A Course on ALGEBRA

by

Ahmet K. Feyzioğlu

I. CİLT

A Course on

ALGEBRA

by

Ahmet K. Feyzioğlu

I. CİLT

Dedicated to the memory of my father

SABAHATTİN FEYZİOĞLU

Cataloging - in - publication Data

Feyzioğlu, Ahmet K.

A course on algebra
Bibliographyip

1. Algebra 1. Title
512

ISBN: 975-518-014-1

Boğaziçi University Publication No:496
printed in Turkey
Boğaziçi University Printing Office
Bebek, İstanbul 80815 Turkey

Preface

The present book is a text-book in a course on algebra, offered to sophmores at Boğaziçi University. It contains much more material than could be reasonably covered in a year. Depending on the level of class, taste of the instructor and other factors, some paragraphs may be omitted. For this reason, I tried to keep the paragraphs independent.

This book is written, of course, to teach algebra, but this is not my chief concern. My principle objective in writing this book is to introduce the student to mathematical reasoning. This book is addressed to students who are earnestly willing to acquire mathematical maturity, and a course on algebra is a very convenient means to gain it. Hence I assume very little on the reader as far as mathematical background is concerned, but I suppose the reader is genuinely ready to undertake some necessary trouble and toil in order to gain mathematical maturity. I am perfectly aware that this might be a heavier prerequisite than demanding a strong mathematical background.

Mathematical maturity cannot be gained by merely wishing it, nor can anything in mathematics be learned at all by reading a text-book or by attending certain courses. He (or she) who wants to learn mathematics should be actively doing mathematics: give lectures, prepare seminars, write articles or reports, solve problems, etc. At the sophmore level, a modest way of doing active mathematics might be solving exercises. For this reason, I included numerous exercises at the end of each paragraph (except §47) ranging from routine computations to extensions of the ideas to more general situations. Some exercises are asked several times. I tried to choose instructive exercises and avoid tricky ones. I would strongly suggest that the you attempt to solve them. This is the way to learn mathematics. But do not be discouraged when you cannot solve an exercise. They are not enemies to be defeated. Even if one cannot solve a problem, one gains a better grasp of the theory. Learning where and why a certain method fails also deepens and solidifies your understanding of the matter.

For reasons of method, the text is independent of the exercises. Although a few of the exercises are mentioned in some examples, these examples may safely be omitted. Exercises are never needed in proofs of theorems. When I need a result contained in an exercise, I prove it, even in extremely easy cases. Hence one can read the proofs without bothering to solve any exercise. But I warn again: reading and following a proof is certainly not the same thing as understanding it.

The prerequisites are collected in the first chapter. As the reader is assumed to be familiar with the contents of Chapter 1, I allowed myself certain deviations from the strictly logical order there. The reader will probably find certain proofs in Chapter 1 unduly long. They might be long, but not without reason. As a rule, I do not give the shortest or the simplest proof of a theorem, but the proof that lends itself most easily to generalization. For instance, Example2.3(f) could be shorter when we used rational numbers, but the given argument is used again in Lemma 31.2 in a much more general situation, and the shorter proof would not work there. Proofs in Chapter 1 are presented in such a way as to work also in broader contexts in later parts of the book.

What is said about Chapter 1 in relation to the rest of the book is also true for the whole book in relation to higher algebra. Whenever possible, I gave proofs that work also in more general situations. For example, writing the groups additively and reading "R-module" in place of "abelian group" in §28, one gets structure theorems of R-modules over a principal ideal domain R.

The book really begins with Chapter 2. Here we introduce groups. After we learned subgroups and Lagrange's theorem, we present some important examples of groups (§11-§17). This makes it possible to see some genuine examples of homomorphisms. Then subgroup structure and homomorphisms are examined (§18-21). The remaining paragraphs of Chapter 2 are not heavily needed in later parts of the book. An exception is the material on solvable groups (§27), which may be deferred until §59.

Chapter 3 is about rings and Chapter 4 about vector spaces. The study of the subring structure of a ring (§30) and the subspace structure of a vector space (§41) depends heavily on the subgroup structure of a group. Here of course there can be no question about independence of para-

graphs. We study divisibility theory in integral domains and in polynomial domains in Chapter 3. The theory of vector spaces in Chapter 4 is quite general, but is developed as far as needed for Chapter 5. In particular, computational aspects are neglected and we do not give methods for solving numerical linear equations.

The book reaches its climax in Chapter 5. Groups, rings, vector spaces come here in a happy interplay. We focus our attention on algebraic field extensions and give an exposition of Galois theory. I did not restrict myself to fields of characteristic 0, but confined the discussion of separability to algebraic extensions.

I might be criticized for my pace, which is indeed extremely slow. But this book is primarily written for beginners, and it suits them better to develope the material slowly. At the same time, I hope the book will have something to say to the experts as well.

I would like to close with thanks to some people. Foremost among them is my friend and colleague Yalçın Koç from the department of Philosophy and now the dean of our faculty. I thank him for his friendly encouragement and for his help that has been most invaluable. Without his material help, this book would appear at a much later date, in a much less satisfactory form. It is a pleasent duty to extend my deep thanks to him. I also thank Attila Aşkar, chairman of our department, and Gürol Irzık from the department of Philosophy for their encouraging remarks. Finally, I would like to thank all the lovely nice students of the mathematics department. It is a great pleasure to teach them.

Ahmet K. Feyzioğlu

November 1990
Istanbul

Table of Contents

Chapter 3
Rings

CHAPTER 1

Preliminaries

§1
Set Theory

We assume that the reader is familiar with basic set theory. In this paragraph, we want to recall the relevant definitions and fix the notation.

Our approach to set theory will be informal. For our purposes, a *set* is a collection of objects, taken as a whole. "Set" is therefore a collective term like "family", "flock", "species", "army", "club", "team" etc. The objects which make up a set are called the *elements* of that set. We write

$$x \in S$$

to denote that the object x is an element of the set S. This can be read "x is an element of S", or "x is a member of S", or "x belongs to S", or "x is in S", or "x is contained in S", or "S contains x". If x is not an element of S, we write

$$x \notin S.$$

For technical reasons, we agree to have a unique set that has no elements at all. This set is called the *empty set* and is denoted by $\varnothing$.

A set S is called a *subset* of a set T if every element of S is also an element of T. The notation

$$S \subseteq T$$

means that S is a subset of T. This is read "S is a subset of T", or "S is included in T", or "S is contained in T". By convention, the empty set $\varnothing$ is a subset of any set. If S is not a subset of T, we write

$$S \not\subseteq T.$$

This means there is at least one element of T which does not belong to S.

If $S \subseteq T$ and $T \subseteq S$, then S and T have exactly the same elements. In this case, S and T are said to be *identical* or *equal*. We write

$$S = T$$

if S and T are equal sets. Whenever we want to prove that two sets S and T are equal, we must show that S is included in T and that T is included in S. If S and T are not equal, we put

$$S \neq T.$$

If $S \subseteq T$ but $T \neq S$, then S is said to be a *proper subset* of T. So S is a proper subset of T if and only if every element of S is an element of T but T contains at least one element which does not belong to S. The notation

$$S \subset T$$

means that S is a proper subset of T. This is read "S is a proper subset of T", or "S is properly included in T", or "S is properly contained in T". By convention, the empty set $\varnothing$ is a proper subset of every set except itself.

Some authors write $S \subset T$ to mean that S is a subset of T, the possibility $S = T$ being included, and $S \subsetneq T$ to mean that S is a proper subset of T. The reader should be careful about the meaning of the symbol "$\subset$" he or she uses. In this book, "$\subset$" denotes proper inclusion.

Sets are sometimes written by displaying their elements within braces (roster notation). Hence

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

is the set whose elements are the numbers 1,2,3,4 and 5. Obviously, only those sets which have a small number of elements can be written in this

way. In many cases, the elements of a set S are characterized by a property P and the set is then written

$$\{x : x \text{ has property } P\}.$$

In this book, $\mathbb{N} = \{1,2,3, \ldots\}$ is the set of natural numbers, $\mathbb{Z} = \{0,\mp1,\mp2, \ldots\}$ is the set of integers, $\mathbb{Q} = \{\frac{a}{b} : a,b \in \mathbb{Z}, b \neq 0\}$ is the set of rational numbers, $\mathbb{R}$ is the set of real numbers, $\mathbb{C}$ is the set of complex numbers. These notations are standard. Some authors regard 0 as a natural number, but we agree that $0 \notin \mathbb{N}$ in this book.

Given two sets S and T, we consider those objects which belong to S or to T. Such objects will make up a new set. This set is called the *union of S and T* and is denoted by $S \cup T$. We remark here that 'or' in the definition of a union is the logical 'or'. Let us recall that

'p or q' is true in case 'p' is true, 'q' is true;
'p' is true, 'q' is false;
'p' is false, 'q' is true;

and

'p or q' is false in case 'p' is false, 'q' is false.

Thus we have

$$S \cup T = \{x : x \in S \text{ or } x \in T\}.$$

In particular, $S \cup T = T \cup S$.

If we have sets $S_1, S_2, \ldots, S_n$, their union $S_1 \cup S_2 \cup \ldots \cup S_n$ is given by

$$S_1 \cup S_2 \cup \ldots \cup S_n = \{x : x \in S_1 \text{ or } x \in S_2 \text{ or } \ldots \text{ or } x \in S_n\}.$$

We usually contract this notation into $\bigcup_{i=1}^{n} S_i$, just like we write $\sum_{i=1}^{n} a_i$ instead of $a_1 + a_2 + \cdots + a_n$. More generally, if we have sets S_i, indexed by a set I, then their union $\bigcup_{i \in I} S_i$ is the set

$$\bigcup_{i \in I} S_i = \{x : x \in S_i \text{ for at least one } i \in I\}.$$

Given two sets S and T, we consider those objects which belong to S and to T. Such objects will make up a new set. This set is called the *intersection of S and T* and is denoted by $S \cap T$. We remark here that 'and' in the definition of a intersection is the logical 'and'. Let us recall that

'p and q' is true in case 'p' is true, 'q' is true;

and

'p and q' is false in case 'p' is true, 'q' is false;
'p' is false, 'q' is true;
'p' is false, 'q' is false.

Thus we have

$$S \cap T = \{x : x \in S \text{ and } x \in T\}.$$

In particular, $S \cap T = T \cap S$.

If we have sets $S_1, S_2, \ldots, S_n$, their intersection $S_1 \cap S_2 \cap \ldots \cap S_n$ is given by

$$S_1 \cap S_2 \cap \ldots \cap S_n = \{x : x \in S_1 \text{ and } x \in S_2 \text{ and } \ldots \text{ and } x \in S_n\}.$$

We usually contract this notation into $\bigcap_{i=1}^{n} S_i$. More generally, if we have sets S_i, indexed by a set I, then their intersection $\bigcap_{i \in I} S_i$ is the set

$$\bigcap_{i \in I} S_i = \{x : x \in S_i \text{ for all } i \in I\}.$$

Two sets S and T are said to be *disjoint* if their intersection is empty: $S \cap T = \varnothing$. Given a family of sets S_i, indexed by a set I, the sets S_i are called *mutually disjoint* if any two distinct of them are disjoint:

$$S_{i_1} \cap S_{i_2} = \varnothing \text{ for all } i_1, i_2 \in I,\ S_{i_1} \neq S_{i_2}.$$

The sets we consider in a particular discussion are usually subsets of a set U. This set U is called the *universal set*. Given a set S, which is a subset of a universal set U, those elements of U that do not belong to S make up a new set, called the *complement* of S and denoted by S' or S^c or $C_U(S)$. Hence

$$S' = \{x : x \in U \text{ and } x \notin S\}.$$

More generally, we write

$$T \setminus S = \{x : x \in T \text{ and } x \notin S\}.$$

and call this set the *relative complement of S in T*, or the *difference set T minus S*. The set S may or may not be a subset of T. Note that

$$T \setminus S = T \cap S'.$$

According to our definition of equality, the sets $\{a,b\}$ and $\{b,a\}$ are equal. Frequently, we want to distinguish between a,b and b,a. To this end, we define ordered pairs. An *ordered pair* is a pair of objects a,b, enclosed

within parentheses and separated by a comma. Thus (a,b) is an ordered pair. The adjective "ordered" is used to emphasize that the objects have a status of being first and being second. a is called the *first component* of the ordered pair (a,b), and b is called its *second component*. Two ordered pairs are declared equal if their first components are equal and their second components are equal. Thus (a,b) and (c,d) are equal if and only if $a = c$ and $b = d$, in which case we write $(a,b) = (c,d)$. Notice that we have $(a,b) \neq (b,a)$ unless $a = b$ (here $\neq$ means the negation of equality).

The set of all ordered pairs, whose first components are the elements of a set S and whose second components are the elements of a set T, is called the *cartesian product of S and T*, and is denoted by $S \times T$. Hence

$$S \times T = \{(a,b): a \in S \text{ and } b \in T\}.$$

We can also define ordered triples (a,b,c), ordered quadruples (a,b,c,d), more generally ordered n-tuples $(a_1,a_2,\ldots,a_n)$. Equality of ordered n-tuples will mean the equality of their corresponding components. The set of all ordered n-tuples, whose i-th components are the elements of a set S_i, is called the *cartesian product of* $S_1,S_2,\ldots,S_n$ and is denoted by $S_1 \times S_2 \times \ldots \times S_n$. Hence

$$S_1 \times S_2 \times \ldots \times S_n = \{(a_1,a_2,\ldots,a_n): a_1 \in S_1, a_2 \in S_2,\ldots,a_n \in S_n\}.$$

It is possible to define the cartesian product of infinitely many sets, too. We do not give this definition, for we will not need it.

