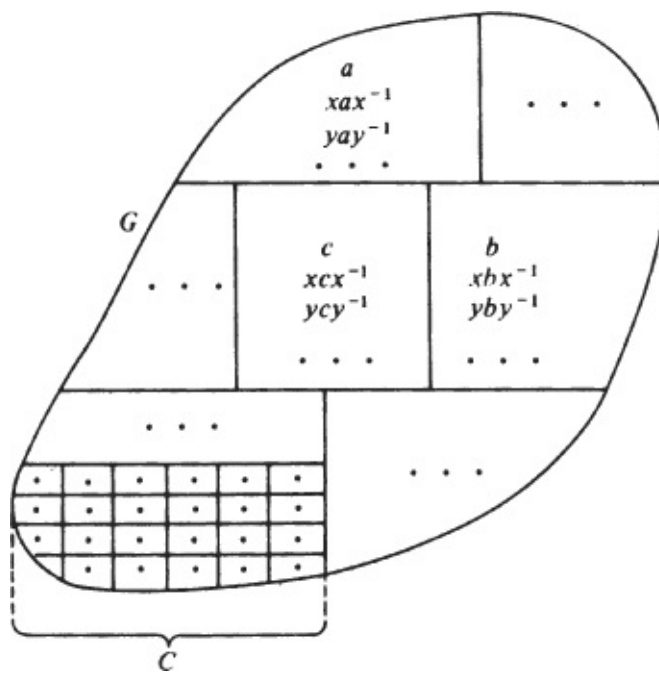
The *center* of a group G is the normal subgroup C of G consisting of all those elements of G which commute with every element of G . Suppose the quotient group G/C is a cyclic group; say it is generated by the element Ca of G/C . Prove parts 1-3:

- 1 For every $x \in G$, there is some integer m such that $Cx = Ca^m$.
- 2 For every $x \in G$, there is some integer m such that $x = ca^m$, where $c \in C$.
- 3 For any two elements x and y in G , $xy = yx$. (HINT: Use part 2 to write $x = ca^m$, $y = c'a^n$, and remember that $c, c' \in C$.)
- 4 Conclude that if G/C is cyclic, then G is abelian.

† G. Using the Class Equation to Determine the Size of the Center

{Prerequisite: [Chapter 13, Exercise I](#).)

Let G be a finite group. Elements a and b in G are called *conjugates* of one another (in symbols, $a \sim b$) iff $a = xbx^{-1}$ for some $x \in G$ (this is the same as $b = x^{-1}ax$). The relation \sim is an equivalence relation in G ; the equivalence class of any element a is called its *conjugacy class*. Hence G is partitioned into conjugacy classes (as shown in the diagram); the size of each conjugacy class divides the order of G . (For these facts, see [Chapter 13, Exercise I](#).)



“Each element of the center C is alone in its conjugacy class.”

Let S_1, S_2, \dots, S_t be the distinct conjugacy classes of G , and let k_1, k_2, \dots, k_t be their sizes. Then $|G| = k_1 + k_2 + \dots + k_t$ (This is called the *class equation* of G .)

Let G be a group whose order is a power of a prime p , say $|G| = p^k$. Let C denote the center of G . Prove parts 1-3:

- 1 The conjugacy class of a contains a (and no other element) iff $a \in C$.
 - 2 Let c be the order of C . Then $|G| = c + k_s + k_{s+1} + \dots + k_t$, where k_s, \dots, k_t are the sizes of all the distinct conjugacy classes of elements $x \notin C$.
 - 3 For each $i \in \{s, s+1, \dots, t\}$, k_i is equal to a power of p . (See [Chapter 13, Exercise I6.](#))
 - 4 Solving the equation $|G| = c + k_s + \dots + k_t$ for c , explain why c is a multiple of p .
- We may conclude from part 4 that C must contain more than just the one element e ; in fact, $|C|$ is a multiple of p .
- 5 Prove: If $|G| = p^2$, G must be abelian. (Use the preceding Exercise F.)
- # 6 Prove: If $|G| = p^2$, then either $G \cong \mathbb{Z}_{p^2}$ or $G \cong \mathbb{Z}_p \times \mathbb{Z}_p$.

† H. Induction on $|G|$: An Example

Many theorems of mathematics are of the form “ $P(n)$ is true for every positive integer n .” [Here, $P(n)$ is used as a symbol to denote some statement involving n .] Such theorems can be proved by induction as follows:

- (a) Show that $P(n)$ is true for $n = 1$.
- (b) For any fixed positive integer k , show that, if $P(n)$ is true for every $n < k$, then $P(n)$ must also be true for $n = k$.

If we can show (a) and (b), we may safely conclude that $P(n)$ is true for all positive integers n .

Some theorems of algebra can be proved by induction on the order n of a group. Here is a classical example: Let G be a finite abelian group. We will show that G must contain at least one element of order

p , for every prime factor p of $|G|$. If $|G| = 1$, this is true by default, since no prime p can be a factor of 1. Next, let $|G| = k$, and suppose our claim is true for every abelian group whose order is less than k . Let p be a prime factor of k .

Take any element $a \neq e$ in G . If $\text{ord}(a) = p$ or a multiple of p , we are done!

- 1 If $\text{ord}(a) = tp$ (for some positive integer t), what element of G has order p ?
- 2 Suppose $\text{ord}(a)$ is *not equal to a multiple of p* . Then $G/\langle a \rangle$ is a group having fewer than k elements. (Explain why.) The order of $G/\langle a \rangle$ is a multiple of p . (Explain why.)
- 3 Why must $G/\langle a \rangle$ have an element of order p ?
- 4 Conclude that G has an element of order p . (HINT: Use Exercise El.)



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – III – ALGEBRA-I – SMT1501

Unit-III Rings and Ideals

In presenting scientific knowledge it is elegant as well as enlightening to begin with the simple and move toward the more complex. If we build upon a knowledge of the simplest things, it is easier to understand the more complex ones. In the first part of this book we dedicated ourselves to the study of groups—surely one of the simplest and most fundamental of all algebraic systems. We will now move on, and, using the knowledge and insights gained in the study of groups, we will begin to examine algebraic systems which have *two* operations instead of just one.

The most basic of the two-operational systems is called a *ring*: it will be defined in a moment. The surprising fact about rings is that, despite their having *two* operations and being more complex than groups, their fundamental properties follow exactly the pattern already laid out for groups. With remarkable, almost compelling ease, we will find two-operational analogs of the notions of subgroup and quotient group, homomorphism and isomorphism—as well as other algebraic notions—and we will discover that rings behave just like groups with respect to these notions.

The two operations of a ring are traditionally called *addition* and *multiplication*, and are denoted as usual by $+$ and \cdot , respectively. We must remember, however, that the elements of a ring are not necessarily numbers (for example, there are rings of functions, rings of switching circuits, and so on); and therefore “addition” does not necessarily refer to the conventional addition of numbers, nor does multiplication necessarily refer to the conventional operation of multiplying numbers. In fact, $+$ and \cdot are nothing more than symbols denoting the two operations of a ring.

By a ring we mean a set A with operations called addition and multiplication which satisfy the following axioms:

- (i) *A with addition alone is an abelian group.*
- (ii) *Multiplication is associative.*
- (iii) *Multiplication is distributive over addition. That is, for all a, b , and c in A ,*

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

and

$$(b + c)a = ba + ca$$

Since A with addition alone is an abelian group, there is in A a neutral element for addition: it is called the *zero* element and is written 0 . Also, every element has an additive inverse called its *negative*; the

negative of a is denoted by $-a$. Subtraction is defined by

$$a - b = a + (-b)$$

The easiest examples of rings are the traditional number systems. The set \mathbf{Z} of the integers, with conventional addition and multiplication, is a ring called the *ring of the integers*. We designate this ring simply with the letter \mathbf{Z} . (The context will make it clear whether we are referring to the *ring* of the integers or the additive *group* of the integers.)

Similarly, \mathbf{Q} is the ring of the rational numbers, \mathbf{R} the ring of the real numbers, and \mathbf{C} the ring of the complex numbers. In each case, the operations are conventional addition and multiplication.

Remember that $\mathcal{F}(\mathbf{R})$ represents the set of all the functions from \mathbf{R} to \mathbf{R} ; that is, the set of all real-valued functions of a real variable. In calculus we learned to add and multiply functions: if f and g are any two functions from \mathbf{R} to \mathbf{R} , their sum $f + g$ and their *product* fg are defined as follows:

$$[f + g](x) = f(x) + g(x) \quad \text{for every real number } x$$

and

$$[fg](x) = f(x)g(x) \quad \text{for every real number } x$$

$\mathcal{F}(\mathbf{R})$ with these operations for adding and multiplying functions is a ring called the *ring of real functions*. It is written simply as $\mathcal{F}(\mathbf{R})$. On page 46 we saw that $\mathcal{F}(\mathbf{R})$ with only addition of functions is an abelian group. It is left as an exercise for you to verify that multiplication of functions is associative and distributive over addition of functions.

The rings \mathbf{Z} , \mathbf{Q} , \mathbf{R} , \mathbf{C} , and $\mathcal{F}(\mathbf{R})$ are all *infinite rings*, that is, rings with infinitely many elements. There are also *finite rings*: rings with a finite number of elements. As an important example, consider the group \mathbf{Z}_n , and define an operation of multiplication on \mathbf{Z}_n by allowing the product ab to be the remainder of the usual product of integers a and b after division by n . (For example, in \mathbf{Z}_5 , $2 \cdot 4 = 3$, $3 \cdot 3 = 4$, and $4 \cdot 3 = 2$.) This operation is called *multiplication modulo n* . \mathbf{Z}_n with addition and multiplication modulo n is a ring: the details are given in [Chapter 19](#).

Let A be any ring. Since A with addition alone is an abelian group, everything we know about abelian groups applies to it. However, it is important to remember that A with addition is an abelian group *in additive notation* and, therefore, before applying theorems about groups to A , these theorems must be translated into additive notation. For example, [Theorems 1, 2, and 3 of Chapter 4](#) read as follows when the notation is additive and the group is abelian:

$$a + b = a + c \quad \text{implies} \quad b = c \quad (1)$$

$$a + b = 0 \quad \text{implies} \quad a = -b \quad \text{and} \quad b = -a \quad (2)$$

$$-(a + b) = -(a) + -(b) \quad \text{and} \quad -(-a) = a \quad (3)$$

Therefore Conditions (1), (2), and (3) are true in every ring.

What happens in a ring when we multiply elements by zero? What happens when we multiply elements by the *negatives* of other elements? The next theorem answers these questions.

Theorem 1 *Let a and b be any elements of a ring A .*

$$(i) \ a0 = 0 \quad \text{and} \quad 0a = 0$$

$$(ii) \ a(-b) = -(ab) \quad \text{and} \quad (-a)b = -(ab)$$

$$(iii) \ (-a)(-b) = ab$$

Part (i) asserts that multiplication by zero always yields zero, and parts (ii) and (iii) state the familiar rules of signs.

PROOF: To prove (i) we note that

$$\begin{aligned} aa + 0 &= aa \\ &= a(a + 0) && \text{because } a = a + 0 \\ &= aa + a0 && \text{by the distributive law} \end{aligned}$$

Thus, $aa + 0 = aa + a0$. By Condition (1) above we may eliminate the term aa on both sides of this equation, and therefore $0 = a0$.

To prove (ii), we have

$$\begin{aligned} a(-b) + ab &= a[(-b) + b] && \text{by the distributive law} \\ &= a0 \\ &= 0 && \text{by part (i)} \end{aligned}$$

Thus, $a(-b) + ab = 0$. By Condition (2) above we deduce that $a(-b) = -(ab)$. The twin formula $(-a)b = -(-ab)$ is deduced analogously.

We prove part (iii) by using part (ii) twice:

$$(-a)(-b) = -[a(-b)] = -[-(ab)] = ab \blacksquare$$

The general definition of a ring is sparse and simple. However, particular rings may also have “optional features” which make them more versatile and interesting. Some of these options are described next.

By definition, addition is commutative in every ring but mutiplication is not. When multiplication *also* is commutative in a ring, we call that ring a *commutative ring*.

A ring A does not necessarily have a neutral element for multiplication. If there *is* in A a neutral element for multiplication, it is called the *unity* of A , and is denoted by the symbol 1 . Thus, $a \cdot 1 = a$ and $1 \cdot a = a$ for every a in A . If A has a unity, we call A a *ring with unity*. The rings $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$, and $\mathcal{F}(\mathbb{R})$ are all examples of commutative rings with unity.

Incidentally, a ring whose only element is 0 is called a *trivial ring*; a ring with more than one element is *nontrivial*. In a nontrivial ring with unity, necessarily $1 \neq 0$. This is true because if $1 = 0$ and x is any element of the ring, then

$$x = x1 = x0 = 0$$

In other words, if $1 = 0$ then every element of the ring is equal to 0 ; hence 0 is the only element of the ring.

If A is a ring with unity, there may be elements in A which *have a multiplicative inverse*. Such elements are said to be *invertible*. Thus, an element a is invertible in a ring if there is some x in the ring such that

$$ax = xa = 1$$

For example, in \mathbb{R} every nonzero element is invertible: its multiplicative inverse is its reciprocal. On the other hand, in \mathbb{Z} the only invertible elements are 1 and -1 .

Zero is never an invertible element of a ring except if the ring is trivial; for if zero had a multiplicative inverse x , we would have $0x = 1$, that is, $0 = 1$.

If A is a *commutative ring with unity in which every nonzero element is invertible*, A is called a *field*. Fields are of the utmost importance in mathematics; for example, \mathbb{Q} , \mathbb{R} , and \mathbb{C} are fields. There are also *finite* fields, such as \mathbb{Z}_5 (it is easy to check that every nonzero element of \mathbb{Z}_5 is invertible). Finite fields have beautiful properties and fascinating applications, which will be examined later in this book.

In elementary mathematics we learned the commandment that if the product of two numbers is equal to zero, say

$$ab = 0$$

then one of the two factors, either a or b (or both) must be equal to zero. This is certainly true if the numbers are real (or even complex) numbers, but the rule is *not* inviolable in every ring. For example, in \mathbb{Z}_6 ,

$$2 \cdot 3 = 0$$

even though the factors 2 and 3 are both nonzero. Such numbers, when they exist, are called *divisors of zero*.

*In any ring, a nonzero element a is called a **divisor of zero** if there is a nonzero element b in the ring such that the product ab or ba is equal to zero.*

(Note carefully that *both* factors have to be nonzero.) Thus, 2 and 3 are divisors of zero in \mathbb{Z}_6 ; 4 is also a divisor of zero in \mathbb{Z}_6 , because $4 \cdot 3 = 0$. For another example, let $\mathcal{M}_2(\mathbb{R})$ designate the set of all 2×2 matrices of real numbers, with addition and multiplication of matrices as described on page 8. The simple task of checking that $\mathcal{M}_2(\mathbb{R})$ satisfies the ring axioms is assigned as [Exercise C1](#) at the end of this chapter. $\mathcal{M}_2(\mathbb{R})$ is rampant with examples of divisors of zero. For instance,

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

hence

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

are both divisors of zero in $\mathcal{M}_2(\mathbb{R})$.

Of course, there are rings which have no divisors of zero at all! For example, \mathbb{Z} , \mathbb{Q} , \mathbb{R} , and \mathbb{C} do not have any divisors of zero. It is important to note carefully what it means for a ring to have *no divisors of zero*: it means that *if the product of two elements in the ring is equal to zero, at least one of the factors is zero*. (Our commandment from elementary mathematics!)

It is also decreed in elementary algebra that a nonzero number a may be canceled in the equation $ax = ay$ to yield $x = y$. While undeniably true in the number systems of mathematics, this rule is not true in every ring. For example, in \mathbb{Z}_6 ,

yet we cannot cancel the common factor 2. A similar example involving 2×2 matrices may be seen on page 9. When cancellation *is* possible, we say the ring has the “cancellation property.”

*A ring is said to have the **cancellation property** if*

$$ab = ac \quad \text{or} \quad ba = ca \quad \text{implies} \quad b = c$$

for any elements a , b , and c in the ring if $a \neq 0$.

There is a surprising and unexpected connection between the cancellation property and divisors of zero:

Theorem 2 *A ring has the cancellation property iff it has no divisors of zero.*

PROOF: The proof is very straightforward. Let A be a ring, and suppose first that A has the cancellation property. To prove that A has no divisors of zero we begin by letting $ab = 0$, and show that a or b is equal to 0. If $a = 0$, we are done. Otherwise, we have

$$ab = 0 = a0$$

so by the cancellation property (cancelling a), $b=0$.

Conversely, assume A has no divisors of zero. To prove that A has the cancellation property, suppose $ab = ac$ where $a \neq 0$. Then

$$ab - ac = a(b - c) = 0$$

Remember, there are no divisors of zero! Since $a \neq 0$, necessarily $b - c = 0$, so $b = c$. ■

An *integral domain* is defined to be a commutative ring with unity having the cancellation property. By [Theorem 2](#), an integral domain may also be defined as a commutative ring with unity having no divisors of zero. It is easy to see that every field is an integral domain. The converse, however, is not true: for example, \mathbb{Z} is an integral domain but not a field. We will have a lot to say about integral domains in the following chapters.

EXERCISES

A. Examples of Rings

In each of the following, a set A with operations of addition and multiplication is given. *Prove that A satisfies all the axioms to be a commutative ring with unity. Indicate the zero element, the unity, and the negative of an arbitrary a .*

1 A is the set \mathbb{Z} of the integers, with the following “addition” \oplus and “multiplication” \odot :

$$a \oplus b = a + b - 1 \qquad a \odot b = ab - (a + b) + 2$$

2 A is the set \mathbb{Q} of the rational numbers, and the operations are \oplus and \odot defined as follows:

$$a \oplus b = a + b + 1 \qquad a \odot b = ab + a + b$$

3 A is the set $\mathbb{Q} \times \mathbb{Q}$ of ordered pairs of rational numbers, and the operations are the following addition

\oplus and multiplication \odot :

$$(a, b) \oplus (c, d) = (a + c, b + d)$$

$$(a, b) \odot (c, d) = (ac - bd, ad + bc)$$

4 $A = \{x + y\sqrt{2} : x, y \in \mathbb{Z}\}$ with conventional addition and multiplication.

5 Prove that the ring in part 1 is an integral domain.

6 Prove that the ring in part 2 is a field, and indicate the multiplicative inverse of an arbitrary nonzero element.

7 Do the same for the ring in part 3.

B. Ring of Real Functions

1 Verify that $\mathcal{F}(\mathbb{R})$ satisfies all the axioms for being a commutative ring with unity. Indicate the zero and unity, and describe the negative of any f .

2 Describe the divisors of zero in $\mathcal{F}(\mathbb{R})$.

3 Describe the invertible elements in $\mathcal{F}(\mathbb{R})$.

4 Explain why $\mathcal{F}(\mathbb{R})$ is neither a field nor an integral domain.

C. Ring of 2×2 Matrices

Let $\mathcal{M}_2(\mathbb{R})$ designate the set of all 2×2 matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

whose entries are real numbers a, b, c , and d , with the following addition and multiplication:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} r & s \\ t & u \end{pmatrix} = \begin{pmatrix} a + r & b + s \\ c + t & d + u \end{pmatrix}$$

and

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} r & s \\ t & u \end{pmatrix} = \begin{pmatrix} ar + bt & as + bu \\ cr + dt & cs + du \end{pmatrix}$$

1 Verify that $\mathcal{M}_2(\mathbb{R})$ satisfies the ring axioms.

2 Show that $\mathcal{M}_2(\mathbb{R})$ is not commutative and has a unity.

3 Explain why $\mathcal{M}_2(\mathbb{R})$ is not an integral domain or a field.

D. Rings of Subsets of a Set

If D is a set, then the power set of D is the set P_D of all the subsets of D . Addition and multiplication are defined as follows: If A and B are elements of P_D (that is, subsets of D), then

$$A + B = (A - B) \cup (B - A) \quad \text{and} \quad AB = A \cap B$$

It was shown in [Chapter 3, Exercise C](#), that P_D with addition alone is an abelian group.

- # **1** Prove: P_D is a commutative ring with unity. (You may assume \cap is associative; for the distributive law, use the same diagram and approach as was used to prove that addition is associative in [Chapter 3, Exercise C.](#))
- 2** Describe the divisors of zero in P_D .
- 3** Describe the invertible elements in P_D .
- 4** Explain why P_D is neither a field nor an integral domain. (Assume D has more than one element.)
- 5** Give the tables of P_3 , that is, P_D where $D = \{a, b, c\}$.

E. Ring of Quaternions

A *quaternion* (in matrix form) is a 2×2 matrix of complex numbers of the form

$$\alpha = \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix}$$

- 1** Prove that the set of all the quaternions, with the matrix addition and multiplication explained on pages 7 and 8, is a ring with unity. This ring is denoted by the symbol \mathfrak{Q} . Find an example to show that \mathfrak{Q} is not commutative. (You may assume matrix addition and multiplication are associative and obey the distributive law.)
- 2** Let

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \quad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

Show that the quaternion α , defined previously, may be written in the form

$$\alpha = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$$

(This is the standard notation for quaternions.)

3 Prove the following formulas:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -\mathbf{1} \quad \mathbf{ij} = -\mathbf{ji} = \mathbf{k} \quad \mathbf{jk} = -\mathbf{kj} = \mathbf{i} \quad \mathbf{ki} = -\mathbf{ik} = \mathbf{j}$$

4 The *conjugate* of α is

$$\bar{\alpha} = \begin{pmatrix} a - bi & -c - di \\ c - di & a + bi \end{pmatrix}$$

The *norm* of α is $a^2 + b^2 + c^2 + d^2$, and is written $\|\alpha\|$. Show directly (by matrix multiplication) that

$$\bar{\alpha}\alpha = \alpha\bar{\alpha} = \begin{pmatrix} t & 0 \\ 0 & t \end{pmatrix} \quad \text{where } t = \|\alpha\|$$

Conclude that the multiplicative inverse of α is $(1/t)\bar{\alpha}$.

5 A *skew field* is a (not necessarily commutative) ring with unity in which every nonzero element has a multiplicative inverse. Conclude from parts 1 and 4 that \mathfrak{Q} is a skew field.

F. Ring of Endomorphisms

Let G be an abelian group in additive notation. An *endomorphism* of G is a homomorphism from G to G . Let $\text{End}(G)$ denote the set of all the endomorphisms of G , and define addition and multiplication of endomorphisms as follows:

$$\begin{aligned}[f + g](x) &= f(x) + g(x) && \text{for every } x \text{ in } G \\ fg &= f \circ g && \text{the composite of } f \text{ and } g\end{aligned}$$

- 1 Prove that $\text{End}(G)$ with these operations is a ring with unity.
- 2 List the elements of $\text{End}(\mathbf{z}_4)$, then give the addition and multiplication tables for $\text{End}(\mathbf{z}_4)$.

REMARK: The endomorphisms of \mathbf{z}_4 are easy to find. Any endomorphisms of \mathbf{z}_4 will carry 1 to either 0, 1, 2, or 3. For example, take the last case: if

$$1 \xrightarrow{f} 3$$

then necessarily

$$1 + 1 \xrightarrow{f} 3 + 3 = 2 \quad 1 + 1 + 1 \xrightarrow{f} 3 + 3 + 3 = 1 \quad \text{and} \quad 0 \xrightarrow{f} 0$$

hence f is completely determined by the fact that

$$1 \xrightarrow{f} 3$$

G. Direct Product of Rings

If A and B are rings, their *direct product* is a new ring, denoted by $A \times B$, and defined as follows: $A \times B$ consists of all the ordered pairs (x, y) where x is in A and y is in B . Addition in $A \times B$ consists of adding corresponding components:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2)$$

Multiplication in $A \times B$ consists of multiplying corresponding components:

$$(x_1, y_1) \cdot (x_2, y_2) = (x_1 x_2, y_1 y_2)$$

- 1 If A and B are rings, verify that $A \times B$ is a ring.
- 2 If A and B are commutative, show that $A \times B$ is commutative. If A and B each has a unity, show that $A \times B$ has a unity.
- 3 Describe carefully the divisors of zero in $A \times B$.
- # 4 Describe the invertible elements in $A \times B$.
- 5 Explain why $A \times B$ can never be an integral domain or a field. (Assume A and B each have more than one element.)

H. Elementary Properties of Rings

Prove parts 1–4:

- 1 In any ring, $a(b - c) = ab - ac$ and $(b - c)a = ba - ca$.
- 2 In any ring, if $ab = -ba$, then $(a + b)^2 = (a - b)^2 = a^2 + b^2$.

- 3 In any integral domain, if $a^2 = b^2$, then $a = \pm b$.
- 4 In any integral domain, only 1 and -1 are their own multiplicative inverses. (Note that $x = x^{-1}$ iff $x^2 = 1$.)
- 5 Show that the commutative law for addition need not be assumed in defining a ring with unity: it may be proved from the other axioms. [HINT: Use the distributive law to expand $(a + b)(1 + 1)$ in two different ways.]
- # 6 Let A be any ring. Prove that if the additive group of A is cyclic, then A is a commutative ring.
- 7 Prove: In any integral domain, if $a^n = 0$ for some integer n , then $a = 0$.

I. Properties of Invertible Elements

Prove that parts 1–5 are true in a nontrivial ring with unity.

- 1 If a is invertible and $ab = ac$, then $b = c$.
- 2 An element a can have no more than *one* multiplicative inverse.
- 3 If $a^2 = 0$ then $a + 1$ and $a - 1$ are invertible.
- 4 If a and b are invertible, their product ab is invertible.
- 5 The set S of all the invertible elements in a ring is a multiplicative group.
- 6 By part 5, the set of all the nonzero elements in a field is a multiplicative group. Now use Lagrange's theorem to prove that in a finite field with m elements, $x^{m-1} = 1$ for every $x \neq 0$.
- 7 If $ax = 1$, x is a *right inverse* of a ; if $ya = 1$, y is a *left inverse* of a . Prove that if a has a right inverse y and a left inverse x , then a is invertible, and its inverse is equal to x and to y . (First show that $yaxa = 1$.)
- 8 Prove: In a commutative ring, if ab is invertible, then a and b are both invertible.

J. Properties of Divisors of Zero

Prove that each of the following is true in a nontrivial ring.

- 1 If $a \neq \pm 1$ and $a^2 = 1$, then $a + 1$ and $a - 1$ are divisors of zero.
- # 2 If ab is a divisor of zero, then a or b is a divisor of zero.
- 3 In a commutative ring with unity, a divisor of zero cannot be invertible.
- 4 Suppose $ab \neq 0$ in a commutative ring. If either a or b is a divisor of zero, so is ab .
- 5 Suppose a is neither 0 nor a divisor of zero. If $ab = ac$, then $b = c$.
- 6 $A \times B$ always has divisors of zero.

K. Boolean Rings

A ring A is a boolean ring if $a^2 = a$ for every $a \in A$. Prove that parts 1 and 2 are true in any boolean ring A .

- 1 For every $a \in A$, $a = -a$. [HINT: Expand $(a + a)^2$.]
- 2 Use part 1 to prove that A is a commutative ring. [HINT: Expand $(a + b)^2$.]

In parts 3 and 4, assume A has a unity and prove:

- 3 Every element except 0 and 1 is a divisor of zero. [Consider $x(x - 1)$.]
- 4 1 is the only invertible element in A .
- 5 Letting $a \vee b = a + b + ab$ we have the following in A :

$$a \vee bc = (a \vee b)(a \vee c) \quad a \vee (1 + a) = 1 \quad a \vee a = a \quad a(a \vee b) = a$$

L. The Binomial Formula

An important formula in elementary algebra is the binomial expansion formula for an expression $(a + b)^n$. The formula is as follows:

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$$

where the binomial coefficient

$$\binom{n}{k} = \frac{n(n-1)(n-2) \cdots (n-k+1)}{k!}$$

This theorem is true in every commutative ring. (If K is any positive integer and a is an element of a ring, ka refers to the sum $a + a + \cdots + a$ with k terms, as in elementary algebra.) The proof of the binomial theorem in a commutative ring is no different from the proof in elementary algebra. We shall review it here.

The proof of the binomial formula is by induction on the exponent n . The formula is trivially true for $n = 1$. In the induction step, we *assume* the expansion for $(a + b)^n$ is as above, and we must prove that

$$(a + b)^{n+1} = \sum_{k=0}^{n+1} \binom{n+1}{k} a^{n+1-k} b^k$$

Now,

$$\begin{aligned} (a + b)^{n+1} &= (a + b)(a + b)^n \\ &= (a + b) \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k \\ &= \sum_{k=0}^n \binom{n}{k} a^{n+1-k} b^k + \sum_{k=0}^n \binom{n}{k} a^{n-k} b^{k+1} \end{aligned}$$

Collecting terms, we find that the coefficient of $a^{n+1-k} b^k$ is

$$\binom{n}{k} + \binom{n}{k-1}$$

By direct computation, show that

$$\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}$$

It will follow that $(a + b)^{n+1}$ is as claimed, and the proof is complete.

M. Nilpotent and Unipotent Elements

An element a of a ring is *nilpotent* if $a^n = 0$ for some positive integer n .

1 In a ring with unity, prove that if a is nilpotent, then $a + 1$ and $a - 1$ are both invertible. [HINT: Use the

factorization

$$1 - a^n = (1 - a)(1 + a + a^2 + \dots + a^{n-1})$$

for $1 - a$, and a similar formula for $1 + a$.]

2 In a commutative ring, prove that any product xa of a nilpotent element a by any element x is nilpotent.

3 In a commutative ring, prove that the sum of two nilpotent elements is nilpotent. (HINT: You must use the binomial formula; see [Exercise L](#).)

An element a of a ring is *unipotent* iff $1 - a$ is nilpotent.

4 In a commutative ring, prove that the product of two unipotent elements a and b is unipotent. [HINT: Use the binomial formula to expand $1 - ab = (1 - a) + a(1 - b)$ to power $n + m$.]

5 In a ring with unity, prove that every unipotent element is invertible. (HINT: Use Part 1.)

IDEALS AND HOMOMORPHISMS

We have already seen several examples of smaller rings contained within larger rings. For example, \mathbb{Z} is a ring inside the larger ring \mathbb{Q} , and \mathbb{Q} itself is a ring inside the larger ring \mathbb{R} . When a ring B is part of a larger ring A , we call B a *subring* of A . The notion of subring is the precise analog for rings of the notion of subgroup for groups. Here are the relevant definitions:

Let A be a ring, and B a nonempty subset of A . If the sum of any two elements of B is again in B , then B is *closed with respect to addition*. If the negative of every element of B is in B , then B is *closed with respect to negatives*. Finally, if the product of any two elements of B is again in B , then B is *closed with respect to multiplication*. B is called a *subring* of A if B is closed with respect to addition, multiplication, and negatives. Why is B then called a subring of A ? Quite elementary:

If a nonempty subset $B \subseteq A$ is closed with respect to addition, multiplication, and negatives, then B with the operations of A is a ring.

This fact is easy to check: If a , b , and c are any three elements of B , then a , b , and c are also elements of A because $B \subseteq A$. But A is a ring, so

$$a + (b + c) = (a + b) + c$$

$$a(bc) = (ab)c$$

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

and

$$(b + c)a = ba + ca$$

Thus, in B addition and multiplication are associative and the distributive law is satisfied. Now, B was assumed to be nonempty, so there is an element $b \in B$ but B is closed with respect to negatives, so $-b$ is also in B . Finally, B is closed with respect to addition; hence $b + (-b) \in B$. That is, 0 is in B . Thus, B satisfies all the requirements for being a ring.

For example, \mathbb{Q} is a subring of \mathbb{R} because the sum of two rational numbers is rational, the product of two rational numbers is rational, and the negative of every rational number is rational.

By the way, if B is a nonempty subset of A , there is a more compact way of checking that B is a subring of A :

B is a subring of A if and only if B is closed with respect to subtraction and multiplication.

The reason is that B is *closed with respect to subtraction* iff B is *closed with respect to both addition and negatives*. This last fact is easy to check, and is given as an exercise.

Awhile back, in our study of groups, we singled out certain special subgroups called *normal subgroups*. We will now describe certain special subrings called *ideals* which are the counterpart of normal subgroups: that is, ideals are in rings as normal subgroups are in groups.

Let A be a ring, and B a nonempty subset of A . We will say that B *absorbs products in A* (or, simply, B *absorbs products*) if, whenever we multiply an element in B by an element in A (regardless of whether the latter is inside B or outside B), their product is always in B . In other words,

$$\text{for all } b \in B \text{ and } x \in A, xb \text{ and } bx \text{ are in } B.$$

A nonempty subset B of a ring A is called an ideal of A if B is closed with respect to addition and negatives, and B absorbs products in A .

A simple example of an ideal is the set \mathbb{E} of the even integers. \mathbb{E} is an ideal of \mathbb{Z} because the sum of two even integers is even, the negative of any even integer is even, and, finally, the product of an even integer by *any* integer is always even.

In a commutative ring with unity, the simplest example of an ideal is the set of all the multiples of a fixed element a by all the elements in the ring. In other words, the set of all the products

$$xa$$

as a remains fixed and x ranges over all the elements of the ring. This set is obviously an ideal because

$$xa + ya = (x + y)a$$

$$-(xa) = (-x)a$$

and

$$y(xa) = (yx)a$$

This ideal is called the *principal ideal generated by a* , and is denoted by

$$\langle a \rangle$$

As in the case of subrings, if B is a nonempty subset of A , there is a more compact way of checking that B is an ideal of A :

B is an ideal of A if and only if B is closed with respect to subtraction and B absorbs products in A .

We shall see presently that ideals play an important role in connection with homomorphism

Homomorphisms are almost the same for rings as for groups.

A homomorphism from a ring A to a ring B is a function $f: A \rightarrow B$ satisfying the identities

$$f(x_1 + x_2) = f(x_1) + f(x_2)$$

and

$$f(x_1x_2) = f(x_1)f(x_2)$$

There is a longer but more informative way of writing these two identities:

1. If $f(x_1) = y_1$ and $f(x_2) = y_2$ then $f(x_1 + x_2) = y_1 + y_2$.
2. If $f(x_1) = y_1$ and $f(x_2) = y_2$ then $f(x_1 x_2) = y_1 y_2$

In other words, if f happens to carry x_1 to y_1 and x_2 to y_2 , then, necessarily, it must carry $x_1 + x_2$ to $y_1 + y_2$ and $x_1 x_2$ to $y_1 y_2$. Symbolically,

If $x_1 \xrightarrow{f} y_1$ and $x_2 \xrightarrow{f} y_2$, then necessarily

$$x_1 + x_2 \xrightarrow{f} y_1 + y_2 \quad \text{and} \quad x_1 x_2 \xrightarrow{f} y_1 y_2$$

One can easily confirm for oneself that a function f with this property will transform the addition and multiplication tables of its domain into the addition and multiplication tables of its range. (We may imagine infinite rings to have “nonterminating” tables.) Thus, a homomorphism from a ring A onto a ring B is a function which transforms A into B .

For example, the ring \mathbf{z}_6 is transformed into the ring \mathbf{z}_3 by

$$f = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 2 & 0 & 1 & 2 \end{pmatrix}$$

as we may verify by comparing their tables. The addition tables are compared on page 136, and we may do the same with their multiplication tables:

\cdot	0	1	2	3	4	5		\cdot	0	1	2	0	1	2
0	0	0	0	0	0	0	Replace x by $f(x)$ \longrightarrow	0	0	0	0	0	0	0
1	0	1	2	3	4	5		1	0	1	2	0	1	2
2	0	2	4	0	2	4		2	0	2	1	0	2	1
3	0	3	0	3	0	3		0	0	0	0	0	0	0
4	0	4	2	0	1	2		1	0	1	2	0	1	2
5	0	5	4	3	2	1		2	0	2	1	0	2	1

Eliminate duplicate information \longrightarrow (For example, $2 \cdot 2 = 1$ appears four separate times in table above.)			\cdot	0	1	2
			0	0	0	0
			1	0	1	2
			2	0	2	1

If there is a homomorphism from A onto B , we call B a *homomorphic image of A* . If f is a homomorphism from a ring A to a ring B , not necessarily *onto*, the range of f is a subring of B . (This fact is routine to verify.) Thus, the range of a ring homomorphism is always a ring. And obviously, the range of a homomorphism is always a homomorphic image of its domain.

Intuitively, if B is a homomorphic image of A , this means that certain features of A are faithfully

preserved in B while others are deliberately lost. This may be illustrated by developing further an example described in [Chapter 14](#). The *parity ring* P consists of two elements, e and o , with addition and multiplication given by the tables

$+$	e	o
e	e	o
o	o	e

and

\cdot	e	o
e	e	e
o	e	o

We should think of e as “even” and o as “odd,” and the tables as describing the rules for adding and multiplying odd and even integers. For example, even + odd = odd, even *times* odd = even, and so on.

The function $f: \mathbf{z} \rightarrow P$ which carries every even integer to e and every odd integer to o is easily seen to be a homomorphism from \mathbf{z} to P this is made clear on page 137. Thus, P is a homomorphic image of \mathbf{z} . Although the ring P is very much smaller than the ring \mathbf{z} , and therefore few of the features of \mathbf{z} can be expected to reappear in P , nevertheless *one* aspect of the structure of \mathbf{z} is retained absolutely intact in P , namely, the structure of odd and even numbers. As we pass from \mathbf{z} to P , the *parity* of the integers (their being even or odd), with its arithmetic, is faithfully preserved while all else is lost. Other examples will be given in the exercises.

If f is a homomorphism from a ring A to a ring B , the *kernel* of f is the set of all the elements of A which are carried by f onto the zero element of B . In symbols, the kernel of f is the set

$$K = \{x \in A : f(x) = 0\}$$

It is a very important fact that the *kernel of f is an ideal of A* . (The simple verification of this fact is left as an exercise.)

If A and B are rings, an *isomorphism* from A to B is a homomorphism which is a one-to-one correspondence from A to B . In other words, it is an injective and surjective homomorphism. If there *is* an isomorphism from A to B we say that A is *isomorphic to B* , and this fact is expressed by writing

$$A \cong B$$

EXERCISES

A. Examples of Subrings

Prove that each of the following is a subring of the indicated ring:

- 1 $\{x + \sqrt{3}y : x, y \in \mathbf{z}\}$ is a subring of \mathbf{R} .
- 2 $\{x + 2^{1/3}y + 2^{2/3}z : x, y, z \in \mathbf{z}\}$ is a subring of \mathbf{R} .
- 3 $\{x2^y : x, y \in \mathbf{z}\}$ is a subring of \mathbf{R} .
- # 4 Let $\mathcal{C}(\mathbf{R})$ be the set of all the functions from \mathbf{R} to \mathbf{R} which are continuous on $(-\infty, \infty)$ and let $\mathcal{D}(\mathbf{R})$ be the set of all the functions from \mathbf{R} to \mathbf{R} which are differentiable on $(-\infty, \infty)$. Then $\mathcal{C}(\mathbf{R})$ and $\mathcal{D}(\mathbf{R})$ are subrings of $\mathcal{F}(\mathbf{R})$.
- 5 Let $\mathcal{U}(\mathbf{z})$ be the set of all functions from \mathbf{R} to \mathbf{R} which are continuous on the interval $[0,1]$. Then $\mathcal{U}(\mathbf{R})$ is a subring of $\mathcal{F}(\mathbf{R})$, and $\mathcal{C}(\mathbf{R})$ is a subring of $\mathcal{U}(\mathbf{R})$.
- 6 The subset of $\mathcal{M}_2(\mathbf{R})$ consisting of all matrices of the form

$$\begin{pmatrix} 0 & 0 \\ 0 & x \end{pmatrix}$$

is a subring of $M_2(\mathbb{R})$.

B. Examples of Ideals

1 Identify which of the following are ideals of $\mathbb{Z} \times \mathbb{Z}$, and explain: $\{(n, n) : n \in \mathbb{Z}\}$; $\{(5n, 0) : n \in \mathbb{Z}\}$; $\{(n, m) : n + m \text{ is even}\}$; $\{(n, m) : nm \text{ is even}\}$; $\{(2n, 3m) : n, m \in \mathbb{Z}\}$.

2 List all the ideals of \mathbb{Z}_{12} .

3 Explain why every subring of \mathbb{Z}_n is necessarily an ideal.

4 Explain why the subring of [Exercise A6](#) is not an ideal.

5 Explain why $\mathcal{C}(\mathbb{R})$ is not an ideal of $\mathcal{F}(\mathbb{R})$.

6 Prove that each of the following is an ideal of $\mathcal{F}(\mathbb{R})$:

(a) The set of all f such that $f(x) = 0$ for every rational x .

(b) The set of all f such that $f(0) = 0$.

7 List all the ideals of P_3 . (P_3 is defined in [Chapter 17, Exercise D](#).)

8 Give an example of a subring of P_3 which is not an ideal.

9 Give an example of a subring of $\mathbb{Z}_3 \times \mathbb{Z}_3$ which is not an ideal.

C. Elementary Properties of Subrings

Prove parts 1–6:

1 A nonempty subset B of a ring A is closed with respect to addition and negatives iff B is closed with respect to subtraction.

2 Conclude from part 1 that B is a subring of A iff B is closed with respect to subtraction and multiplication.

3 If A is a finite ring and B is a subring of A , then the order of B is a divisor of the order of A .

4 If a subring B of an integral domain A contains 1, then B is an integral domain. (B is then called a *subdomain* of A .)

5 Every subring containing the unity of a field is an integral domain.

6 If a subring B of a field F is closed with respect to multiplicative inverses, then B is a field. (B is then called a *subfield* of F .)

7 Find subrings of \mathbb{Z}_{18} which illustrate each of the following:

(a) A is a ring with unity, B is a subring of A , but B is not a ring with unity.

(b) A and B are rings with unity, B is a subring of A , but the unity of B is not the same as the unity of A .

8 Let A be a ring, $f: A \rightarrow A$ a homomorphism, and $B = \{x \in A : f(x) = x\}$. Prove that B is a subring of A .

9 The *center* of a ring A is the set of all the elements $a \in A$ such that $ax = xa$ for every $x \in A$. Prove that the center of A is a subring of A .

D. Elementary Properties of Ideals

Let A be a ring and J a nonempty subset of A .

- 1 Using [Exercise C1](#), explain why J is an ideal of A iff J is closed with respect to subtraction and J absorbs products in A .
- 2 If A is a ring with unity, prove that J is an ideal of A iff J is closed with respect to addition and J absorbs products in A .
- 3 Prove that the intersection of any two ideals of A is an ideal of A .
- 4 Prove that if J is an ideal of A and $1 \in J$, then $J = A$.
- 5 Prove that if J is an ideal of A and J contains an invertible element a of A , then $J = A$.
- 6 Explain why a field F can have no nontrivial ideals (that is, no ideals except $\{0\}$ and F).

E. Examples of Homomorphisms

Prove that each of the functions in parts 1–6 is a homomorphism. Then describe its kernel and its range.

1 $\phi : \mathcal{F}(\mathbb{R}) \rightarrow \mathbb{R}$ given by $\phi(f) = f(0)$.

2 $h : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ given by $h(x, y) = x$.

3 $h : \mathbb{R} \rightarrow \mathcal{M}_2(\mathbb{R})$ given by

$$h(x) = \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$$

4 $h : \mathbb{R} \times \mathbb{R} \rightarrow \mathcal{M}_2(\mathbb{R})$ given by

$$h(x, y) = \begin{pmatrix} x & 0 \\ 0 & y \end{pmatrix}$$

5 Let A be the set $\mathbb{R} \times \mathbb{R}$ with the usual addition and the following “multiplication”:

$$(a, b) \odot (c, d) = (ac, bc)$$

Granting that A is a ring, let $f : A \rightarrow \mathcal{M}_2(\mathbb{R})$ be given by

$$f(x, y) = \begin{pmatrix} x & 0 \\ y & 0 \end{pmatrix}$$

6 $h : P_c \rightarrow P_c$ given by $h(A) = A \cap D$, where D is a fixed subset of C .

7 List all the homomorphisms from \mathbb{Z}_2 to \mathbb{Z}_4 ; from \mathbb{Z}_3 to \mathbb{Z}_6 .

F. Elementary Properties of Homomorphisms

Let A and B be rings, and $f : A \rightarrow B$ a homomorphism. Prove each of the following:

1 $f(A) = \{f(x) : x \in A\}$ is a subring of B .

2 The kernel of f is an ideal of A .

3 $f(0) = 0$, and for every $a \in A$, $f(-a) = -f(a)$.

4 f is injective iff its kernel is equal to $\{0\}$.

5 If B is an integral domain, then either $f(1) = 1$ or $f(1) = 0$. If $f(1) = 0$, then $f(x) = 0$ for every $x \in A$. If $f(1) = 1$, the image of every invertible element of A is an invertible element of B .

6 Any homomorphic image of a commutative ring is a commutative ring. Any homomorphic image of a field is a field.

7 If the domain A of the homomorphism f is a field, and if the range of f has more than one element, then f is injective. (HINT: Use [Exercise D6](#).)

G. Examples of Isomorphisms

1 Let A be the ring of [Exercise A2](#) in [Chapter 17](#). Show that the function $f(x) = x - 1$ is an isomorphism from \mathbb{Q} to A hence $\mathbb{Q} \cong A$.

2 Let \mathcal{S} be the following subset of $M_2(\mathbb{R})$:

$$\mathcal{S} = \left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} : a, b \in \mathbb{R} \right\}$$

Prove that the function

$$f(a + bi) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

is an isomorphism from \mathbb{C} to \mathcal{S} . [REMARK: You must begin by checking that f is a well-defined function; that is, if $a + bi = c + di$, then $f(a + bi) = f(c + di)$. To do this, note that if $a + bi = c + di$ then $a - c = (d - b)i$; this last equation is impossible unless both sides are equal to zero, for otherwise it would assert that a given real number is equal to an imaginary number.]

3 Prove that $\{(x, x) : x \in \mathbb{Z}\}$ is a subring of $\mathbb{Z} \times \mathbb{Z}$, and show $\{(x, x) : x \in \mathbb{Z}\} \cong \mathbb{Z}$.

4 Show that the set of all 2×2 matrices of the form

$$\begin{pmatrix} 0 & 0 \\ 0 & x \end{pmatrix}$$

is a subring of $M_2(\mathbb{R})$, then prove this subring is isomorphic to \mathbb{R} .

For any integer k , let $k\mathbb{Z}$ designate the subring of \mathbb{Z} which consists of all the multiples of k .

5 Prove that $\mathbb{Z} \not\subseteq 2\mathbb{Z}$ then prove that $2\mathbb{Z} \not\subseteq 3\mathbb{Z}$. Finally, explain why if $k \neq l$, then $k\mathbb{Z} \not\subseteq l\mathbb{Z}$. (REMEMBER: How do you show that two rings, or groups, are *not* isomorphic?)

H. Further Properties of Ideals

Let A be a ring, and let J and K be ideals of A .

Prove parts 1-4. (In parts 2-4, assume A is a commutative ring.)

1 If $J \cap K = \{0\}$, then $jk = 0$ for every $j \in J$ and $k \in K$.

2 For any $a \in A$, $I_a = \{ax + j + k : x \in A, j \in J, k \in K\}$ is an ideal of A .

3 The *radical* of J is the set $\text{rad } J = \{a \in A : a^n \in J \text{ for some } n \in \mathbb{Z}\}$. For any ideal J , $\text{rad } J$ is an ideal of A .

4 For any $a \in A$, $\{x \in A : ax = 0\}$ is an ideal (called the *annihilator* of a).

Furthermore, $\{x \in A : ax = 0 \text{ for every } a \in A\}$ is an ideal (called the *annihilating ideal* of A). If A is a ring with unity, its annihilating ideal is equal to $\{0\}$.

5 Show that $\{0\}$ and A are ideals of A . (They are *trivial* ideals; every other ideal of A is a *proper* ideal.)

A proper ideal J of A is called *maximal* if it is not strictly contained in any strictly larger proper ideal: that is, if $J \subseteq K$, where K is an ideal containing some element not in J , then necessarily $K = A$. Show that the following is an example of a maximal ideal: In $\mathcal{F}(\mathbb{R})$, the ideal $J = \{f : f(0) = 0\}$. [HINT: Use [Exercise D5](#). Note that if $g \in K$ and $g(0) \neq 0$ (that is, $g \notin J$), then the function $h(x) = g(x) - g(0)$ is in J hence $h(x) - g(x) \in K$. Explain why this last function is an invertible element of $\mathcal{F}(\mathbb{R})$.]

I. Further Properties of Homomorphisms

Let A and B be rings. Prove each of the following:

- 1 If $f: A \rightarrow B$ is a homomorphism from A onto B with kernel K , and J is an ideal of A such that $K \cap J$ then $f(J)$ is an ideal of B .
- 2 If $f: A \rightarrow B$ is a homomorphism from A onto B , and B is a *field*, then the kernel of f is a maximal ideal. (HINT: Use part 1, with [Exercise D6](#). Maximal ideals are defined in [Exercise H5](#).)
- 3 There are no nontrivial homomorphisms from \mathbb{Z} to \mathbb{Z} . [The trivial homomorphisms are $f(x) = 0$ and $f(x) = x$.]
- 4 If n is a multiple of m , then \mathbb{Z}_m is a homomorphic image of \mathbb{Z}_n .
- 5 If n is odd, there is an injective homomorphism from \mathbb{Z}_2 into \mathbb{Z}_{2n} .

† J. A Ring of Endomorphisms

Let A be a commutative ring. Prove each of the following:

- 1 For each element a in A , the function π_a defined by $\pi_a(x) = ax$ satisfies the identity $\pi_a(x + y) = \pi_a(x) + \pi_a(y)$. (In other words, π_a is an endomorphism of the additive group of A .)
- 2 π_a is injective iff a is not a divisor of zero. (Assume $a \neq 0$.)
- 3 π_a is surjective iff a is invertible. (Assume A has a unity.)
- 4 Let \mathcal{A} denote the set $\{\pi_a : a \in A\}$ with the two operations

$$[\pi_a + \pi_b](x) = \pi_a(x) + \pi_b(x) \quad \text{and} \quad \pi_a \pi_b = \pi_a \circ \pi_b$$

Verify that \mathcal{A} is a ring.

- 5 If $\phi: A \rightarrow \mathcal{A}$ is given by $\phi(a) = \pi_a$, then ϕ is a homomorphism.
- 6 If A has a unity, then ϕ is an isomorphism. Similarly, if A has no divisors of zero then ϕ is an isomorphism.

QUOTIENT RINGS

We continue our journey into the elementary theory of rings, traveling a road which runs parallel to the familiar landscape of groups. In our study of groups we discovered a way of actually *constructing* all the homomorphic images of any group G . We constructed quotient groups of G , and showed that every quotient group of G is a homomorphic image of G . We will now imitate this procedure and construct *quotient rings*.

We begin by defining cosets of rings:

Let A be a ring, and J an ideal of A . For any element $a \in A$, the symbol $J + a$ denotes the set of all sums $j + a$, as a remains fixed and j ranges over J . That is,

$$J + a = \{j + a : j \in J\}$$

*$J + a$ is called a **coset** of J in A .*

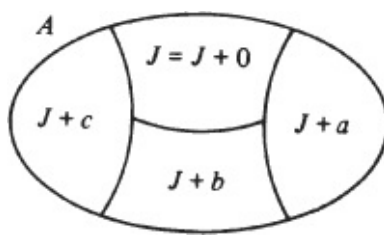
It is important to note that, if we provisionally ignore multiplication, A with addition alone is an abelian group and J is a subgroup of A . Thus, the cosets we have just defined are (if we ignore multiplication) *precisely the cosets of the subgroup J in the group A , with the notation being additive*. Consequently, everything we already know about group cosets continues to apply in the present case—only, care must be taken to translate known facts about group cosets into *additive notation*. For example, Property (1) of [Chapter 13](#), with [Theorem 5](#) of [Chapter 15](#), reads as follows in additive notation:

$$a \in J + b \quad \text{iff} \quad J + a = J + b \quad (1)$$

$$J + a = J + b \quad \text{iff} \quad a - b \in J \quad (2)$$

$$J + a = J \quad \text{iff} \quad a \in J \quad (3)$$

We also know, by the reasoning which leads up to Lagrange's theorem, that the family of all the cosets $J + a$, as a ranges over A , is a partition of A .



There is a way of *adding* and *multiplying cosets* which works as follows:

$$(J + a) + (J + b) = J + (a + b)$$

$$(J + a)(J + b) = J + ab$$

In other words, the sum of the coset of a and the coset of b is the coset of $a + b$; the product of the coset of a and the coset of b is the coset of ab .

It is important to know that the sum and product of cosets, defined in this fashion, are determined without ambiguity. Remember that $J + a$ may be the same coset as $J + c$ [by Condition (1) this happens iff c is an element of $J + a$], and, likewise, $J + b$ may be the same coset as $J + d$. Therefore, we have the equations

$$\begin{array}{ccc} (J + a) + (J + b) = J + (a + b) & & (J + a)(J + b) = J + ab \\ \parallel & & \parallel \\ (J + c) + (J + d) = J + (c + d) & \text{and} & (J + c)(J + d) = J + cd \end{array}$$

Obviously we must be absolutely certain that $J + (a + b) = J + (c + d)$ and $J + ab = J + cd$. The next theorem provides us with this important guarantee.

Theorem 1 *Let J be an ideal of A . If $J + a = J + c$ and $J + b = J + d$, then*

- (i) $J + (a + b) = J + (c + d)$, and
- (ii) $J + ab = J + cd$.

PROOF: We are given that $J + a = J + c$ and $J + b = J + d$; hence by Condition (2),

$$a - c \in J \quad \text{and} \quad b - d \in J$$

Since J is closed with respect to addition, $(a - c) + (b - d) = (a + b) - (c + d)$ is in J . It follows by Condition (2) that $J + (a + b) = J + (c + d)$, which proves (i). On the other hand, since J absorbs products in A ,

$$\underbrace{(a - c)b}_{ab - cb} \in J \quad \text{and} \quad \underbrace{c(b - d)}_{cb - cd} \in J$$

and therefore $(ab - cb) + (cb - cd) = ab - cd$ is in J . It follows by Condition (2) that $J + ab = J + cd$. This proves (ii). ■

Now, think of the set which consists of *all the cosets of J in A* . This set is conventionally denoted by the symbol A/J . For example, if $J + a, J + b, J + c, \dots$ are cosets of J , then

$$A/J = \{J + a, J + b, J + c, \dots\}$$

We have just seen that coset addition and multiplication are valid operations on this set. In fact,

Theorem 2 A/J with coset addition and multiplication is a ring.

PROOF: Coset addition and multiplication are associative, and multiplication is distributive over addition. (These facts may be routinely checked.) The zero element of A/J is the coset $J = J + 0$, for if $J + a$ is any coset,

$$(J + a) + (J + 0) = J + (a + 0) = J + a$$

Finally, the negative of $J + a$ is $J + (-a)$, because

$$(J + a) + (J + (-a)) = J + (a + (-a)) = J + 0 \blacksquare$$

The ring A/J is called the *quotient ring* of A by J .

And now, the crucial connection between quotient rings and homomorphisms :

Theorem 3 A/J is a homomorphic image of A .

Following the plan already laid out for groups, the *natural homomorphism* from A onto A/J is the function f which carries every element to its own coset, that is, the function f given by

$$f(x) = J + x$$

This function is very easily seen to be a homomorphism.

Thus, when we construct quotient rings of A , we are, in fact, constructing homomorphic images of A . The quotient ring construction is useful because it is a way of actually manufacturing homomorphic images of any ring A .

The quotient ring construction is now illustrated with an important example. Let \mathbf{z} be the ring of the integers, and let $\langle 6 \rangle$ be the ideal of \mathbf{z} which consists of all the multiples of the number 6. The elements of the quotient ring $\mathbf{z}/\langle 6 \rangle$ are all the cosets of the ideal $\langle 6 \rangle$, namely:

$$\langle 6 \rangle + 0 = \{ \dots, -18, -12, -6, 0, 6, 12, 18, \dots \} = \bar{0}$$

$$\langle 6 \rangle + 1 = \{ \dots, -17, -11, -5, 1, 7, 13, 19, \dots \} = \bar{1}$$

$$\langle 6 \rangle + 2 = \{ \dots, -16, -10, -4, 2, 8, 14, 20, \dots \} = \bar{2}$$

$$\langle 6 \rangle + 3 = \{ \dots, -15, -9, -3, 3, 9, 15, 21, \dots \} = \bar{3}$$

$$\langle 6 \rangle + 4 = \{ \dots, -14, -8, -2, 4, 10, 16, 22, \dots \} = \bar{4}$$

$$\langle 6 \rangle + 5 = \{ \dots, -13, -7, -1, 5, 11, 17, 23, \dots \} = \bar{5}$$

We will represent these cosets by means of the simplified notation $\bar{0}, \bar{1}, \bar{2}, \bar{3}, \bar{4}, \bar{5}$. The rules for adding and multiplying cosets give us the following tables:

+	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{0}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{1}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$
$\bar{2}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$
$\bar{3}$	$\bar{3}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$
$\bar{4}$	$\bar{4}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$
$\bar{5}$	$\bar{5}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$

·	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$	$\bar{0}$
$\bar{1}$	$\bar{0}$	$\bar{1}$	$\bar{2}$	$\bar{3}$	$\bar{4}$	$\bar{5}$
$\bar{2}$	$\bar{0}$	$\bar{2}$	$\bar{4}$	$\bar{0}$	$\bar{2}$	$\bar{4}$
$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{0}$	$\bar{3}$	$\bar{0}$	$\bar{3}$
$\bar{4}$	$\bar{0}$	$\bar{4}$	$\bar{2}$	$\bar{0}$	$\bar{4}$	$\bar{2}$
$\bar{5}$	$\bar{0}$	$\bar{5}$	$\bar{4}$	$\bar{3}$	$\bar{2}$	$\bar{1}$

One cannot fail to notice the analogy between the quotient ring $\mathbf{Z}/\langle 6 \rangle$ and the ring \mathbf{Z}_6 . In fact, we will regard them as one and the same. More generally, for every positive integer n , we consider \mathbf{Z}_n to be the same as $\mathbf{Z}/\langle n \rangle$. In particular, this makes it clear that \mathbf{Z}_n is a homomorphic image of \mathbf{Z} .

By [Theorem 3](#), any quotient ring A/J is a homomorphic image of A . Therefore the quotient ring construction is a way of actually producing homomorphic images of any ring A . In fact, as we will now see, it is a way of producing *all* the homomorphic images of A .

Theorem 4 *Let $f: A \rightarrow B$ be a homomorphism from a ring A onto a ring B , and let K be the kernel of f . Then $B \cong A/K$.*

PROOF: To show that A/K is isomorphic with B , we must look for an isomorphism from A/K to B . Mimicking the procedure which worked successfully for groups, we let ϕ be the function from A/K to B which matches each coset $K + x$ with the element $f(x)$; that is,

$$\phi(K + x) = f(x)$$

Remember that if we ignore multiplication for just a moment, A and B are groups and f is a group homomorphism from A onto B , with kernel K . Therefore we may apply [Theorem 2](#) of [Chapter 16](#): ϕ is a well-defined, bijective function from A/K to B . Finally,

$$\begin{aligned} \phi((K + a) + (K + b)) &= \phi(K + (a + b)) = f(a + b) \\ &= f(a) + f(b) = \phi(K + a) + \phi(K + b) \end{aligned}$$

and

$$\begin{aligned} \phi((K + a)(K + b)) &= \phi(K + ab) = f(ab) \\ &= f(a)f(b) = \phi(K + a)\phi(K + b) \end{aligned}$$

Thus, ϕ is an isomorphism from A/K onto B . ■

[Theorem 4](#) is called the *fundamental homomorphism theorem* for rings. [Theorems 3](#) and [4](#) together assert that every quotient ring of A is a homomorphic image of A , and, conversely, every homomorphic image of A is isomorphic to a quotient ring of A . Thus, for all practical purposes, quotients and homomorphic images of a ring are the same.

As in the case of groups, there are many practical instances in which it is possible to select an ideal J of A so as to “factor out” unwanted traits of A , and obtain a quotient ring A/J with “desirable” features.

As a simple example, let A be a ring, not necessarily commutative, and let J be an ideal of A which contains all the differences

$$ab - ba$$

as a and b range over A . It is quite easy to show that the quotient ring A/J is then commutative. Indeed, to say that A/J is commutative is to say that for any two cosets $J + a$ and $J + b$,

$$(J + a)(J + b) = (J + b)(J + a) \quad \text{that is} \quad J + ab = J + ba$$

By Condition (2) this last equation is true iff $ab - ba \in J$. Thus, if every difference $ab - ba$ is in J , then any two cosets commute.

A number of important quotient ring constructions, similar in principle to this one, are given in the exercises.

An ideal J of a commutative ring is said to be a *prime ideal* if for any two elements a and b in the ring,

$$\text{If } ab \in J \text{ then } a \in J \text{ or } b \in J$$

Whenever J is a prime ideal of a commutative ring with unity A , the quotient ring A/J is an integral domain. (The details are left as an exercise.)

An ideal of a ring is called *proper* if it is not equal to the whole ring. A proper ideal J of a ring A is called a *maximal ideal* if there exists no proper ideal K of A such that $J \subseteq K$ with $J \neq K$ (in other words, J is not contained in any strictly larger proper ideal). It is an important fact that if A is a commutative ring with unity, then J is a maximal ideal of A iff A/J is a field.