A set can have finitely many or infinitely many elements. The number of elements in a set S is called the *cardinality* or the *cardinal number of* S. The cardinality of S is denoted by $|S|$. The set S is said to be *finite* if $|S|$ is a finite number. S is said to be *infinite* if S is not finite. A rigorous definition of finite and infinite sets must be based on the notion of one-to-one correspondence between sets, which will be introduced in §3. However, we will not make any attempt to give a rigorous definition of finite and infinite sets. We shall be content with the suggestive description above.

Exercises

1. Show that, if R is a subset of S and S is a subset of T, then R is a subset of T.

2. Show that $(R \cup S) \cup T = R \cup (S \cup T)$
and $(R \cap S) \cap T = R \cap (S \cap T)$.

3. Prove: $S \cap T = S$ if and only if $S \subseteq T$, and $S \subseteq T$ if and only if $S \cup T = T$.

4. Prove the distributivity of union over intersection and of intersection over union:

$$R \cup (S \cap T) = (R \cup S) \cap (R \cup T),$$
$$R \cap (S \cup T) = (R \cap S) \cup (R \cap T).$$

5. Prove the deMorgan laws:

$$(S \cup T)' = S' \cap T' \quad \text{and} \quad (S \cap T)' = S' \cup T'$$

for any subsets S,T of a universal set U.

6. Show that $T \setminus \varnothing = T$ and $T \setminus T = \varnothing$ for any set T.

7. Prove: $(S \setminus T) \cup (T \setminus S) = (S \cup T) \setminus (T \cap S)$. This set is called the *symmetric difference of S and T*. It is denoted by $S \Delta T$.

8. With the notation of Ex. 7, prove that

$$(R \Delta S) \Delta T = R \Delta (S \Delta T)$$
$$S \Delta \varnothing = S$$
$$S \Delta S = \varnothing$$
$$S \Delta T = T \Delta S.$$

9. Let S and T be finite sets. Prove the following assertions.
   a) If $S \cap T = \varnothing$, then $|S \cup T| = |S| + |T|$.
   b) $|S \cup T| = |S| + |T| - |S \cap T|$. (Hint: $S \cup T = S \cup (T \setminus S)$.)

10. Find all subsets of $\varnothing$, $\{1\}$, $\{1,2\}$,$\{1,2,3\}$,$\{1,2,3,4\}$.

11. Prove: if S is a finite set, then S has exactly $2^{|S|}$ subsets.

§2
Equivalence Relations

In mathematics, we often investigate relationships between certain objects (numbers, functions, sets, figures, etc.). If an element a of a set A is related to an element b of a set B, we might write

$$a \text{ is related to } b$$

or shortly

$$a \text{ related } b$$

or even more shortly

$$a\ R\ b.$$

The essential point is that we have two objects, a and b, that are related in some way. Also, we say "a is related to b", not "b is related to a", so the *order* of a and b is important. In other words, the ordered pair (a,b) is distinguished by the relation. This observation suggests the following formal definition of a relation.

2.1 Definition: Let A and B be two sets. A *relation R from A into B* is a subset of the cartesian product $A \times B$.

If A and B happen to be equal, we speak of a relation *on* A instead of using the longer phrase "a relation from A into A".

Equivalence relations constitute a very important type of relations on a set.

2.2 Definition: Let A be a nonempty set. A relation R on A (that is, a subset R of $A \times A$) is called an *equivalence relation on* A if the following hold.

(i) $(a,a) \in R$ for all $a \in A$,

(ii) if $(a,b) \in R$, then $(b,a) \in R$ (for all $a,b \in A$).

(iii) if $(a,b) \in R$ and $(b,c) \in R$, then $(a,c) \in R$

(for all $a,b,c \in R$).

This definition presents the logical structure of an equivalence relation very clearly, but we will almost never use this notation. We prefer to write $a \sim b$, or $a \approx b$, or $a \equiv b$ or some similar symbolism instead of $(a,b) \in R$ in order to express that a,b are related by an equivalence relation R. Here $a \sim b$ can be read "a is equivalent to b". Our definition then assumes the form below.

2.2 Definition: Let A be a nonempty set. A relation R on A (that is, a subset R of $A \times A$) is called an *equivalence relation on* A if the following hold.

(i) $a \sim a$ for all $a \in A$,
(ii) if $a \sim b$, then $b \sim a$ (for all $a,b \in A$),
(iii) if $a \sim b$ and $b \sim c$ then $a \sim c$ (for all $a,b,c \in A$).

A relation $\sim$ that satisfies the first condition (i) is called a *reflexive* relation, one that satisfies the second condition (ii) is called a *symmetric* relation, one that satisfies the third condition (iii) is called a *transitive* relation. An equivalence relation is therefore a relation which is reflexive, symmetric and transitive. Notice that symmetry and transitivity requirements involve conditional statements (if ..., then ...). In order to show that $\sim$ is symmetric, for example, we must make the hypothesis $a \sim b$ and use this hypothesis to establish $b \sim a$. On the other hand, in order to show that $\sim$ is reflexive, we have to establish $a \sim a$ for all $a \in A$, without any further assumption.

2.3 Examples: (a) Let A be a nonempty set of numbers and let equality $=$ be our relation. Then $=$ is certainly an equivalence relation on A since

(i) $a = a$ for all $a \in A$,
(ii) if $a = b$, then $b = a$ (for all $a,b \in A$),
(iii) if $a = b$ and $b = c$, then $a = c$ (for all $a,b,c \in A$).

(b) Let A be the set of all points in the plane except the origin. For any two points P and R in A, let us put $P \sim R$ if R lies on the line through the origin and P.

(i) $P \sim P$ for all points P in A since any point lies on the line through the origin and itself. Thus $\sim$ is reflexive.

(ii) If $P \sim R$, then R lies on the line through the origin and P; therefore the origin, P, R lie on one and the same line; therefore P lies on the line through the origin and R; and $R \sim P$. Thus $\sim$ is symmetric.

(iii) If $P \sim R$ and $R \sim T$, then the line through the origin and R contains the points P and T, so T lies on the line through the origin and P, so we get $P \sim T$. Thus $\sim$ is transitive.

This proves that $\sim$ is an equivalence relation on A.

(c) Let S be the set of all straight lines in the plane. Let us put $m // n$ if the line m is parallel to the line n. It is easily seen that $//$ (parallelism) is an equivalence relation on S.

(d) Let $\mathbb{Z}$ be the set of integers. For any two numbers a,b in $\mathbb{Z}$, let us put $a \equiv b$ if $a - b$ is even (divisible by 2).

(i) $a \equiv a$ for all $a \in \mathbb{Z}$ since $a - a = 0$ is an even number.

(ii) If $a \equiv b$, then $a - b$ is even, then $b - a = -(a - b)$ is also even, so $b \equiv a$.

(iii) If $a \equiv b$ and $b \equiv c$, then $a - b$ and $b - c$ are even. Their sum is also an even number. So $a - c = (a - b) + (b - c)$ is even and $a \equiv c$.

We see that $\equiv$ is an equivalence relation on $\mathbb{Z}$.

(e) The last example may be generalized. We fix a whole number $n \neq 0$ (n is called the *modulus* in this context). For any two numbers a,b in $\mathbb{Z}$, let us put $a \equiv b$ if $a - b$ is divisible by n.

(i) $a \equiv a$ for all $a \in \mathbb{Z}$ since $a - a = 0 = n0$ is divisible by n.

(ii) If $a \equiv b$, then $a - b = nm$ for some $m \in \mathbb{Z}$, so $b - a = -(a-b) = n(-m)$ is divisible by n, and $b \equiv a$.

(iii) If $a \equiv b$ and $b \equiv c$, then $a - b = nm$ and $b - c = nk$ for some $m,k \in \mathbb{Z}$, so $a - c = (a - b) + (b - c) = nm + nk = n(m + k)$ is divisible by n, and so $a \equiv b$.

Therefore, $\equiv$ is an equivalence relation on $\mathbb{Z}$. This relation is called *congruence*. For each nonzero integer n, there is a congruence relation. In order to distinguish between them, we write, when n is the modulus, $a \equiv b \pmod{n}$ rather than $a \equiv b$.

(f) Let $S = \mathbb{Z} \times (\mathbb{Z}\backslash\{0\})$. Thus S is the set of all ordered pairs of integers whose second components are distinct from zero. Let us write $(a,b) \approx (c,d)$ for $(a,b), (c,d) \in S$ if $ad = bc$.

(i) $(a,b) \approx (a,b)$ for all $(a,b) \in S$, since $ab = ba$ for all $a \in \mathbb{Z}$, $b \in \mathbb{Z}\backslash\{0\}$.

(ii) If $(a,b) \approx (c,d)$, then $ad = bc$, then $da = cb$, then $cb = da$, so $(c,d) \approx (a,b)$.

(iii) If $(a,b) \approx (c,d)$ and $(c,d) \approx (e,f)$, then

$$ad = bc \quad \text{and} \quad cf = de$$
$$adf = bcf \quad \text{and} \quad bcf = bde$$
$$adf = bde$$
$$d(af - be) = 0$$
$$af - be = 0 \qquad (\text{since } d \neq 0)$$
$$af = be$$
$$(a,b) \approx (e,f).$$

Thus $\approx$ is an equivalence relation on S.

(g) Let T be the set of all triangles in the Euclidean plane. Congruence of triangles is an equivalence relation on T.

(h) Let S be the set of all continuous functions defined on the closed interval [0,1]. For any two functions f,g in S, let us write $f \overset{A}{=} g$ if

$$\int_0^1 f(x)dx = \int_0^1 g(x)dx.$$

Then $\overset{A}{=}$ is an equivalence relation on S.

An equivalence relation is a weak form of equality. Suppose we have various objects, which are similar in one respect and dissimilar in certain other respects. We may wish to ignore their dissimilarity and focus our attention on their similar behaviour. Then there is no need to distinguish between our various objects that behave in the same way. We may regard them as equal or identical. Of course, "equal" or "identical" are poor words to employ here, for the objects are not absolutely identical, they are equal only in one respect that we wish to investigate more closely. So we employ the word "equivalent". That a and b are equivalent means, then, a and b are equal, not in every respect, but rather as far as a particular property is concerned. An equivalence relation is a formal tool for disregarding differences between various objects and treating them as equals.

Let us examine our examples under this light. In Example 2.3(b), the points P and R may be different, but the lines they determine with the origin are equal. In Example 2.3(c), the lines may be different, but their directions are equal. In Example 2.3(d), the integers may be different, but their parities are equal. In Example 2.3(e), the integers may be

different, but their remainders, when they are divided by n, are equal. In Example 2.3(f), the pairs may be different, but the ratio of their components are equal. In Example 2.3(g), the triangles may have different locations in the plane, but their geometrical properties are the same. In Example 2.3(h), the functions may be different, but the "areas under their curves" are equal.

An equivalence relation $\sim$ on a set A gives rise to a partition of A into disjoint subsets. This means that A is a union of certain subsets of A and that the distinct subsets here are mutually disjoint. The converse is also true: whenever we have a partition of a nonempty set A into pairwise disjoint subsets, there is an equivalence relation on A. Before proving this important result, we introduce a definition.

2.4 Definition: Let $\sim$ be an equivalence relation on a nonempty set A, and let a be an element of A. The *equivalence class of* a is defined to be the set of all elements of A that are equivalent to a.

The equivalence class of a will be denoted by $[a]$ (or by class(a), cl(a), $\bar{a}$ or by a similar symbol): $[a] = \{ x \in A : x \sim a \}$.

An element of an equivalence class $X \subseteq A$ is called a *representative of* X. Notice that $x \in [a]$ and $x \sim a$ have exactly the same meaning. In particular, we have $a \in [a]$ by reflexivity. So any $a \in A$ is a representative of its own equivalence class.

The equivalence classes $[a]$ are subsets of A. The set of all equivalence classes is sometimes denoted by $A/\sim$. It will be a good exercise for the reader to find the equivalence classes in Example 2.3.

We now state and prove the result we promised.

2.5 Theorem: *Let A be a nonempty set and let $\sim$ be an equivalence relation on A. Then the equivalence classes form a partition of A. In other words, A is the union of the equivalence classes and the distinct equivalence classes are disjoint:*

$$A = \bigcup_{a \in A} [a] \quad \text{and} \quad \text{if } [a] \neq [b], \text{ then } [a] \cap [b] = \varnothing.$$

Conversely, let

$$A = \bigcup_{a \in A} P_i, \quad P_i \cap P_j = \varnothing \quad \text{if } i \neq j$$

be a union of nonempty, mutually disjoint sets P_i indexed by I. Then there is an equivalence relation on A such that the P_i's are the equivalence classes under this relation.

Proof: First we prove $A = \bigcup_{a \in A} [a]$. For any $a \in A$, we have $[a] \subseteq A$, hence $\bigcup_{a \in A} [a] \subseteq A$. Also, if $a \in A$, then $a \in [a]$ by reflexivity, so $a \in \bigcup_{a \in A} [a]$ and $A \subseteq \bigcup_{a \in A} [a]$. So $A = \bigcup_{a \in A} [a]$.

Now we must prove that distinct equivalence classes are disjoint. We prove its contrapositive, which is logically the same: if two equivalence classes are not disjoint, then they are identical. Suppose that the equivalence classes $[a]$ and $[b]$ are not disjoint. This means there is a c in A such that $c \in [a]$ and $c \in [b]$. Hence

	$c \sim a$ and $c \sim b$	
	$a \sim c$ and $c \sim b$	(by symmetry)
(1)	$a \sim b$	(by transitivity)
(2)	$b \sim a$	(by symmetry).