To prove this assertion, let J be a maximal ideal of A . If A is a commutative ring with unity, it is easy to see that A/J is one also. In fact, it should be noted that the unity of A/J is the coset $J + 1$, because if $J + a$ is any coset, $(J + a)(J + 1) = J + a1 = J + a$. Thus, to prove that A/J is a field, it remains only to show that if $J + a$ is any nonzero coset, there is a coset $J + x$ such that $(J + a)(J + x) = J + 1$.

The zero coset is J . Thus, by Condition (3), to say that $J + a$ is *not* zero, is to say that $a \notin J$. Now, let K be the set of all the sums

$$xa + j$$

as x ranges over A and j ranges over J . It is easy to check that K is an ideal. Furthermore, K contains a because $a = 1a + 0$, and K contains every element $j \in J$ because j can be written as $0a + j$. Thus, K is an ideal which contains J and is strictly larger than J (for remember that $a \in K$ but $a \notin J$). But J is a maximal ideal! Thus, K must be the whole ring A .

It follows that $1 \in K$, so $1 = xa + j$ for some $x \in A$ and $j \in J$. Thus, $1 - xa = j \in J$, so by Condition (2), $J + 1 = J + xa = (J + x)(J + a)$. In the quotient ring A/J , $J + x$ is therefore the multiplicative inverse of $J + a$.

The converse proof consists, essentially, of “unraveling” the preceding argument; it is left as an entertaining exercise.

EXERCISES

A. Examples of Quotient Rings

In each of the following, A is a ring and J is an ideal of A . List the elements of A/J , and then write the addition and multiplication tables of A/J .

Example $A = \mathbb{Z}_6$, $J = \{0, 3\}$.

The elements of A/J are the three cosets $J = J + 0 = \{0,3\}$, $J + 1 = \{1,4\}$, and $J + 2 = \{2,5\}$. The tables for A/J are as follows:

+	J	J + 1	J + 2	·	J	J + 1	J + 2
J	J	J + 1	J + 2	J	J	J	J
J + 1	J + 1	J + 2	J	J + 1	J	J + 1	J + 2
J + 2	J + 2	J	J + 1	J + 2	J	J + 2	J + 1

- $A = \mathbb{Z}_{10}$, $J = \{0,5\}$.
- $A = P_3$, $J = \{0, \{a\}\}$. (P_3 is defined in [Chapter 17, Exercise D](#).)
- $A = \mathbb{Z}_2 \times \mathbb{Z}_6$; $J = \{(0,0), (0,2), (0,4)\}$.

B. Examples of the Use of the FHT

In each of the following, use the FHT (fundamental homomorphism theorem) to prove that the two given groups are isomorphic. Then display their tables.

Example \mathbb{Z}_2 and $\mathbb{Z}_6/\langle 2 \rangle$.

The following function is a homomorphism from \mathbb{Z}_6 onto \mathbb{Z}_2 :

$$f = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

- (Do not prove that J is a homomorphism.)
The kernel of f is $\{0, 2, 4\} = (2)$. Thus:

$$\mathbb{Z}_6 \xrightarrow{\langle 2 \rangle} \mathbb{Z}_2$$

It follows by the FHT that $\mathbb{Z}_2 \cong \mathbb{Z}_6/\langle 2 \rangle$.

- \mathbb{Z}_5 and $\mathbb{Z}_{20}/\langle 5 \rangle$.
- \mathbb{Z}_3 and $\mathbb{Z}_6/\langle 3 \rangle$.
- P_2 and P_3/K , where $K = \{0, \{c\}\}$. [HINT: See [Chapter 18, Exercise E6](#). Consider the function $f(X) = X \cap \{a, b\}$.]
- \mathbb{Z}_2 and $\mathbb{Z}_2 \times \mathbb{Z}_2/K$, where $K = \{(0, 0), (0,1)\}$.

C. Quotient Rings and Homomorphic Images in $\mathcal{F}(\mathbb{R})$

- Let ϕ be the function from $\mathcal{F}(\mathbb{R})$ to $\mathbb{R} \times \mathbb{R}$ defined by $\phi(f) = (f(0), f(1))$. Prove that ϕ is a homomorphism from $\mathcal{F}(\mathbb{R})$ onto $\mathbb{R} \times \mathbb{R}$, and describe its kernel.
- Let J be the subset of $\mathcal{F}(\mathbb{R})$ consisting of all f whose graph passes through the points $(0,0)$ and $(1,0)$. Referring to part 1, explain why J is an ideal of $\mathcal{F}(\mathbb{R})$, and $\mathcal{F}(\mathbb{R})/J \cong \mathbb{R} \times \mathbb{R}$.
- Let ϕ be the function from $\mathcal{F}(\mathbb{R})$ to $\mathcal{F}(\mathbb{Q}, \mathbb{R})$ defined as follows:

$$\phi(f) = f_{\mathbb{Q}} = \text{the restriction of } f \text{ to } \mathbb{Q}$$

(NOTE: The domain of $f_{\mathbb{Q}}$ is \mathbb{Q} and on this domain $f_{\mathbb{Q}}$ is the same function as f .) Prove that ϕ is a homomorphism from $\mathcal{F}(\mathbb{R})$ onto $\mathcal{F}(\mathbb{Q}, \mathbb{R})$, and describe the kernel of ϕ . [$\mathcal{F}(\mathbb{Q}, \mathbb{R})$ is the ring of functions from \mathbb{Q} to \mathbb{R} .]

4 Let J be the subset of $\mathcal{F}(\mathbb{R})$ consisting of all f such that $f(x) = 0$ for every rational x . Referring to part 3, explain why J is an ideal of $\mathcal{F}(\mathbb{R})$ and $\mathcal{F}(\mathbb{R})/J \cong \mathcal{F}(\mathbb{Q})$.

D. Elementary Applications of the Fundamental Homomorphism Theorem

In each of the following let A be a commutative ring. If $a \in A$ and n is a positive integer, the notation na will stand for

$$a + a + \cdots + a \quad (n \text{ terms})$$

1 Suppose $2x = 0$ for every $x \in A$. Prove that $(x + y)^2 = x^2 + y^2$ for all x and y in A . Conclude that the function $h(x) = x^2$ is a homomorphism from A to A . If $J = \{x \in A : x^2 = 0\}$ and $B = \{x^2 : x \in A\}$, explain why J is an ideal of A , B is a subring of A , and $A/J \cong B$.

2 Suppose $6x = 0$ for every $x \in A$. Prove that the function $h(x) = 3x$ is a homomorphism from A to A . If $J = \{x : 3x = 0\}$ and $B = \{3x : x \in A\}$, explain why J is an ideal of A , B is a subring of A , and $A/J \cong B$.

3 If a is an idempotent element of A (that is, $a^2 = a$), prove that the function $\pi_a(x) = ax$ is a homomorphism from A into A . Show that the kernel of π_a is I_a , the annihilator of a (defined in [Exercise H4](#) of [Chapter 18](#)). Show that the range of π_a is $\langle a \rangle$. Conclude by the FHT that $A/I_a = \langle a \rangle$.

4 For each $a \in A$, let π_a be the function given by $\pi_a(x) = ax$. Define the following addition and multiplication on $\bar{A} = \{\pi_a : a \in A\}$:

$$\pi_a + \pi_b = \pi_{a+b} \text{ and } \pi_a \pi_b = \pi_{ab}$$

(\bar{A} is a ring; however, do not prove this.) Show that the function $\phi(a) = \pi_a$ is a homomorphism from A onto \bar{A} . Let I designate the annihilating ideal of A (defined in [Exercise H4](#) of [Chapter 18](#)). Use the FHT to show that $A/I \cong \bar{A}$.

E. Properties of Quotient Rings A/J in Relation to Properties of J

Let A be a ring and J an ideal of A . Use Conditions (1), (2), and (3) of this chapter. Prove each of the following:

1 Every element of A/J has a square root iff for every $x \in A$, there is some $y \in A$ such that $x - y^2 \in J$.

2 Every element of A/J is its own negative iff $x + x \in J$ for every $x \in A$.

3 A/J is a boolean ring iff $x^2 - x \in J$ for every $x \in A$. (A ring S is called a boolean ring iff $s^2 = s$ for every $s \in S$.)

4 If J is the ideal of all the nilpotent elements of a commutative ring A , then A/J has no nilpotent elements (except zero). (Nilpotent elements are defined in [Chapter 17](#), [Exercise M](#); by M2 and M3 they form an ideal.)

5 Every element of A/J is nilpotent iff J has the following property: for every $x \in A$, there is a positive integer n such that $x^n \in J$.

6 A/J has a unity element iff there exists an element $a \in A$ such that $ax - x \in J$ and $xa - x \in J$ for every $x \in A$.

F. Prime and Maximal Ideals

Let A be a commutative ring with unity, and J an ideal of A . Prove each of the following:

- 1 A/J is a commutative ring with unity.
- 2 J is a prime ideal iff A/J is an integral domain.
- 3 Every maximal ideal of A is a prime ideal. (HINT: Use the fact, proved in this chapter, that if J is a maximal ideal then A/J is a field.)
- 4 If A/J is a field, then J is a maximal ideal. (HINT: See [Exercise I2](#) of [Chapter 18](#).)

G. Further Properties of Quotient Rings in Relation to Their Ideals

Let A be a ring and J an ideal of A . (In parts 1–3 and 5 assume that A is a commutative ring with unity.)

- # 1 Prove that A/J is a field iff for every element $a \in A$, where $a \notin J$, there is some $b \in A$ such that $ab - 1 \in J$.
- 2 Prove that every nonzero element of A/J is either invertible or a divisor of zero iff the following property holds, where $a, x \in A$: For every $a \notin J$, there is some $x \notin J$ such that either $ax \in J$ or $ax - 1 \in J$.
- 3 An ideal J of a ring A is called *primary* iff for all $a, b \in A$, if $ab \in J$, then either $a \in J$ or $b^n \in J$ for some positive integer n . Prove that every zero divisor in A/J is nilpotent iff J is primary.
- 4 An ideal J of a ring A is called *semiprime* iff it has the following property: For every $a \in A$, if $a^n \in J$ for some positive integer n , then necessarily $a \in J$. Prove that J is semiprime iff A/J has no nilpotent elements (except zero).
- 5 Prove that an integral domain can have no nonzero nilpotent elements. Then use part 4, together with [Exercise F2](#), to prove that every prime ideal in a commutative ring is semiprime.

H. \mathbf{Z}_n as a Homomorphic Image of \mathbf{Z}

Recall that the function

$$f(a) = \bar{a}$$

is the natural homomorphism from \mathbf{Z} onto \mathbf{Z}_n . If a polynomial equation $p = 0$ is satisfied in \mathbf{Z} , necessarily $f(p) = f(0)$ is true in \mathbf{Z}_n . Let us take a specific example; there are integers x and y satisfying $11x^2 - 8y^2 + 29 = 0$ (we may take $x = 3$ and $y = 4$). It follows that there must be elements \bar{x} and \bar{y} in \mathbf{Z}_6 which satisfy $\overline{11} \bar{x}^2 - \overline{8} \bar{y}^2 + \overline{29} = \bar{0}$ in \mathbf{Z}_6 , that is, $\bar{5} \bar{x}^2 - \bar{2} \bar{y}^2 + \bar{5} = \bar{0}$. (We take $\bar{x} = \bar{3}$ and $\bar{y} = \bar{4}$.) The problems which follow are based on this observation.

- 1 Prove that the equation $x^2 - 7y^2 - 24 = 0$ has no integer solutions. (HINT: If there are integers x and y satisfying this equation, what equation will \bar{x} and \bar{y} satisfy in \mathbf{Z}_7 ?)
- 2 Prove that $x^2 + (x + 1)^2 + (x + 2)^2 = y^2$ has no integer solutions.
- 3 Prove that $x^2 + 10y^2 = n$ (where n is an integer) has no integer solutions if the last digit of n is 2, 3, 7, or 8.
- 4 Prove that the sequence 3, 8, 13, 18, 23, ... does not include the square of any integer. (HINT: The image

of each number on this list, under the natural homomorphism from \mathbf{Z} to \mathbf{Z}_5 , is 3.)

5 Prove that the sequence 2, 10, 18, 26, ... does not include the cube of any integer.

6 Prove that the sequence 3, 11, 19, 27, ... does not include the sum of two squares of integers.

7 Prove that if n is a product of two consecutive integers, its units digit must be 0, 2, or 6.

8 Prove that if n is the product of three consecutive integers, its units digit must be 0, 4, or 6.



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – IV – ALGEBRA-I – SMT1501

Unit-IV Quotient Rings and Fields

Let us recall that an integral domain is a commutative ring with unity having the cancellation property, that is,

$$\text{if } a \neq 0 \text{ and } ab = ac \text{ then } b = c \quad (1)$$

At the end of [Chapter 17](#) we saw that an integral domain may also be defined as a commutative ring with unity having no divisors of zero, which is to say that

$$\text{if } ab = 0 \text{ then } a = 0 \text{ or } b = 0 \quad (2)$$

for as we saw, (1) and (2) are equivalent properties in any commutative ring.

The system \mathbf{Z} of the integers is the exemplar and prototype of integral domains. In fact, the term “integral domain” means a system of algebra (“domain”) having *integerlike* properties. However, \mathbf{Z} is not the only integral domain: there are a great many integral domains different from \mathbf{Z} .

Our first few comments will apply to rings generally. To begin with, we introduce a convenient notation for multiples, which parallels the *exponent notation* for powers. Additively, the sum

$$a + a + \cdots + a$$

of n equal terms is written as $n \cdot a$. We also define $0 \cdot a$ to be 0, and let $(-n) \cdot a = -(n \cdot a)$ for all positive integers n . Then

$$m \cdot a + n \cdot a = (m + n) \cdot a \quad \text{and} \quad m \cdot (n \cdot a) = (mn) \cdot a$$

for every element a of a ring and all integers m and n . These formulas are the translations into additive notation of the *laws of exponents* given in [Chapter 10](#).

If A is a ring, A with addition alone is a group. Remember that in additive notation the *order* of an element a in A is the least positive integer n such that $n \cdot a = 0$. If there is no such positive integer n , then a is said to have order infinity. To emphasize the fact that we are referring to the order of a in terms of

addition, we will call it the *additive order* of a .

In a ring with unity, if 1 has additive order n , we say the ring has “characteristic n .” In other words, if A is a ring with unity,

*the **characteristic** of A is the least positive integer n such that*

$$\underbrace{1 + 1 + \cdots + 1}_{n \text{ times}} = 0$$

*If there is no such positive integer n , A has **characteristic 0**.*

These concepts are especially simple in an integral domain. Indeed,

Theorem 1 *All the nonzero elements in an integral domain have the same additive order.*

PROOF: That is, every $a \neq 0$ has the same additive order as the additive order of 1. The truth of this statement becomes transparently clear as soon as we observe that

$$n \cdot a = a + a + \cdots + a = 1a + \cdots + 1a = (1 + \cdots + 1)a = (n \cdot 1)a$$

hence $n \cdot a = 0$ iff $n \cdot 1 = 0$. (Remember that in an integral domain, if the product of two factors is equal to 0, at least one factor must be 0.) ■

It follows, in particular, that if the characteristic of an integral domain is a positive integer n , then

$$n \cdot x = 0$$

for every element x in the domain.

Furthermore,

Theorem 2 *In an integral domain with nonzero characteristic, the characteristic is a prime number.*

PROOF: If the characteristic were a composite number mn , then by the distributive law,

$$(m \cdot 1)(n \cdot 1) = \underbrace{(1 + \cdots + 1)}_{m \text{ terms}} \underbrace{(1 + \cdots + 1)}_{n \text{ terms}} = \underbrace{1 + 1 + \cdots + 1}_{mn \text{ terms}} = (mn) \cdot 1 = 0$$

Thus, either $m \cdot 1 = 0$ or $n \cdot 1 = 0$, which is impossible because mn was chosen to be the *least* positive integer such that $(mn) \cdot 1 = 0$. ■

A very interesting rule of arithmetic is valid in integral domains whose characteristic is not zero.

Theorem 3 *In any integral domain of characteristic p ,*

$$(a + b)^p = a^p + b^p \quad \text{for all elements } a \text{ and } b$$

PROOF: This formula becomes clear when we look at the binomial expansion of $(a + b)^p$. Remember that by the binomial formula,

$$(a + b)^p = a^p + \binom{p}{1} \cdot a^{p-1}b + \cdots + \binom{p}{p-1} \cdot ab^{p-1} + b^p$$

where the binomial coefficient

$$\binom{p}{k} = \frac{p(p-1)(p-2) \cdots (p-k+1)}{k!}$$

It is demonstrated in [Exercise L](#) of [Chapter 17](#) that the binomial formula is correct in every commutative ring.

Note that if p is a prime number and $0 < k < p$, then

$$\binom{p}{k} \text{ is a multiple of } p$$

because every factor of the denominator is less than p , hence p does not cancel out. Thus, each term of the binomial expansion above, except for the first and last terms, is of the form px , which is equal to 0 because the domain has characteristic p . Thus, $(a + b)^p = a^p + b^p$. ■

It is obvious that every field is an integral domain: for if $a \neq 0$ and $ax = ay$ in a field, we can multiply both sides of this equation by the multiplicative inverse of a to cancel a . However, not every integral domain is a field: for example, \mathbf{Z} is not a field. Nevertheless,

Theorem 4 *Every finite integral domain is a field.*

List the elements of the integral domain in the following manner:

$$0, 1, a_1, a_2, \dots, a_n$$

In this manner of listing, there are $n + 2$ elements in the domain. Take any a_i , and show that it is invertible: to begin with, note that the products

$$a_i 0, a_i 1, a_i a_1, a_i a_2, \dots, a_i a_n$$

are all distinct: for if $a_i x = a_i y$, then $x = y$. Thus, there are $n + 2$ *distinct* products $a_i x$; but there are exactly $n + 2$ elements in the domain, so every element in the domain is equal to one of these products. In particular, $1 = a_i x$ for some x ; hence a_i is invertible.

OPTIONAL

The integral domain \mathbf{Z} is not a field because it does not contain the quotients m/n of integers. However, \mathbf{Z} can be *enlarged* to a field by adding to it all the quotients of integers; the resulting field, of course, is \mathbf{Q} , the field of the rational numbers. \mathbf{Q} consists of all quotients of integers, and it contains \mathbf{Z} (or rather, an isomorphic copy of \mathbf{Z}) when we identify each integer n with the quotient $n/1$. We say that \mathbf{Q} is *the field of quotients* of \mathbf{Z} .

It is a fascinating fact that the method for constructing \mathbf{Q} from \mathbf{Z} can be applied to any integral domain. Starting from any integral domain A , it is possible to construct a field which contains A : a field of quotients of A . This is not merely a mathematical curiosity, but a valuable addition to our knowledge. In applications it often happens that a system of algebra we are dealing with lacks a needed property, but is contained in a larger system which *has* that property—and that is almost as good! In the present case, A is not a field but may be enlarged to one.

Thus, if A is any integral domain, we will proceed to construct a field A^* consisting of all the quotients of elements in A ; and A^* will contain A , or rather an isomorphic copy of A , when we identify each element a of A with the quotient $a/1$. The construction will be carefully outlined and the busy work left as an exercise.

Given A , let S denote the set of all ordered pairs (a, b) of elements of A , where $b \neq 0$. That is,

$$S = \{(a, b) : a, b \in A \text{ and } b \neq 0\}$$

In order to understand the next step, we should think of (a, b) as a/b . [It is too early in the proof to introduce fractional notation; nevertheless each ordered pair (a, b) should be *thought of* as a fraction a/b .] Now a problem of representation arises here, because it is obvious that the quotient xa/xb is equal to the quotient a/b ; to put the same fact differently, the quotients a/b and c/d are equal whenever $ad = bc$. That is, if $ad = bc$, then a/b and c/d are two different ways of writing the *same* quotient. Motivated by this observation, we *define* $(a, b) \sim (c, d)$ to mean that $ad = bc$, and easily verify that \sim is an equivalence relation on the set S . (Equivalence relations are explained in [Chapter 12](#).) Then we let $[a, b]$ denote the *equivalence class* of (a, b) , that is,

$$[a, b] = \{(c, d) \in S : (c, d) \sim (a, b)\}$$

Intuitively, all the pairs which represent a given quotient are lumped together in one equivalence class; thus, *each quotient is represented by exactly one equivalence class*.

Let us recapitulate the formal details of our construction up to this point: Given the set S of ordered pairs of elements in A , we define an equivalence relation — in S by letting $(a, b) \sim (c, d)$ iff $ad = bc$. We let $[a, b]$ designate the equivalence class of (a, b) , and finally, we let A^* denote the set of all the *equivalence classes* $[a, b]$. The elements of A^* will be called *quotients*.

Before going on, observe carefully that

$$[a, b] = [r, s] \quad \text{iff} \quad (a, b) \sim (r, s) \quad \text{iff} \quad as = br \quad (3)$$

As our next step, we define operations of addition and multiplication in A^* :

$$[a, b] + [c, d] = [ad + bc, bd]$$

and

$$[a, b] \cdot [c, d] = [ac, bd]$$

To understand these definitions, simply remember the formulas

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

We must make certain these definitions are unambiguous; that is, if $[a, b] = [r, s]$ and $[c, d] = [t, u]$, we have equations

$$\begin{array}{ccc} [a, b] + [c, d] = [ad + bc, bd] & & [a, b] \cdot [c, d] = [ac, bd] \\ \parallel & \text{and} & \parallel \\ [r, s] + [t, u] = [ru + st, su] & & [r, s] \cdot [t, u] = [ru, su] \end{array}$$

and we must therefore verify that $[ad + bc, bd] = [ru + st, su]$ and $[ac, bd] = [rt, su]$. This is left as an exercise. It is also left for the student to verify that addition and multiplication are associative and commutative and the distributive law is satisfied.

The zero element is $[0, 1]$, because $[a, b] + [0, 1] = [a, b]$. The negative of $[a, b]$ is $[-a, b]$, for $[a, b] + [-a, b] = [0, b^2] = [0, 1]$. [The last equation is true because of [Equation \(3\)](#).] The unity is $[1, 1]$, and the multiplicative inverse of $[a, b]$ is $[b, a]$, for $[a, b] \cdot [b, a] = [ab, ab] = [1, 1]$. Thus, A^* is a field!

Finally, if A' is the subset of A^* which contains every $[a, 1]$, we let ϕ be the function from A to A' defined by $\phi(a) = [a, 1]$. This function is injective because, by Equation (3), if $[a, 1] = [b, 1]$ then $a = b$. It is obviously surjective and is easily shown to be a homomorphism. Thus, ϕ is an isomorphism from A to A' so A^* contains an isomorphic copy A' of A .

EXERCISES

A. Characteristic of an Integral Domain

Let A be a finite integral domain. Prove each of the following:

- 1 Let a be any nonzero element of A . If $n \cdot a = 0$, where $n \neq 0$, then n is a multiple of the characteristic of A .
- 2 If A has characteristic zero, $n \neq 0$, and $n \cdot a = 0$, then $a = 0$.
- 3 If A has characteristic 3, and $5 \cdot a = 0$, then $a = 0$.
- 4 If there is a nonzero element a in A such that $256 \cdot a = 0$, then A has characteristic 2.
- 5 If there are distinct nonzero elements a and b in A such that $125 \cdot a = 125 \cdot b$, then A has characteristic 5.
- 6 If there are nonzero elements a and b in A such that $(a + b)^2 = a^2 + b^2$, then A has characteristic 2.
- 7 If there are nonzero elements a and b in A such that $10a = 0$ and $14b = 0$, then A has characteristic 2.

B. Characteristic of a Finite Integral Domain

Let A be an integral domain. Prove each of the following:

- 1 If A has characteristic q , then q is a divisor of the order of A .
- 2 If the order of A is a prime number p , then the characteristic of A must be equal to p .
- 3 If the order of A is p^m , where p is a prime, the characteristic of A must be equal to p .
- 4 If A has 81 elements, its characteristic is 3.
- 5 If A , with addition alone, is a cyclic group, the order of A is a prime number.

C. Finite Rings

Let A be a finite commutative ring with unity.

- 1 Prove: Every nonzero element of A is either a divisor of zero or invertible. (HINT: Use an argument analogous to the proof of [Theorem 4](#).)
- 2 Prove: If $a \neq 0$ is not a divisor of zero, then some positive power of a is equal to 1. (HINT: Consider a, a^2, a^3, \dots . Since A is finite, there must be positive integers $n < m$ such that $a^n = a^m$.)
- 3 Use part 2 to prove: If a is invertible, then a^{-1} is equal to a positive power of a .

D. Field of Quotients of an Integral Domain

The following questions refer to the construction of a field of quotients of A , as outlined on pages 203 to 205.

- 1 If $[a, b] = [r, s]$ and $[c, d] = [t, u]$, prove that $[a, b] + [c, d] = [r, s] + [t, u]$.
- 2 If $[a, b] = [r, s]$ and $[c, d] = [t, u]$, prove that $[a, b][c, d] = [r, s][t, u]$.
- 3 If $(a, b) \sim (c, d)$ means $ad = bc$, prove that \sim is an equivalence relation on S .
- 4 Prove that addition in A^* is associative and commutative.
- 5 Prove that multiplication in A^* is associative and commutative.
- 6 Prove the distributive law in A^* .
- 7 Verify that $\phi: A \rightarrow A'$ is a homomorphism.

E. Further Properties of the Characteristic of an Integral Domain

Let A be an integral domain. Prove parts 1-4:

- 1 Let $a \in A$. If A has characteristic p , and $n \cdot a = 0$ where n is *not* a multiple of p , then $a = 0$.
 - 2 If p is a prime, and there is a nonzero element $a \in A$ such that $p \cdot a = 0$, then A has characteristic p .
 - 3 If p is a prime, and there is a nonzero element $a \in A$ such that $p^m \cdot a = 0$ for some integer m , then A has characteristic p .
 - 4 If A has characteristic p , then the function $f(a) = a^p$ is a homomorphism from A to A .
- # 5 Let A have order p , where p is a prime. Explain why

$$A = \{0, 1, 2 \cdot 1, 3 \cdot 1, \dots, (p-1) \cdot 1\}$$

Prove that $A \cong \mathbb{Z}_p$.

- # 6 If A has characteristic p , prove that for any positive integer n ,

$$(a) (a + b)^{p^n} = a^{p^n} + b^{p^n}$$

$$(b) (a_1 + a_2 + \dots + a_r)^{p^n} = a_1^{p^n} + \dots + a_r^{p^n}$$

- 7 Let $A \subseteq B$ where A and B are integral domains. Prove: A has characteristic p iff B has characteristic p .

F. Finite Fields

By [Theorem 4](#), “finite integral domain” and “finite field” are the same.

- 1 Prove: Every finite field has nonzero characteristic.
- 2 Prove that if A is a finite field of characteristic p , the function $f(a) = a^p$ is an automorphism of A ; that is, an isomorphism from A to A . (HINT: Use [Exercise E4](#) above and [Exercise F7](#) of [Chapter 18](#). To show that f is surjective, compare the number of elements in the domain and in the range of f .)

*The function $f(a) = a^p$ is called the **Frobenius automorphism**.*

- 3 Use part 2 to prove: In a finite field of characteristic p , every element has a p -th root.

There are two possible ways of describing the system of the integers.

On the one hand, we may attempt to describe it concretely.

On the other hand, we may find a list of axioms from which it is possible to deduce all the properties of the integers, so the *only* system which has all these properties is the system of the integers.

The second of these two ways is the way of mathematics. It is elegant, economical, and simple. We select as axioms only those particular properties of the integers which are absolutely necessary in order to prove further properties of the integers. And we select a sufficiently *complete* list of axioms so that, using them, one can prove all the properties of the integers needed in mathematics.

We have already seen that the integers are an integral domain. However, there are numerous examples of integral domains which bear little resemblance to the set of the integers. For example, there are finite integral domains such as \mathbb{Z}_5 , fields (remember that every field is an integral domain) such as \mathbb{Q} and \mathbb{R} , and others. Thus, in order to pin down the integers — that is, in order to find a list of axioms which applies to the integers and *only* the integers—we must select some additional axioms and add them to the axioms of integral domains. This we will now proceed to do.

Most of the traditional number systems have two aspects. One aspect is their algebraic structure: they are integral domains or fields. The other aspect—which we have not yet touched upon—is that their elements can be *ordered*. That is, if a and b are distinct elements, we can say that a is less than b or b is less than a . This second aspect—the ordering of elements—will now be formalized.

*An **ordered integral domain** is an integral domain A with a relation, symbolized by $<$, having the following properties:*

1. *For any a and b in A , exactly one of the following is true:*

$$a = b \quad a < b \quad \text{or} \quad b < a$$

Furthermore, for any a , b , and c in A ,

2. *If $a < b$ and $b < c$, then $a < c$.*
3. *If $a < b$, then $a + c < b + c$.*

4. If $a < b$, then $ac < bc$ on the condition that $0 < c$.

The relation $<$ is called an *order relation* on A . The four conditions which an order relation must fulfill are familiar to everyone. Properties 1 and 2 require no comment. Property 3 asserts that we are allowed to add any given c to both sides of an inequality. Property 4 asserts that we may multiply both sides of an inequality by any c , on the condition that c is greater than zero.

As usual, $a > b$ has the same meaning as $b < a$. Furthermore, $a \leq b$ means “ $a < b$ or $a = b$,” and $b \geq a$ means the same as $a \leq b$.

In an ordered integral domain A , an element a is called *positive* if $a > 0$. If $a < 0$ we call a *negative*. Note that if a is positive then $-a$ is negative. (Proof: Add $-a$ to both sides of the inequality $a > 0$.) Similarly, if a is negative, then $-a$ is positive.

In any ordered integral domain, *the square of every nonzero element is positive*. Indeed, if c is nonzero, then either $c > 0$ or $c < 0$. If $c > 0$, then, multiplying both sides of the inequality $c > 0$ by c ,

$$cc > c0 = 0$$

so $c^2 > 0$. On the other hand, if $c < 0$, then

$$(-c) > 0$$

hence

$$(-c)(-c) > 0(-c) = 0$$

But $(-c)(-c) = c^2$, so once again, $c^2 > 0$.

In particular, since $1 = 1^2$, 1 is always positive.

From the fact that $1 > 0$, we immediately deduce that $1 + 1 > 1$, $1 + 1 + 1 > 1 + 1$, and so on. In general, for any positive integer n ,

$$(n + 1) \cdot 1 > n \cdot 1$$

where $n \cdot 1$ designates the unity element of the ring A added to itself n times. Thus, in any ordered integral domain A , *the set of all the multiples of 1 is ordered as in \mathbb{Z}* : namely

$$\dots < (-2) \cdot 1 < (-1) \cdot 1 < 0 < 1 < 2 \cdot 1 < 3 \cdot 1 < \dots$$

The set of all the positive elements of A is denoted by A^+ . An ordered integral domain A is called an *integral system* if every nonempty subset of A^+ has a least element. In other words, if *every nonempty set of positive elements of A has a least element*. This property is called the *well-ordering property* for A^+ .