We want to prove $[a] = [b]$. To this end, we have to prove $[a] \subseteq [b]$ and also $[b] \subseteq [a]$. Let us prove $[a] \subseteq [b]$. If $x \in [a]$, then $x \sim a$ and $a \sim b$ by (1), then $x \sim b$ by transitivity, then $x \in [b]$, so $[a] \subseteq [b]$. Similarly, if $y \in [b]$, then $y \sim b$, then $y \sim b$ and $b \sim a$ by (2), then $y \sim a$ by transitivity, then $y \in [a]$, so $[b] \subseteq [a]$. Hence $[a] = [b]$ if $[a]$ and $[b]$ are not disjoint. This completes the proof of the first assertion.

Now the converse. Let $A = \bigcup_{a \in A} P_i$, where any two distinct P_i's are disjoint. We want to define an equivalence relation on A and want the P_i's to be the equivalence classes. How do we accomplish this? Well, if the P_i are to be the equivalence classes, we had better call two elements equivalent if they belong to one and the same P_{i_0}.

Let $a \in A$. Since $A = \bigcup_{a \in A} P_i$, we see that $a \in P_{i_0}$ for some $i_0 \in I$. This index i_0 is uniquely determined by a. That is to say, a cannot belong to

two or more of the subsets P_i, for then P_i would not be mutually disjoint. So each element of A belongs to one and only one of the subsets P_i.

Let a,b be elements of A and suppose $a \in P_{i_0}$ and $b \in P_{i_1}$. We put $a \approx b$ if $P_{i_0} = P_{i_1}$, i.e., we put $a \approx b$ if the sets P_i to which a and b belong are identical. We show that $\approx$ is an equivalence relation.

(i) For any $a \in A$, of course a belongs to the set P_{i_0} it belongs to, and so $a \approx a$ and $\approx$ is reflexive.

(ii) Let $a \approx b$. This means a and b belong to the same set P_{i_0}, say, so b and a belong to the same set P_{i_0}, hence $b \approx a$. So $\approx$ is symmetric.

(iii) Let $a \approx b$ and $b \approx c$. Then the set P_i to which b belongs contains a and c. Thus a and c belong to the same set P_i and $a \approx c$. This proves that $\approx$ is transitive.

We showed that $\approx$ is indeed an equivalence relation on A. It remains to prove that P_i are the equivalence classes under $\approx$. For any $a \in A$, we have, if $a \in P_{i_1}$,

$$\begin{aligned}[a] &= \{x \in A : x \approx a\} \\ &= \{x \in A : x \text{ belongs to } P_{i_1}\} \\ &= \{x \in \bigcup_{i \in I} P_i : x \text{ belongs to } P_{i_1}\} \\ &= P_{i_1}.\end{aligned}$$

This proves that P_i are the equivalence classes under $\approx$. □

Exercises

1. On $\mathbb{N} \times \mathbb{N}$, define a relation $\equiv$ by declaring $(a,b) \equiv (c,d)$ if and only if $a + d = b + c$. Show that $\equiv$ is an equivalence relation on $\mathbb{N} \times \mathbb{N}$.

2. Determine whether the relation $\sim$ on $\mathbb{R}$ is an equivalence relation on $\mathbb{R}$, when $\sim$ is defined by declaring $x \sim y$ for all $x,y \in \mathbb{R}$ if and only if

(a) there are integers a,b,c,d such that $ad - bc = \pm 1$ and $x = \dfrac{ay+b}{cy+d}$;

(b) $|x - y| < 0.000001$;

(c) $|x| = |y|$;

(d) $x - y$ is an integer;

(e) $x - y$ is an even integer;

(f) there are natural numbers n,m such that $x^n = y^m$;

(g) there are natural numbers n,m such that $nx = my$;

(h) $x \geqslant y$.

3. Let $\sim$ and $\approx$ be two equivalence relations on a set A. We define $\equiv$ by declaring $a \equiv b$ if and only if $a \sim b$ and $a \approx b$; and we define $\cong$ by declaring $a \cong b$ if and only if $a \sim b$ or $a \approx b$. Determine whether $\equiv$ and $\cong$ are equivalence relations on A.

4. If a relation on A is symmetric and transitive, then it is also reflexive. Indeed, let $\sim$ be the relation and let $a \in A$. Choose an element $b \in A$ such that $a \sim b$. Then $b \sim a$ by symmetry, and from $a \sim b, b \sim a$, it follows that $a \sim a$, by transitivity. So $a \sim a$ for any $a \in A$ and $\sim$ is reflexive.

This argument is wrong. Why?

§ 3
Mappings and Operations

Functions, also called mappings, build a very important type of relations. Let us recall that a relation from A into B is a subset of $A \times B$. Under special circumctances, a relation will be called a function or a mapping. These two terms will be used interchangably.

3.1 Definition: Let A and B be nonempty sets. A relation f from A into B is called a *function from A into B*, or a *mapping from A into B* if every element of A is the first component of a single ordered pair in $f \subseteq A \times B$.

This definition embraces two conditions. First, every element a of A will appear as the first component of at least one ordered pair $(a,*)$ in f, that is, the first components of the ordered pairs in f should make up the whole A. No element of A can be left out. There should be no element of A which is not the first component of any pair in f. Second, for any $a \in A$, there can be only one ordered pair in f whose first component is a. In other words, if (a,b) and (a,b') are both in f, these pairs should be identical, which means $b = b'$. A relation f from A into B is a mapping if and only if every element of A is the first component of one and only one ordered pair in f.

If f is a mapping from A into B, then A is called the *domain* of f, and B is called the *range* of f. A function f from A into B must be thought of as a rule or mechanism by which elements of A are assigned to certain elements of B. The first condition, that every element of A is the first component of at least one ordered pair in f, is a formal way of expressing that elements of A, not of any other set, in particular not of any proper subset of A, are the objects that are assigned (to some elements of B). The second condition, that every element of A is the first component of at most one ordered pair in f, is a formal way of expressing that no element of A is assigned to two, three or more elements of B.

We introduce some notation. We write $f: A \to B$ to mean that f is a mapping from A into B. Occasionally, we write $A \xrightarrow{f} B$. The reader probably

expects that we write $f(a) = b$ in place of $(a,b) \in f$. This is the symbolism that the reader is accustomed to, and reminds us of a mapping rule that assigns b to a. However, we will rarely write $f(a) = b$. We prefer to write $(a)f = b$ or $af = b$, with the function symbol f on the right side of the element a. This might seem odd, and the reader might wonder about this strange order of elements and functions. It takes some time to get accustomed to this way of writing functions on the right, but the advantages of this notation will far outweigh the little trouble it causes at first. This will be amply clear in the sequel. We remark that not every algebraist conforms to this usage, and an isolated notation will have different meanings according as whether the functions are written on the right or on the left. We will point out these differences as occasions arise.

Suppose f is a mapping from A into B and $a \in A$ and $b \in B$ are such that $af = b$ (in this case, we sometimes write $a \to b$ or $a \xrightarrow{f} b$ and say that f *maps* a to b). Then b is called the *image of* a *under* f. We also say a is a *preimage* or an *inverse image of* b *under* f. Please mark the articles: b is the image of a, since a has one and only one image, but a is a preimage of b, for b may have many preimages.

3.2 Examples:(a) Let A be a nonempty set and let $\iota = \{(a,a): a \in A\} \subseteq A \times A$. Then ι is a function from A into A. In our second notation, this reads $a\iota = a$. This function is called the *identity mapping on* A. When we want to point out the set A, we write ι_A instead of ι.

Now let $A \subseteq B$ and put $\mu = \{(a,a) \in A \times B : a \in A\} \subseteq A \times B$. Then μ is a function from A into B. In our second notation, this reads $a\mu = a$. This function is called the *inclusion mapping from* A *into* B. Writing $a\mu$ for a is a formal way of recalling $A \subseteq B$ and $a \in B$.

(b) Let $A = \{1,2,3,4,5\}$ and $B = \{a,b,c,d\}$. Consider

$$f = \{(1,b), (2,a), (4,d), (5,d)\}.$$

Then f is not a function from A into B since $3 \in A$ is not the first component of any ordered pair in $f \subseteq A \times B$. Consider

$$g = \{(1,b), (2,a), (3,a), (3,b), (4,c), (5,d)\}.$$

Then g is not a function from A into B since $3 \in A$ is the first component of two distinct ordered pairs in $g \subseteq A \times B$.

(c) Let A and B be two nonempty sets and let $b \in B$ be a fixed element of B. Then f, defined by

$$af = b \text{ for all } a \in A, \text{ i.e., } f = \{(a,b) \in A \times B: a \in A\}$$

is a mapping from A into B. This is sometimes called the *constant function* b.

(d) For any $(a,b) \in \mathbb{R} \times \mathbb{R}$, put $(a,b)s = a + b$. Then s is a function from $\mathbb{R} \times \mathbb{R}$ into $\mathbb{R}$. This s may be called the *sum* function. It is an example of a binary operation. We will examine binary operations later in this paragraph.

(e) Let $A = \{u,x,y,z\}$ and $B = \{1,2,3\}$, and put

$$uf = 1, xf = 2, yf = 2, zf = 1.$$

Then f is a function from A into B.

(f) Let A be a nonempty set and let S be the set of all subsets of A. For any $a \in A$, put $af = \{a\} \in S$. Then f is a function from A into S.

(g) Put $xf = x^2$ for all $x \in \mathbb{R}$. Then f is a function from $\mathbb{R}$ into $\mathbb{R}$.

(h) Consider $f = \{(x,y) \in \mathbb{R} \times \mathbb{R}: x^2 = y^2\}$. Then f is not a function from $\mathbb{R}$ into $\mathbb{R}$, since 1, for example, is the first component of two distinct ordered pairs $(1,1)$ and $(1,-1)$ in f. On the other hand, if $\mathbb{R}^+$ denotes the set of positive real numbers, then $g = \{(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+: x^2=y^2\}$ is a function from $\mathbb{R}^+$ into $\mathbb{R}^+$. In fact, g is the identity function on $\mathbb{R}^+$.

(i) Let $f: A \to B$ be a mapping from A into B and let A_1 be a nonempty subset of A. For any $a \in A_1$, we put $ag = af$. Then g is a mapping from A_1 into B. In terms of ordered pairs, we have

$$g = f \cap (A_1 \times B).$$

g is called the *restriction of* f *to* A_1. We usually write f_{A_1} or $f_{|A_1}$ to denote the restriction of f to A_1. If g is a restriction of f to a subset of the domain of f, then f is called an *extension of* g.

(j) Let A be a nonempty set and let B be a fixed subset of A. For any a in A, we put

$$\chi_B = \begin{cases} 0 \text{ if } a \notin B, \\ 1 \text{ if } a \in B. \end{cases}$$

Then χ_B is a function from A into $\{0,1\}$. It is called the *characteristic function of* B. Here we wrote the function on the left.

(k) For any $x \in \mathbb{R}$, we put

$$f(x) = \begin{cases} 0 \text{ if } x \text{ is irrational,} \\ 1 \text{ if } x \text{ is rational.} \end{cases}$$

Then f is a function from $\mathbb{R}$ into $\mathbb{R}$. In fact, f is the characteristic function of the set of rational numbers. The image of some x is not known. For instance, it is not known whether Euler's constant γ is rational or not. Nevertheless, f is a genuine function. This example is due to L. Dirichlet (1805-1859).

(l) Let A be a nonempty set and let $\sim$ be an equivalence relation on A. Let $A/\sim$ be the set of equivalence classes under $\sim$. Then

$$\begin{aligned} v\colon A &\to A/\sim \\ a &\to [a] \end{aligned}$$

is a mapping from A into $A/\sim$. It is called the *natural mapping* or the *canonical mapping from* A *into* $A/\sim$.

3.3 Definition: Let $f\colon A \to B$ and $f_1\colon A_1 \to B$ be two functions. f and f_1 are called *equal* if $A = A_1$ and $af = af_1$ for all $a \in A = A_1$.

So, in order that two functions f and f_1 be equal, their domains must be equal and the images of any element in this common domain under the mappings f and f_1 must be equal, too. In particular, if $f\colon A \to B$ is a function and $B \subseteq C$, then the function $g\colon A \to C$, defined by $ag = af$ for all a in A, is equal to f. The ranges do not play any role in the definition of equality. (In some branches of mathematics, for example in topology, two functions with different ranges are sometimes considered distinct, even if their domains and functional values coincide.)

In the definition of a mapping $f: A \to B$, we required that every element of A be the first component of at least one ordered pair in f and also that every element of A be the first component of at most one ordered pair in f. There was no analogous requirement for the elements of B. If we impose similar conditions on the elements of B, we get special types of functions, which we now introduce.

3.4 Definition: Let $f: A \to B$ be a mapping. If every element of B is the second component of at least one ordered pair in f, then f is called a mapping from A *onto* B.

The reader must be careful about the usage of the prepositions "into" and "onto", for they are used with different meanings. That f is a function from A onto B means that every element of B is the image of some element of A. For an arbitrary mapping $f: A \to B$, an element of B has perhaps no preimage at all, but if f is a mapping from A onto B, then each element of B has at least one preimage in A.

The range should be specified whenever the term "onto" is used. A function is not "onto" by itself, it is only onto a specific set. We shall frequently treat the word "onto" as an adjective, but it will be always clear from the context which range set is meant.

3.5 Examples: (a) The mapping $f: \mathbb{R} \to \mathbb{R}$, given by $f(x) = x^2$ for all $x \in \mathbb{R}$, is not onto, since $-1 \in \mathbb{R}$, for instance, has no preimage under f.

(b) Let $\mathbb{R}^+$ denote the set of all positive real numbers. Then the mapping $f: \mathbb{R}^+ \to \mathbb{R}^+$, given by $f(x) = x^2$ for all $x \in \mathbb{R}$, is onto.

(c) The mapping $g: \{1,2,3,4,5\} \to \{a,b,c\}$, given by

$$1g = a,\ 2g = a,\ 3g = a,\ 4g = b,\ 5g = c$$

is onto.

(d) Let A be any nonempty set. Then $\iota_A : A \to A$ is onto, for any $a \in A$ has a preimage a in A under ι_A since $a\iota_A = a$.

3.6 Definition: Let $f: A \to B$ be a mapping. If every element of B is the second component of at most one ordered pair in f, then f is called a *one-to-one* mapping from A into B.

A function $f: A \to B$ is therefore one-to-one if an arbitrary element of B has either no preimage in A or exactly one preimage: any two preimages of $b \in B$ (if b has a preimage at all) must be equal. So the necessary and sufficient condition for a mapping $f: A \to B$ to be one-to-one is

$$af = b \text{ and } a_1 f = b \implies a = a_1 \qquad (a, a_1 \in A, b \in B)$$

or, more shortly

$$af = a_1 f \implies a = a_1 \qquad (a, a_1 \in A),$$

whose contrapositive reads

$$a \neq a_1 \implies af \neq a_1 f \qquad (a, a_1 \in A).$$

A one-to-one mapping is a mapping by which different elements in the domain are matched with different elements in the range. Being a one-to-one function is the negation of being a "many-to-one" function, by which many elements in the domain are matched with one and the same element in the range.

3.7 Examples: (a) $\{(x,y): x^2 = y\} \subseteq \mathbb{R} \times \mathbb{R}$ is not a one-to-one function from $\mathbb{R}$ into $\mathbb{R}$, for two distinct elements x and $-x$ (if $x \neq 0$) have the same image.

(b) Let $\mathbb{R}^+$ denote the set of all positive real numbers. Then the mapping $\{(x,y): x^2 = y\} \subseteq \mathbb{R}^+ \times \mathbb{R}^+$ is a one-to-one function from $\mathbb{R}^+$ into $\mathbb{R}^+$.

(c) The mapping $g: \{1,2,3\} \to \{a,b,c,d\}$, given by

$$1g = b,\ 2g = d,\ 3g = a,$$

is one-to-one.

(d) Let A be a nonempty set. Then $\iota_A: A \to A$ is one-to-one, for if $a\iota_A = b\iota_A$, then $a = b$ from the definition of ι_A.

Suppose we have two functions $f: A \to B$ and $g: B \to C$. For any $a \in A$, we find $af = b \in B$ and then apply g to this element $af = b$ of B. We get an element $c = bg$ of C. In this way, the element a of A is assigned to an element c of C. Here $af = b$ is uniquely determined by f (since f is a mapping) and $bg = c$ is uniquely determined by g (since g is a mapping). So c is uniquely determined: we have a a mapping from A into C.

3.8 Definition: Let $f: A \to B$ and $g: B \to C$ be two functions. Then

$$h = \{(a,(af)g) \in A \times C : a \in A\} \subseteq A \times C,$$

which is a function from A into C, is called the *composition of f with g*, or the *product of f by g*.

We write $h = f \circ g$ or more simply $h = fg$. Thus $a(fg)$ is defined as $(af)g$.

In order to compose two functions f and g, we must make sure that the range of the first function f is a subset of the domain of the second function g. Otherwise, their composition is not defined. Note the order of the functions f and g. We apply f first, then g; and we write first f, then g in the composition notation fg. One of the advantages of writing the functions on the right becomes evident here. If we had written the functions on the left, then fg would have meant: first apply g, then f [as in the calculus, where $(f \circ g)(x) = f(g(x))$] and we would have been reading backwards. Notice also that the domain of fg is the domain of f.

3.9 Examples: (a) Let $f: A \to B$ be a mapping. Then it is easily seen that $f\iota_B = f$ and $\iota_A f = f$. Indeed, the domains of $f\iota_B, f, \iota_A f$ are all equal to A and

$$a(f\iota_B) = (af)\iota_B = af \quad \text{and} \quad a(\iota_A f) = (a\iota_A)f = af$$

for all $a \in A$. In particular, if $g: A \to A$ is a mapping, then $g\iota_A = g = \iota_A g$.

(b) Let $f: \{1,2,3,4\} \to \{a,b,c,d\}$ and $g: \{a,b,c,d\} \to \{5,x,U,\xi,\eta\}$ be given by

$$\begin{array}{ll} 1f = a & ag = U \\ 2f = c & bg = x \\ 3f = d & cg = \eta \\ 4f = b & dg = 5 \end{array}$$

Then we have

$$\begin{aligned} 1(fg) &= (1f)g &&= ag = U \\ 2(fg) &= (2f)g &&= cg = \eta \\ 3(fg) &= (3f)g &&= dg = 5 \\ 4(fg) &= (4f)g &&= bg = x. \end{aligned}$$

(c) Given $f: \{1,2,3\} \to \{a,b\}$ and $g: \{a,b\} \to \{x,y,z\}$, where

$$f: \begin{aligned} 1 &\to a \\ 2 &\to a \\ 3 &\to b \end{aligned} \quad \text{and} \quad g: \begin{aligned} a &\to y \\ b &\to z \\ c &\to z, \end{aligned}$$

we have

$$fg: \begin{aligned} 1 &\to y \\ 2 &\to y \\ 3 &\to z. \end{aligned}$$

Notice that gf is not defined.

(d) Given $f: \mathbb{R} \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$, we have
$x \to \sin x$ $\quad$ $x \to x^2$

$$x(fg) = (xf)g = (\sin x)g = (\sin x)^2 = \sin^2 x,$$

$$x(gf) = (xg)f = (x^2)f = \sin(x^2).$$

(e) Given $f: \mathbb{R} \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$, we have
$x \to x^2-1$ $\quad$ $x \to x^2+1$

$$x(fg) = (xf)g = (x^2-1)g = (x^2-1)^2 + 1 = x^4 - 2x^2 + 2$$

$$x(gf) = (xg)f = (x^2+1)f = (x^2+1)^2 - 1 = x^4 + 2x^2.$$

Given two functions $f: A \to B$ and $g: B \to C$, we might be tempted to ask whether $fg = gf$. Example 3.9(b) and Example 3.9(c) tell us that this question is meaningless, for, although fg is defined in these examples, gf is not even defined, let alone is equal to fg. Example 3.9(d) and Example 3.9(e) show that the two functions fg and gf, even if they both exist, are not necessarily equal. We have $fg \neq gf$ in general: the composition of mappings is not commutative.

However, it is associative.

3.10 Theorem: *Let $f: A \to B$, $g: B \to C$, $h: C \to D$ be three functions. Then $(fg)h = f(gh)$.*

Proof: We must prove that the domains of $(fg)h$ and $f(gh)$ are equal and that an arbitrary element in the common domain is assigned to the same element by $(fg)h$ and by $f(gh)$.

The domain of $(fg)h$ is the domain of fg, which is the domain of f, which is A. The domain of $f(gh)$ is the domain of f, which is A. So the domains of $(fg)h$ and $f(gh)$ coincide.

Now let a be an arbitrary element of A. Then

$$\begin{aligned} a((fg)h) &= (a(fg))h && \text{(by the definition of } (fg)h\text{; forget that } fg \text{ is a composition itself)} \\ &= ((af)g)h && \text{(recall now that } fg \text{ is a composition, applied to an element } a) \\ &= (af)(gh) && \text{(definition of } gh\text{, applied to an element } af) \\ &= a(f(gh)) && \text{(definition of } f(gh)), \end{aligned}$$

which yields $(fg)h = f(gh)$. □

Onto mappings and one-to-one mappings behave very nicely when they are composed.

3.11 Theorem: *Let $f: A \to B$, $g: B \to C$ be two functions and let $fg: A \to C$ be their composition.*
(1) *If f is onto and g is onto, then fg is onto.*
(2) *If f is one-to-one and g is one-to-one, then fg is one-to-one.*

Proof: (1) Suppose f and g are onto. For any $c \in C$, we must find a preimage of c under fg. The only thing we know about C is that C is the range of g. Now g is onto, so c has a preimage in B under g. Let $b \in B$ be such that $bg = c$. Since $b \in B$ and B is the range of f, and f is onto, b has a preimage $a \in A$ under f, so that $af = b$. Then we get $a(fg) = (af)g = bg = c$. So a is a preimage of c under fg. This proves that fg is onto. (Summary: a preimage of a preimage is a preimage that works.)

(2) Now suppose f and g are one-to-one. We must prove $a = a_1$ whenever $a(fg) = a_1(fg)$, for all $a, a_1 \in A$. Indeed, if

$$a(fg) = a_1(fg),$$

then $(af)g = (a_1f)g$,

$af = a_1f$ (since g is one-to-one),

$a = a_1$ (since f is one-to-one).

This proves that fg is one-to-one. □

The converse of Theorem 3.11 is wrong. If $f: A \to B$ and $g: B \to C$ are two functions and if $fg: A \to C$ is onto, it does not always follow that both f and g are onto. Also, if $f: A \to B$ and $g: B \to C$ are two functions and if $fg: A \to C$ is one-to-one, it does not always follow that both f and g are one-to-one. This can be read off from the functions displayed below.

Here fg is onto, but f is not onto; and f_1g_1 is one-to-one, but g_1 is not one-to-one.

However, we have a partial result in this direction. Observe that g is onto and f_1 is one-to-one in these examples. This is not a coincidence.

3.12 Lemma: *Let $f: A \to B$, $g: B \to C$ be two functions and let $fg: A \to C$ be their composition.*
(1) *If fg is onto, then g is onto.*
(2) *If fg is one-to-one, then f is one-to-one.*

Proof: (1) Assume fg is onto. For any $c \in C$, we must find a preimage of c in B under g. Now any $c \in C$ has a preimage in A under fg. Let $c = a(fg)$, where $a \in A$. Then $c = (af)g$. So $af \in B$ is a preimage of c in B under g. This proves that g is onto. (Summary: the image of a preimage is a preimage that works.)

(2) Assume fg is one-to-one. We wish to prove that f is one-to-one. Suppose that $af = a_1f$, where $a, a_1 \in A$. Applying g to both sides of this equation, we get $(af)g = (a_1f)g$, therefore $a(fg) = a_1(fg)$. Since fg is one-to-one by hypothesis, we get $a = a_1$. This proves that $af = a_1f$ implies $a = a_1$. Thus f is one-to-one. □

In view of its importance, we record the most important corollary of Theorem 3.11 as a separate theorem.

3.13 Theorem: *Let $f: A \to B$, $g: B \to C$ be one-to-one and onto. Then the composition $fg: A \to C$ is one-to-one and onto.* □

Assume we have a mapping $f: A \to B$. We want to define a new mapping $g: B \to A$ by inverting the order of the components of the ordered pairs in f. In other words, we want to define g by putting $(b,a) \in g$ if and only if $(a,b) \in f$. This g is a relation from B into A. The question arises: when is g in fact a mapping from B into A?

The necessary and sufficient condition for g to be a mapping is that each element of B be the first component of at least one and at most one ordered pair in g. By the definition of g, this is equivalent to the condition that each element of B be the second component of at least one ordered pair in f (i.e., f be onto) and also of at most one ordered pair in f (i.e., f be one-to-one). Let us observe that the mapping g is then uniquely determined by

$$bg = a \text{ if and only if } af = b.$$

We proved the

3.14 Theorem: *Let $f: A \to B$ be a mapping. The following assertions are equivalent.*

(i) *f is one-to-one and onto.*

(ii) *There is a unique mapping $g: B \to A$ such that*

$$bg = a \text{ if and only if } af = b \quad (a \in A, b \in B)$$ □

3.15 Definition: The mapping g of Theorem 3.14 is called the *inverse mapping of f*, or simply the *inverse of f*. It is denoted by f^{-1}.

3.16 Theorem: *Let $f: A \to B$ be one-to-one and onto, and let $f^{-1}: B \to A$ be its inverse. Then $ff^{-1} = \iota_A$ and $f^{-1}f = \iota_B$.*

Proof: We must show that the domains and functional values coincide. The domain of ff^{-1} is the domain of f, which is A, and A is the domain of ι_A. Further, for any $a \in A$, we have $a(ff^{-1}) = (af)f^{-1} = a = a\,\iota_A$ by the definition of f^{-1}. This proves $ff^{-1} = \iota_A$.

The domain of $f^{-1}f$ is the domain of f^{-1}, which is B, and B is the domain of ι_B. Further, for any $b \in B$, we have

$b(f^{-1}f) = (bf^{-1})f = af$ (where a is the unique element of A with $af = b$)
$= b = b\iota_B$.

This proves $f^{-1}f = \iota_B$. □

3.17 Theorem: (1) *Let $f: A \to B$ be one-to-one and onto. Then $f^{-1}: B \to A$ is one-to-one and onto.*

(2) *Let $f: A \to B$ be a mapping. If there is a mapping $g: B \to A$ such that $fg = \iota_A$ and $gf = \iota_B$, then f is one-to-one and onto (and therefore g is the the inverse of f).*

Proof: (1) We have $f^{-1}f = \iota_B$ by Theorem 3.16. Since ι_B is one-to-one (Example 3.7(d)), f^{-1} is one-to-one by Lemma 3.12(2). Also, we have $ff^{-1} = \iota_A$ by Theorem 3.16. Since ι_A is onto (Example 3.5(d)), f^{-1} is onto by Lemma 3.12(1).