It is obvious that \mathbb{Z} is an integral system, for every nonempty set of positive integers contains a least number. For example, the smallest element of the set of all the positive even integers is 2. Note that \mathbb{Q} and \mathbb{R} are *not* integral systems. For although both are ordered integral domains, they contain sets of positive numbers, such as $\{x: 0 < x < 1\}$, which have no least element.

In any integral system, *there is no element between 0 and 1*. For suppose A is an integral system in which there are elements x between 0 and 1. Then the set $\{x \in A: 0 < x < 1\}$ is a nonempty set of positive members of A , so by the well-ordering property it has a least element c . That is,

$$0 < c < 1$$

and c is the least element of A with this property. But then (multiplying by c),

$$0 < c^2 < c$$

Thus, c^2 is between 0 and 1 and is less than c , which is impossible.

Thus, there is no element of A between 0 and 1.

Finally, in any integral system, *every element is a multiple of 1*. If that were not the case, we could use the well-ordering principle to pick the least positive element of A which is *not* a multiple of 1 : call it b . Now, $b > 0$ and there are no elements of A between 0 and 1, so $b > 1$. (Remember that b cannot be equal to 1 because b is not a multiple of 1.) Since $b > 1$, it follows that $b - 1 > 0$. But $b - 1 < b$ and b is the *least* positive element which is not a multiple of 1, so $b - 1$ is a multiple of 1. Say

$$b - 1 = n \cdot 1$$

But then $b = n \cdot 1 + 1 = (n + 1) \cdot 1$, which is impossible.

Thus, in any integral system, all the elements are multiples of 1 and these are ordered exactly as in \mathbf{z} . It is now a mere formality to prove that *every integral system is isomorphic to \mathbf{z}* . This is left as [Exercise D](#) at the end of this chapter.

Since every integral system is isomorphic to \mathbf{z} , any two integral systems are isomorphic to each other. Thus \mathbf{z} is, up to isomorphism, the *only* integral system. We have therefore succeeded in giving a complete axiomatic characterization of \mathbf{z} .

Henceforward we consider \mathbf{z} to be defined by the fact that it is an integral system.

The theorem which follows is the basis of proofs by mathematical induction. It is intuitively clear and easy to prove.

Theorem 1 *Let K represent a set of positive integers. Consider the following two conditions’.*

- (i) *1 is in K .*
- (ii) *For any positive integer k , if $k \in K$, then also $k + 1 \in K$.*

If K is any set of positive integers satisfying these two conditions, then K consists of all the positive integers.

PROOF: Indeed, if K does not contain all the positive integers, then by the well-ordering principle, the set of all the positive integers which are *not* in K has a least element. Call it b ; b is the least positive integer *not* in K . By Condition (i), $b \neq 1$, hence $b > 1$.

Thus, $b - 1 > 0$, and $b - 1 \in K$. But then, by Condition (ii), $b \in K$, which is impossible. ■

Let the symbol S_n represent any statement about the positive integer n . For example, S_n might stand for “ n is odd,” or “ n is a prime,” or it might represent an equation such as $(n - 1)(n + 1) = n^2 - 1$ or an inequality such as $n \leq n^2$. If, let us say, S_n stands for $n \leq n^2$, then S_1 asserts that $1 \leq 1^2$, S_2 asserts that $2 \leq 2^2$, S_3 asserts that $3 \leq 3^2$, and so on.

Theorem 2: Principle of mathematical induction *Consider the following conditions :*

- (i) *S_1 is true.*

(ii) For any positive integer k , if S_k is true, then also S_{k+1} is true.

If Conditions (i) and (ii) are satisfied, then S_n is true for every positive integer n .

PROOF: Indeed, if K is the set of all the positive integers k such that S_k is true, then K complies with the conditions of [Theorem 1](#). Thus, K contains all the positive integers. This means that S_n is true for every n . ■

As a simple illustration of how the principle of mathematical induction is applied, let S_n be the statement that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

that is, the sum of the first n positive integers is equal to $n(n+1)/2$. Then S_1 is simply

$$1 = \frac{1 \cdot 2}{2}$$

which is clearly true. Suppose, next, that k is any positive integer and that S_k is true. In other words,

$$1 + 2 + \cdots + k = \frac{k(k+1)}{2}$$

Then, by adding $k+1$ to both sides of this equation, we obtain

$$1 + 2 + \cdots + k + (k+1) = \frac{k(k+1)}{2} + (k+1)$$

that is,

$$1 + 2 + \cdots + (k+1) = \frac{(k+1)(k+2)}{2}$$

However, this last equation is exactly S_{k+1} . We have therefore verified that whenever S_k is true, S_{k+1} also is true. Now, the principle of mathematical induction allows us to conclude that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

for every positive integer n .

A variant of the principle of mathematical induction, called the *principle of strong induction*, asserts that S_n is true for every positive integer n on the conditions that

(i) S_1 is true, and

(ii) For any positive integer k , if S_i is true for every $i < k$, then S_k is true.

The details are outlined in [Exercise H](#) at the end of this chapter.

One of the most important facts about the integers is that any integer m may be divided by any positive integer n to yield a quotient q and a positive remainder r . (The remainder is less than the divisor n .) For example, 25 may be divided by 8 to give a quotient of 3 and a remainder of 1:

$$\underbrace{25}_m = \underbrace{8}_n \times \underbrace{3}_q + \underbrace{1}_r$$

This process is known as the *division algorithm*. It is stated in a precise manner as follows:

Theorem 3: Division algorithm *If m and n are integers and n is positive, there exist unique integers q and r such that*

$$m = nq + r \quad \text{and} \quad 0 \leq r < n$$

We call q the quotient, and r the remainder, in the division of m by n .

PROOF: We begin by showing a simple fact:

$$\text{There exists an integer } x \text{ such that } xn \leq m. \quad (*)$$

Remember that n is positive; hence $n \geq 1$. As for m , either $m \geq 0$ or $m < 0$. We consider these two cases separately:

Suppose $m \geq 0$. Then

$$0 \leq m \quad \text{hence} \quad \underbrace{(0)n}_x \leq m$$

Suppose $m < 0$. We may multiply both sides of $n \geq 1$ by the positive integer $-m$ to get $(-m)n \geq -m$. Adding $mn + m$ to both sides yields

$$\underbrace{mn}_x \leq m$$

Thus, regardless of whether m is positive or negative, there is some integer x such that $xn \leq m$.

Let W be the subset of \mathbb{Z} consisting of all the nonnegative integers which are expressible in the form $m - xn$, where x is any integer. By $(*)$ W is not empty; hence by the well-ordering property, W contains a least integer r . Because $r \in W$, r is nonnegative and is expressible in the form $m - nq$ for some integer q . That is,

$$r \geq 0$$

and

$$r = m - nq$$

Thus, we already have $m = nq + r$ and $0 \leq r$. It remains only to verify that $r < n$. Suppose not: suppose $n \leq r$, that is, $r - n \geq 0$. But

$$r - n = (m - nq) - n = m - n(q + 1)$$

and clearly $r - n < r$. This means that $m - n(q + 1)$ is an element of W less than r , which is impossible because r is the least element of W . We conclude that $n \leq r$ is impossible; hence $r < n$.

The verification that q and r are unique is left as an exercise. ■

EXERCISES

A. Properties of Order Relations in Integral Domains

Let A be an ordered integral domain. Prove the following, for all a , b , and c in A :

- 1 If $a \leq b$ and $b \leq c$, then $a \leq c$.
- 2 If $a \leq b$, then $a + c \leq b + c$.
- 3 If $a \leq b$ and $c \geq 0$, then $ac \leq bc$.
- 4 If $a < b$ and $c < 0$, then $bc < ac$.
- 5 $a < b$ iff $-b < -a$.
- 6 If $a + c < b + c$, then $a < b$.
- 7 If $ac < bc$ and $c > 0$, then $a < b$.
- 8 If $a < b$ and $c < d$, then $a + c < b + d$.

B. Further Properties of Ordered Integral Domains

Let A be an ordered integral domain. Prove the following, for all a , b , and c in A :

- 1 $a^2 + b^2 \geq 2ab$
- 2 $a^2 + b^2 \geq ab$ and $a^2 + b^2 \geq -ab$
- 3 $a^2 + b^2 + c^2 \geq ab + bc + ac$
- 4 $a^2 + b^2 > 0$, if $a^2 + b^2 \neq 0$
- 5 $a + b < ab + 1$, if $a, b > 1$
- 6 $ab + ac + bc + 1 < a + b + c + abc$, if $a, b, c > 1$

C. Uses of Induction

Prove parts 1–7, using the principle of mathematical induction. (Assume n is a positive integer.)

- 1 $1 + 3 + 5 + \dots + (2n - 1) = n^2$ (The sum of the first n odd integers is n^2 .)
- 2 $1^3 + 2^3 + \dots + n^3 = (1 + 2 + \dots + n)^2$
- 3 $1^2 + 2^2 + \dots + (n - 1)^2 < \frac{n^3}{3} < 1^2 + 2^2 + \dots + n^2$
- 4 $1^3 + 2^3 + \dots + (n - 1)^3 < \frac{n^4}{4} < 1^3 + 2^3 + \dots + n^3$
- 5 $1^2 + 2^2 + \dots + n^2 = \frac{1}{6} n(n + 1)(2n + 1)$
- 6 $1^3 + 2^3 + \dots + n^3 = \frac{1}{4} n^2(n + 1)^2$
- 7 $\frac{1}{2!} + \frac{2}{3!} + \dots + \frac{n}{(n + 1)!} = \frac{(n + 1)! - 1}{(n + 1)!}$
- 8 The *Fibonacci sequence* is the sequence of integers F_1, F_2, F_3, \dots defined as follows: $F_1 = 1$; $F_2 = 1$; $F_{n+2} = F_{n+1} + F_n$ for all positive integers n . (That is, every number, after the second one, is the sum of the two preceding ones.) Use induction to prove that for all $n > 0$,

$$F_{n+1} F_{n+2} - F_n F_{n+3} = (-1)^n$$

D. Every Integral System Is Isomorphic to \mathbf{z}

Let A be an integral system. Let $h: \mathbf{z} \rightarrow A$ be defined by: $h(n) = n \cdot 1$. The purpose of this exercise is to prove that h is an isomorphism, from which it follows that $A \cong \mathbf{z}$

- 1 Prove: For every positive integer n , $n \cdot 1 > 0$. What is the characteristic of A
- 2 Prove that h is injective and surjective.
- 3 Prove that h is an isomorphism.

E. Absolute Values

In any ordered integral domain, define $|a|$ by

$$|a| = \begin{cases} a & \text{if } a \geq 0 \\ -a & \text{if } a < 0 \end{cases}$$

Using this definition, prove the following:

- 1 $|-a| = |a|$
- 2 $a \leq |a|$
- 3 $a \leq -|a|$
- 4 If $b > 0$, $|a| \leq b$ iff $-b \leq a \leq b$
- 5 $|a+b| \leq |a| + |b|$
- 6 $|a-b| \leq |a| + |b|$
- 7 $|ab| = |a| \cdot |b|$
- 8 $|a| - |b| \leq |a - b|$
- 9 $||a| - |b|| \leq |a - b|$

F. Problems on the Division Algorithm

Prove parts 1–3, where k , m , n , q , and r designate integers.

- 1 Let $n > 0$ and $k > 0$. If q is the quotient and r is the remainder when m is divided by n , then q is the quotient and kr is the remainder when km is divided by kn .
- # 2 Let $n > 0$ and $k > 0$. If q is the quotient when m is divided by n , and q_1 is the quotient when q is divided by k , then q_1 is the quotient when m is divided by nk .
- 3 If $n \neq 0$, there exist q and r such that $m = nq + r$ and $0 \leq r < |n|$. (Use Theorem 3, and consider the case when $n < 0$.)
- 4 In Theorem 3, suppose $m = nq_1 + r_1 = nq_2 + r_2$ where $0 \leq r_1, r_2 < n$. Prove that $r_1 - r_2 = 0$. [HINT: Consider the difference $(nq_1 + r_1 - (nq_2 + r_2))$.]
- 5 Use part 4 to prove that $q_1 - q_2 = 0$. Conclude that the quotient and remainder, in the division algorithm, are unique.
- 6 If r is the remainder when m is divided by n , prove that $m = r$ in \mathbf{z}_n .

G. Laws of Multiples

The purpose of this exercise is to give rigorous proofs (using induction) of the basic identities involved in the use of exponents or multiples. If A is a ring and $a \in A$, we define $n \cdot a$ (where n is any positive integer) by the pair of conditions:

$$(i) \ 1 \cdot a = a, \quad \text{and} \quad (ii) \ (n + 1) \cdot a = n \cdot a + a$$

Use mathematical induction (with the above definition) to prove that the following are true for all positive integers n and all elements $a, b \in A$:

- 1 $n \cdot (a + b) = n \cdot a + n \cdot b$
- 2 $(n + m) \cdot a = n \cdot a + m \cdot a$
- 3 $(n \cdot a)b = a(n \cdot b) = n \cdot (ab)$
- 4 $m \cdot (n \cdot a) = (mn) \cdot a$
- 5 $n \cdot a = (n \cdot 1)a$ where 1 is the unity element of A
- 6 $(n \cdot a)(m \cdot b) = (nm) \cdot ab$ (Use parts 3 and 4.)

H. Principle of Strong Induction

Prove the following in \mathbb{Z} :

- 1 Let K denote a set of positive integers. Consider the following conditions:

- (i) $1 \in K$.
- (ii) For any positive integer k , if every positive integer less than k is in K , then $k \in K$.

If K satisfies these two conditions, prove that K contains all the positive integers.

- 2 Let S_n represent any statement about the positive integer n . Consider the following conditions:

- (i) S_1 is true.
- (ii) For any positive integer k , if S_i is true for every $i < k$, S_k is true.

If Conditions (i) and (ii) are satisfied, prove that S_n is true for every positive integer n .

RINGS OF POLYNOMIALS

In elementary algebra an important role is played by polynomials in an unknown x . These are expressions such as

$$2x^3 - \frac{1}{2}x^2 + 3$$

whose terms are grouped in powers of x . The exponents, of course, are positive integers and the coefficients are real or complex numbers.

Polynomials are involved in countless applications—applications of every kind and description. For example, polynomial functions are the easiest functions to compute, and therefore one commonly attempts to approximate arbitrary functions by polynomial functions. A great deal of effort has been expended by mathematicians to find ways of achieving this.

Aside from their uses in science and computation, polynomials come up very naturally in the general study of rings, as the following example will show:

Suppose we wish to enlarge the ring \mathbf{z} by adding to it the number π . It is easy to see that we will have to adjoin to \mathbf{z} other new numbers besides just π ; for the enlarged ring (containing π as well as all the integers) will also contain such things as $-\pi$, $\pi + 7$, $6\pi^2 - 11$, and so on.

As a matter of fact, any ring which contains \mathbf{z} as a subring and which also contains the number π will have to contain every number of the form

$$a\pi^n + b\pi^{n-1} + \dots + k\pi + l$$

where a, b, \dots, k, l are integers. In other words, it will contain *all the polynomial expressions in π with integer coefficients*.

But the set of all the polynomial expressions in π with integer coefficients is a ring. (It is a subring of \mathbb{R} because it is obvious that the sum and product of any two polynomials in π is again a polynomial in π .) This ring contains \mathbf{z} because every integer a is a polynomial with a constant term only, and it also contains π .

Thus, if we wish to enlarge the ring \mathbf{z} by adjoining to it the new number π , it turns out that the “next

largest” ring after \mathbf{z} which contains \mathbf{z} as a subring and includes π , is exactly the ring of all the polynomials in π with coefficients in \mathbf{z} .

As this example shows, aside from their practical applications, polynomials play an important role in the scheme of ring theory because they are precisely what we need when we wish to enlarge a ring by adding new elements to it.

In elementary algebra one considers polynomials whose coefficients are real numbers, or in some cases, complex numbers. As a matter of fact, the properties of polynomials are pretty much independent of the exact nature of their coefficients. All we need to know is that the coefficients are contained in some ring. For convenience, we will assume this ring is a commutative ring with unity.

Let A be a commutative ring with unity. Up to now we have used letters to denote elements or sets, but now we will use the letter x in a different way. In a polynomial expression such as $ax^2 + bx + c$, where $a, b, c \in A$, we do not consider x to be an element of A , but rather x is a symbol which we use in an entirely formal way. Later we will allow the substitution of other things for x , but at present x is *simply a placeholder*.

Notationally, the terms of a polynomial may be listed in either ascending or descending order. For example, $4x^3 - 3x^2 + x + 1$ and $1 + x - 3x^2 + 4x^3$ denote the same polynomial. In elementary algebra descending order is preferred, but for our purposes ascending order is more convenient.

Let A be a commutative ring with unity, and x an arbitrary symbol. Every expression of the form

$$a_0 + a_1x + a_2x^2 + \dots + a_nx^n$$

*is called a **polynomial in x with coefficients in A** , or more simply, a **polynomial in x over A** . The expressions a_kx^k , for $k \in \{1, \dots, n\}$, are called the **terms** of the polynomial.*

Polynomials in x are designated by symbols such as $a(x)$, $b(x)$, $q(x)$, and so on. If $a(x) = a_0 + a_1x + \dots + a_nx^n$ is any polynomial and a_kx^k is any one of its terms, a_k is called the *coefficient* of x^k . By the *degree* of a polynomial $a(x)$ we mean the *greatest n such that the coefficient of x^n is not zero*. In other words, if $a(x)$ has degree n , this means that $a_n \neq 0$ but $a_m = 0$ for every $m > n$. The degree of $a(x)$ is symbolized by

$$\deg a(x)$$

For example, $1 + 2x - 3x^2 + x^3$ is a polynomial degree 3.

The polynomial $0 + 0x + 0x^2 + \dots$ all of whose coefficients are equal to zero is called the *zero polynomial*, and is symbolized by 0. It is the only polynomial whose *degree is not defined* (because it has no nonzero coefficient).

If a nonzero polynomial $a(x) = a_0 + a_1x + \dots + a_nx^n$ has degree n , then a_n is called its *leading coefficient*: it is the last nonzero coefficient of $a(x)$. The term a_nx^n is then called its *leading term*, while a_0 is called its *constant term*.

If a polynomial $a(x)$ has degree zero, this means that its constant term a_0 is its only nonzero term: $a(x)$ is a *constant polynomial*. Beware of confusing a polynomial of degree zero with the zero polynomial.

Two polynomials $a(x)$ and $b(x)$ are *equal* if they have the same degree and corresponding coefficients are equal. Thus, if $a(x) = a_0 + \dots + a_nx^n$ is of degree n , and $b(x) = b_0 + \dots + b_mx^m$ is of degree

m , then $a(x) = b(x)$ iff $n = m$ and $a_k = b_k$ for each k from 0 to n .

The familiar sigma notation for sums is useful for polynomials. Thus,

$$a(x) = a_0 + a_1x + \cdots + a_nx^n = \sum_{k=0}^n a_kx^k$$

with the understanding that $x^0 = 1$.

Addition and multiplication of polynomials is familiar from elementary algebra. We will now define these operations formally. Throughout these definitions we let $a(x)$ and $b(x)$ stand for the following polynomials:

$$\begin{aligned} a(x) &= a_0 + a_1x + \cdots + a_nx^n \\ b(x) &= b_0 + b_1x + \cdots + b_nx^n \end{aligned}$$

Here we do *not* assume that $a(x)$ and $b(x)$ have the same degree, but allow ourselves to insert zero coefficients if necessary to achieve uniformity of appearance.

We add polynomials by adding corresponding coefficients. Thus,

$$a(x) + b(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n$$

Note that the degree of $a(x) + b(x)$ is less than or equal to the *higher* of the two degrees, $\deg a(x)$ and $\deg b(x)$.

Multiplication is more difficult, but quite familiar:

$$\begin{aligned} a(x)b(x) &= a_0b_0 + (a_0b_1 + b_0a_1)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots + a_nb_nx^{2n} \end{aligned}$$

In other words, the product of $a(x)$ and $b(x)$ is the polynomial

$$c(x) = c_0 + c_1x + \cdots + c_{2n}x^{2n}$$

whose k th coefficient (for any k from 0 to $2n$) is

$$c_k = \sum_{i+j=k} a_ib_j$$

This is the sum of all the a_ib_j for which $i + j = k$. Note that $\deg [a(x)b(x)] \leq \deg a(x) + \deg b(x)$.

If A is any ring, the symbol

$$A[x]$$

designates the set of all the polynomials in x whose coefficients are in A , with addition and multiplication of polynomials as we have just defined them.

Theorem 1 *Let A be a commutative ring with unity. Then $A[x]$ is a commutative ring with unity.*

PROOF: To prove this theorem, we must show systematically that $A[x]$ satisfies all the axioms of a

commutative ring with unity. Throughout the proof, let $a(x)$, $b(x)$, and $c(x)$ stand for the following polynomials:

$$\begin{aligned} a(x) &= a_0 + a_1x + \cdots + a_nx^n \\ b(x) &= b_0 + b_1x + \cdots + b_nx^n \\ \text{and} \quad c(x) &= c_0 + c_1x + \cdots + c_nx^n \end{aligned}$$

The axioms which involve only addition are easy to check: for example, addition is commutative because

$$\begin{aligned} a(x) + b(x) &= (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n \\ &= (b_0 + a_0) + (b_1 + a_1)x + \cdots + (b_n + a_n)x^n \\ &= b(x) + a(x) \end{aligned}$$

The associative law of addition is proved similarly, and is left as an exercise. The zero polynomial has already been described, and the negative of $a(x)$ is

$$-a(x) = (-a_0) + (-a_1)x + \cdots + (-a_n)x^n$$

To prove that multiplication is associative requires some care. Let $b(x)c(x) = d(x)$, where $d(x) = d_0 + d_1x + \cdots + d_{2n}x^{2n}$. By the definition of polynomial multiplication, the k th coefficient of $b(x)c(x)$ is

$$d_k = \sum_{i+j=k} b_i c_j$$

Then $a(x)[b(x)c(x)] = a(x)d(x) = e(x)$, where $e(x) = e_0 + e_1x + \cdots + e_{3n}x^{3n}$. Now, the l th coefficient of $a(x)d(x)$ is

$$e_l = \sum_{h+k=l} a_h d_k = \sum_{h+k=l} a_h \left(\sum_{i+j=k} b_i c_j \right)$$

It is easy to see that the sum on the right consists of all the terms $a_h b_i c_j$ such that $h + i + j = l$. Thus,

$$e_l = \sum_{h+i+j=l} a_h b_i c_j$$

For each l from 0 to $3n$, e_l is the l th coefficient of $a(x)[b(x)c(x)]$.

If we repeat this process to find the l th coefficient of $[a(x) b(x)]c(x)$, we discover that it, too, is e_l . Thus,

$$a(x)[b(x)c(x)] = [a(x)b(x)]c(x)$$

To prove the distributive law, let $a(x)[b(x) + c(x)] = d(x)$ where $d(x) = d_0 + d_1x + \cdots + d_{2n}x^{2n}$. By the definitions of polynomial addition and multiplication, the k th coefficient $a(x)[b(x) + c(x)]$ is

$$\begin{aligned}
 d_k &= \sum_{i+j=k} a_i(b_j + c_j) = \sum_{i+j=k} (a_i b_j + a_i c_j) \\
 &= \sum_{i+j=k} a_i b_j + \sum_{i+j=k} a_i c_j
 \end{aligned}$$

But $\sum_{i+j=k} a_i b_j$ is exactly the k th coefficient of $a(x) b(x)$, and $\sum_{i+j=k} a_i c_j$ is the k th coefficient of $a(x)c(x)$, hence d_k is equal to the k th coefficient of $a(x) b(x) + a(x)c(x)$. This proves that

$$a(x)[b(x) + c(x)] = a(x)b(x) + a(x)c(x)$$

The commutative law of multiplication is simple to verify and is left to the student. Finally, the unity polynomial is the constant polynomial 1. ■

Theorem 2 *If A is an integral domain, then $A[x]$ is an integral domain.*

PROOF: If $a(x)$ and $b(x)$ are nonzero polynomials, we must show that their product $a(x) b(x)$ is not zero. Let a_n be the leading coefficient of $a(x)$, and b_m the leading coefficient of $b(x)$. By definition, $a_n \neq 0$, and $b_m \neq 0$. Thus $a_n b_m \neq 0$ because A is an integral domain. It follows that $a(x) b(x)$ has a nonzero coefficient (namely, $a_n b_m$), so it is not the zero polynomial. ■

If A is an integral domain, we refer to $A[x]$ as a *domain of polynomials*, because $A[x]$ is an integral domain. Note that by the preceding proof, if a_n and b_m are the leading coefficients of $a(x)$ and $b(x)$, then $a_n b_m$ is the leading coefficient of $a(x) b(x)$. Thus, $\deg a(x)b(x) = n + m$: *In a domain of polynomials $A[x]$, where A is an integral domain,*

$$\deg[a(x) \cdot b(x)] = \deg a(x) + \deg b(x)$$

In the remainder of this chapter we will look at a property of polynomials which is of special interest when all the coefficients lie in a field. Thus, from this point forward, let F be a field, and let us consider polynomials belonging to $F[x]$.

It would be tempting to believe that if F is a field then $F[x]$ also is a field. However, this is not so, for one can easily see that the multiplicative inverse of a polynomial is not generally a polynomial. Nevertheless, by [Theorem 2](#), $F[x]$ is an integral domain.

Domains of polynomials over a *field* do, however, have a very special property: any polynomial $a(x)$ may be divided by any nonzero polynomial $b(x)$ to yield a quotient $q(x)$ and a remainder $r(x)$. The remainder is either 0, or if not, its degree is less than the degree of the divisor $b(x)$. For example, x^2 may be divided by $x - 2$ to give a quotient of $x + 2$ and a remainder of 4:

$$\begin{array}{ccccccc}
 x^2 & = & (x-2)(x+2) & + & 4 \\
 \underbrace{}_{a(x)} & & \underbrace{}_{b(x)} \underbrace{}_{q(x)} & & \underbrace{}_{r(x)}
 \end{array}$$

This kind of polynomial division is familiar to every student of elementary algebra. It is customarily set up as follows:

$$\begin{array}{rcl}
 \text{Divisor} \longrightarrow x-2 & \overline{) \begin{array}{r} x+2 \\ x^2-2x \\ 2x-4 \\ 4 \end{array}} & \longleftarrow \begin{array}{l} \text{Quotient } q(x) \\ \text{Dividend } b(x) \end{array} \\
 a(x) & & \\
 & & \longleftarrow \text{Remainder } r(x)
 \end{array}$$

The process of polynomial division is formalized in the next theorem.

Theorem 3: Division algorithm for polynomials *If $a(x)$ and $b(x)$ are polynomials over a field F , and $b(x) \neq 0$, there exist polynomials $q(x)$ and $r(x)$ over F such that*

$$a(x) = b(x)q(x) + r(x)$$

and

$$r(x) = 0 \quad \text{or} \quad \deg r(x) < \deg b(x)$$

PROOF: Let $b(x)$ remain fixed, and let us show that every polynomial $a(x)$ satisfies the following condition:

There exist polynomials $q(x)$ and $r(x)$ over F such that $a(x) = b(x)q(x) + r(x)$, and $r(x) = 0$ or $\deg r(x) < \deg b(x)$.

We will assume there are polynomials $a(x)$ which do *not* fulfill the condition, and from this assumption we will derive a contradiction. Let $a(x)$ be a polynomial of *lowest degree* which fails to satisfy the conditions. Note that $a(x)$ cannot be zero, because we can express 0 as $0 = b(x) \cdot 0 + 0$, whereby $a(x)$ would satisfy the conditions. Furthermore, $\deg a(x) \geq \deg b(x)$, for if $\deg a(x) < \deg b(x)$ then we could write $a(x) = b(x) \cdot 0 + a(x)$, so again $a(x)$ would satisfy the given conditions.

Let $a(x) = a_0 + \dots + a_n x^n$ and $b(x) = b_0 + \dots + b_m x^m$. Define a new polynomial

$$\begin{aligned}
 A(x) &= a(x) - \frac{a_n}{b_m} x^{n-m} b(x) \\
 &= a(x) - \left(b_0 \frac{a_n}{b_m} x^{n-m} + b_1 \frac{a_n}{b_m} x^{n-m+1} + \dots + b_m \frac{a_n}{b_m} x^{n-m+m} \right)
 \end{aligned} \tag{1}$$

This expression is the difference of two polynomials both of degree n and both having the same leading term $a_n x^n$. Because $a_n x^n$ cancels in the subtraction, $A(x)$ has degree *less than* n .

Remember that $a(x)$ is a polynomial of *least degree* which fails to satisfy the given condition; hence $A(x)$ *does satisfy* it. This means there are polynomials $p(x)$ and $r(x)$ such that

$$A(x) = b(x)p(x) + r(x)$$

where $r(x) = 0$ or $\deg r(x) < \deg b(x)$. But then

$$a(x) = A(x) + \frac{a_n}{b_m} x^{n-m} b(x) \quad \text{by Equation (1)}$$

$$= b(x)p(x) + r(x) + \frac{a_n}{b_m} x^{n-m} b(x)$$

$$= b(x) \left[p(x) + \frac{a_n}{b_m} x^{n-m} \right] + r(x)$$

If we let $p(x) + (a_n/b_m)x^{n-m}$ be renamed $q(x)$, then $a(x) = b(x)q(x) + r(x)$, so $a(x)$ fulfills the given condition. This is a contradiction, as required. ■

EXERCISES

A. Elementary Computation in Domains of Polynomials

REMARK ON NOTATION: In some of the problems which follow, we consider polynomials with coefficients in \mathbf{z}_n for various n . To simplify notation, we denote the elements of \mathbf{z}_n by $1, 2, \dots, n-1$ rather than the more correct $\overline{1}, \overline{2}, \dots, \overline{n-1}$.

1 Let $a(x) = 2x^2 + 3x + 1$ and $b(x) = x^3 + 5x^2 + x$. Compute $a(x) + b(x)$, $a(x) - b(x)$ and $a(x)b(x)$ in $\mathbf{z}[x]$, $\mathbf{z}_5[x]$, $\mathbf{z}_6[x]$, and $\mathbf{z}_7[x]$.

2 Find the quotient and remainder when $x^3 + x^2 + x + 1$ is divided by $x^2 + 3x + 2$ in $\mathbf{z}[x]$ and in $\mathbf{z}_5[x]$.