(2) We use the same reasoning. $fg = \iota_A$ is one-to-one, so f is one-to-one, and $gf = \iota_B$ is onto, so f is onto. □

A mapping $f: A \to B$ is said to be a *one-to-one correspondence between A and B* in case f is one-to-one and onto. If f is a one-to-one correspondence between A and B, then f^{-1} is a one-to-one correspondence between B and A by Theorem 3.17(1).

*

* *

We now introduce binary operations. They constitute a generalization of the four elementary operations addition, subtraction, multiplication and division that everybody learns in the primary school. Consider addition, for example. Given any two numbers a and b, their sum is a uniquely determined number. This is the core of the operation concept: given two objects a and b, associate with them a unique object of the same kind. More precisely, we have the

3.18 Definition: Let S be a nonempty set. A *binary operation on* S is a mapping from $S \times S$ into S.

The important thing about a binary operation λ is that it is defined for all ordered pairs $(a,b) \in S \times S$ and that the result of the operation, $(a,b)\lambda$, is an element of S.

Although a binary operation λ is a mapping, we will not employ the functional notation $(a,b)\lambda$. As in the case of the elementary operations, we write a sign like "+", "−", "$\circ$", "$\oplus$", "$\otimes$" between the elements a and b to denote the image of (a,b) under λ. So the image of (a,b) will be denoted by $a + b$, $a - b$, $a \circ b$, $a \oplus b$, $a \otimes b$ or by a similar symbol.

3.19 Examples: (a) The elementary operations addition, subtraction, multiplication are binary operations on $\mathbb{R}$. Subtraction is not a binary operation on $\mathbb{N}$, since $1 - 2$, for instance, is not an element of $\mathbb{N}$ (although 1 and 2 are).

(b) Let M be a set and let S be the set of all subsets of M. Taking union and taking intersection are binary operations on S. The usual notation "$A \cup B$", "$A \cap B$" conforms to the remarks above.

(c) Let F be the set of all functions from a set A into A. The usual composition of functions is a binary operation on F.

(d) Let us write $x \circ y = x + y^2$ and $x \Delta y = x^2+x+1$ for real numbers x,y. Then $\circ$ and Δ are binary operations on $\mathbb{R}$. Here y does not enter into $x \Delta y$ in any way, but this does not preclude Δ from being a binary operation.

(e) Let V be the set of all vectors in the three space $\mathbb{R}^3$. Taking dot product of two vectors is not a binary operation on V, since the result is

a scalar (real number), not a vector. On the other hand, taking cross product is a binary operation on V, since the result is a uniquely determined vector in V.

(f) For any natural numbers m,n, let $m \cdot n$ denote their (positive) greatest common divisor. Then $\cdot$ is a binary operation on $\mathbb{N}$.

(g) Let S be the set of all students in a classroom. For any students a,b in S, let $a \cdot b$ be that student who sits in front of a. Then $\cdot$ is not a binary operation on S, for $a \cdot b$ is not defined if a happens to sit in the foremost row. Remember that a binary operation on S has to be defined for all pairs in $S \times S$.

(h) For any ordered pairs (a,b), (c,d) of real numbers, we put

$$(a,b) + (c,d) = (a + c, b + d),$$
$$(a,b).(c,d) = (ac - bd, ad + bc).$$

Then $+$ and $.$ are binary operations on $\mathbb{R} \times \mathbb{R}$. Notice that one and the same symbol "+" stands for two different binary operations, one on $\mathbb{R}$, and one on $\mathbb{R} \times \mathbb{R}$.

Exercises

1. Let $f: A \to B$ be a mapping. Prove that f is one-to-one if and only if there is a mapping $g: B \to A$ such that $fg = \iota_A$; prove that f is onto if and only if there is a mapping $h: B \to A$ such that $hf = \iota_B$.

2. Let $f: A \to B$ be a mapping. For any subset A_1 of A, we put

$$f(A_1) = \{f(a) \in B: a \in A_1\}$$

and for any subset B_1 of B, we put

$$f^{\leftarrow}(B_1) = \{a \in A : f(a) \in B_1\}.$$

($f(A_1)$ is called the *image of* A_1, and $f^{\leftarrow}(B_1)$ is called the *preimage of* B_1. Most people refer to $f(A)$ as the range of f. Here we wrote the functions on the left.) Prove that

$$f(A_1 \cap A_2) \subseteq f(A_1) \cap f(A_2), \qquad f(A_1 \cup A_2) = f(A_1) \cup f(A_2)$$
$$f^{\leftarrow}(B_1 \cap B_2) = f^{\leftarrow}(B_1) \cap f^{\leftarrow}(B_2), \qquad f^{\leftarrow}(B_1 \cup B_2) = f^{\leftarrow}(B_1) \cup f^{\leftarrow}(B_2)$$
$$A_1 \subseteq f^{\leftarrow}(f(A_1)), \qquad f(f^{\leftarrow}(B_1)) \subseteq B_1$$

for any subsets A_1, A_2 of A and for any subsets B_1, B_2 of B.

3. Keep the notation of Ex. 2. Prove that f is one-to-one if and only if

$f(A_1 \cap A_2) = f(A_1) \cap f(A_2)$ for any subsets A_1, A_2 of A.

4. Keep the notation of Ex. 2. Assume that f is one-to-one and onto, and let $f^{-1}: B \to A$ be its inverse. Show that

$$f^{\leftarrow}(B_1) = f^{-1}(B_1) \qquad \text{and} \qquad (f^{-1})^{\leftarrow}(A_1) = f(A_1)$$

for any subsets B_1 and A_1 of B and A, respectively.

§4
Mathematical Induction

Examine the propositions

$$2^n \geqslant n \qquad \text{for all } n \in \mathbb{N},$$

$$1+2+\cdots+n = \frac{n(n+1)}{2} \qquad \text{for all } n \in \mathbb{N},$$

$$n^{n+1} \geqslant (n+1)^n \qquad \text{for all } n \in \mathbb{N}.$$

How do we prove them? They are statements involving a variable n running through the infinite set $\mathbb{N}$. Strictly speaking, each one of these propositions above is a collection of infinitely many propositions. We can verify them for a finite number of cases where n assumes some specific values. Thus we might verify $2^n \geqslant n$ for $n = 1,2,3, \ldots, 1\,000\,000$ and convince ourselves of the truth of this statement, but this is far from a proof. On the other hand, we cannot check the truth of infinitely many statements within finite time. So we must resort to some other means.

In order to prove propositions about all natural numbers, an axiom is introduced. It is the fifth Peano axiom about $\mathbb{N}$ (Giuseppe Peano (1858-1932), an Italian mathematician and logician). It is called the axiom of mathematical induction.

4.1 Axiom (of mathematical induction): If S is a subset of $\mathbb{N}$ such that

I. $1 \in S$,

II. for all $k \in \mathbb{N}$, if $k \in S$, then $k+1 \in \mathbb{N}$,

then S is the whole of $\mathbb{N}$, i.e., $S = \mathbb{N}$.

We can use this axiom to prove statements of the form 'p_n for all $n \in \mathbb{N}$' as follows. We let $S \subseteq \mathbb{N}$ be the set of all natural numbers n for which p_n is true. First we verify $1 \in S$, that is, we verify that p_1 is true. Second, we *assume* that $k \in S$ and under this hypothesis, which is called the *induction hypothesis*, we prove that p_{k+1} is true. So we show that $k \in S$ implies $k+1 \in S$. By the axiom of mathematical induction, $S = \mathbb{N}$, so the

statement p_n is true for all $n \in \mathbb{N}$. We formulate the axiom as an operational procedure.

4.2 Principle of mathematical induction: Let p_n be a statement involving a natural number n. We can prove the proposition

$$\text{for all } n \in \mathbb{N},\ p_n$$

by establishing that

I. p_1 is true,

II. for all $k \in \mathbb{N}$, if p_k is true, then p_{k+1} is true.

Proofs by the principle of mathematical induction consist of two steps. In the first step, we show that p_1 is true. In practice, this is often quite easy, but we should not neglect it. In the second step, we assume that p_k is true. This assumption is the inductive hypothesis. Using this hypothesis, we prove that p_{k+1} is true. A proof by induction will not be complete (and valid) if we carry out the first step but not the second, or if we carry out the second step but not the first.

4.3 Examples: (a) Prove that $1 + 2 + \cdots + n = \dfrac{n(n+1)}{2}$ for all $n \in \mathbb{N}$.

We use the principle of mathematical induction.

I. $1 = \dfrac{1(1+1)}{2}$, so the formula is true for $n = 1$.

II. Make the inductive hypothesis that $1 + 2 + \cdots + k = \dfrac{k(k+1)}{2}$.

We want to establish $1 + 2 + \cdots + k + (k+1) = \dfrac{(k+1)((k+1)+1)}{2}$. We have

$$\begin{aligned} 1 + 2 + \cdots + k + (k+1) &= \frac{k(k+1)}{2} + (k+1) \qquad \text{(by inductive hyp.)} \\ &= \left(\frac{k}{2} + 1\right)(k+1) \\ &= \frac{(k+1)(k+2)}{2}, \end{aligned}$$

so the formula is true for $n = k + 1$ if it is true for $n = k$. Hence

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2} \quad \text{for all } n \in \mathbb{N}.$$

(b) Prove that $2 + 2^2 + 2^3 + \cdots + 2^n = 2^{n+1} - 2$ for all $n \in \mathbb{N}$.

I. We have $2 = 2^{1+1} - 2$, which proves the assertion for $n = 1$.

II. Assume $2 + 2^2 + 2^3 + \cdots + 2^k = 2^{k+1} - 2$. Now we must prove $2 + 2^2 + 2^3 + \cdots + 2^k + 2^{k+1} = 2^{(k+1)+1} - 2$. We have

$$\begin{aligned} 2 + 2^2 + 2^3 + \cdots + 2^k + 2^{k+1} &= (2^{k+1} - 2) + 2^{k+1} \quad \text{(by inductive hyp.)} \\ &= 2(2^{k+1}) - 2 \\ &= 2^{k+2} - 2, \end{aligned}$$

so the assertion is true for $n = k + 1$ if it is true for $n = k$. Thus

$$2 + 2^2 + 2^3 + \cdots + 2^n = 2^{n+1} - 2 \text{ for all } n \in \mathbb{N}.$$

(c) Let $h > -1$ be a fixed real number. Prove that $(1 + h)^n \geqslant 1 + nh$ for all $n \in \mathbb{N}$.

I. We have $(1 + h)^1 \geqslant 1 + 1h$, so the inequality is true for $n = 1$.

II. Let us assume $(1 + h)^k \geqslant 1 + kh$. We want to prove that $(1 + h)^{k+1} \geqslant 1 + (k + 1)h$. We have

$$\begin{aligned} (1 + h)^{k+1} &= (1 + h)^k (1 + h) \\ &\geqslant (1 + kh)(1 + h) \quad \text{(by inductive hyp. and } 1 + h > 0) \\ &= 1 + h + kh + kh^2 \\ &\geqslant 1 + h + kh + 0 \\ &= 1 + (k + 1)h, \end{aligned}$$

so the inequality is true for $n = k + 1$ if it is true for $n = k$. By the principle of mathematical induction,

$$(1 + h)^n \geqslant 1 + nh \text{ for all } n \in \mathbb{N}.$$

Sometimes it is convenient to use the principle of mathematical induction in a slightly different form. We assume (not only q_k, but rather) each one of $q_1, q_2, q_3, \ldots, q_k$ is true and then conclude that q_{k+1} is true. This establishes the truth of q_n for all $n \in \mathbb{N}$, as the following lemma shows.

4.4 Lemma: *Let q_n be a statement involving a natural number n. Assume that*

i. *q_1 is true,*

ii. *for all $k \in \mathbb{N}$, if $q_1, q_2, q_3, \ldots, q_k$ are true, then q_{k+1} is true.*

Then q_n is true for all $n \in \mathbb{N}$.

Proof: We prove the lemma by the principle of mathematical induction. We put

$p_1 = q_1$

$p_k = q_1$ and q_2 and ... and q_k (for all $k \in \mathbb{N}, k \geqslant 2$).

Now induction.

I. p_1 is true (by the hypothesis i.)

II. Make the inductive hypothesis that p_k is true. Then

q_1 and q_2 and ... and q_k is true (definition of p_k)

$q_1, q_2, \ldots, q_k$ are all true (truth value of conjunction)

q_{k+1} is true (by the hypothesis ii.)

$q_1, q_2, \ldots, q_k, q_{k+1}$ are all true

q_1 and q_2 and ... and q_k and q_{k+1} is true

p_{k+1} is true.

Hence, for all $k \in \mathbb{N}$, if p_k is true, then p_{k+1} is true. By the principle of mathematical induction, p_n is true for all $n \in \mathbb{N}$. So

q_1 and q_2 and ... and q_n is true for all $n \in \mathbb{N}$.

In particular, q_n is true for all $n \in \mathbb{N}$. This completes the proof. □

We can now formulate a new form of the principle of mathematical induction. This form will be used many times in the sequel.

4.5 Principle of mathematical induction: Let q_n be a statement involving a natural number n. We can prove the proposition

for all $n \in \mathbb{N}$, q_n

by establishing that

i. q_1 is true,

ii. for all $k \in \mathbb{N}$, if $q_1, q_2, \ldots, q_k$ are true, then q_{k+1} is true.

The statement '$2^n \geqslant n^2$' is not true for all natural numbers n, but true for all natural numbers $n \geqslant 5$. The principle of mathematical induction can be used to prove this and similar propositions. Let a be a fixed integer (positive, negative or zero) and let p_n be a statement involving an integer $n \geqslant a$. We prove the truth of p_n for all $n \geqslant a$ by showing that

1. p_a is true

2. for all $k \geqslant a$, if p_k is true, then p_{k+1} is true.

This is easily seen when we put $q_n = p_{n+a-1}$ for $n \in \mathbb{N}$ and use Principle 4.2 with q_n in place of p_n. There is a similar modification of Principle 4.5.

Exercises

Prove the assertions in Ex. 1-5 for all $n \in \mathbb{N}$ by the principle of mathematical induction.

1. $1 + 3 + \cdots + (2n + 1) = n^2$.

2. $1 + 4 + 7 + \cdots + (3n - 2) = \dfrac{n(3n - 1)}{2}$.

3. $1^2 + 2^2 + \cdots + n^2 = \dfrac{n(n + 1)(2n + 1)}{6}$.

4. $1^3 + 2^3 + \cdots + n^3 = \dfrac{n^2(n + 1)^2}{4}$.

5. $1^4 + 2^4 + \cdots + n^4 = \dfrac{n(n + 1)(2n + 1)(3n^2 + 3n - 1)}{30}$.

6. Prove that $2^n \geqslant n^2$ for all $n \geqslant 5$, $n \in \mathbb{N}$.

7. Prove that $n^3 + 3n^2 + 2n + 1 \geqslant 0$ for all $n \geqslant -2$, $n \in \mathbb{Z}$.

8. Prove that, for any $n \in \mathbb{N}$ and for any positive real numbers $a_1, a_2, \ldots, a_{2^n}$,

$$\sqrt[2^n]{a_1 a_2 \cdots a_{2^n}} \leqslant \frac{a_1 + a_2 + \cdots + a_{2^n}}{2^n}.$$

9. Prove that, for any $n \in \mathbb{N}$ and for any positive real numbers $a_1, a_2, \ldots, a_n$,

$$\sqrt[n]{a_1 a_2 \cdots a_n} \leqslant \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

(Hint: if $m \neq 2^n$, then choose n so that $2^{n-1} < m < 2^n$. Put $b = (a_1 + a_2 + \cdots + a_m)/m$. Then use Ex. 8 with $a_1, a_2, \ldots, a_m, a_{m+1}, \ldots, a_{2^n}$, where $a_{m+1} = \ldots = a_{2^n} = b$.)

§5
Divisibility

In this paragraph, we remind the reader of certain properties of integers concerning divisibility. First we recall the definition.

5.1 Definition: Let $a, b \in \mathbb{Z}$. If $a \neq 0$ and if there is a $c \in \mathbb{Z}$ such that $ac = b$, then a is called a *divisor* or a *factor* of b, and b is said to be divisible by a. We also say a *divides* b.

We write $a|b$ to express that a divides b. Whenever we employ the notation $a|b$, it will be assumed of course $a \neq 0$. We shall write $a \nmid b$ when $a \neq 0$ and b is not divisible by a. Thus we have $3|6$, $3|9$, $2|8$, $5 \nmid 9$, $5|10$, $3 \nmid 7$, $4|8$, $-3|6$, $2|-4$, $-2|-4$. The notations $0|b$ and $0 \nmid b$ are meaningless: not true or false, simply undefined.

Some basic properties of divisibility are collected below.

5.2 Lemma: *Let* $a,b,c,m,n,m_1,m_2, \ldots m_s,b_1,b_2, \ldots ,b_s$, *be integers.*

(1) *If* $a|b$, *then* $a|-b$, $-a|-b$, $-a|b$.
(2) *If* $a|b$ *and* $b|c$, *then* $a|c$.
(3) *If* $a|b$ *and* $c \neq 0$, *then* $ac|bc$.
(4) *If* $ac|bc$, *then* $a|b$.
(5) *If* $a|b$ *and* $a|c$, *then* $a|b + c$.
(6) *If* $a|b$ *and* $a|c$, *then* $a|b - c$.
(7) *If* $a|b$ *and* $a|c$, *then* $a|mb + nc$.
(8) *If* $a|b_1, a|b_2, \ldots ,a|b_s$, *then* $a| m_1b_1 + m_2b_2 + \cdots + m_sb_s$.
(9) *If* $a \neq 0$, *then* $a|0$.
(10) $1|a$ *and* $-1|a$.
(11) *If* $a|b$ *and* $b \neq 0$, *then* $|a| \leqslant |b|$.
(12) *If* $a|b$ *and* $b|a$, *then* $|a| = |b|$.

Proof: (1) If $a|b$, then $a \neq 0$ and $ak = b$ for some $k \in \mathbb{Z}$. So $a \neq 0$ and $a(-k) = -b$; $a \neq 0$ and $(-a)k = -b$; $a \neq 0$ and $(-a)(-k) = b$; with $k, -k \in \mathbb{Z}$. Hence $a|-b$, $-a|-b$, $-a|b$.

(2) If $a|b$ and $b|c$, then $a \neq 0 \neq b$ and $ak = b$ and $bh = c$ for some $k,h \in \mathbb{Z}$. So $a(kh) = bh = c$ and, since $kh \in \mathbb{Z}$, we obtain $a|c$.

(3) If $a|b$, then $a \neq 0$ and $ak = b$ for some $k \in \mathbb{Z}$. So $(ac)k = bc$. From $a \neq 0$, $c \neq 0$, we conclude $ac \neq 0$. Hence $ac|bc$.

(4) If $ac|bc$, then $ac \neq 0$ and $(ac)k = bc$ for some $k \in \mathbb{Z}$. From $ac \neq 0$, we obtain $a \neq 0$ and $c \neq 0$. Since $c \neq 0$, we have $ak = b$. Since $a \neq 0$, we can write $a|b$.

(5) If $a|b$ and $a|c$, then $a \neq 0$, and $ak = b$, $ah = c$ for some $k,h \in \mathbb{Z}$. So $a(k+h) = b + c$. Since $k+h \in \mathbb{Z}$ and $a \neq 0$, we have $a|b+c$.

(6) This can be proved in the same way as (5). We might also observe that $a|-c$ if $a|c$ by (1), hence $a|b+(-c)$ by (5), so $a|b-c$.

(7) If $a|b$ and $a|c$, then $a \neq 0$, and $ak = b$, $ah = c$ for some $k,h \in \mathbb{Z}$. So $a(km+hn) = ak.m + ah.n = bm + cn = mb + nc$. Since $km + hn \in \mathbb{Z}$ and $a \neq 0$, we have $a|mb+nc$.

(8) This can be proved by a simple application of the principle of mahematical induction.

(9) $a0 = 0$ for any $a \in \mathbb{Z}$. If $a \neq 0$, we can write $a|0$.

(10) $a = 1.a = (-1)(-a)$. Hence $1|a$ and $-1|a$.

(11) If $a|b$, then $a \neq 0$ and $ak = b$ for some $k \in \mathbb{Z}$. So $|a||k| = |b|$ Since $b \neq 0$, we have $|k| \geqslant 1$. Thus $|b| = |a||k| \geqslant |a|$.

(12) If $a|b$ and $b|a$, then $a \neq 0$ and $b \neq 0$, so we may apply (11) to get $|a| \leqslant |b|$ and $|b| \leqslant |a|$. Thus $|a| = |b|$. □

5.3 Theorem (Division algorithm): *Let $a,b \in \mathbb{Z}$, $b > 0$. Then there are unique integers $q,r \in \mathbb{Z}$ such that*

$$a = qb + r, \quad 0 \leqslant r < b.$$

(The integer q is called the *quotient*, and r is called the *remainder* obtained when a is divided by b.)

Proof: There are two claims in this theorem: (1) that there are integers q,r, with the stated properties and (2) that these are unique, that is, the pair of integers q,r is the only one which has the stated properties. The proof of this theorem will accordingly consist of two parts. In the first part, we prove the existence of q,r, in the second part, their uniqueness.

Existence. Consider the set $T = \{a - ub: u \in \mathbb{Z}\} \subseteq \mathbb{Z}$. This set T contains nonnegative integers (for example, $a - (-|a|)b$ is nonnegative). We choose the smallest nonnegative integer in T. Let it be called r. Thus $r \geqslant 0$ and, by the very definition of T, we infer $r = a - qb$ for some $q \in \mathbb{Z}$. We claim $r < b$. If we had $r \geqslant b$, then, since $b > 0$, we would get

$$r > r - b = a - (q + 1)b \geqslant 0$$

and $r - b$ would be a nonnegative integer in T, smaller than the smallest nonnegative integer in T, which is absurd. So $r \geqslant b$ is impossible and $r < b$. Hence there are integers q,r such that

$$a = qb + r, \qquad 0 \leqslant r < b.$$

Uniqueness. Let $a = qb + r$, $0 \leqslant r < b$, and $a = q_1 b + r_1$, $0 \leqslant r_1 < b$, where q,r,q_1,r_1 are integers. We wish to prove $q_1 = q$ and $r_1 = r$. It suffices to prove $q_1 = q$, for then we would get $r_1 = a - q_1 b = a - qb = r$ also. Suppose, by way of contradiction, that $q \neq q_1$. Then there are two possibilities: $q > q_1$ or $q < q_1$. Interchanging q,r with q_1,r_1 if necessary, we may assume $q > q_1$ without loss of generality (make sure that you understand this reasoning). From $q > q_1$, we get $q - q_1 \geqslant 1$, hence

$$r_1 = r_1 - 0 \geqslant r_1 - r = (a - q_1 b) - (a - qb) = (q - q_1)b \geqslant 1.b = b,$$

a contradiction. So $q_1 = q$ and $r_1 = r$. □

This theorem formalizes what everybody learns at primary school: when we divide a by b, we get a quotient, and a remainder smaller than b. At primary school, one learns it in the case a is positive, but here a can be negative. Also, division is carried out by successive subtractions

$$\begin{array}{ll} a & \mid b \\ \cdots & \quad q \\ \cdots & \\ r & \end{array}$$

We subtract b from a until we get a number r smaller than b. This is exactly what happens when we perform division, and this is essentially the proof of Theorem 5.3.

Given any two integers a,b, an integer d is said to be a *common divisor of a and b* if $d|a$ and $d|b$. Using the division algorithm, we can show that any two integers have a greatest common divisor, provided only that not both of them are equal to zero.

5.4 Theorem: *Let $a,b \in \mathbb{Z}$, not both zero. Then there is a unique integer d such that*

(i) $d|a$ *and* $d|b$,
(ii) *for all* $d_1 \in \mathbb{Z}$, *if* $d_1|a$ *and* $d_1|b$, *then* $d_1|d$,
(iii) $d > 0$.

Proof: The proof will be similar to the proof of Theorem 5.3. We consider the set $U = \{ax - by \in \mathbb{Z}: x,y \in \mathbb{Z}\}$. Now U contains positive integers. (For example, $a(\mp 1) - b0$ is positive when $a \neq 0$ and the sign is chosen suitably. When $a = 0$, $a1 - b(\mp 1) = \pm b$ is positive, provided we choose the sign appropriately, since $b \neq 0$ when $a = 0$ by hypothesis.) We choose the smallest positive integer in U. Let it be called d. So $d > 0$ and d satisfies (iii). Moreover, if $d_1|a$ and $d_1|b$, then $d_1|ax - by$ for any $x,y \in \mathbb{Z}$ by Lemma 5.2(7), so d_1 divides every element of U. In particular, $d_1|d$. Thus (ii) is satisfied. It remains to prove (i).

By the very definition of U, we have $d = ax_0 - by_0$ for some $x_0,y_0 \in \mathbb{Z}$. We want to prove $d|a$ and $d|b$. Using the division algorithm, we write $a = qd + r$, where q and r are integers and $0 \leqslant r < d$. Then

$$\begin{aligned} a &= q(ax_0 - by_0) + r, \\ r &= a - q(ax_0 - by_0) \\ &= a(1 - qx_0) - b(-y_0), \text{ with } 1 - qx_0, -y_0 \in \mathbb{Z}, \end{aligned}$$

so r is an element of U and $0 \leqslant r < d$. Since d is the smallest positive integer in U and $r < d$, we have necessarily $r = 0$. This gives $a = qd$, so $d|a$. The proof of $d|b$ is similar and will be omitted.

Now the uniqueness of d. Suppose d' satisfies the conditions (i), (ii), (iii), too. Then $d'|a$, $d'|b$ by (i), and so $d'|d$ by (ii). Also, $d|a$, $d|b$ by (i), and so

$d|d'$ by (ii). By Lemma 5.2(12), we obtain $|d| = |d'|$. From (iii), we get $d > 0$, $d' > 0$, which yields $d = d'$. Thus d is unique. □

5.5 Definition: Let $a,b \in \mathbb{Z}$, not both zero. The unique integer d in Theorem 5.4 is called the *greatest common divisor of a and b.*

The greatest common divisor of a and b will be denoted by (a,b). This notation is standard. The reader should not confuse it with an ordered pair. The greatest common divisor of a and b is a natural number, not an ordered pair.