3 Find the quotient and remainder when $x^3 + 2$ is divided by $2x^2 + 3x + 4$ in $\mathbf{z}[x]$, in $\mathbf{z}_3[x]$, and in $\mathbf{z}_5[x]$.

We call $b(x)$ a *factor* of $a(x)$ if $a(x) = b(x)q(x)$ for some $q(x)$, that is, if the remainder when $a(x)$ is divided by $b(x)$ is equal to zero.

4 Show that the following is true in $A[x]$ for any ring A : For any odd n ,

(a) $x + 1$ is a factor of $x^n + 1$.

(b) $x + 1$ is a factor of $x^n + x^{n-1} + \dots + x + 1$.

5 Prove the following: In $\mathbf{z}_3[x]$, $x + 2$ is a factor of $x^m + 2$, for all m . In $\mathbf{z}_n[x]$, $x + (x - 1)$ is a factor of $x^m + (n - 1)$, for all m and n .

6 Prove that there is no integer m such that $3x^2 + 4x + m$ is a factor of $6x^4 + 50$ in $\mathbf{z}[x]$.

7 For what values of n is $x^2 + 1$ a factor of $x^5 + 5x + 6$ in $\mathbf{z}_n[x]$?

B. Problems Involving Concepts and Definitions

1 Is $x^8 + 1 = x^3 + 1$ in $\mathbf{z}_5[x]$? Explain your answer.

2 Is there any ring A such that in $A[x]$, some polynomial of degree 2 is equal to a polynomial of degree 4? Explain.

3 Write all the quadratic polynomials in $\mathbf{z}_5[x]$. How many are there? How many cubic polynomials are there in $\mathbf{z}_5[x]$? More generally, how many polynomials of degree m are there in $\mathbf{z}_n[x]$?

4 Let A be an integral domain; prove the following:

If $(x + 1)^2 = x^2 + 1$ in $A[x]$, then A must have characteristic 2.

If $(x + 1)^4 = x^4 + 1$ in $A[x]$, then A must have characteristic 2.

If $(x + 1)^6 = x^6 + 2x^3 + 1$ in $A[x]$, then A must have characteristic 3.

5 Find an example of each of the following in $\mathbf{Z}_8[x]$: a divisor of zero, an invertible element. (Find nonconstant examples.)

6 Explain why x cannot be invertible in any $A[x]$, hence no domain of polynomials can ever be a field.

7 There are rings such as P_3 in which every element $\neq 0, 1$ is a divisor of zero. Explain why this cannot happen in any ring of polynomials $A[x]$, even when A is *not* an integral domain.

8 Show that in every $A[x]$, there are elements $\neq 0, 1$ which are not idempotent, and elements $\neq 0, 1$ which are not nilpotent.

C. Rings $A[x]$ Where A Is Not an Integral Domain

1 Prove: If A is not an integral domain, neither is $A[x]$.

2 Give examples of divisors of zero, of degrees 0, 1, and 2, in $\mathbf{Z}_4[x]$.

3 In $\mathbf{Z}_{10}[x]$, $(2x + 2)(2x + 2) = (2x + 2)(5x^3 + 2x + 2)$, yet $(2x + 2)$ cannot be canceled in this equation. Explain why this is possible in $\mathbf{Z}_{10}[x]$, but not in $\mathbf{Z}_5[x]$.

4 Give examples in $\mathbf{Z}_4[x]$, in $\mathbf{Z}_6[x]$, and in $\mathbf{Z}_9[x]$ of polynomials $a(x)$ and $b(x)$ such that $\deg a(x)b(x) < \deg a(x) + \deg b(x)$.

5 If A is an integral domain, we have seen that in $A[x]$,

$$\deg a(x)b(x) = \deg a(x) + \deg b(x)$$

Show that if A is *not* an integral domain, we can always find polynomials $a(x)$ and $b(x)$ such that $\deg a(x)b(x) < \deg a(x) + \deg b(x)$.

6 Show that if A is an integral domain, the only invertible elements in $A[x]$ are the constant polynomials with inverses in A . Then show that in $\mathbf{Z}_4[x]$ there are invertible polynomials of all degrees.

7 Give all the ways of factoring x^2 into polynomials of degree 1 in $\mathbf{Z}_9[x]$; in $\mathbf{Z}_5[x]$. Explain the difference in behavior.

8 Find all the square roots of $x^2 + x + 4$ in $\mathbf{Z}_5[x]$. Show that in $\mathbf{Z}_8[x]$, there are infinitely many square roots of 1.

D. Domains $A[x]$ Where A Has Finite Characteristic

In each of the following, let A be an integral domain:

1 Prove that if A has characteristic p , then $A[x]$ has characteristic p .

2 Use part 1 to give an example of an infinite integral domain with finite characteristic.

3 Prove: If A has characteristic 3, then $x + 2$ is a factor of $x^m + 2$ for all m . More generally, if A has characteristic p , then $x + (p - 1)$ is a factor of $x^m + (p - 1)$ for all m .

4 Prove that if A has characteristic p , then in $A[x]$, $(x + c)^p = x^p + c^p$. (You may use essentially the same argument as in the proof of [Theorem 3, Chapter 20](#).)

5 Explain why the following “proof of part 4 is not valid: $(x + c)^p = x^p + c^p$ in $A[x]$ because $(a + c)^p = a^p + c^p$ for all $a, c \in A$. (Note the following example: in \mathbb{Z}_2 , $a^2 + 1 = a^4 + 1$ for every a , yet $x^2 + 1 \neq x^4 + 1$ in $\mathbb{Z}_2[x]$.)

6 Use the same argument as in part 4 to prove that if A has characteristic p , then $[a(x) + b(x)]^p = a(x)^p + b(x)^p$ for any $a(x), b(x) \in A[x]$. Use this to prove:

$$(a_0 + a_1x + \cdots + a_nx^n)^p = a_0^p + a_1^p x^p + \cdots + a_n^p x^{np}$$

E. Subrings and Ideals in $A[x]$

1 Show that if B is a subring of A , then $B[x]$ is a subring of $A[x]$.

2 If B is an *ideal* of A , $B[x]$ is an ideal of $A[x]$.

3 Let S be the set of all the polynomials $a(x)$ in $A[x]$ for which every coefficient a_i for *odd* i is equal to zero. Show that S is a subring of $A[x]$. Why is the same not true when “odd” is replaced by “even”?

4 Let J consist of all the elements in $A[x]$ whose constant coefficient is equal to zero. Prove that J is an ideal of $A[x]$.

5 Let J consist of all the polynomials $a_0 + a_1x + \cdots + a_nx^n$ in $A[x]$ such that $a_0 + a_1 + \cdots + a_n = 0$. Prove that J is an ideal of $A[x]$.

6 Prove that the ideals in both parts 4 and 5 are *prime* ideals. (Assume A is an integral domain.)

F. Homomorphisms of Domains of Polynomials

Let A be an integral domain.

1 Let $h : A[x] \rightarrow A$ map every polynomial to its constant coefficient; that is,

$$h(a_0 + a_1x + \cdots + a_nx^n) = a_0$$

Prove that h is a homomorphism from $A[x]$ onto A , and describe its kernel.

2 Explain why the kernel of h in part 1 consists of all the products $xa(x)$, for all $a(x) \in A[x]$. Why is this the same as the principal ideal $\langle x \rangle$ in $A[x]$?

3 Using parts 1 and 2, explain why $A[x]/\langle x \rangle \cong A$.

4 Let $g : A[x] \rightarrow A$ send every polynomial to the sum of its coefficients. Prove that g is a surjective homomorphism, and describe its kernel.

5 If $c \in A$, let $h : A[x] \rightarrow A[x]$ be defined by $h(a(x)) = a(cx)$, that is,

$$h(a_0 + a_1x + \cdots + a_nx^n) = a_0 + a_1cx + a_2c^2x^2 + \cdots + a_nc^n x^n$$

Prove that h is a homomorphism and describe its kernel.

6 If h is the homomorphism of part 5, prove that h is an automorphism (isomorphism from $A[x]$ to itself) iff c is invertible.

G. Homomorphisms of Polynomial Domains Induced by a Homomorphism of the Ring of Coefficients

Let A and B be rings and let $h : A \rightarrow B$ be a homomorphism with kernel K . Define $\bar{h} : A[x] \rightarrow B[x]$ by

$$\bar{h}(a_0 + a_1x + \cdots + a_nx^n) = h(a_0) + h(a_1)x + \cdots + h(a_n)x^n$$

(We say that \bar{h} is *induced by* h .)

- 1 Prove that \bar{h} is a homomorphism from $A[x]$ to $B[x]$.
- 2 Describe the kernel \bar{K} of \bar{h} .
- # 3 Prove that \bar{h} is surjective iff h is surjective.
- 4 Prove that \bar{h} is injective iff h is injective.
- 5 Prove that if $a(x)$ is a factor of $b(x)$, then $\bar{h}(a(x))$ is a factor of $\bar{h}(b(x))$.
- 6 If $h : \mathbb{Z} \rightarrow \mathbb{Z}_n$ is the natural homomorphism, let $\bar{h} : \mathbb{Z}[x] \rightarrow \mathbb{Z}_n[x]$ be the homomorphism induced by h . Prove that $\bar{h}(a(x)) = 0$ iff n divides every coefficient of $a(x)$.
- 7 Let \bar{h} be as in part 6, and let n be a prime. Prove that if $a(x)b(x) \in \ker \bar{h}$, then either $a(x)$ or $b(x)$ is in $\ker \bar{h}$. (HINT: Use [Exercise F2](#) of [Chapter 19](#).)

H. Polynomials in Several Variables

$A[x_1, x_2]$ denotes the ring of all the polynomials in *two letters* x_1 and x_2 with coefficients in A . For example, $x^2 - 2xy + y^2 + x - 5$ is a quadratic polynomial in $\mathbb{Q}[x, y]$. More generally, $A[x_1, \dots, x_n]$ is the ring of the polynomials in n *letters* x_1, \dots, x_n with coefficients in A . Formally it is defined as follows: Let $A[x_1]$ be denoted by A_1 ; then $A_1[x_2]$ is $A[x_1, x_2]$. Continuing in this fashion, we may adjoin one new letter x_i at a time, to get $A[x_1, \dots, x_n]$.

- 1 Prove that if A is an integral domain, then $A[x_1, \dots, x_n]$ is an integral domain.
- 2 Give a reasonable definition of the *degree* of any polynomial $p(x, y)$ in $A[x, y]$ and then list all the polynomials of degree ≤ 3 in $\mathbb{Z}_3[x, y]$.

Let us denote an arbitrary polynomial $p(x, y)$ in $A[x, y]$ by $\sum a_{ij}x^i y^j$ where Σ ranges over *some* pairs i, j of nonnegative integers.

- 3 Imitating the definitions of sum and product of polynomials in $A[x]$, give a definition of sum and product of polynomials in $A[x, y]$.
- 4 Prove that $\deg a(x, y)b(x, y) = \deg a(x, y) + \deg b(x, y)$ if A is an integral domain.

I. Fields of Polynomial Quotients

Let A be an integral domain. By the closing part of [Chapter 20](#), every integral domain can be extended to a “field of quotients.” Thus, $A[x]$ can be extended to a field of polynomial quotients, which is denoted by $A(x)$. Note that $A(x)$ consists of all the fractions $a(x)/b(x)$ for $a(x)$ and $b(x) \neq 0$ in $A[x]$, and these fractions are added, subtracted, multiplied, and divided in the customary way.

- 1 Show that $A(x)$ has the same characteristic as A .
- 2 Using part 1, explain why there is an infinite field of characteristic p , for every prime p .
- 3 If A and B are integral domains and $h : A \rightarrow B$ is an isomorphism, prove that h determines an

isomorphism $\bar{h} : A(x) \rightarrow B(x)$.

J. Division Algorithm: Uniqueness of Quotient and Remainder

In the division algorithm, prove that $q(x)$ and $r(x)$ are uniquely determined. [HINT: Suppose $a(x) = b(x)q_1(x) + r_1(x) = b(x)q_2(x) + r_2(x)$, and subtract these two expressions, which are both equal to $a(x)$.]

FACTORING POLYNOMIALS

Just as every integer can be factored into primes, so every polynomial can be factored into “irreducible” polynomials which cannot be factored further. As a matter of fact, polynomials behave very much like integers when it comes to factoring them. This is especially true when the polynomials have all their coefficients in a *field*.

Throughout this chapter, we let F represent some field and we consider polynomials over F . It will be found that $F[x]$ has a considerable number of properties in common with \mathbf{Z} . To begin with, all the ideals of $F[x]$ are principal ideals, which was also the case for the ideals of \mathbf{Z} .

Note carefully that in $F[x]$, the principal ideal generated by a polynomial $a(x)$ consists of all the products $a(x)s(x)$ as $a(x)$ remains fixed and $s(x)$ ranges over all the members of $F[x]$.

Theorem 1 *Every ideal of $F[x]$ is principal.*

PROOF: Let J be any ideal of $F[x]$. If J contains nothing but the zero polynomial, J is the principal ideal generated by 0. If there are nonzero polynomials in J , let $b(x)$ be any polynomial of *lowest degree* in J . We will show that $J = \langle (b(x)) \rangle$, which is to say that every element of J is a polynomial multiple $b(x)q(x)$ of $b(x)$.

Indeed, if $a(x)$ is any element of J , we may use the division algorithm to write $a(x) = b(x)q(x) + r(x)$, where $r(x) = 0$ or $\deg r(x) < \deg b(x)$. Now, $r(x) = a(x) - b(x)q(x)$; but $a(x)$ was chosen in J , and $b(x) \in J$; hence $b(x)q(x) \in J$. It follows that $r(x)$ is in J .

If $r(x) \neq 0$, its degree is less than the degree of $b(x)$. But this is impossible because $b(x)$ is a polynomial of lowest degree in J . Therefore, of necessity, $r(x) = 0$.

Thus, finally, $a(x) = b(x)q(x)$; so every member of J is a multiple of $b(x)$, as claimed. ■

It follows that every ideal J of $F[x]$ is principal. In fact, as the proof above indicates, J is generated by any one of its members of lowest degree.

Throughout the discussion which follows, remember that we are considering polynomials in a fixed domain $F[x]$ where F is a *field*.

Let $a(x)$ and $b(x)$ be in $F[x]$. We say that $b(x)$ is a *multiple* of $a(x)$ if

$$b(x) = a(x)s(x)$$

for some polynomial $s(x)$ in $F[x]$. If $b(x)$ is a multiple of $a(x)$, we also say that $a(x)$ is a *factor* of $b(x)$, or

that $a(x)$ divides $b(x)$. In symbols, we write

$$a(x)|b(x)$$

Every nonzero constant polynomial divides every polynomial. For if $c \neq 0$ is constant and $a(x) = a_0 + \dots + a_n x^n$, then

$$a_0 + a_1 x + \dots + a_n x^n = c \left(\frac{a_0}{c} + \frac{a_1}{c} x + \dots + \frac{a_n}{c} x^n \right)$$

hence $c | a(x)$. A polynomial $a(x)$ is invertible iff it is a divisor of the unity polynomial 1. But if $a(x)b(x) = 1$, this means that $a(x)$ and $b(x)$ both have degree 0, that is, are constant polynomials: $a(x) = a$, $b(x) = b$, and $ab = 1$. Thus,

the invertible elements of $F[x]$ are all the nonzero constant polynomials.

A pair of nonzero polynomials $a(x)$ and $b(x)$ are called *associates* if they divide one another: $a(x)|b(x)$ and $b(x)|a(x)$. That is to say,

$$a(x) = b(x)c(x) \quad \text{and} \quad b(x) = a(x)d(x)$$

for some $c(x)$ and $d(x)$. If this happens to be the case, then

$$a(x) = b(x)c(x) = a(x)d(x)c(x)$$

hence $d(x)c(x) = 1$ because $F[x]$ is an integral domain. But then $c(x)$ and $d(x)$ are constant polynomials, and therefore $a(x)$ and $b(x)$ are constant multiples of each other. Thus, in $F[x]$,

$a(x)$ and $b(x)$ are associates iff they are constant multiples of each other.

If $a(x) = a_0 + \dots + a_n x^n$, the associates of $a(x)$ are all its nonzero constant multiples. Among these multiples is the polynomial

$$\frac{a_0}{a_n} + \frac{a_1}{a_n} x + \dots + x^n$$

which is equal to $(1/a_n)a(x)$, and which has 1 as its leading coefficient. Any polynomial whose leading coefficient is equal to 1 is called *monic*. Thus, *every nonzero polynomial $a(x)$ has a unique monic associate*. For example, the monic associate of $3 + 4x + 2x^3$ is $\frac{3}{2} + 2x + x^3$.

A polynomial $d(x)$ is called a *greatest common divisor* of $a(x)$ and $b(x)$ if $d(x)$ divides $a(x)$ and $b(x)$, and is a multiple of any other common divisor of $a(x)$ and $b(x)$; in other words,

- (i) $d(x)|a(x)$ and $d(x)|b(x)$, and
- (ii) For any $u(x)$ in $F[x]$, if $u(x)|a(x)$ and $u(x)|b(x)$, then $u(x)|d(x)$. According to this definition, two different gcd's of $a(x)$ and $b(x)$ divide each other, that is, are associates. Of all the possible gcd's of $a(x)$ and $b(x)$, we select the monic one, call it *the* gcd of $a(x)$ and $b(x)$, and denote it by $\gcd[a(x), b(x)]$.

It is important to know that any pair of polynomials always *has* a greatest common divisor.

Theorem 2 *Any two nonzero polynomials $a(x)$ and $b(x)$ in $F[x]$ have a gcd $d(x)$. Furthermore, $d(x)$ can be expressed as a “linear combination”*

$$d(x) = r(x)a(x) + s(x)b(x)$$

where $r(x)$ and $s(x)$ are in $F[x]$.

PROOF: The proof is analogous to the proof of the corresponding theorem for integers. If J is the set of all the linear combinations

$$u(x)a(x) + v(x)b(x)$$

as $u(x)$ and $v(x)$ range over $F[x]$, then J is an ideal of $F[x]$, say the ideal $\langle d(x) \rangle$ generated by $d(x)$. Now $a(x) = 1a(x) + 0b(x)$ and $b(x) = 0a(x) + 1b(x)$, so $a(x)$ and $b(x)$ are in J . But every element of J is a multiple of $d(x)$, so

$$d(x) \mid a(x) \quad \text{and} \quad d(x) \mid b(x)$$

If $k(x)$ is any common divisor of $a(x)$ and $b(x)$, this means there are polynomials $f(x)$ and $g(x)$ such that $a(x) = k(x)f(x)$ and $b(x) = k(x)g(x)$. Now, $d(x) \in J$, so $d(x)$ can be written as a linear combination

$$\begin{aligned} d(x) &= r(x)a(x) + s(x)b(x) \\ &= r(x)k(x)f(x) + s(x)k(x)g(x) \\ &= k(x)[r(x)f(x) + s(x)g(x)] \end{aligned}$$

hence $k(x) \mid d(x)$. This confirms that $d(x)$ is the gcd of $a(x)$ and $b(x)$. ■

Polynomials $a(x)$ and $b(x)$ in $F[x]$ are said to be *relatively prime* if their gcd is equal to 1. (This is equivalent to saying that their only common factors are constants in F .)

A polynomial $a(x)$ of positive degree is said to be *reducible over F* if there are polynomials $b(x)$ and $c(x)$ in $F[x]$, both of positive degree, such that

$$a(x) = b(x)c(x)$$

Because $b(x)$ and $c(x)$ both have positive degrees, and the sum of their degrees is $\deg a(x)$, *each has degree less than $\deg a(x)$* .

A polynomial $p(x)$ of positive degree in $F[x]$ is said to be *irreducible over F* if it cannot be expressed as the product of two polynomials of positive degree in $F[x]$. Thus, $p(x)$ is irreducible iff it is not reducible.

When we say that a polynomial $p(x)$ is irreducible, it is important that we specify *irreducible over the field F* . A polynomial may be irreducible over F , yet reducible over a larger field E . For example, $p(x) = x^2 + 1$ is irreducible over \mathbb{R} ; but over \mathbb{C} it has factors $(x + i)(x - i)$.

We next state the analogs for polynomials of Euclid's lemma and its corollaries. The proofs are almost identical to their counterparts in \mathbb{Z} ; therefore they are left as exercises.

Euclid's lemma for polynomials Let $p(x)$ be irreducible. If $p(x) \mid a(x)b(x)$, then $p(x) \mid a(x)$ or $p(x) \mid b(x)$.

Corollary 1 Let $p(x)$ be irreducible. If $p(x) \mid a_1(x)a_2(x) \cdots a_n(x)$, then $p(x) \mid a_i(x)$ for one of the factors $a_i(x)$ among $a_1(x), \dots, a_n(x)$.

Corollary 2 Let $q_1(x), \dots, q_r(x)$ and $p(x)$ be monic irreducible polynomials. If $p(x) \mid q_1(x) \cdots q_r(x)$, then $p(x)$ is equal to one of the factors $q_1(x), \dots, q_r(x)$.

Theorem 3: Factorization into irreducible polynomials *Every polynomial $a(x)$ of positive degree in $F[x]$ can be written as a product*

$$a(x) = kp_1(x)p_2(x) \dots p_r(x)$$

where k is a constant in F and $p_1(x), \dots, p_r(x)$ are monic irreducible polynomials of $F[x]$.

If this were not true, we could choose a polynomial $a(x)$ of lowest degree among those which cannot be factored into irreducibles. Then $a(x)$ is reducible, so $a(x) = b(x)c(x)$ where $b(x)$ and $c(x)$ have lower degree than $a(x)$. But this means that $b(x)$ and $c(x)$ can be factored into irreducibles, and therefore $a(x)$ can also.

Theorem 4: Unique factorization *If $a(x)$ can be written in two ways as a product of monic irreducibles, say*

$$a(x) = kp_1(x) \dots p_r(x) = lq_1(x) \dots q_s(x)$$

then $k = l$, $r = s$, and each $p_i(x)$ is equal to a $q_j(x)$.

The proof is the same, in all major respects, as the corresponding proof for \mathbf{z} ; it is left as an exercise.

In the next chapter we will be able to improve somewhat on the last two results in the special cases of $\mathbf{R}[x]$ and $\mathbf{C}[x]$. Also, we will learn more about factoring polynomials into irreducibles.

EXERCISES

A. Examples of Factoring into Irreducible Factors

1 Factor $x^4 - 4$ into irreducible factors over \mathbf{Q} , over \mathbf{R} , and over \mathbf{C} .

2 Factor $x^6 - 16$ into irreducible factors over \mathbf{Q} , over \mathbf{R} , and over \mathbf{C} .

3 Find all the irreducible polynomials of degree ≤ 4 in $\mathbf{z}_2[x]$.

4 Show that $x^2 + 2$ is irreducible in $\mathbf{z}_5[x]$. Then factor $x^4 - 4$ into irreducible factors in $\mathbf{z}_5[x]$. (By Theorem 3, it is sufficient to search for monic factors.)

5 Factor $2x^3 + 4x + 1$ in $\mathbf{z}_5[x]$. (Factor it as in Theorem 3.)

6 In $\mathbf{z}_6[x]$, factor each of the following into two polynomials of degree 1 : x , $x + 2$, $x + 3$. Why is this possible?

B. Short Questions Relating to Irreducible Polynomials

Let F be a field. Explain why each of the following is true in $F[x]$:

1 Every polynomial of degree 1 is irreducible.

2 If $a(x)$ and $b(x)$ are distinct monic polynomials, they cannot be associates.

3 Any two distinct irreducible polynomials are relatively prime.

4 If $a(x)$ is irreducible, any associate of $a(x)$ is irreducible.

5 If $a(x) \neq 0$, $a(x)$ cannot be an associate of 0.

6 In $\mathbf{z}_p[x]$, every nonzero polynomial has exactly $p - 1$ associates.

7 $x^2 + 1$ is reducible in $\mathbf{z}_p[x]$ iff $p = a + b$ where $ab \equiv 1 \pmod{p}$.

C. Number of Irreducible Quadratics over a Finite Field

1 Without finding them, determine *how many* reducible monic quadratics there are in $\mathbf{z}_5[x]$. [HINT: Every reducible monic quadratic can be uniquely factored as $(x + a)(x + b)$.]

2 How many reducible quadratics are there in $\mathbf{z}_5[x]$? How many irreducible quadratics?

3 Generalize: How many irreducible quadratics are there over a finite field of n elements?

4 How many irreducible cubics are there over a field of n elements?

D. Ideals in Domains of Polynomials

Let F be a field, and let J designate any ideal of $F[x]$. Prove parts 1–4.

1 Any two generators of J are associates.

2 J has a *unique* monic generator $m(x)$. An arbitrary polynomial $a(x) \in F[x]$ is in J iff $m(x) \mid a(x)$.

3 J is a prime ideal iff it has an irreducible generator.

4 If $p(x)$ is irreducible, then $\langle p(x) \rangle$ is a *maximal* ideal of $F[x]$. (See [Chapter 18, Exercise H5](#).)

5 Let S be the set of all polynomials $a_0 + a_1x + \cdots + a_nx^n$ in $F[x]$ which satisfy $a_0 + a_1 + \cdots + a_n = 0$. It has been shown ([Chapter 24, Exercise E5](#)) that S is an ideal of $F[x]$. Prove that $x - 1 \in S$, and explain why it follows that $S = \langle x - 1 \rangle$.

6 Conclude from part 5 that $F[x]/\langle x - 1 \rangle \cong F$. (See [Chapter 24, Exercise F4](#).)

7 Let $F[x, y]$ denote the domain of all the polynomials $\sum a_{ij}x^i y^j$ in *two* letters x and y , with coefficients in F . Let J be the ideal of $F[x, y]$ which contains all the polynomials whose constant coefficient is zero. Prove that J is not a principal ideal. Conclude that [Theorem 1](#) is not true in $F[x, y]$.

E. Proof of the Unique Factorization Theorem

1 Prove Euclid's lemma for polynomials.

2 Prove the two corollaries of Euclid's lemma.

3 Prove the unique factorization theorem for polynomials.

F. A Method for Computing the gcd

Let $a(x)$ and $b(x)$ be polynomials of positive degree. By the division algorithm, we may divide $a(x)$ by $b(x)$:

$$a(x) = b(x)q_1(x) + r_1(x)$$

1 Prove that every common divisor of $a(x)$ and $b(x)$ is a common divisor of $b(x)$ and $r_1(x)$.

It follows from part 1 that the gcd of $a(x)$ and $b(x)$ is the same as the gcd of $b(x)$ and $r_1(x)$. This procedure can now be repeated on $b(x)$ and $r_1(x)$; divide $b(x)$ by $r_1(x)$:

$$b(x) = r_1(x)q_2(x) + r_2(x)$$

Next

$$r_1(x) = r_2(x)q_3(x) + r_3(x)$$

$$\vdots$$

Finally,

$$r_{n-1}(x) = r_n(x)q_{n+1}(x) + 0$$

In other words, we continue to divide each remainder by the succeeding remainder. Since the remainders continually decrease in degree, there must ultimately be a zero remainder. But we have seen that

$$\gcd[a(x), b(x)] = \gcd[b(x), r_1(x)] = \dots = \gcd[r_{n-1}(x), r_n(x)]$$

Since $r_n(x)$ is a divisor of $r_{n-1}(x)$, it must be the gcd of $r_n(x)$ and r_{n-1} . Thus,

$$r_n(x) = \gcd[a(x), b(x)]$$

This method is called the *euclidean algorithm* for finding the gcd.

2 Find the gcd of $x^3 + 1$ and $x^4 + x^3 + 2x^2 + x - 1$. Express this gcd as a linear combination of the two polynomials.

3 Do the same for $x^{24} - 1$ and $x^{15} - 1$.

4 Find the gcd of $x^3 + x^2 + x + 1$ and $x^4 + x^3 + 2x^2 + 2x$ in $\mathbb{Z}_3[x]$.

G. A Transformation of $F[x]$

Let G be the subset of $F[x]$ consisting of all polynomials whose constant term is nonzero. Let $h : G \rightarrow G$ be defined by

$$h(a_0 + a_1x + \dots + a_nx^n) = a_n + a_{n-1}x + \dots + a_0x^n$$

Prove parts 1–3:

1 h preserves multiplication, that is, $h[a(x)b(x)] = h[a(x)]h[b(x)]$.

2 h is injective and surjective and $h \circ h = \varepsilon$.

3 $a_0 + a_1x + \dots + a_nx^n$ is irreducible iff $a_n + a_{n-1}x + \dots + a_0x^n$ is irreducible.

4 Let $a_0 + a_1x + \dots + a_nx^n = (b_0 + \dots + b_mx^m)(c_0 + \dots + c_qx^q)$. Factor

$$a_n + a_{n-1}x + \dots + a_0x^n$$

5 Let $a(x) = a_0 + a_1x + \dots + a_nx^n$ and $\hat{a}(x) = a_n + a_{n-1}x + \dots + a_0x^n$. If $c \in F$, prove that $a(c) = 0$ iff $\hat{a}(1/c) = 0$.



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SCHOOL OF SCIENCE AND HUMANITIES

DEPARTMENT OF MATHEMATICS

UNIT – V – ALGEBRA-I – SMT1501

Unit-5 Euclidean Rings

In this section, we investigate the role that prime numbers play in the integers in greater generality. Recall that every nonzero integer can be written as plus or minus a product of distinct prime powers, and these prime powers are unique. Note that the units in \mathbb{Z} are ± 1 , so we can say that every nonzero integer can be written as a product of prime powers times a unit. In this section, we investigate this property for a larger class of integral domains.

First, we introduce an analogue of prime numbers.

DEFINITION 5.1.1. Let R be an integral domain. A nonunit and nonzero element $p \in R$ is said to be an *irreducible element* if for every $a, b \in R$ with $p = ab$, either a or b is a unit.

DEFINITION 5.1.2. Two elements a and b of a nonzero commutative ring R with unity are said to *associate* if $a = ub$ with $u \in R^\times$.

Of course, the property of being associate is an equivalence relation on an integral domain R . The equivalence class of 0 is $\{0\}$ and that of 1 is R^\times . We have the following simple lemma, which tells us that the equivalence class of an irreducible element consists of irreducible elements.

LEMMA 5.1.3. If R is an integral domain, and $p \in R$ is irreducible, then so is every associate of p .

PROOF. That is, if $u \in R^\times$ and $up = ab$ for $a, b \in R$, then $p = (u^{-1}a)b$. As p is irreducible, either $u^{-1}a \in R^\times$ or $b \in R^\times$. Finally, if $u^{-1}a \in R^\times$, then $a \in R^\times$. \square

EXAMPLES 5.1.4.

a. The irreducible elements of \mathbb{Z} are $\pm p$ for prime numbers p . The elements p and $-p$ are associates.

b. The irreducible elements of $F[x]$, for a field F , are the irreducible polynomials of F , since the units of $F[x]$ are the nonzero constant polynomials. Every nonzero polynomial has a unique associate with leading coefficient equal to 1.

c. In the subring of \mathbb{C} that is

$$\mathbb{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} \mid a, b \in \mathbb{Z}\},$$

a number of prime integers are no longer irreducible. For instance $2 = -(\sqrt{-2})^2$, and $\sqrt{-2}$ is not a unit. Also, $3 = (1 + \sqrt{-2})(1 - \sqrt{-2})$, and neither $1 + \sqrt{-2}$ nor $1 - \sqrt{-2}$ is a unit, for if, e.g., $u \in \mathbb{Z}[\sqrt{-2}]$ with $u(1 + \sqrt{-2}) = 1$, then $3u = 1 - \sqrt{-2}$, which is clearly impossible. On the other hand, it turns out that 5 is irreducible, though we do not prove this now.

DEFINITION 5.1.5. An integral domain R is a *unique factorization domain*, or a *UFD*, if every nonzero, nonunit element $a \in R$ can be written as a product

$$a = p_1 p_2 \cdots p_r$$

with p_1, p_2, \dots, p_r irreducible elements of R for some $r \geq 1$, and moreover, this expression is unique in the sense that if

$$a = q_1 q_2 \cdots q_s$$

with q_1, q_2, \dots, q_s irreducible for some $s \geq 1$, then $s = r$ and there exists a permutation $\sigma \in S_r$ such that $q_{\sigma(i)}$ and p_i are associates for all $1 \leq i \leq r$.

REMARK 5.1.6. If one wants to allow units, one can rephrase Definition 5.1.5 to read that every nonzero element $a \in R$ can be written as $a = up_1 \cdots p_r$ with $u \in R^\times$ and p_1, p_2, \dots, p_r irreducible in R for some $r = 0$ in a unique manner such that any such decomposition of $a = vq_1 \cdots q_s$ has $s = r$ and, after a reordering of the irreducibles, each q_i is an associate of p_i .

EXAMPLE 5.1.7. The ring \mathbb{Z} is a unique factorization domain.

As we shall see later, $F[x]$ for a field F is a unique factorization domain as well.

EXAMPLE 5.1.8. Consider the subring $F[x^2, xy, y^2]$ of $F[x, y]$. It consists exactly of the polynomials in $F[x, y]$ that can be written as polynomials in x^2, xy , and y^2 . These latter three elements are irreducible in $F[x^2, xy, y^2]$, but we have

$$x^2 \cdot y^2 = xy \cdot xy,$$

so factorization is not unique.

A more standard example is the following.

EXAMPLE 5.1.9. Consider the subring $\mathbb{Z}[\sqrt{-5}]$ of \mathbb{C} . We have

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

The element 2 divides only elements of the form $a + b\sqrt{-5}$ with $a, b \in \mathbb{Z}$ even, so it does not divide $1 + \sqrt{-5}$ or $1 - \sqrt{-5}$. On the other hand, 2 is irreducible since if $a + b\sqrt{-5}$ divides 2, then so does its complex conjugate, and then

$$(a + b\sqrt{-5})(a - b\sqrt{-5}) = a^2 + 5b^2$$

divides 2, which happens only if $a = \pm 1$ and $b = 0$. Therefore, $\mathbb{Z}[\sqrt{-5}]$ is not a unique factorization domain.

One advantage of unique factorization domains is that they allow us to define a concept of greatest common divisor.

DEFINITION 5.1.10. Let R be a UFD. Let $a_1, a_2, \dots, a_r \in R$ be nonzero. A principal ideal (d) for $d \in R$ is said to be the *greatest common divisor*, or *GCD*, of a_1, a_2, \dots, a_r if d divides a_i for each $1 \leq i \leq r$ and if d' also divides each a_i , then d' divides d .

The element d in the definition of GCD, if it exists, is only defined up to unit. On the other hand, (d) is independent of this choice.

LEMMA 5.1.11. *Let R be a UFD. Then every collection a_1, a_2, \dots, a_r of nonzero elements of R has a GCD.*

PROOF. We sketch the proof. Factor each a_i into a unit times a product of irreducibles. If there exists an irreducible element p_1 that divides each a_i , an associate of it is one of the irreducibles appearing in the factorization of a_i . We then have $b_i \in R$ with $a_i = pb_i$ for each i , and the factorization of b_i has one fewer irreducible element than that of a_i . We repeat this process until the collection no longer has a common irreducible divisor, obtaining irreducibles p_1, p_2, \dots, p_k such that $d = p_1 p_2 \cdots p_k$ divides every a_i .

We claim that (d) is the GCD of a_1, a_2, \dots, a_k . If not, then there exists d' that does not divide d which divides every a_i . This means that there exists an irreducible element $q \in R$ and some $n \geq 1$ such that q^n divides d' but not d . Then q^n divides every a_i , which means since q^n does not divide d that q actually divides each c_i such that $a_i = dc_i$, in contradiction to the definition of d . \square

One advantage of having the notion of a GCD is that in quotient fields, it allows us to talk about fractions being in lowest terms.

DEFINITION 5.1.12. Let R be a UFD, and let $a, b \in R$ with $b \neq 0$. We say that the fraction $\frac{a}{b}$ is reduced, or in lowest terms, if the GCD of a and b is (1) .

LEMMA 5.1.13. *Let R be a UFD. Every fraction in $Q(R)$ may be written in lowest terms.*

PROOF. Let $a, b \in R$ with $b \neq 0$. Let (d) be the GCD of a and b . Then there exist $a', b' \in R$ with $a = da'$ and $b = db'$, and we have that the GCD of a' and b' is (1) . We therefore have that $\frac{a'}{b'} = \frac{a}{b}$, and the former form of the fraction is in lowest terms. \square

Let us study factorization in principal ideal domains.

DEFINITION 5.1.14. Let X be a set, and let \leq be a partial ordering on X .

a. An *ascending chain* in X is a sequence $(a_i)_{i \geq 1}$ of elements of X such that $a_i \leq a_{i+1}$ for all $i \geq 1$.

b. We say that X satisfies the *ascending chain condition*, or *ACC*, if every ascending chain $(a_i)_{i \geq 1}$ in X is eventually constant: i.e., there exists $j \geq 1$ such that $a_i = a_j$ for all $i \geq j$.

The following is an equivalent characterization of the ACC.

PROPOSITION 5.1.15. *A nonempty set X with a partial ordering \leq satisfies the ACC if and only if every subset of X contains a maximal element.*

PROOF. If every subset of X contains a maximal element, then clearly ascending chains are eventually constant: i.e., their underlying sets are finite. For the other direction, it suffices to show that if X satisfies the ACC, then it contains a maximal element. Let C be a nonempty chain in X , and suppose it does not have an upper bound. For each $x \in C$, there exists $y \in C$ with $y > x$, as otherwise x would be an upper bound. We may therefore recursively pick $a_i \in C$ with $a_i < a_{i+1}$ for each i , but this is impossible. Thus C has an upper bound, and therefore X has a maximal element by Zorn's lemma. \square

DEFINITION 5.1.16. We say that a commutative ring R is *noetherian* if the set of its ideals satisfies the ascending chain condition with respect to containment of ideals.

REMARK 5.1.17. We may rephrase the condition that R be noetherian by saying that if $(I_n)_{n \geq 1}$ is an ascending chain of ideals, then there exists $m \geq 1$ such that the union I of the I_n with $n \geq 1$ equals I_i for all $i \geq m$.

REMARK 5.1.18. One may define a noncommutative ring to be left noetherian (resp., right noetherian rings) if it satisfies the ACC on left ideals (resp., right ideals). In general, a noetherian ring is taken to be one that is both left and right noetherian.

THEOREM 5.1.19. *A commutative ring R is noetherian if and only if every ideal of R is finitely generated.*

PROOF. Suppose that every ideal of R is finitely generated. Let $(I_n)_{n \geq 1}$ be a chain of ideals of R . Let I be the union of the I_n for $n \geq 1$, which is an ideal by Lemma 3.11.10. Since I is finitely generated, $I = (a_1, a_2, \dots, a_r)$, with $a_k \in I$ with $1 \leq k \leq r$ for some $r \geq 1$. For each k , there exists $m_k \geq 1$ with $a_k \in I_{m_k}$, and if we let m be the maximum of the m_k , then $a_k \in I_m$ for every a_k . Since I is the smallest ideal of R containing each a_k , we have $I \subseteq I_m$, which forces $I = I_m$.

Conversely, suppose R is noetherian, and let I be an ideal of R . Let $x_1 \in I$, and suppose inductively that we have constructed $x_1, x_2, \dots, x_n \in R$ with the property that if we set $I_k = (x_1, x_2, \dots, x_k)$ for every $1 \leq k \leq n$, then $I_k \subseteq I_{k+1}$ for every $1 \leq k \leq n-1$. If $I_n \neq I$, then let $x_{n+1} \in I$ with $x_{n+1} \notin I_n$. Then $I_{n+1} = (x_1, x_2, \dots, x_{n+1})$ properly contains I_n . If this process repeats indefinitely, then we have constructed an ascending chain $(I_n)_{n \geq 1}$ that is not eventually constant, which would contradict the assumption that R is noetherian. Therefore, there exists $m \geq 1$ such that $I_m = I$, and so $I = (a_1, a_2, \dots, a_m)$ is finitely generated. \square

COROLLARY 5.1.20. *Every principal ideal domain is noetherian.*

PROPOSITION 5.1.21. *Let R be a principal ideal domain. Then every nonzero, nonunit $a \in R$ may be written as $a = p_1 p_2 \cdots p_r$ with the $p_i \in R$ irreducible for all $1 \leq i \leq r$ and some $r \geq 1$.*

PROOF. We claim first that every nonunit $a \in R$ is divisible by an irreducible element of R . If a is not irreducible, set $a_0 = a$ and write $a = a_1 b_1$ with $a_1, b_1 \notin R^\times$. Suppose that a_i divides a_{i-1} for some $i \geq 1$, which implies recursively that a_i divides a . If a_i is irreducible, then we have the claim. If not, then write $a_i = a_{i+1} b_{i+1}$ for some nonunits $a_{i+1}, b_{i+1} \in R^\times$. Since a_{i+1} properly divides a_i , we have that $(a_i) \subsetneq (a_{i+1})$. By Corollary 5.1.20, this process must terminate, which is to say that some a_m is eventually irreducible, and therefore a is divisible by an irreducible element.

Next, we construct another sequence out of our reducible element a . That is, we write $a = a_1 b_1$ with a_1 irreducible, and assume inductively that we have written

$$a = a_1 a_2 \dots a_n b_n$$

with $a_1, a_2, \dots, a_n \in R$ irreducible and nonunit $b_n \in R$ for some $n \geq 0$. If b_n is irreducible for any n , we are done. Otherwise, we obtain a sequence of elements $(b_i)_{i \geq 1}$ with $b_i = a_{i+1} b_{i+1}$ for all $i \geq 1$, which means that $(b_i) \subsetneq (b_{i+1})$ for each i . Again this would contradict the fact that R is noetherian, so eventually the process does terminate, and we have written a as a product of irreducible elements. \square

LEMMA 5.1.22. *Let R be a PID, and let $a \in R$ be nonzero. Then (a) is maximal if and only if a is an irreducible element.*

PROOF. Clearly, a cannot be a unit for either condition to hold. If $a = bc$ with b and c non-units, then $(a) \subsetneq (b) \subsetneq R$, so (a) is not maximal. And if (a) is not maximal, then there exists an proper ideal $I = (c)$ of R properly containing (a) , so we may write $a = bc$ with $b \in R$. Since the containment is proper, b is not a unit, and c is not a unit by definition. Therefore, a is reducible. \square

In a principal ideal domain, irreducible elements play the role that prime numbers play in \mathbb{Z} .

LEMMA 5.1.23. *Let R be a PID, and let $p \in R$ be irreducible. If $a, b \in R$ are such that $p \mid ab$, then $p \mid a$ or $p \mid b$.*

PROOF. Let $a, b \in R$ with $p \mid ab$. Then $ab \in (p)$, and (p) is maximal by Lemma 5.1.22. Since every maximal ideal of R is prime, we have that (p) is prime, and therefore either $a \in (p)$ or $b \in (p)$. \square

We now prove a key theorem.

THEOREM 5.1.24. *Every principal ideal domain is a unique factorization domain.*

PROOF. Let $a \in R$ be a nonzero, nonunit element. By Proposition 5.1.21, we may write

$$a = p_1 p_2 \cdots p_r$$

with p_1, p_2, \dots, p_r irreducible. We have only to show that this decomposition is unique in the appropriate sense. So, suppose that

$$a = q_1 q_2 \cdots q_s$$

with q_1, q_2, \dots, q_s irreducible. If $r = 1$, then a is irreducible, so $s = 1$ and $p_1 = q_1$. Suppose by induction we have proven uniqueness whenever there is a decomposition of a with fewer than $r \geq 2$ irreducibles. In particular, we may assume that $s \geq r$.

As a consequence of Lemma 5.1.23, we have that p_r divides some q_i for some $1 \leq i \leq s$. Since q_i is irreducible, this means that $q_i = wp_r$ with $w \in R^\times$. Since R is an integral domain, we then have

$$p_1 p_2 \cdots p_{r-1} = w q_1 q_2 \cdots q_{i-1} q_{i+1} \cdots q_s.$$

As $s \geq 2$ by assumption, note that $w q_1$ is an associate to q_1 and the expression on the right is a product of $s - 1$ irreducible elements. By induction, we have $r = s$, and there exists a bijective function

$$\sigma: \{1, 2, \dots, r-1\} \rightarrow \{1, 2, \dots, i-1, i+1, \dots, r\}$$

with $q_{\sigma(i)}$ and p_i associates for each $1 \leq i \leq r-1$. We may extend σ to an element of S_r by setting $\sigma(r) = i$, and then $q_{\sigma(r)} = q_i$ is an associate of p_r as well, proving uniqueness. \square

Given that every polynomial ring over a field is a PID, we have the following corollary. It is an interesting exercise to prove it directly.

COROLLARY 5.1.25. *For any field F , the ring $F[x]$ is a unique factorization domain.*

Corollary 3.10.2 tells us what we may already have known from experience, that we can factor one-variable polynomials into irreducible factors over a field, and there is only one way to do this.

Polynomial rings over UFDs

Now that we know that every PID is a UFD, the question arises: is every UFD also a PID? The answer, in fact, is no. For this, let us examine polynomial rings over integral domains in a bit more detail.

DEFINITION 5.2.1. Let R be an integral domain. A polynomial $f \in R[x]$ is said to be *primitive* if the only elements of R that divide all of the coefficients of f are units.

In a UFD, we can actually talk about the GCD of the coefficients of a polynomial.

DEFINITION 5.2.2. Let R be a UFD. The *content* of the a polynomial in $R[x]$ is the GCD of its coefficients.

REMARK 5.2.3. If R is a UFD, then a polynomial in $R[x]$ is primitive if and only if the GCD of its coefficients is (1).

DEFINITION 5.2.4. A polynomial in $R[x]$ for a nonzero ring R with unity is said to be *monic* if its leading coefficient is 1.

REMARK 5.2.5. Monic polynomials in $R[x]$, where R is a UFD, are primitive.

LEMMA 5.2.6. Let R be a UFD. If (c) is the content of $f \in R[x]$ for, then there exists a primitive polynomial $g \in R[x]$ with $f = cg$.

PROOF. By definition, c divides each coefficient of f , so $f = cg$ for some $g \in R[x]$. Let $d \in R$ be such that (d) is the content of g . Then $g = dh$ for some $h \in R[x]$, so we have $f = cdh$. But this implies that cd divides every coefficient of f , so cd divides the content c , forcing d to be a unit. Therefore, g is primitive. \square

EXAMPLE 5.2.7. The polynomial $f = 25x^2 + 10x - 15$ in $\mathbb{Z}[x]$ has content 5, and so it is not primitive. In fact, $f = 5g$, where $g = 5x^2 + 2x - 3$, and g is primitive.

LEMMA 5.2.8 (Gauss's Lemma). Let R be a UFD. Then the product of any two primitive polynomials in $R[x]$ is primitive.

PROOF. Let

$$f = \sum_{i=0}^n a_i x^i \quad \text{and} \quad g = \sum_{j=0}^m b_j x^j$$

be primitive polynomials in $R[x]$. The k th coefficient of fg is $c_k = \sum_{i=0}^k a_i b_{k-i}$. If p is an irreducible element of R , then since f and g are primitive, there exist minimal nonnegative integers r and s such that $p \nmid a_r$ and $p \nmid b_s$. Since $p \mid a_i$ for $i < r$ and $p \mid b_j$ for $j < s$, which is to say that $p \mid b_{r+s-i}$ for $r < i \leq r+s$, we have that p divides every term of c_{r+s} except $a_r b_s$, which it does not divide. Therefore, p does not divide c_{r+s} . Since p was arbitrary, fg is primitive. \square

Note that we can speak about polynomials being irreducible in $R[x]$ for any integral domain R , since we have a notion of irreducible element in such a ring. For a field F , this coincides with the usual notion of an irreducible polynomial.

PROPOSITION 5.2.9. *Let R be an integral domain, and let $F = Q(R)$.*

a. If $f \in R[x]$ is a primitive polynomial that is irreducible as an element of $F[x]$, then f is irreducible in $R[x]$. In particular, if f cannot be written as a product of two nonconstant polynomials in $R[x]$, then it is irreducible in $R[x]$.

b. Suppose that R is a UFD. If $f \in R[x]$ is irreducible, then it is irreducible as an element of $F[x]$ as well. In fact, if $f \in R[x]$ and $f = gh$ for nonconstant $g, h \in F[x]$, then there exists $\alpha \in F^\times$ such that $g' = \alpha g$ and $h' = \alpha^{-1}h$ are in $R[x]$ and therefore $f = g'h'$ in $R[x]$.

PROOF. First, we treat part a. If $f \in R[x]$ is primitive and $f \in R[x]$ is reducible (which is to say, not irreducible and not a unit or zero), then we can write $f = gh$ for nonunits $g, h \in R[x]$. If g or h is constant, then f is not primitive, so neither is constant, and therefore f is reducible in $F[x]$.

Next, we turn to part b. Suppose that $f \in R[x]$ can be written as $f = gh$ with $g, h \in F[x]$ nonconstant. Let (d) (resp., (e)) be a multiple of all of the denominators of the coefficients of g (resp., h), written in lowest terms. Then $def = g'h'$, where $g', h' \in R[x]$ are nonconstant. The content of def is contained in (de) , so the content of $g'h'$ is as well. By unique factorization in R , we may write $de = d'e'$, where $d' \in R$ divides the content of g' and e' divides the content of h' , and we may then divide g' by d' and h' by e' to obtain g'' and h'' in $R[x]$ such that $f = g''h''$. Therefore, f is reducible in $R[x]$, and the remaining statement of the lemma holds as well. \square

We are now ready to prove the following.

THEOREM 5.2.10. *If R is a UFD, then $R[x]$ is a UFD as well.*

PROOF. Let $f \in R[x]$ be a nonzero element that is not a unit. Write

$$f = f_1 f_2 \cdots f_r$$

with $f_i \in R[x]$ nonconstant, where r is maximal such that this can be done. Note that such a maximal r exists as the degree of f is finite. For $1 \leq i \leq r$, let (c_i) be the content of f_i , and define $g_i \in R[x]$ by $f_i = c_i g_i$. Set $c = c_1 c_2 \cdots c_r$, and set $g = g_1 g_2 \cdots g_r$. Now, if any g_i were not irreducible in $F[x]$ for $F = Q(R)$, then it would not be irreducible in $R[x]$ by Proposition 5.2.9b. Moreover, since g_i is primitive, it would then be written as a product of two nonconstant polynomials in $R[x]$, which would contradict the maximality of r . Therefore, each g_i is irreducible. Since R is a UFD, we may also write $c = p_1 p_2 \cdots p_k$ with $p_i \in R$ irreducible for $1 \leq i \leq k$ and some $k \geq 0$, and so

$$f = p_1 p_2 \cdots p_k g_1 g_2 \cdots g_r$$

is a factorization of f into irreducibles in $R[x]$.

Now, if

$$f = q_1 q_2 \cdots q_l h_1 h_2 \cdots h_s$$

with $q_i \in R$ irreducible and $h_i \in R[x]$ irreducible and nonconstant, then $(q_1 q_2 \cdots q_l)$ is the content of f by Gauss's lemma, and so $q_1 q_2 \cdots q_l$ agrees with c up to unit in R . Since R is a UFD, it follows that $l = k$ and there exists $\sigma \in S_k$ such that each $q_{\sigma(i)}$ is an associate of p_i . Next, we have

$$g_1 g_2 \cdots g_r = u h_1 h_2 \cdots h_s$$

for some unit $u \in R^\times$, and by uniqueness of factorization in $F[x]$, we have that $s = r$, and there exists $\tau \in S_r$ such that $h_{\tau(i)} = v_i g_i$ for some $v_i \in F^\times$ for each $1 \leq i \leq r$. But the content of each g_i and each h_j is (1), since these elements are irreducible in $R[x]$, and therefore writing $v_i = \frac{a_i}{b_i}$ with $a_i, b_i \in R$, the fact that $b_i h_{\tau(i)} = a_i g_i$ implies that $(a_i) = (b_i)$, since both sides must have the same content. In other words, $v_i \in R^\times$, and so $h_{\tau(i)}$ and g_i are associates in $R[x]$, finishing the proof of uniqueness. \square

EXAMPLES 5.2.11.

- Since \mathbb{Z} is a UFD, so is $\mathbb{Z}[x]$. However, $\mathbb{Z}[x]$ is not a PID, since (p, x) is not principal.
- Since $\mathbb{Q}[x]$ is a UFD, so is $\mathbb{Q}[x, y]$. Again, $\mathbb{Q}[x, y]$ is not a PID, since (x, y) is not principal.
- If R is any UFD, then $R[x_1, x_2, \dots, x_n]$ is a UFD for any $n \geq 1$.

5.3. Irreducibility of polynomials

In this section, we investigate criteria for determining if a polynomial is irreducible or not.

DEFINITION 5.3.1. Let R be an integral domain. We say that a polynomial $f = \sum_{i=0}^n a_i x^i$ be a polynomial in $R[x]$ that satisfies $a_n \notin \mathfrak{p}$, $a_i \in \mathfrak{p}$ for all $0 \leq i \leq n-1$, and $a_0 \notin \mathfrak{p}^2$ for some $n \geq 1$ and prime ideal \mathfrak{p} in R is an Eisenstein polynomial (with respect to \mathfrak{p}).

THEOREM 5.3.2 (Eisenstein criterion). *Let R be an integral domain, and let $f \in R[x]$ be an Eisenstein polynomial.*

- If R is a UFD, then f is irreducible in $Q(R)[x]$.*
- If f is primitive, then it is irreducible in $R[x]$.*

PROOF. Suppose $f = \sum_{i=0}^n a_i x^i$ is of degree n and Eisenstein with respect to a prime ideal \mathfrak{p} of R . By Proposition 5.2.9, it suffices for each part to show that f is not a product of two nonconstant polynomials in $R[x]$. So, let $g = \sum_{i=0}^s b_i x^i$ and $h = \sum_{j=0}^t c_j x^j$ be polynomials in $R[x]$ with $f = gh$, where $s + t = n$. We then have

$$a_k = \sum_{i=0}^k b_i c_{k-i}$$

for all $0 \leq k \leq n$. In particular, $a_0 = b_0 c_0$ is an element of \mathfrak{p} but not \mathfrak{p}^2 . Since \mathfrak{p} is prime, at least one of b_0 and c_0 lies in \mathfrak{p} , but as $a_0 \notin \mathfrak{p}^2$, at least one does not lie in \mathfrak{p} as well.

Without loss of generality, suppose that $b_0 \in \mathfrak{p}$ and $c_0 \notin \mathfrak{p}$. As $a_n = b_s c_t \notin \mathfrak{p}$, we have $b_s \notin \mathfrak{p}$. Let $k \geq 1$ be minimal such that $b_k \notin \mathfrak{p}$. If $k < n$, then $a_k \in \mathfrak{p}$ and $b_i \in \mathfrak{p}$ for $i < k$, so we have $b_k c_0 \in \mathfrak{p}$, which therefore forces $c_0 \in \mathfrak{p}$ by the primality of \mathfrak{p} . Therefore, $k = n$, which means that h is constant, proving the result. \square

We will most commonly be concerned with the Eisenstein criterion in the case that $R = \mathbb{Z}$.

EXAMPLE 5.3.3. For any prime number p and integer $n \geq 1$, the polynomial $x^n - p$ is irreducible by the Eisenstein criterion. That is, we take our prime ideal to be (p) in the ring \mathbb{Z} .

EXAMPLE 5.3.4. For a prime number p , set

$$\Phi_p = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + 1.$$

This polynomial has as its roots in \mathbb{C} the distinct p th roots of unity that are not equal to 1. Over \mathbb{Q} , we claim it is irreducible. For this, consider the polynomial

$$\Phi_p(x+1) = \frac{(x+1)^p - 1}{x} = \sum_{i=0}^{p-1} \binom{p}{i+1} x^i,$$

which has coefficients divisible by p but not p^2 except for its leading coefficient a_{p-1} , which is 1. Therefore, $\Phi_p(x+1)$ is Eisenstein, hence irreducible. But if Φ_p were to factor into g and h , then $\Phi_p(x+1)$ would factor into $g(x+1)$ and $h(x+1)$, which have the same leading coefficients as g and h , and hence are nonconstant if and only if g and h are. In other words, Φ_p is irreducible as well.

REMARK 5.3.5. The condition in the Eisenstein criterion that the constant coefficient not lie in the square of the prime ideal is in general necessary. For instance, $x^2 - p^2 \in \mathbb{Z}[x]$ is never irreducible for a prime p .

Often, we can tell if a polynomial is irreducible by considering its reductions modulo ideals.

PROPOSITION 5.3.6. *Let R be an integral domain, and let \mathfrak{p} be a prime ideal of R . Let $f \in R[x]$ with leading coefficient not in \mathfrak{p} . Let \bar{f} denote the image of f in $(R/\mathfrak{p})[x]$ given by reducing its coefficients modulo \mathfrak{p} .*

- If R is a UFD and \bar{f} is irreducible in $Q(R/\mathfrak{p})[x]$, then f is irreducible in $Q(R)[x]$.*
- If f is primitive and \bar{f} is irreducible in $R/\mathfrak{p}[x]$, then f is irreducible in $R[x]$.*

PROOF. If R is a UFD and f is reducible in $Q(R)[x]$, then by Proposition 5.2.9, we have that $f = gh$ for some nonconstant $g, h \in R[x]$. Similarly, if f is primitive and reducible in $R[x]$, then $f = gh$ for nonconstant $g, h \in R[x]$. In either case, since the leading coefficient of f is not in \mathfrak{p} and \mathfrak{p} is prime, we have that the leading coefficients of g and h are not in \mathfrak{p} as well. That is, the images of g and h in $(R/\mathfrak{p})[x]$ are nonconstant, which means that \bar{f} is a product of two nonconstant polynomials, hence reducible in $Q(R/\mathfrak{p})[x]$. \square

REMARK 5.3.7. For $R = \mathbb{Z}$, Proposition 5.3.6 tells us in particular that if $f \in \mathbb{Z}[x]$ is monic and its reduction $\bar{f} \in \mathbb{F}_p[x]$ modulo p is irreducible for any prime p , then f is irreducible.

EXAMPLE 5.3.8. Let $f = x^4 + x^3 + 1001 \in \mathbb{Z}[x]$. We claim that f is irreducible in $\mathbb{Q}[x]$. For this, consider its reduction modulo 2. The polynomial $\bar{f} = x^4 + x^3 + 1 \in (\mathbb{Z}/2\mathbb{Z})[x]$ is either irreducible, has a root in $(\mathbb{Z}/2\mathbb{Z})[x]$, or is a product of two irreducible polynomials of degree 2. But $\bar{f}(0) = \bar{f}(1) = 1$, and $x^2 + x + 1$ is the only irreducible polynomial of degree 2 in $(\mathbb{Z}/2\mathbb{Z})[x]$,

and $(x^2 + x + 1)^2 = x^4 + x^2 + 1 \neq \bar{f}$, so \bar{f} is irreducible. By Proposition 5.3.6, f is irreducible in $\mathbb{Q}[x]$.

EXAMPLE 5.3.9. The converse to Proposition 5.3.6 does not hold. For instance, $x^2 + x + 1$ is irreducible in $\mathbb{Q}[x]$, but it has a root in $(\mathbb{Z}/3\mathbb{Z})[x]$.

We also have the following simple test for the existence of roots of polynomials over UFDs.

PROPOSITION 5.3.10. *Let R be a UFD and $f = \sum_{i=0}^n a_i x^i \in R[x]$ with $a_0, a_n \neq 0$. Suppose that $\alpha \in Q(R)$ is a root of f , and write α in reduced form as $\alpha = \frac{c}{d}$ for some $c, d \in R$. Then c divides a_0 and d divides a_n in R .*

PROOF. Since $x - \frac{c}{d}$ divides f in $Q(R)[x]$ and $\frac{c}{d}$ is in reduced form, it follows from Proposition 5.2.9 that $f = (dx - c)g$ for some $g \in R[x]$. Writing $g = \sum_{i=0}^{n-1} b_i x^i$, we see that $a_0 = -cb_0$ and $a_n = db_{n-1}$. \square

EXAMPLE 5.3.11. Let $f = 2x^3 - 3x + 5 \in \mathbb{Z}[x]$. We check that $f(1) = 4$, $f(-1) = 6$, $f(5) \equiv -10 \pmod{25}$, $f(-5) \equiv 20 \pmod{25}$, and $f(\frac{1}{2})$, $f(-\frac{1}{2})$, $f(\frac{5}{2})$, and $f(-\frac{5}{2})$ are all represented by reduced fractions with denominators equal to 4. Proposition 5.3.10 therefore tells us that f has no roots in \mathbb{Q} , hence is irreducible, being of degree 3.

Euclidean domains

DEFINITION 5.4.1. A *norm* f on an ring R is a function $f: R \rightarrow \mathbb{Z}_{\geq 0}$ with $f(0) = 0$. We say that f is *positive* if the only $a \in R$ for which $f(a) = 0$ is $a = 0$.

DEFINITION 5.4.2. Let R be an integral domain. A *Euclidean norm* v on R is a norm on R such that for all nonzero $a, b \in R$, one has

- i. $v(a) \leq v(ab)$, and
- ii. there exist $q, r \in R$ with $a = qb + r$ and either $v(r) < v(b)$ or $r = 0$.

REMARK 5.4.3. Property (ii) of Definition 5.4.2 is known as the division algorithm.