Definition 5.5 and the proof of Theorem 5.4 enables us to write the

5.6 Theorem: *Let $a,b \in \mathbb{Z}$, not both zero. Then (a,b) is the smallest positive integer in the set $\{ax - by \in \mathbb{Z}: x,y \in \mathbb{Z}\}$.* □

Theorem 5.4 is a typical existence theorem. It tells us that the greatest common divisor (a,b) of any pair of integers a,b exists (provided a and b are not both zero), but gives no method for finding it. If a and b are small in absolute value, we might try to find the smallest positive integer in the set $\{ax - by \in \mathbb{Z}: x,y \in \mathbb{Z}\}$. This is not very satisfactory, of course. Also, it is almost impossible if a and b are rather large. We propose to give a systematic method for finding (a,b) for any pair of integers a,b, not both zero. This method will prove anew the existence of (a,b) and in addition will give us a systematic method of finding integers x,y such that $(a,b) = ax - by$. It is Proposition 2 in Euclid's *Elements*, Book VII. (in algebraic notation) and is known as the Euclidean algorithm.

We first observe that the set U in Theorem 5.6 does not change if we write $-a$ in place of a or $-b$ in place of b. This yields

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

for all a,b, not both zero. Hence $(a,b) = (|a|,|b|)$ and, when we want to find (a,b), we may assume $a \geqslant 0$, $b \geqslant 0$ (the case $a = 0$, $b = 0$ is excluded) without loss of generality. Moreover, the set U in Theorem 5.6 remains unaltered if we interchange a and b. Thus

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

Therefore, when we want to find (a,b), we may assume $a \geq b$ without loss of generality. (Instead of appealing to Theorem 5.6, we could use the definition to obtain $(a,b) = (-a,b) = (-a,-b) = (a,-b) = (b,a)$.)

The greatest common divisor of $a \in \mathbb{Z}$ $(a \neq 0)$ and 0 is easily found. We have $(a,0) = |a|$, as follows from Theorem 5.6 or immediately from Theorem 5.4.

Suppose now $a \geq b > 0$ and we want to find (a,b). We divide a by b and get

$$a = q_1 b + r_1, \quad 0 \leq r_1 < b.$$

Here r_1 may be zero. If $r_1 \neq 0$, we divide b by r_1 and get

$$b = q_2 r_1 + r_2, \quad 0 \leq r_2 < r_1.$$

Here r_2 may be zero. If $r_2 \neq 0$, we divide r_1 by r_2 and get

$$r_1 = q_3 r_2 + r_3, \quad 0 \leq r_3 < r_2.$$

We proceed in this way. We have $b > r_1 > r_2 > r_3 > \cdots$. Since the r_j's are nonnegative integers and b is a finite positive integer, this process cannot go on indefinitely. Sooner or later, we will meet a division in which the remainder is zero, say at the $(k+1)$-st step $(k \geq 0)$:

$$r_{k-2} = q_k r_{k-1} + r_k, \quad 0 \leq r_k < r_{k-1}.$$
$$r_{k-1} = q_{k+1} r_k + r_{k+1}, \quad 0 = r_{k+1}.$$

We claim that r_k, the last nonzero remainder, is the greatest common divisor of a and b, and that it can be written in the form $ax - by$, where x,y are integers.

5.7 Theorem: *Let $a \geq b > 0$ be integers and let*

$$\begin{aligned}
a &= q_1 b + r_1, && 0 \leq r_1 < b,\\
b &= q_2 r_1 + r_2, && 0 \leq r_2 < r_1,\\
r_1 &= q_3 r_2 + r_3, && 0 \leq r_3 < r_2,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
r_{i-1} &= q_{i+1} r_i + r_{i+1}, && 0 \leq r_{i+1} < r_i,\\
&\cdots\cdots\cdots\cdots\cdots\cdots\\
r_{k-2} &= q_k r_{k-1} + r_k, && 0 \leq r_k < r_{k-1},\\
r_{k-1} &= q_{k+1} r_k
\end{aligned}$$

be the equations we obtain when we use the division algorithm (Theorem 5.3) *succesively until we get a nonzero remainder.* (This chain of equations is known as the *Euclidean algorithm.*) *Then the last nonzero remainder* r_k *is the greatest common divisor of* a *and* b. *Moreover,* r_k *can be written in the form* $r_{i-1}x - r_i y$; $x,y \in \mathbb{Z}$ *for* $i = k-1, k-2, \ldots, 2, 1, 0$ (we put $r_0 = b$, $r_{-1} = a$). *In particular, there are integers* x_0, y_0 *such that* $(a,b) = ax_0 - by_0$, *and eliminating* $r_1, r_2, \ldots, r_{k-1}$ *from the equations above gives a systematic way of finding the integers* x_0, y_0.

Proof: We must show that r_k satisfies the conditions (i),(ii),(iii) of Theorem 5.4. We know $r_k > 0$ from the k-th equation in the Euclidean algorithm, so (iii) of Theorem 5.4 is satisfied.

We prove (i) of Theorem 5.4, namely that $r_k|a$ and $r_k|b$. We start from the last equation in the algorithm and go up through the algorithm. From the $(k+1)$-st equation, we get $r_k|r_{k-1}$. Using Lemma 5.2, we get $r_k|r_{k-2}$ from the k-th equation. So $r_k|r_{k-1}$ and $r_k|r_{k-2}$. From the $(k-1)$-st equation, we get $r_k|r_{k-3}$, so $r_k|r_{k-2}$ and $r_k|r_{k-3}$. In general, if we have $r_k|r_{i+1}$ and $r_k|r_i$, the $(i+1)$-st equation gives $r_k|r_{i-1}$, so we have $r_k|r_i$ and $r_k|r_{i-1}$. Going through the equations in this way, we finally get $r_k|r_0$ and $r_k|r_{-1}$, that is, we get $r_k|b$ and $r_k|a$. This proves (i) of Theorem 5.4.

Now (ii) of Theorem 5.4. Assume $e|a$ and $e|b$. We must prove $e|r_k$. We start from the first equation in the algorithm and go down through the algorithm. From the first equation, we get $e|a - q_1 b$ and $e|r_1$ by Lemma 5.2. So $e|b$ and $e|r_1$. From the second equation, we get $e|b - q_2 r_1$ and $e|r_2$. So $e|r_1$ and $e|r_2$. In general, if we have $e|r_{i-1}$ and $e|r_i$, the $(i+1)$-st equation gives $e|r_{i-1} - q_{i+1}r_i$ and $e|r_{i+1}$. So $e|r_i$, and $e|r_{i+1}$. Going through the equations in this way, we finally get $e|r_k$. This proves (ii) of Theorem 5.4.

Hence r_k is the greatest common divisor of a and b.

Finally, we show the representability of r_k in terms of r_{i-1}, r_i as described. We start from the penultimate equation in the algorithm and go up through the algorithm. From the k-th equation, we obtain $r_k = r_{k-2} - r_{k-1}q_k$, so r_k can be represented as $r_{k-2}x - r_{k-1}y$, namely with $x = 1$, $y = q_k$. Substituting $r_{k-3} - q_{k-1}r_{k-2}$ for r_{k-1} in this equation, we get

$$\begin{aligned} r_k = r_{k-2} - r_{k-1}q_k &= r_{k-2} - (r_{k-3} - q_{k-1}r_{k-2})q_k \\ &= r_{k-3}(-q_k) + r_{k-2}(1 + q_{k-1}q_k), \end{aligned}$$

so r_k can be represented as $r_{k-3}x - r_{k-2}y$, namely with $x = -q_k$, $y = -(1 + q_{k-1}q_k)$. In general, if r_k can be written in the form

$$r_i x - r_{i+1} y, \qquad x,y \in \mathbb{Z},$$

we get, using the $(i+1)$-st equation in the Euclidean algorithm,

$$\begin{aligned} r_k &= r_i x - r_{i+1} y \\ &= r_i x - (r_{i-1} - q_{i+1} r_i) y \\ &= r_{i-1}(-y) + r_i(x + q_{i+1} y), \end{aligned}$$

which shows that r_k can be written also in the form $r_{i-1}x_1 - r_i y_1$, namely with $x_1 = -y$, $y_1 = -(x + q_{i+1}y)$. Going through the equations in this way, we finally obtain

$$r_k = ax_0 - b y_0$$

for some $x_0, y_0 \in \mathbb{Z}$. This completes the proof. □

5.8 Example: To find the greatest common divisor of 14732 and 37149, and to express it in the form $14732x - 37149y$, with $x,y \in \mathbb{Z}$.

We have

$$\begin{aligned} 37149 &= 2.14732 + 7685 \\ 14732 &= 1.7685 + 7047 \\ 7685 &= 1.7047 + 638 \\ 7047 &= 11.638 + 29 \\ 638 &= 22.29 \end{aligned}$$

and the last nonzero divisor is 29. So $(14732,37149) = 29$. Also

$$\begin{aligned} 29 &= 7047 - 11.638 \\ &= 7047 - 11(7685 - 1.7047) \\ &= 12.7047 - 11.7685 \\ &= 12(14732 - 1.7685) - 11.7685 \\ &= 12.14732 - 23.7685 \\ &= 12.14732 - 23(37149 - 2.14732) \\ &= 58.14732 - 23.37149, \end{aligned}$$

so $29 = 14732x - 37149y$ with $x = 58$, $y = 23$.

5.9 Definition: Let a,b be integers, not both zero. a is said to be *relatively prime to* b if $(a,b) = 1$.

Since $(a,b) = (b,a)$, b is relatively prime to a in case a is relatively prime to b. This observation enables us to use a symmetric phrase in this case. We say a and b are relatively prime if $(a,b) = 1$.

5.10 Lemma: *Let a,b be integers, not both zero. Then a and b are relatively prime if and only if there are integers x_0,y_0 such that $ax_0 - by_0 = 1$.*

Proof: If $(a,b) = 1$, then there are integers x_0,y_0 such that $ax_0 - by_0 = 1$ by Theorem 5.6 or also by Theorem 5.7. Conversely, if there are integers x_0,y_0 with $ax_0 - by_0 = 1$, then 1 is certainly the smallest positive integer in the set $\{ax - by \in \mathbb{Z}: x,y \in \mathbb{Z}\}$, hence $(a,b) = 1$ by Theorem 5.6. □

5.11 Lemma: *Let a,b be integers, not both zero, and let $d = (a,b)$. Then a/d and b/d are relatively prime.*

Proof: a/d, b/d are integers, not both of them zero. We have $ax - by = d$ for suitable integers $x,y \in \mathbb{Z}$ by Theorem 5.7. Dividing both sides of this equation by $d > 0$, we get

$$(a/d)x - (b/d)y = 1,$$

and so $(a/d, b/d) = 1$ by Lemma 5.10 □

Using Lemma 5.10, we prove an important result that will be crucial in the proof of the fundamental theorem of arithmetic.

5.12 Theorem: *Let a, b,c be integers. If $a|bc$ and $(a,b) = 1$, then $a|c$.*

Proof: Since $(a,b) = 1$, we have $ax - by = 1$ with some $x,y \in \mathbb{Z}$. Multiplying both sides of this equation by c, we obtain $acx - bcy = c$. Now $a|acx$ and, since $a|bc$ by hypothesis, $a|bcx$; hence $a|acx - bcy$ by Lemma 5.2. So $a|c$. □

We separate $\mathbb{Z}\backslash\{0\}$ into three subsets: (1) units, (2) prime numbers, (3) composite numbers. The numbers 1 and -1 will be called *units*. The units divide every integer by Lemma 5.2(10). Any other integer a has at least four divisors: $\mp 1, \mp a$. These are called the *trivial divisors* of a. A divisor of a, which is not one of the four trivial divisors of a, is called a *proper divisor* of a. If a nonzero integer a is not a unit and has no proper divisors, then a is called a *prime* number. Thus 2, -3, 5, 7, -11 are prime numbers. A nonzero integer, which is neither a unit nor a prime number,

will be called a *composite* number. So $a \in \mathbb{Z}\backslash\{0\}$ is a composite number if and only if there is a $d \in \mathbb{Z}$ with $1 < |d| < |a|$ and $d|a$.

Prime numbers are the building blocks of integers in the following sense.

5.13 Theorem: *Any nonzero integer, which is not a unit, is either a prime number or a product of prime numbers.*

Proof: Take an integer $n \neq 0$, and assume that n is not a unit. If n is prime, there is nothing to prove. If n is composite, then $n = n_1 n_2$ for some $n_1, n_2 \in \mathbb{Z}$, $1 < |n_1| < |n|$, $1 < |n_2| < |n|$. If n_1 and n_2 are prime, we are through. Otherwise, factor n_1 and n_2 into two numbers. Keep factoring until you get down to prime numbers. Since the factors get smaller and smaller in absolute value, we will reach prime numbers at the end. This is the basic idea and we make this reasoning into a rigorous proof by induction.

We use Principle 4.5. Let q_n be the statement that $n \in \mathbb{N}$ is a prime number or a product of prime numbers. We begin induction at $n = 2$. Since 2 is a prime number, q_2 is true. q_3 is also true, for 3 is prime. q_4 is true, for $4 = 2.2$ is a product of the prime numbers 2 and 2.

Suppose now $q_2, q_3, q_4, \ldots, q_{k-1}$ are true, so that $2,3,4, \ldots, k-1$ are either prime numbers or products of prime numbers. We want to prove that k is a prime number or a product of prime numbers. If k is prime, we are done. If k is not prime, we have $k = k_1 k_2$, $1 < k_1 < k$, $1 < k_2 < k$, for some integers k_1, k_2. Since q_{k_1} and q_{k_2} are true by the induction hypothesis, each of k_1, k_2 is either a prime number or a product of prime numbers:

$$k_1 = p_1 p_2 \ldots p_r \qquad k_2 = p'_1 p'_2 \ldots p'_s$$

where $p_1, p_2, \ldots, p_r, p'_1, p'_2, \ldots, p'_s$ are prime numbers ($r = 1$ or $s = 1$ is possible, in which case $k_1 = p_1$ or $k_2 = p'_1$ are prime numbers), and so

$$k = k_1 k_2 = p_1 p_2 \ldots p_r p'_1 p'_2 \ldots p'_s,$$

is a product of prime numbers. Hence q_k is true.