DEFINITION 5.4.4. A *Euclidean domain* R is an integral domain such that there exists a Euclidean norm on R .

EXAMPLES 5.4.5.

- a. The integers \mathbb{Z} are a Euclidean domain with Euclidean norm $v(a) = |a|$ for any nonzero $a \in \mathbb{Z}$.
- b. Every polynomial ring $F[x]$ over a field F is a Euclidean domain, the degree function providing a Euclidean norm on $F[x]$.

LEMMA 5.4.6. *In a Euclidean domain R with Euclidean norm v , the minimal value of v on all nonzero elements of R is $v(1)$, and $v(u) = v(1)$ for $u \in R$ if and only if $u \in R^\times$.*

PROOF. By the definition of a Euclidean norm, we have $v(1) \leq v(a \cdot 1) = v(a)$ for all nonzero $a \in R$. If $u \in R^\times$, then $v(u) \leq v(u \cdot u^{-1}) = v(1)$, so $v(u) = v(1)$. Conversely, if $b \in R$ with $v(b) = v(1)$, then we may write $1 = qb + r$ for some $q, r \in R$ with either $v(r) < v(1)$ or $r = 0$. By what we have shown, the latter holds, so $qb = 1$, and b is a unit. \square

EXAMPLE 5.4.7. In $F[x]$, the units are exactly the nonzero constant polynomials, i.e., those with degree 0.

While we will explain below that not every PID is a Euclidean domain, it is the case that every Euclidean domain is a PID.

THEOREM 5.4.8. *Every Euclidean domain is a PID.*

PROOF. Let I be a nonzero ideal in a Euclidean domain R with Euclidean norm v . We must show that I is principal. Let $b \in I$ be a nonzero element with minimal norm among all elements of I . For any $a \in I$, we may write $a = qb + r$ with $q, r \in R$ and either $v(r) < v(b)$ or $r = 0$. Note that $a, b \in I$, so $r \in I$ as well, which precludes the possibility of $v(r) < v(b)$, since $v(r)$ is minimal among norms of elements of I . Therefore, we have $r = 0$, so $a \in (b)$. As a was arbitrary and $b \in I$, we have $I = (b)$. \square

The key property of Euclidean domains is the ability to perform the Euclidean algorithm, which we see in the following.

THEOREM 5.4.9 (Euclidean algorithm). *Let R be a Euclidean domain with Euclidean norm v , and let $a, b \in R$ be nonzero elements. Let $r_{-1} = a$ and $r_0 = b$. Suppose recursively that we are given elements $r_j \in R$ for $-1 \leq j \leq i$ and some $i \geq 0$. If $r_i \neq 0$, write*

$$(5.4.1) \quad r_{i-1} = q_{i+1}r_i + r_{i+1}$$

with $q_{i+1}, r_{i+1} \in R$ and either $v(r_{i+1}) < v(r_i)$ or $r_{i+1} = 0$. If $r_{i+1} \neq 0$, repeat the process with i replaced by $i + 1$. The process terminates with $d = r_n \neq 0$ and $r_{n+1} = 0$ for some $n \geq 1$, and (d) is the GCD of a and b . Moreover, we may use the formulas in (5.4.1) and recursion to write d as $d = xa + yb$ for some $x, y \in R$.

PROOF. We note that the process must terminate, as the values of the $v(r_i)$ for $i \geq 0$ are decreasing. Moreover, the result $d = r_n$ satisfies $r_{n-1} = q_{n+1}r_n$, so it divides r_{n-1} by definition, and then we see by downward recursion using (5.4.1) that d divides every r_{i-1} . Finally, if c is any common divisor of a and b , then it again recursively divides each r_i (this time by upwards recursion and (5.4.1)), so c divides d . Therefore, (d) is the GCD of a and b .

Note that $d = r_{n-2} - q_n r_{n-1}$, and suppose that we may write $d = zr_j + wr_{j+1}$ for some $-1 \leq j \leq n-2$. If $j = -1$, we are done. Otherwise, note that $r_{j+1} = r_{j-1} - q_{j+1}r_j$, so

$$d = zr_j + w(r_{j-1} - q_{j+1}r_j) = wr_{j-1} + (z - q_{j+1}w)r_j,$$

and we have written d as an R -linear combination of r_{j-1} and r_j . Repeat the process for $j - 1$. The final result is the desired R -linear combination of a and b . \square

EXAMPLE 5.4.10. Take \mathbb{Z} and its usual Euclidean norm. We take $a = 550$ and $b = 154$. Then $550 = 3 \cdot 154 + 88$, so we set $r_1 = 88$. Then $154 = 88 + 66$, so we set $r_2 = 66$, and $88 = 66 + 22$, so we set $r_3 = 22$, and $66 = 3 \cdot 22$, so we stop at $d = r_3 = 22$, which is therefore the greatest common divisor of a and b . Working backwards, we obtain

$$22 = 88 - 66 = 88 - (154 - 88) = 2 \cdot 88 - 154 = 2 \cdot (550 - 3 \cdot 154) - 154 = 2 \cdot 550 - 7 \cdot 154.$$

That is, we have written d as $a + (-4)b$.

Often Euclidean norms come in the form of multiplicative norms.

DEFINITION 5.4.11. A *multiplicative norm* $N: R \rightarrow \mathbb{Z}_{\geq 0}$ on a commutative ring R with unity is a positive norm such that for all $N(ab) = N(a)N(b)$ for all $a, b \in R$.

REMARK 5.4.12. Note that the existence of a multiplicative norm N on a commutative ring R with unity forces R to be an integral domain, for if $ab = 0$, then $N(a)N(b) = N(ab) = 0$, so either $N(a) = 0$ or $N(b) = 0$, and therefore either $a = 0$ or $b = 0$.

EXAMPLE 5.4.13. The absolute value on \mathbb{Z} is a multiplicative norm, as well as a Euclidean norm.

EXAMPLE 5.4.14. The function N on the Gaussian integers $\mathbb{Z}[i]$ given by $N(a + bi) = a^2 + b^2$ is a multiplicative norm. Clearly, $a^2 + b^2 = 0$ if and only if $a + bi = 0$. Given $a, b, c, d \in \mathbb{Z}$, we have

$$\begin{aligned} N((a + bi)(c + di)) &= (ac - bd)^2 + (ad + bc)^2 = (ac)^2 + (bd)^2 + (ad)^2 + (bc)^2 \\ &= (a^2 + b^2)(c^2 + d^2) = N(a + bi)N(c + di). \end{aligned}$$

PROPOSITION 5.4.15. The ring $\mathbb{Z}[i]$ of Gaussian integers is a Euclidean domain with respect to the Euclidean norm $N(a + bi) = a^2 + b^2$ for $a, b \in \mathbb{Z}$.

PROOF. Since N is a multiplicative norm, we need only check the division algorithm. Extend N to a function on \mathbb{C} by defining $N(a + bi) = a^2 + b^2$ for $a, b \in \mathbb{R}$. Let $a, b, c, d \in \mathbb{Z}$ with $(c, d) \neq (0, 0)$. Then we have

$$\frac{a + bi}{c + di} = s + ti$$

for some $s, t \in \mathbb{Q}$, and let $e, f \in \mathbb{Z}$ be integers with $|s - e| \leq 1/2$ and $|t - f| \leq 1/2$. Then we have

$$\begin{aligned} N(a + bi - (e + fi)(c + di)) &= N(c + di)N((s - e) + (t - f)i) \\ &\leq N(c + di) \left(\left(\frac{1}{2} \right)^2 + \left(\frac{1}{2} \right)^2 \right) = N(c + di)/2 < N(c + di), \end{aligned}$$

so the division algorithm is satisfied: $a + bi = q(c + di) + r$ with $q = e + fi$ and $N(r) < N(c + di)$ if $r \neq 0$. \square

COROLLARY 5.4.16. The units in $\mathbb{Z}[i]$ are exactly $1, -1, i, -i$.

PROOF. Since N is a Euclidean norm on $\mathbb{Z}[i]$, the units are exactly those nonzero elements of norm $N(1) = 1$. We have $a^2 + b^2 = 1$ if and only if $(a, b) = (\pm 1, 0)$ or $(a, b) = (0, \pm 1)$. \square

LEMMA 5.4.17. If $a, b, c, d \in \mathbb{Z}$ and $c + di$ divides $a + bi$ in $\mathbb{Z}[i]$, then $c - di$ divides $a - bi$ in $\mathbb{Z}[i]$.

PROOF. Write $a + bi = (c + di)(e + fi)$ for some $e, f \in \mathbb{Z}$. Then $a = ce - df$ and $b = cd + de$, so

$$(c - di)(e - fi) = (ce - df) - (cf + de)i = a - bi.$$

\square

We can completely determine the irreducible elements in $\mathbb{Z}[i]$ as follows.

PROPOSITION 5.4.18. *The irreducible elements in $\mathbb{Z}[i]$ are, up to multiplication by a unit, $1+i$, primes $p \in \mathbb{Z}$ with $p \equiv 3 \pmod{4}$, and $a+bi$ for $a, b \in \mathbb{Z}$ such that $p = a^2 + b^2 \equiv 1 \pmod{4}$ is a prime in \mathbb{Z} . Moreover, the primes in \mathbb{Z} that can be written in the form $a^2 + b^2$ are exactly 2 and those that are 1 modulo 4.*

PROOF. First, note that if $a+bi$ divides $c+di$ in $\mathbb{Z}[i]$ for integers a, b, c, d , then $N(a+bi)$ divides $N(c+di)$, since N is multiplicative. So, $1+i$ is irreducible since $N(1+i) = 2$.

Let p be an odd prime in \mathbb{Z} . If p is divisible by some irreducible element $\pi = a+bi$ with $a, b \in \mathbb{Z}$, then since p is prime, only one of two things can happen. Either $ab = 0$, or a and b are relatively prime in \mathbb{Z} , noting Corollary 5.4.16. Suppose $ab \neq 0$. By Lemma 5.4.17, we have that $a-bi$ divides p , and $\bar{\pi} = a-bi$ is irreducible. If $\bar{\pi}$ were associate to π , then π would divide $2a = (a+bi) + (a-bi)$ and $2b = -i((a+bi) - (a-bi))$. Then π divides 2, but that is impossible. Thus, π and $\bar{\pi}$ both dividing p implies that p is divisible by $N(\pi) = a^2 + b^2$. As p is prime, we have $p = a^2 + b^2$.

So, we have shown that either our odd prime p is irreducible in $\mathbb{Z}[i]$ or $p = a^2 + b^2$ for some $a, b \in \mathbb{Z}$. Note that the squares in $\mathbb{Z}/4\mathbb{Z}$ are 0 and 1, so any integer of the form $a^2 + b^2$ is 0, 1, or 2 modulo 4. In particular, if $p \equiv 3 \pmod{4}$, then p is irreducible in $\mathbb{Z}[i]$.

If $p \equiv 1 \pmod{4}$ is prime in \mathbb{Z} , then $(\mathbb{Z}/p\mathbb{Z})^\times$ has order divisible by 4. As $\mathbb{Z}/p\mathbb{Z}$ contains only two roots of $x^2 - 1$, which are -1 and 1 , so $(\mathbb{Z}/p\mathbb{Z})^\times$ contains an element of order 4. In particular, there exists $n \in \mathbb{Z}$ such that $n^2 \equiv -1 \pmod{p}$, which is to say that p divides $n^2 + 1$. If p were irreducible in $\mathbb{Z}[i]$, then p would divide either $n+i$ or $n-i$, but then it would divide both, being an integer. Thus p would divide $2i$, which it does not. So, p is reducible, which means equals $a^2 + b^2$ for some $a, b \in \mathbb{Z}$. \square

LEMMA 5.4.19. *Let N be a multiplicative norm on an integral domain R . Then $N(u) = 1$ for all $u \in R^\times$.*

PROOF. We have $N(1) = N(1)^2$, and R is an integral domain, so $N(1) = 1$. Moreover, since

$$N(u^{-1})N(u) = N(1) = 1,$$

we have that $N(u^{-1}) = N(u)^{-1}$, and therefore $N(u) = 1$. \square

EXAMPLE 5.4.20. Consider the multiplicative norm N on $\mathbb{Z}[\sqrt{-5}]$ given by

$$N(a + b\sqrt{-5}) = |a^2 + 5b^2|.$$

We have $a^2 + 5b^2 = 1$ if and only if $a = \pm 1$ and $b = 0$, so the only units in $\mathbb{Z}[\sqrt{-5}]$ are ± 1 . Now, if $2 = \alpha\beta$ for some nonunits $\alpha, \beta \in \mathbb{Z}[\sqrt{-5}]$, then $4 = N(2) = N(\alpha)N(\beta)$, so $N(\alpha) = 2$, but 2 is clearly not a value of N . Therefore, 2 is irreducible, and so is 3. Also, we have that $N(1 \pm \sqrt{-5}) = 6$, and since 2 and 3 are not values of N , we have that $1 \pm \sqrt{-5}$ is irreducible as well. As these elements are all non-associates, the existence of the two factorizations

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

proves that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.

Not all principal ideal domains are Euclidean. We give most of the outline of how one produces an example.

DEFINITION 5.4.21. An nonzero, non-unit element b of an integral domain R is called a *universal side divisor* if every element $a \in R$ may be written in the form $a = qb + r$ for some $q, r \in R$ with $r = 0$ or $r \in R^\times$.

LEMMA 5.4.22. Let R be a Euclidean domain with Euclidean norm v . Let $b \in R$ be a nonzero, non-unit element such that $v(b)$ is minimal among nonzero, non-unit elements of R . Then b is a universal side divisor of R .

PROOF. Let $a \in R$. By definition of v , we may write $a = qb + r$ with $v(r) < v(b)$ or $r = 0$. By the minimality of $v(b)$, we must have that r is a unit or 0. \square

EXAMPLE 5.4.23. We claim that the ring $R = \mathbb{Z}[(1 + \sqrt{-19})/2]$ is not Euclidean. Suppose by contradiction that it is a Euclidean domain, and let v be a Euclidean norm on R . We also have the multiplicative norm N on R given by

$$(5.4.2) \quad N\left(a + b\frac{1 + \sqrt{-19}}{2}\right) = \left(a + b\frac{1 + \sqrt{-19}}{2}\right)\left(a + b\frac{1 - \sqrt{-19}}{2}\right) = a^2 + ab + 5b^2.$$

Note that if $\alpha \in R - \mathbb{Z}$, then $N(\alpha) \geq 5$, so the only units in R are ± 1 .

Let $\beta \in R$ be a universal side divisor, which exists as R is Euclidean, and write $2 = q\beta + r$ for $q \in R$ and $r \in \{0, 1, -1\}$. We then have that $N(\beta)$ divides $N(2 - r)$ as N is multiplicative, so $N(\beta)$ divides 4 or 9, and this implies $\beta \in \{\pm 2, \pm 3\}$ by the formula for N . Now take $\alpha = (1 + \sqrt{-19})/2$, and set $\alpha = q'\beta + r'$ with $q' \in R$ and $r' \in \{0, 1, -1\}$. We have $N(\alpha) = N(\alpha - 1) = 5$ and $N(\alpha + 1) = 7$, which are not multiples of $N(\beta) \in \{4, 9\}$, so we obtain a contradiction.

DEFINITION 5.4.24. A *Dedekind-Hasse norm* on an integral domain R is a positive norm μ on R such that for every $a, b \in R$, either $a \in (b)$ or there exists a nonzero element $c \in (a, b)$ such that $\mu(c) < \mu(b)$.

PROPOSITION 5.4.25. An integral domain R is a PID if and only if there exists a Dedekind-Hasse norm on R .

PROOF. Suppose first that μ is a Dedekind-Hasse norm on R . Let I be a nonzero ideal of R , and let $b \in I - \{0\}$ with minimal norm under μ . If $a \in I$, then since there does not exist a nonzero element $c \in (a, b) \subseteq I$ with $\mu(c) < \mu(b)$ by the minimality of $\mu(b)$, we have by definition of a Dedekind-Hasse norm that $a \in (b)$. Thus $I = (b)$.

Suppose on the other hand the R is a PID. Define $\mu : R \rightarrow \mathbb{Z}_{\geq 0}$ by $\mu(0) = 0$, $\mu(u) = 1$ for $u \in R^\times$, and $\mu(p_1 p_2 \cdots p_k) = 2^k$ if p_1, \dots, p_k are irreducible elements of R . This is well-defined as R is a UFD. Given $a, b \in R$, we have $(a, b) = (d)$ for some $d \in R$, since R is a PID. Since d divides b , we have $\mu(d) \leq \mu(b)$. If $\mu(d) = \mu(b)$, then a and b have the same number of divisors as d and therefore are associates, so $a \in (b)$. Thus, μ is a Dedekind-Hasse norm. \square

EXAMPLE 5.4.26. We have already seen that $R = \mathbb{Z}[(1 + \sqrt{-19})/2]$ is not a Euclidean domain. To see that R is a PID, it suffices to show that the multiplicative norm N on R given by (5.4.2) is a Dedekind-Hasse norm on R . We outline the standard unenlightening verification.

Let $\alpha, \beta \in R$ with $\alpha \notin (\beta)$. We claim that there exist $s, t \in R$ with $0 < N(s\alpha - t\beta) < N(\beta)$. Note that we can extend N to a map $N: Q(R) \rightarrow \mathbb{Z}_{\geq 0}$ by the formula (5.4.2), allowing $a, b \in \mathbb{Q}$. Our condition that N on R be a Dedekind-Hasse norm is then that $0 < N(s\frac{\alpha}{\beta} - t) < 1$. We will find s and t . For this, write

$$\frac{\alpha}{\beta} = \frac{a + b\sqrt{-19}}{c}$$

for $a, b, c \in \mathbb{Z}$ with no common divisor and $c > 1$.

First one considers the cases with $c \geq 4$. If $c = 2$, then either a or b is odd, then take $s = 1$ and $t = ((a - 1) + b\sqrt{-19})/2$. If $c = 3$, then $a^2 + 19b^2 \not\equiv 0 \pmod{3}$, so $a^2 + 19b^2 = 3q + r$ with $r \in \{1, 2\}$. Take $s = a - b\sqrt{-19}$ and $t = q$. If $c = 4$, then again either a or b is odd. If only one is, then write $a^2 + 19b^2 = 4q + r$ with $1 \leq r \leq 4$, and take $s = a - b\sqrt{-19}$ and $t = q$. If both are, write $a^2 + 19b^2 = 8q + 4$, and take $s = \frac{1}{2}(a - b\sqrt{-19})$ and $t = q$.

Now suppose that $c \geq 5$. Since $(a, b, c) = (1)$, we have $x, y, z \in \mathbb{Z}$ such that $xa + yb + zc = 1$. Write $ay - 19bx = qc + r$, with $q \in \mathbb{Z}$ and $|r| \leq c/2$. Take $s = y + x\sqrt{-19}$ and $t = q - z\sqrt{-19}$. The reader will check that

$$N\left(s\frac{\alpha}{\beta} - t\right) = c^{-2}N\left(s(a + b\sqrt{-19}) - tc\right) = \frac{r^2 + 19}{c^2},$$

which is at most $\frac{1}{4} + \frac{19}{36} = \frac{7}{9}$ if $c \geq 6$ and at most $\frac{4}{25} + \frac{19}{25} = \frac{23}{25}$ if $c = 5$.

5.5. Vector spaces over fields

In this section, we give a very brief discussion of the theory of vector spaces over fields, as it shall be subsumed by the sections that follow it.

DEFINITION 5.5.1. Let F be a field. A *vector space* V over F is an abelian group under addition that is endowed with an operation $\cdot: F \times V \rightarrow V$ of scalar multiplication such that for all $a, b \in F$ and $v, w \in V$, one has

- i. $1 \cdot v = v$,
- ii. $a \cdot (b \cdot v) = (ab) \cdot v$,
- iii. $(a + b) \cdot v = a \cdot v + b \cdot v$,
- iv. $a \cdot (v + w) = a \cdot v + a \cdot w$.

REMARK 5.5.2. In a vector space V over a field F , we typically write av for $a \cdot v$, where $a \in F$ and $v \in V$.

EXAMPLE 5.5.3. If F is a field, then F^n is a vector space over F under the operation

$$a \cdot (\alpha_1, \alpha_2, \dots, \alpha_n) = (a\alpha_1, a\alpha_2, \dots, a\alpha_n)$$

for $a, \alpha_1, \alpha_2, \dots, \alpha_n \in F$.

DEFINITION 5.5.4. An element of a vector space V over a field F is called a *vector*, and the elements of F under the operation \cdot are referred to as *scalars*.

EXAMPLE 5.5.5. In every vector space V , there is an element 0 , and it is called the zero vector.

DEFINITION 5.5.6. The *zero vector space* 0 is the vector space over any field F that is the set $\{0\}$ with the operation $a \cdot 0 = 0$ for all $a \in F$.

EXAMPLE 5.5.7. If F is a field, then $F[x]$ is a vector space over F with $a \cdot f$ for $a \in F$ and $f \in F[x]$ defined to be the usual product of polynomials in $F[x]$. I.e., the operation of scalar multiplication is just multiplication by a constant polynomial.

EXAMPLE 5.5.8. The field \mathbb{C} is an \mathbb{R} -vector space, as well as a \mathbb{Q} -vector space. The field \mathbb{R} is a \mathbb{Q} -vector space. The operations of scalar multiplication are just restrictions of the usual multiplication map on \mathbb{C} .

The reader will easily check the following.

LEMMA 5.5.9. If V is a vector space over a field F , then for $a \in F$ and $v \in V$, we have

- a. $0 \cdot v = 0$,
- b. $a \cdot 0 = 0$,
- c. $-(av) = (-a)v = a(-v)$.

DEFINITION 5.5.10. Let V be a vector space over a field F . A *subspace* W of V is a subset that is closed under the operations of addition and scalar multiplication to W (i.e., to maps $W \times W \rightarrow V$ and $F \times W \rightarrow V$, respectively) and is a vector space with respect to these operations.

The following is easily proven.

LEMMA 5.5.11. A subset W of a vector space V is a subspace if and only if it is a subgroup under addition and closed under scalar multiplication.

EXAMPLES 5.5.12.

- a. The zero subspace $\{0\}$ and V are both subspaces of any vector space V .
- b. The field F is a subspace of $F[x]$.

DEFINITION 5.5.13. Let V be a vector space over a field F , and let S be a subset of V . A *linear combination* of elements of S is any sum

$$\sum_{i=1}^n a_i v_i$$

with v_1, v_2, \dots, v_n distinct vectors in S and $a_1, a_2, \dots, a_n \in F$ for some $n \geq 0$. We say that such a linear combination is *nontrivial* if there exists a j with $1 \leq j \leq n$ and $a_j \neq 0$.

DEFINITION 5.5.14. Let V be a vector space over a field F and S be a set of vectors in V . The subspace *spanned* by S , also known as the *span* of V , is the set of all linear combinations of elements of S , or simply the zero subspace if S is empty.

EXAMPLE 5.5.15. For any vector space V , the set V spans V .

DEFINITION 5.5.16. We say that a set S of vectors in a vector space V over a field F *spans* V if V equals the subspace spanned by S .

That is, S spans an F -vector space V if, for every $v \in V$, there exist $n \geq 0$, $v_i \in V$, and $a_i \in F$ for $1 \leq i \leq n$ such that

$$v = \sum_{i=1}^n a_i v_i.$$

DEFINITION 5.5.17. We say that a set S of vectors in a vector space V over a field F is *linearly independent* if every nontrivial linear combination of vectors in S is nonzero. Otherwise, S is said to be *linearly dependent*.

That is, a set S of vectors in an F -vector space V is linearly independent if whenever $n \geq 1$, $v_i \in V$ and $a_i \in F$ for $1 \leq i \leq n$ and

$$\sum_{i=1}^n a_i v_i = 0,$$

then $a_i = 0$ for all $1 \leq i \leq n$.

LEMMA 5.5.18. Let S be a linearly independent subset of a vector space V over a field F , and let W be the span of S . If $v_0 \in V - W$, then $S \cup \{v_0\}$ is also linearly independent.

PROOF. Let $v_1, v_2, \dots, v_n \in S$ and $c_0, c_1, \dots, c_n \in F$ for some $n \geq 1$, and suppose that

$$\sum_{i=0}^n c_i v_i = 0.$$

We cannot have $c_0 \neq 0$, as then

$$v_0 = -c_0^{-1} \sum_{i=1}^n c_i v_i \in W.$$

On the other hand, the fact that $c_0 = 0$ implies that $c_i = 0$ for all $1 \leq i \leq n$ by the linear independence of S . Thus, $S \cup \{v_0\}$ is linearly independent. \square

EXAMPLE 5.5.19. In any vector space V , the empty set is linearly independent. If $v \in V$ is nonzero, then $\{v\}$ is also a linearly independent set.

DEFINITION 5.5.20. A subset B of a vector space V over a field F is said to be a *basis* of V over F if it is linearly independent and spans V .

EXAMPLE 5.5.21. The set $\{e_1, e_2, \dots, e_n\}$ of F^n , where e_i is the element of F^n that has a 1 in its i th coordinate and 0 in all others, is a basis of F^n .

EXAMPLE 5.5.22. The set $\{x^i \mid i \geq 0\}$ is a basis of $F[x]$. That is, every polynomial can be written as a finite sum of distinct monomials in a unique way.

REMARK 5.5.23. For a field F , it is very hard to write down a basis of $\prod_{i=0}^{\infty} F$. In fact, the proof that it has a basis uses the axiom of choice.

DEFINITION 5.5.24. A vector space V is said to be *finite dimensional* if it has a finite basis (i.e., a basis with finitely many elements). Otherwise V is said to be *infinite dimensional*.

The following theorem employs Zorn's lemma.

THEOREM 5.5.25. *Let V be a vector space over a field F . Every linearly independent subset of V is contained in a basis of V .*

PROOF. Let S be a linearly independent subset of V , and let X denote the set of linearly independent subsets of V that contain S . We order X by containment of subsets. If \mathcal{C} is a chain in X , then its union $U = \bigcup_{T \in \mathcal{C}} T$ is linearly independent since if $v_1, v_2, \dots, v_n \in U$ for some $n \geq 1$, then each v_i is contained in some $T_i \in \mathcal{C}$ for each $1 \leq i \leq n$, and one of the sets T_j contains the others, since \mathcal{C} is a chain. Since T_j is linearly independent, any nontrivial linear combination of the elements v_i with $1 \leq i \leq n$ is nonzero. Therefore, U is linearly independent as well, so is contained in X .

By Zorn's Lemma, X now contains a maximal element B , and we want to show that B spans V , so is a basis of V containing S . Let W denote the span of B . If $v_0 \in V - W$, then $B' = B \cup \{v_0\}$ is linearly independent by Lemma 5.5.18, so an element of X , which contradicts the maximality of B . That is, $V = W$, which is to say that B spans V . \square

In particular, the empty set is contained in a basis of any vector space, so we have the following:

COROLLARY 5.5.26. *Every vector space over a field contains a basis.*

A similar argument yields the following.

THEOREM 5.5.27. *Let V be a vector space over a field F . Every subset of V that spans V contains a basis of V .*

PROOF. Let S be a spanning subset of V . Let X denote the set of linearly independent subsets of S , and order X by containment. As seen in the proof of Theorem 5.5.25, any union of a chain of linearly independent subsets is linearly independent, so has an upper bound. Thus, Zorn's lemma tells us that X contains a maximal element B . Again, we want to show that B spans V , so is a basis. If it were not, then there would exist some element of V which is not in the span of B , but is in the span of S . In particular, there exists an element $v_0 \in S$ that is not in the span of B . The set $B \cup \{v_0\}$ is linearly independent, contradicting the maximality of B . \square

We also have the following, which can be generalized to a statement on cardinality.

THEOREM 5.5.28. *Let V be a vector space over a field F . If V is finite dimensional, then every basis of V contains the same number of elements, and otherwise every basis of V is infinite.*

PROOF. Let $B_1 = \{v_1, v_2, \dots, v_n\}$ be a basis of V with a minimal number n of elements, and let $B = \{w_1, w_2, \dots, w_m\}$ be another basis of V with $m \geq n$. Then B_1 spans V , so w_1 is a nontrivial linear combination of the v_i for $1 \leq i \leq n$:

$$(5.5.1) \quad w_1 = \sum_{i=1}^n a_i v_i$$

for some $a_i \in F$. Letting j be such that $a_j \neq 0$, we may write v_j as a linear combination of w_1 and the v_i with $i \neq j$. In other words, $B_2 = (B_1 - \{v_j\}) \cup \{w_1\}$ spans V . Suppose

$$(5.5.2) \quad c_j w_1 + \sum_{\substack{i=1 \\ i \neq j}}^n c_i v_i = 0$$

for some $c_i \in F$. Using (5.5.1), we may rewrite the sum in (5.5.2) as a linear combination of the v_i , the coefficient of v_j in which is $a_j c_j$, which forces $c_j = 0$ as B_1 is a linearly independent set. But then we see from (5.5.2) that all $c_i = 0$ as $B - \{v_j\}$ is linearly independent. So, B_2 is a basis of V .

Suppose by recursion that, for $k \leq m$, we have found a basis B_k of order n of V that contains only w_1, \dots, w_{k-1} and elements of B . Then w_k is a nontrivial linear combination of the elements of B_k , and the coefficient of some v_l is nonzero in this linear combination by the linear independence of B . We therefore have that $B_{k+1} = (B_k - \{v_l\}) \cup \{w_k\}$ spans V , and a similar argument to the above shows that it is a basis. Finally, we remark that the basis B_{m+1} must be B_1 itself, since it contains B_1 , so we have $m = n$, as desired. \square

DEFINITION 5.5.29. The *dimension* of a finite-dimensional vector space V over a field F is the number of elements in a basis of V over F . We write $\dim_F(V)$ for this dimension.

EXAMPLE 5.5.30. The space F^n is of dimension n over F .

The maps between vector spaces that respect the natural operations on the spaces are called linear transformations.

DEFINITION 5.5.31. A *linear transformation* $T: V \rightarrow W$ of F -vector spaces is a function from V to W satisfying

$$T(v + v') = T(v) + T(v') \quad \text{and} \quad T(av) = aT(v)$$

for all $a \in F$ and $v, v' \in V$

REMARK 5.5.32. In other words, a linear transformation is a homomorphism of the underlying groups that “respects scalar multiplication.”

DEFINITION 5.5.33. A linear transformation $T: V \rightarrow W$ of F -vector spaces is an isomorphism of F -vector spaces if it is there exists an linear transformation $T^{-1}: W \rightarrow V$ that is inverse to it.

Much as with group and ring homomorphisms, we have the following:

LEMMA 5.5.34. A linear transformation is an isomorphism if and only if it is a bijection.

EXAMPLES 5.5.35. Let V and W be F -vector spaces.

- The identity map $\text{id}_V: V \rightarrow V$ is an F -linear transformation (in fact, isomorphism).
- The zero map $0: V \rightarrow W$ is an F -linear transformation.

Modules over rings

DEFINITION 5.6.1. Let R be a ring. A *left R -module*, or *left module over R* , is an abelian group M together with an operation $\cdot : R \times M \rightarrow M$ such that for all $a, b \in R$ and $m, n \in M$, one has

- i. $1 \cdot m = m$,
- ii. $(a \cdot b) \cdot m = (ab) \cdot m$,
- iii. $(a + b) \cdot m = a \cdot m + b \cdot m$,
- iv. $a \cdot (m + n) = a \cdot m + a \cdot n$.

DEFINITION 5.6.2. Let R be a commutative ring. We refer more simply to a *left R -module* as a *R -module*, or *module over R* .

REMARK 5.6.3. When one speaks simply of a module over a ring R , one means by default a left R -module.

NOTATION 5.6.4. When an abelian group M is seen as a left module over a ring R via the extra data of some operation $R \times M \rightarrow M$, we say that this operation endows M with the additional structure of a left R -module.

EXAMPLE 5.6.5. The definition of a module over a field coincides with the definition of a vector space over a field. In other words, to say that M a module over a field F is exactly to say that M is a vector space over F .

EXAMPLE 5.6.6. The modules over \mathbb{Z} are exactly the abelian groups. That is, suppose that A is a \mathbb{Z} -module, which by definition is an abelian group with an additional operation $\cdot : \mathbb{Z} \times A \rightarrow A$. We show that this additional operation satisfies $n \cdot a = na$ for $n \in \mathbb{Z}$ and $a \in A$, where na is the usual element of the abelian group A . So, let $a \in A$. By axiom (i), we have $1 \cdot a = a$, and then the distributivity of axiom (iii) allows us to see that $n \cdot a = na$ for all $n \geq 1$. Using axioms (iv) and (ii), we have

$$0 \cdot a = 0 \cdot (2a - a) = 0 \cdot 2a - 0 \cdot a = (0 \cdot 2) \cdot a - 0 \cdot a = 0 \cdot a - 0 \cdot a = 0,$$

and then finally we have

$$(-n) \cdot a + n \cdot a = (n - n) \cdot a = 0 \cdot a = 0,$$

so $(-n) \cdot a = -na$ for $n \geq 1$.

EXAMPLE 5.6.7. For a ring R and $n \geq 1$, the direct product R^n is a left $M_n(R)$ -module via matrix multiplication $(A, v) \mapsto A \cdot v$ for $A \in M_n(R)$ and $v \in R^n$, viewing elements of R^n as column vectors.

We also have the notion of a right R -module.

DEFINITION 5.6.8. Let R be a ring. A *right R -module*, or *right module over R* , is an abelian group M together with an operation $\cdot : M \times R \rightarrow R$ such that for all $a, b \in R$ and $m, n \in M$, one has

- i. $m \cdot 1 = m$,

- ii. $m \cdot (a \cdot b) = m \cdot (ab)$,
- iii. $m \cdot (a + b) = m \cdot a + n \cdot b$,
- iv. $(m + n) \cdot a = m \cdot a + n \cdot a$.

EXAMPLE 5.6.9. Every left ideal I over a ring R is a left R -module with respect to the restriction $R \times I \rightarrow I$ of the multiplication on R . Every right ideal over R is a right module with respect to the restriction $I \times R \rightarrow I$ of the multiplication on R .

DEFINITION 5.6.10. Let R be a ring. The opposite ring R^{op} to R is the ring that is the abelian group R together with the multiplication $\cdot^{\text{op}}: R \times R \rightarrow R$ given by $a \cdot^{\text{op}} b = ba$, where the latter product is taken in R .

REMARK 5.6.11. The identity map induces an isomorphism $R \rightarrow (R^{\text{op}})^{\text{op}}$ of rings.

The reader will easily check the following.

LEMMA 5.6.12. A right module M over R also has the structure of a left module over R^{op} , where the latter operation $\cdot^{\text{op}}: R^{\text{op}} \times M \rightarrow M$ is given by $a \cdot^{\text{op}} m = ma$, where the latter product is that given by the right R -module structure of M .

EXAMPLE 5.6.13. For a field F , the map $T: M_n(F) \rightarrow M_n(F)$ given by transpose (that is, $A \mapsto A^T$ for $A \in M_n(F)$) is a ring isomorphism between $M_n(F)$ and $M_n(F)^{\text{op}}$.

We also have the notion of a bimodule.

DEFINITION 5.6.14. Let R and S be rings. An abelian group M that is a left R -module and a right S -module is called an R - S -bimodule if

$$(r \cdot m) \cdot s = r \cdot (m \cdot s)$$

for all $r \in R$, $s \in S$, and $m \in M$.

EXAMPLES 5.6.15.

- a. Any left R -module M over a commutative ring R is an R - R -bimodule with respect to given left operation and the (same) right operation $m \cdot r = rm$ for $m \in M$ and $r \in R$.
- b. A two-sided ideal of a ring R is an R - R -bimodule with respect to the operations given by the usual multiplication on R .
- c. For $m, n \geq 1$, the abelian group $M_{mn}(R)$ of m -by- n matrices with entries in R is an $M_m(R)$ - $M_n(R)$ -bimodule for the operations of matrix multiplication.

Let us return our focus to R -modules, focusing on the case of left modules, as right modules are just left modules over the opposite ring by Lemma 5.6.12.

DEFINITION 5.6.16. An R -submodule (or, submodule) N of a left module M over a ring R is a subset of N that is closed under addition and the operation of left R -multiplication and is an R -module with respect to their restrictions $+: N \times N \rightarrow N$ and $\cdot: R \times N \rightarrow N$ to N .

LEMMA 5.6.17. Let R be a ring, M be a left R -module, and N be a subset of M . Then N is an R -submodule of M if and only if it is nonempty, closed under addition, and closed under left R -multiplication.

PROOF. Clearly, it suffices to check that if N is nonempty and closed under addition and left R -multiplication, then it is an R -submodule. The condition of being closed under left R -multiplication assures that 0 and inverses of elements of N lies in N , so N is an abelian group under $+$ on M . The axioms for N to be an R -module under \cdot are clearly satisfied as they are satisfied by elements of the larger set M . \square

EXAMPLES 5.6.18.

- The subspaces of a vector space V over a field F are exactly the F -submodules of V .
- The subgroups of an abelian group are the \mathbb{Z} -submodules of that group.
- Any left ideal I of R is a left R -submodule of R viewed as a left R -module.
- Any intersection of R -submodules is an R -submodule as well.
- For an R -module M and a left ideal I , the abelian group

$$IM = \left\{ \sum_{i=1}^n a_i m_i \mid a_i \in I, m_i \in M \text{ for } 1 \leq i \leq n \right\}$$

is an R -submodule of M .

We also have the following construction.

DEFINITION 5.6.19. Let M be an R -module and $\{N_i \mid i \in I\}$ be a collection of submodules for an indexing set I . The *sum* of the submodules N_i is the submodule $\sum_{i \in I} N_i$ of M with elements $\sum_{i \in I} n_i$ for $n_i \in N_i$ and all but finitely many n_i equal to 0 .

If M is an R -module and N is a submodule, we may speak of the quotient abelian group M/N . It is an R -module under the action $r \cdot (m + N) = rm + N$ for $r \in R$ and $m \in M$. This is well-defined, as a different representative $m + n$ of the coset $m + N$ for $n \in N$ will satisfy $r(m + n) + N = rm + rn + N = rm + N$.

DEFINITION 5.6.20. Let M be a left R -module and N be an R -submodule of M . The *quotient module* M/N of M by N is the abelian group of cosets together with the multiplication $R \times M/N \rightarrow M/N$ given by $r \cdot (nN) = (rn)N$.

EXAMPLE 5.6.21. For an R -module M and a left ideal I , we have the quotient module M/IM . In particular, note that R/I is a left R -module with respect to $r(s + I) = rs + I$, even if it is not a ring (i.e., if I is not two-sided).

We can also speak of homomorphisms of R -modules.

DEFINITION 5.6.22. Let M and N be left modules over a ring R . A left R -module homomorphism $\phi: M \rightarrow N$ is a function such that $\phi(r \cdot m) = r\phi(m)$ and $\phi(m + n) = \phi(m) + \phi(n)$ for all $r \in R$ and $m, n \in M$.

NOTATION 5.6.23. If R is commutative (or it is understood that we are working with left modules), we omit the word “left” and speak simply of R -module homomorphisms.

REMARK 5.6.24. A right R -module homomorphism $\phi: M \rightarrow N$ is just a left R^{op} -module homomorphism.

DEFINITION 5.6.25. Let M and N be left modules over a ring R .

- a. An *isomorphism* $f: M \rightarrow N$ of left R -modules is a bijective homomorphism.
- b. An *endomorphism* of a left R -module M is a homomorphism $f: M \rightarrow M$ of left R -modules.
- c. An *automorphism* of a left R -modules M is an isomorphism $f: M \rightarrow M$ of left R -modules.

NOTATION 5.6.26. Sometimes, we refer to an R -module homomorphism as an R -linear map, and an endomorphism of R -modules as an R -linear endomorphism.

EXAMPLES 5.6.27.

- a. The zero map $0: M \rightarrow M$ and the identity map $\text{id}: M \rightarrow M$ are endomorphisms of an R -module M , with id being an automorphism.
- b. Let V and W be vector spaces over a field F . A left F -module homomorphism $\phi: V \rightarrow W$ is just an F -linear transformation.
- c. Let N be an R -submodule of a left R -module M . The inclusion map $\iota_N: N \rightarrow M$ is an R -module homomorphism, as is the quotient map $\pi_N: M \rightarrow M/N$.
- d. If M is an R - S -bimodule, then right multiplication $\psi_s: M \rightarrow M$ by an element $s \in S$ defines a left R -module endomorphism. In particular, if R is a commutative ring, then multiplication by $r \in R$ defines an R -module endomorphism. Note that if R is noncommutative, then the condition that left multiplication by $r \in R$ be a left module homomorphism $M \rightarrow M$ is that $r(sm) = s(rm)$ for all $r, s \in R$ and $m \in M$, which need not hold.
- e. The identity map $F^n \rightarrow F^n$ provides an isomorphism between F^n viewed as a left $M_n(F)$ -module via $(A, v) \mapsto Av$ for $A \in M_n(F)$ and $v \in F^n$ (viewing v as a column vector) and F^n viewed as a left $M_n(F)^{\text{op}}$ -module via $(A, v) \mapsto v^T A$.

Note that we may speak of the kernel and the image of a left R -module, as an R -module homomorphism is in particular a group homomorphism. The reader will easily verify the following.

LEMMA 5.6.28. Let $\phi: M \rightarrow N$ be a left R -module homomorphism. Then $\ker \phi$ and $\text{im } \phi$ are R -submodules of M and N , respectively.

We also have analogues of all of the isomorphism theorems for groups. Actually, these are virtually immediate consequences of said isomorphism theorems, as the fact that one has isomorphisms of groups follows immediately from them, and then one need only note that these isomorphisms are actually homomorphisms of R -modules.

THEOREM 5.6.29. Let R be a ring. Let $\phi: M \rightarrow N$ be an homomorphism of left R -modules. Then there is an isomorphism $\bar{\phi}: M/\ker \phi \rightarrow \text{im } \phi$ given by $\bar{\phi}(m + \ker \phi) = \phi(m)$.

THEOREM 5.6.30. Let R be a ring, and let N and N' be left R -submodules of an R -module N . Then there is an isomorphism of R -modules

$$M/(M \cap N) \xrightarrow{\sim} (M + N)/N, \quad m + (M \cap N) \mapsto m + N.$$

THEOREM 5.6.31. Let R be a ring, let M be an R -module, and let $Q \subseteq N$ be R -submodules of M . Then there is an isomorphism

$$M/N \xrightarrow{\sim} (M/Q)/(N/Q), \quad m + N \mapsto (m + Q) + (N/Q).$$

We also have the following analogue of Theorems 2.13.10 and 3.8.23.

THEOREM 5.6.32. *Let R be a ring, let M be an R -module, and let N be an R -submodule of M . Then the map $P \mapsto P/N$ gives a bijection between submodules P of M containing N and submodules of M/N . This bijection has inverse $Q \mapsto \pi_N^{-1}(Q)$ on submodules Q of M/N , where $\pi_N: M \rightarrow M/N$ is the quotient map.*

5.7. Free modules and generators

DEFINITION 5.7.1. Let S be a subset of an R -module M .

a. The submodule of M *generated by S* is the intersection of all submodules of M containing S .

b. We say that S *generates M* , or is a *set of generators* or *generating set* of M , if no proper R -submodule of M contains S .

REMARK 5.7.2. The R -submodule of M generated by S consists of the elements $\sum_{i=1}^n a_i m_i$ with $m_i \in S$ and $a_i \in R$ for $1 \leq i \leq n$ and some $n \geq 1$. The proof is much as before.

REMARK 5.7.3. The sum $\sum_{i \in I} N_i$ of submodules N_i of M is the submodule generated by $\cup_{i \in I} N_i$.

NOTATION 5.7.4. The R -submodule of an R -module M generated by for a single element $m \in M$ (or, more precisely, by $\{m\}$) is denoted $R \cdot m$.

DEFINITION 5.7.5. We say that an R -module is *finitely generated* if it has a finite set of generators.

DEFINITION 5.7.6. We say that an R -module is *cyclic* if it can be generated by a single element.

EXAMPLE 5.7.7. A cyclic R -submodule of R is just a principal left ideal.

We can define direct sums and direct products of modules.

DEFINITION 5.7.8. Let $(M_i)_{i \in I}$ be a collection of left modules over a ring R .

a. The *direct product* $\prod_{i \in I} M_i$ is the R -module that is the direct product of the abelian groups M_i together with the left R -multiplication $r \cdot (m_i)_{i \in I} = (rm_i)_{i \in I}$ for $r \in R$ and $m_i \in M_i$ for all $i \in I$.

b. The *direct sum* $\bigoplus_{i \in I} M_i$ is the R -module that is the direct sum of the abelian groups M_i together with the left R -multiplication $r \cdot (m_i)_{i \in I} = (rm_i)_{i \in I}$ for $r \in R$ and $m_i \in M_i$ for all $i \in I$ with all but finitely many $m_i = 0$.

REMARK 5.7.9. If I is a finite set, then the canonical injection

$$\bigoplus_{i \in I} M_i \rightarrow \prod_{i \in I} M_i$$

is an isomorphism. In this case, the two concepts are often used interchangeably.

NOTATION 5.7.10. A direct sum (resp., product) of two R -modules M and N is denoted $M \oplus N$.

DEFINITION 5.7.11. We say that an R -submodule A of an R -module B is a *direct summand* of B if there exists an R -module C such that $B = A \oplus C$. In this case, C is called a *complement* to A in B .

DEFINITION 5.7.12. Let R be a ring.

a. An R -module M is *free* on a subset X of M if for any R -module N and function $\bar{\phi} : X \rightarrow N$ of elements of N , there exists a unique R -module homomorphism $\phi : M \rightarrow N$ such that $\phi(x) = \bar{\phi}(x)$ for all $x \in X$.

b. A *basis* of an R -module M is a subset of M on which it is free.

REMARK 5.7.13. An abelian group A is free on a set X if and only if it is a free \mathbb{Z} -module on X , as follows from Proposition 4.4.11.

In fact, we have the following alternative definition of a free R -module. The proof is nearly identical to Proposition 4.4.11, so omitted.

PROPOSITION 5.7.14. An R -module M is free on a basis X if and only if the set X generates M and, for every $n \geq 1$ and $x_1, x_2, \dots, x_n \in X$, the equality

$$\sum_{i=1}^n c_i x_i = 0$$

for some $c_1, c_2, \dots, c_n \in R$ implies that $c_i = 0$ for all i .

REMARK 5.7.15. We might refer to the property that a set X generates an R -module M as saying that M is the R -span of X . The property that $\sum_{i=1}^n c_i x_i = 0$ implies $c_i = 0$, where $c_i \in R$ and $x_i \in X$ for $1 \leq i \leq n$ and some $n \geq 1$ can be referred to as saying that the set X is R -linearly independent.

COROLLARY 5.7.16. For any set X , the R -module $\bigoplus_{x \in X} R$ is free on the standard basis $\{e_x \mid x \in X\}$, where e_x for $x \in X$ is the element which is nonzero only in its x -coordinate, in which it is 1.

PROOF. The e_x span $\bigoplus_{x \in X} R$ by its definition and are clearly R -linearly independent. □

COROLLARY 5.7.17. Every R -module is a quotient of a free R -module.

PROOF. Let M be an R -module, and choose a generating set X of M (e.g., M itself). Take the unique R -module homomorphism

$$\psi : \bigoplus_{x \in X} R \rightarrow M$$

which satisfies $\psi(e_x) = x$ for all $x \in X$. It is onto as X generates M . □

Noting Corollary 5.5.26, we also have the following.

COROLLARY 5.7.18. Every vector space over a field F is a free F -module.

The following is also a consequence of the universal property. Though we restrict to the finite case, it can be improved to a statement on cardinality.

THEOREM 5.7.19. *Let R be a commutative ring. A free module M on a set X is isomorphic to a free module N on a set Y if and only if X and Y have the same cardinality.*

PROOF. If X and Y have the same cardinality, then any bijection $f: X \rightarrow Y$ gives an injection $X \rightarrow N$ which extends uniquely to a homomorphism $\phi: M \rightarrow N$. Similarly, the inverse of f extends uniquely to a homomorphism $\psi: N \rightarrow M$, and $\psi \circ \phi$ (resp., $\phi \circ \psi$) is then the unique extension to a homomorphism of the inclusion $X \rightarrow M$ (resp., $Y \rightarrow N$), therefore the identity. That is, ϕ and ψ are inverse isomorphisms.

For the converse, we first suppose that Y is infinite and that there is an isomorphism $M \rightarrow N$. Let B denote the image of X in N , which is then necessarily an R -basis of N . Each element $y \in Y$ is contained in the span of a finite subset B_y of B . The union B' of these sets B_y spans Y . For any $v \in B - B'$, the set $B' \cup \{v\}$ is R -linearly dependent, which cannot happen as B is a basis. Thus, $B = B'$. Now, the cardinality $|B|$ of B is at most the cardinality of the disjoint union of the sets B_y for $y \in Y$, each of which is finite. In particular, we have

$$|X| = |B| \leq |Y \times \mathbb{Z}| = |Y|,$$

the latter equality holding as Y is infinite. If X is also infinite, then by reversing the roles of X and Y , this forces $|X| = |Y|$.

Finally, suppose that Y is finite, without loss of generality. Let \mathfrak{m} be a maximal ideal of R . Consider the field $F = R/\mathfrak{m}$, and observe that

$$M/\mathfrak{m}M \cong \left(\bigoplus_{x \in X} R \right) / \mathfrak{m} \left(\bigoplus_{x \in X} R \right) \cong \bigoplus_{x \in X} F,$$

and similarly for Y . An isomorphism $M \xrightarrow{\sim} N$ induces an isomorphism of F -vector spaces $M/\mathfrak{m}M \xrightarrow{\sim} N/\mathfrak{m}N$, which by the above isomorphisms have bases of cardinality $|X|$ and $|Y|$ respectively. Since Y is finite, Theorem 5.5.28 tells us that X must be finite of order $|Y|$. \square

The following is immediate.

COROLLARY 5.7.20. *Let R be a commutative ring, and let M be a free R -module on a set of n elements. Then every basis of M has n elements.*

By Theorem 5.7.22, we may make the following definition.

DEFINITION 5.7.21. The *rank* of a free module M over a commutative ring R is the unique $n \geq 0$ such that $M \cong R^n$ if it exists. Otherwise, M is said to have infinite rank.

For an integral domain, we can do somewhat better with a bit of work. In fact, the following result does not require this assumption, but the proof we give does.

THEOREM 5.7.22. *Let R be an integral domain. Let M be a free R -module on a set of n elements, and let Y be a subset of M . Then:*

- i. *if Y generates M , then Y has at least n elements,*
- ii. *if Y is R -linearly independent, then Y has at most n elements, and*
- iii. *Y is a basis if and only if it generates M and has exactly n elements.*

Moreover, a free module on an infinite set cannot be generated by a finite set of elements.

PROOF. Suppose that M is free on n elements. A choice of basis defines an isomorphism $M \xrightarrow{\sim} R^n$ of R -modules, so we may assume that $M = R^n$. Note that R^n is contained in the $Q(R)$ -module $Q(R)^n$ via the canonical inclusion, and any generating set Y of R^n spans $Q(R)^n$. But by Theorems 5.5.28 and 5.5.27, this forces Y to have at least n elements. If Y has n elements, then Y would similarly be a basis of $Q(R)^n$. So, if we had $\sum_{i=1}^n c_i y_i = 0$ for some $c_i \in R$ and distinct $y_i \in Y$, then each $c_i = 0$, which means that Y is an R -basis of R^n .

On the other hand, if Y has more than n elements, then by Theorem 5.5.25, the set Y cannot be linearly independent in $Q(R)^n$. That is, there exist $\alpha_i \in Q(R)^n$ and distinct $y_i \in Y$ for $1 \leq i \leq m$ and $m \geq 1$ with $\sum_{i=1}^m \alpha_i y_i = 0$ and not all $\alpha_i = 0$. For each i , write $\alpha_i = c_i d_i^{-1}$ with $c_i, d_i \in R$ and $d_i \neq 0$. Taking d to be the product of the d_i , we then have $a_i = d \alpha_i \in R$ and not all $a_i = 0$. Since $\sum_{i=1}^m a_i y_i = 0$, it follows that Y is not a basis.

Finally, if N is a free module on an infinite set X , then $N \cong \bigoplus_{x \in X} R$, and so we take N to be the latter module. We then have that $\bigoplus_{x \in X} Q(R)$ is a $Q(R)$ -vector space with an infinite basis. But then Theorem 5.5.28 tells us that every basis is infinite, which by Theorem 5.5.27 tells us that a finite set cannot span. \square

REMARK 5.7.23. The full analogues of Theorems 5.5.25 and 5.5.27 do not hold for modules over arbitrary rings, over even abelian groups. That is, take the free \mathbb{Z} -module \mathbb{Z} . The set $\{2\}$ does not span it and is not contained in a basis of \mathbb{Z} , and the set $\{2, 3\}$ does span it and does not contain a basis.

EXAMPLE 5.7.24. The polynomial ring $R[x]$ is a free R -module on the basis $\{x^i \mid i \in \mathbb{Z}_{\geq 0}\}$.

REMARK 5.7.25. Consider the ideal $I = (2, x)$ of $\mathbb{Z}[x]$. It is not a free $\mathbb{Z}[x]$ -module. To see this, first note that it is not a principal ideal so cannot be generated by a single element. As I can be generated by the two elements 2 and x , if I were free, then it would follow from Theorem 5.7.22 that $\{2, x\}$ would be a basis for I . On the other hand, $x \cdot 2 - 2 \cdot x = 0$, which would contradict Proposition 5.7.14.

PROPOSITION 5.7.26. Let M be an R -module, and let $\pi: M \rightarrow F$ be a surjective R -module homomorphism, where F is R -free. Then there exists an injective R -module homomorphism $\iota: F \rightarrow M$ such that $\pi \circ \iota = \text{id}_F$. Moreover, we have $M = \ker(\pi) \oplus \iota(F)$.

PROOF. Let X be an R -basis of F , and for each $x \in X$, choose $m_x \in M$ with $\pi(m_x) = x$. We take $\iota: F \rightarrow M$ to be the unique R -module homomorphism with $\iota(x) = m_x$ for all $x \in X$, which exists as F is free. Then $\pi \circ \iota(x) = x$ for all $x \in X$, so $\pi \circ \iota = \text{id}_F$ by uniqueness, and ι must be injective.

Finally, let $A = \ker \pi$. Note that any $m \in M$ satisfies $m - \iota \circ \pi(m) \in A$, so $M = A + \iota(F)$. If $m \in A \cap \iota(F)$, then $m = \iota(n)$ for some $n \in F$ and $n = \pi \circ \iota(n) = \pi(m) = 0$, so $m = 0$. In other words, we have $M = A \oplus \iota(F)$. \square

In particular, every free quotient of an R -module M is isomorphic to a direct summand of M .

We work in this section with (nonzero) homomorphisms of free modules over a ring R . Most of the time, the case of interest is that of linear transformations of vector spaces over fields, but there is no additional restriction caused by working in full generality.

LEMMA 5.8.1. *Let R be a ring. Let $A \in M_{mn}(R)$ be a matrix for some $m, n \geq 1$. Then there is a unique R -module homomorphism $T: R^n \rightarrow R^m$ satisfying $T(v) = Av$ for all $v \in R^n$, where Av is matrix multiplication, viewing elements of R^m and R^n as column vectors.*

PROOF. Define $T(e_j) = \sum_{i=1}^m a_{ij}f_i$, where e_j (resp., f_i) is the j th (resp., i th) standard basis element of R^n (resp., R^m). If $v = \sum_{j=1}^n c_j e_j$ for some $c_j \in R$ with $1 \leq j \leq n$, then

$$T(v) = \sum_{j=1}^n c_j T(e_j) = \sum_{i=1}^m \left(\sum_{j=1}^n a_{ij} c_j \right) f_i = Av.$$

The uniqueness follows from the fact that R^n is free, so any R -module homomorphism from it is determined by its values on a basis □

DEFINITION 5.8.2. An *ordered basis* is a basis of a free R -module together with a total ordering on the basis.

REMARK 5.8.3. We refer to a finite (ordered) basis on a free R -module as a set $\{v_1, v_2, \dots, v_n\}$ and take this implicitly to mean that the set has cardinality n and that the basis is ordered in the listed order (i.e., by the ordering $v_i \leq v_{i+1}$ for all $1 \leq i < n$).

EXAMPLE 5.8.4. The standard basis $\{e_1, e_2, \dots, e_n\}$ on R^n is ordered in the order of positions of the nonzero coordinate of its elements.

NOTATION 5.8.5. If $B = \{v_1, v_2, \dots, v_n\}$ is an ordered basis of a free R -module V , then we let $\varphi_B: R^n \rightarrow V$ denote the R -module isomorphism satisfying $\varphi_B(e_i) = v_i$ for all i .

Given ordered bases of free R -modules V and W , an R -module homomorphism $T: V \rightarrow W$ can be described by a matrix.

DEFINITION 5.8.6. Let V and W be free modules over a ring R with ordered bases $B = \{v_1, v_2, \dots, v_n\}$ and $C = \{w_1, w_2, \dots, w_m\}$, respectively. Let $T: V \rightarrow W$ be an R -module homomorphism. We say that a matrix $A = (a_{ij}) \in M_{nm}(R)$ *represents T with respect to the bases B and C* if

$$T(v_j) = \sum_{i=1}^m a_{ij} w_i$$

for all $1 \leq j \leq n$.

REMARK 5.8.7. Given ordered bases $B = \{v_1, \dots, v_n\}$ of a free module V and $C = \{w_1, \dots, w_m\}$ of a free module W , the composition

$$\varphi_C^{-1} \circ T \circ \varphi_B: R^n \xrightarrow{\varphi_B} V \xrightarrow{T} W \xrightarrow{\varphi_C^{-1}} R^m,$$

is given by multiplication by a matrix A by Lemma 5.8.1. This A is the matrix representing T with respect to B and C .

TERMINOLOGY 5.8.8. Let V be a free R -module with finite basis B , and let $T: V \rightarrow V$ be an R -module homomorphism. We say that a matrix A *represents T with respect to B* if A represents T with respect to B and B . If $V = R^n$ and B is the standard basis, we simply say that A *represents T* .

LEMMA 5.8.9. Let $T': U \rightarrow V$ and $T: V \rightarrow W$ be homomorphisms of finite rank free R -modules. Let B, C , and D be bases of U, V , and W , respectively. Suppose that A' represents T' with respect to B and C and that A represents T with respect to C and D . Then AA' represents $T \circ T': U \rightarrow W$ with respect to B and D .

PROOF. We have that A represents $\varphi_D^{-1} \circ T \circ \varphi_C$ and A' represents $\varphi_C^{-1} \circ T' \circ \varphi_B$. In other words, the maps are left multiplication by the corresponding matrices. The map

$$\varphi_D^{-1} \circ T \circ T' \circ \varphi_B = (\varphi_D^{-1} \circ T \circ \varphi_C) \circ (\varphi_C^{-1} \circ T' \circ \varphi_B),$$

is then left multiplication by AA' , which is to say that it is represented by AA' . \square

DEFINITION 5.8.10. Let $B = \{v_1, \dots, v_n\}$ and $B' = \{v'_1, \dots, v'_n\}$ be ordered bases of a free R -module V . The *change-of-basis matrix* from B to B' is the matrix $Q_{B,B'} = (q_{ij})$ that represents the R -module homomorphism $T_{B,B'}: V \rightarrow V$ with $T_{B,B'}(v_i) = v'_i$ for $1 \leq i \leq n$ with respect to B .

REMARK 5.8.11. If $v'_j = \sum_{i=1}^n q_{ij}v_i$ for all i , then the change-of-basis matrix $Q_{B,B'}$ of Definition 5.8.10 is the matrix (q_{ij}) . It is invertible, and $Q_{B',B} = Q_{B,B'}^{-1}$.

REMARK 5.8.12. Let V be free of rank n with bases B and B' . By definition, the change-of-basis matrix $Q_{B,B'}$ represents $\varphi_B^{-1} \circ T_{B,B'} \circ \varphi_B$. On the other hand, we also have that $\varphi_{B'} = T_{B,B'} \circ \varphi_B$. Thus, see that

$$\varphi_B^{-1} \circ \varphi_{B'} = \varphi_B^{-1} \circ T_{B,B'} \circ \varphi_B,$$

is represented by $Q_{B,B'}$.

THEOREM 5.8.13 (Change of basis theorem). Let $T: V \rightarrow W$ be a linear transformation of free R -modules of finite rank. Let B and B' be ordered bases of V and C and C' be ordered bases of W . If A is the matrix representing T with respect to B and C , then $Q_{C,C'}^{-1}AQ_{B,B'}$ is the matrix representing T with respect to B' and C' .

PROOF. We have that A represents $\varphi_C^{-1} \circ T \circ \varphi_B$, and we wish to compute the matrix representing $\varphi_{C'}^{-1} \circ T \circ \varphi_{B'}$. We have

$$\varphi_{C'}^{-1} \circ T \circ \varphi_{B'} = (\varphi_{C'}^{-1} \circ \varphi_C) \circ (\varphi_C^{-1} \circ T \circ \varphi_B) \circ (\varphi_B^{-1} \circ \varphi_{B'}),$$

and these three matrices are represented by $Q_{C,C'}^{-1}$, A , and $Q_{B,B'}$, respectively. \square