This proves the theorem for positive integers. For a negative integer $-n$, where $-n$ is not a unit, we have

$$n = p_1 p_2 \cdots p_t$$

for some prime numbers $p_1, p_2, \ldots, p_t$ by what we proved above (possibly $t = 1$). Hence

$$-n = (-p_1)p_2 \cdots p_t$$

is prime or is a product of prime numbers. □

After reading the proof of this theorem, it will be clear to the reader that an abbreviation of the phrase "prime number or a product of prime numbers" will be very useful. When we speak of a product, we mean a product of two, three, or more terms. We now extend this to one factor. A single term will be called a product of one factor (or of one factors). A prime number is also a product of prime numbers with this convention. Our theorem reads now more shortly as follows.

5.13 Theorem: *Any nonzero integer, which is not a unit, is a product of prime numbers.* □

Now that we know any integer, which is not zero or a unit, can be expressed as a product of prime numbers, we ask if it can be written as a product of prime numbers in different ways. By way of example, let us begin decomposing 60 into prime numbers as in the proof of Theorem 5.13. We can begin from any decomposition of 60 into factors. For instance,

$$60 = 10.6 \qquad 60 = 15.4$$

Now we are to decompose each one of the factors 10,6,15,4 into smaller factors until we get prime numbers. Will we reach the same prime numbers if we use the two different decompositions as our starting point? We know of course that further decomposition

$$60 = (2.5)(2.3) \qquad 60 = (3.5)(2.2)$$

yields the same prime numbers 2,2,3,5 (aside from order). Nevertheless, our question should not be taken lightly. It is a very pertinent question. We remark that Theorem 5.13 says nothing in this regard. Theorem 5.13 says that, after enough factorizations, the factors will be prime. As it is, the prime numbers we obtain may very well be distinct if we start with different factorizations. Indeed, if you start with different things, why

on earth should you end with the same things? If 60 can be written as a product of factors in two different ways, as above, why should it not be written as a a product of prime factors in two different ways? The readers experience with the uniqueness of prime factors of integers should not mislead him (or her) to believe the uniqueness is obvious. It is anything but obvious.

Let us clarify what we mean by uniqueness. The two decompositions

2.5.2.3 3.5.2.2

of 60 involve the same prime numbers. Their order in the two decompositions are different, but nobody would consider these decompositions as very distinct. After all, multiplication of integers is commutative, and we can permute the factors without changing the value of the product. It would be foolish to regard two factorizations as different when they consist of the same prime numbers in different orders.

Moreover, we have $(-2)(5)(2)(-3) = 2.5.2.3$, where the numbers appearing are prime. These decompositions of 60 are not essentially distinct, of course. Given two nonzero integers a, b, we say a is *associate to* b if $a = b$ or $a = -b$. Then b is associate to a as well. Hence we may also say that a and b are associate. This means $a|b$ and $b|a$. It is clear from the definition that, whenever p and q are associate and p is prime, then q is a prime number, too. When we say uniqueness, we shall mean that the prime numbers in the decompositions of an integer are associate; we shall not mean that they are identical.

With this understanding, we will prove that any integer ($\neq 0,1,-1$) has a unique decomposition into prime numbers. We need some lemmas.

5.14 Lemma: *Let a be an integer and p be a prime number. If $p \nmid a$, then* $(a,p) = 1$.

Proof: Let $d = (a,p)$. Then $d|p$. Since p is a prime number and $d > 0$, either $d = \pm p$ or $d = 1$. From $d|a, p \nmid a$, we conclude $d \neq \pm p$. So $d = 1$. □

Our proof will depend heavily on the following corollary to Theorem 5.12. It is Proposition 30 in Euclid's *Elements*, Book VII. We shall refer to it as Euclid's lemma.

5.15 Lemma (Euclid's lemma): *Let a,b,p be integers. If p is prime and $p \mid ab$, then $p \mid a$ or $p \mid b$.*

Proof: If $p \mid a$, the lemma is proved. If $p \nmid a$, then $(a,p) = 1$ by Lemma 5.14 and, since $p \mid ab$, we get $p \mid b$ by Theorem 5.12 (with p,a,b in place of a,b,c, respectively). □

5.16 Lemma: *Let $a_1,a_2,\ldots,a_n,p$ be integers. If p is prime and $p \mid a_1a_2\ldots a_n$, then $p \mid a_1$ or $p \mid a_2$ or $\ldots$ or $p \mid a_n$.*

Proof: This follows from Euclid's lemma by a routine induction argument. The details are left to the reader. □

We can now prove uniqueness aside from trivial variations.

5.17 Theorem (Fundamental theorem of arithmetic): *Every integer, which is not zero or a unit, can be expressed as a product of prime numbers in a unique way, apart from the order of the factors and ambiguity of associate numbers.*

Proof: Let $n \in \mathbb{Z}$, $n \neq 0$, $n \neq$ unit. By Theorem 5.13, n can be expressed as a product of prime numbers. We must show uniqueness. This will be done by induction on $|n| \in \mathbb{N}$. Given two decompositions

$$p_1p_2\ldots p_r = n = q_1q_2\ldots q_s$$

of n into prime factors, we have to show that

$$r = s$$

and that

$p_1,p_2,\ldots,p_r$ are, in some order, associate to $q_1,q_2,\ldots,q_s$.

Assume first $|n| = 2$. Then $n = 2$ or $n = -2$ is prime and $n = \mp 2$ is the unique representation of n as a product of prime numbers (having only one factor). So the theorem is true for n if $|n| = 2$.

Now we make the inductive hypothesis that $|n| > 2$ and that the theorem is true for all $k \in \mathbb{Z}$ with $2 \leqslant |k| \leqslant n - 1$, and prove it for n.

If n is a prime number and

$$p_1p_2\ldots p_r = n = q_1q_2\ldots q_s \quad (r,s \in \mathbb{N};\ p\text{'s and } q\text{'s are prime}),$$

then necessarily $r = 1$, $s = 1$, $|p_1| = |n| = |q_1|$, so $p_1 = \mp q_1$. So p_1 and q_1 are associate and the decomposition is unique.

Assume now that

$$p_1p_2\ldots p_r = n = q_1q_2\ldots q_s \quad (r,s \in \mathbb{N};\ p\text{'s and } q\text{'s are prime}),$$

and that n is not a prime number. From $p_r|p_1p_2\ldots p_r$, we get $p_r|q_1q_2\ldots q_s$. By Lemma 5.16, $p_r|q_i$ for some $i = 1,2,\ldots,s$. Changing the order of the q's if necessary, we may assume $p_r|q_s$. The divisors of q_s are ∓ 1, $\mp q_s$. Since p_r is prime, so not a unit, and since $p_r|q_s$, we obtain $p_r = q_s$ or $p_r = -q_s$. Let $p_r = \varepsilon q_s$, with the appropriate unit $\varepsilon = \mp 1 \in \mathbb{Z}$. Then we get

$$\begin{aligned}\varepsilon p_1p_2\ldots p_r = \varepsilon n &= q_1q_2\ldots(\varepsilon q_s)\\ &= q_1q_2\ldots q_{s-1}p_r,\end{aligned}$$

so

$$\varepsilon p_1p_2\ldots p_{r-1} = \varepsilon n/p_r = q_1q_2\ldots q_{s-1}$$

as two decompositions of $\varepsilon n/p_r$ into prime numbers. Since n is not a prime number and $|p_r| > 1$, we have $1 < |\varepsilon n/p_r| < n$. The induction hypothesis tells us that the two decompositions

$$\varepsilon p_1p_2\ldots p_{r-1} = q_1q_2\ldots q_{s-1}$$

of $\varepsilon n/p_r$ are essentially the same:

$$r - 1 = s - 1$$

and $\varepsilon p_1, p_2, \ldots, p_{r-1}$ are, in some order, associate to $q_1, q_2, \ldots, q_{s-1}$. Then

$$r = s$$

and $p_1, p_2, \ldots, p_{r-1}$ are, in some order, associate to $q_1, q_2, \ldots, q_{s-1}$; and p_r is associate to q_s. This completes the proof. □

5.18 Remarks: Collecting the same prime divisors of $n \in \mathbb{N}$ $(n > 1)$ in a single prime power, we can write

$$n = p_1^{a_1}p_2^{a_2}\ldots p_r^{a_r}, \qquad (*)$$

where $0 < p_1 < p_2 < \ldots < p_r$ are the distinct prime divisors of n, and $a_1, a_2, \ldots, a_r$ positive integers. Then $(*)$ is called the *canonical decomposition of* n into prime numbers.

Sometimes it is convenient to relax the condition that the exponents a_i be all positive to the condition that they be nonnegative. For example, the divisors of $n \in \mathbb{N}$, whose canonical decomposition is (*), are exactly the numbers

$$\mp p_1^{b_1} p_2^{b_2} \ldots p_r^{b_r},$$

where $0 \leqslant b_i \leqslant a_i$ for all $i = 1,2, \ldots, r$. If m and n are two natural numbers and

$$m = p_1^{c_1} p_2^{c_2} \ldots p_r^{c_r}, \qquad n = p_1^{e_1} p_2^{e_2} \ldots p_r^{e_r},$$

where $p_1, p_2, \ldots, p_r$ are distinct prime numbers and $c_i \geqslant 0$, $e_i \geqslant 0$ for all $i = 1,2, \ldots, r$, then (m,n) is given by

$$(m,n) = p_1^{t_1} p_2^{t_2} \ldots p_r^{t_r}$$

with $t_i = \min\{c_i, e_i\}$ for all $i = 1,2, \ldots, r$. Here $\min\{x,y\}$ denotes the smaller (minimum) of x and y when $x \neq y$ and denotes x when $x = y$.

It can be shown that $((a,b),c) = (a,(b,c))$ for any $a,b,c \in \mathbb{Z}$, provided a,b are not both equal to zero and b,c are not both equal to zero. The positive number $((a,b),c)$ is called the *greatest common divisor* of a,b,c, and is denoted shortly by (a,b,c). One proves easily that (a,b,c) is the unique integer d such that

(i) $d|a, d|b, d|c$,
(ii) for all $d_1 \in \mathbb{Z}$, if $d_1|a, d_1|b, d_1|c$, then $d_1|d$,
(iii) $d > 0$,

and that there are integers x,y,z satisfying

$$ax + by + cz = (a,b,c).$$

Inductively, if the greatest common divisor of $n-1$ integers $a_1, a_2, \ldots, a_{n-1}$ has already been defined and denoted as $(a_1, a_2, \ldots, a_{n-1})$, then the greatest common divisor $(a_1, a_2, \ldots, a_{n-1}, a_n)$ of n integers $a_1, a_2, \ldots, a_{n-1}, a_n$ is defined to be $((a_1, a_2, \ldots, a_{n-1}), a_n)$. One can show that their greatest common divisor $(a_1, a_2, \ldots, a_{n-1}, a_n)$ is the unique integer d such that

(i) $d|a_1, d|a_2, \ldots, d|a_{n-1}, d|a_n$
(ii) for all $d_1 \in \mathbb{Z}$, if $d_1|a_1, d_1|a_2, \ldots, d_1|a_{n-1}, d_1|a_n$, then $d_1|d$.
(iii) $d > 0$.

In addition, one proves that there are integers $x_1, x_2, \ldots, x_{n-1}, x_n$ such that

$$a_1x_1 + a_2x_2 + \cdots + a_{n-1}x_{n-1} + a_nx_n = (a_1, a_2, \ldots, a_{n-1}, a_n).$$

If $(a_1, a_2, \ldots, a_{n-1}, a_n) = 1$, we say that $a_1, a_2, \ldots, a_{n-1}, a_n$ are relatively prime. In this case, there are integers $x_1, x_2, \ldots, x_{n-1}, x_n$ satisfying

$$a_1x_1 + a_2x_2 + \cdots + a_{n-1}x_{n-1} + a_nx_n = 1.$$

The proofs of these assertions are left to the reader.

Our final topic in this paragraph will be the least common multiple of two nonzero integers. If $a,b \in \mathbb{Z}$ and $a|b$, we say that b is a *multiple of* a.

5.19 Theorem: *Let* $a,b \in \mathbb{Z}$, *neither of them zero* (i.e., $a \neq 0 \neq b$). *Then there is a unique integer such that*

(i) $a|m$ *and* $b|m$,
(ii) *for all* $m_1 \in \mathbb{Z}$, *if* $a|m_1$ *and* $b|m_1$, *then* $m|m_1$,
(iii) $m > 0$.

Proof: The proof will be similar to that of Theorem 5.4. We consider the set $V = \{n \in \mathbb{N}: a|n \text{ and } b|n\}$. This set is not empty, since, for example, $|ab|$ is in V (here we use the hypothesis $a \neq 0 \neq b$). We choose the smallest positive integer in V. Let it be called m. Thus $m > 0$ and m satisfies (iii). Also, $a|m$ and $b|m$ since $m \in V$, and m satisfies (i). It remains to show that m satisfies (ii).

Suppose $m_1 \in \mathbb{Z}$, and $a|m_1$ and $b|m_1$. We divide m_1 by m and get, say, $m_1 = qm + r$, where $q,r \in \mathbb{Z}$ and $0 \leqslant r < m$. Since $a|m$, $a|m_1$ and $b|m$, $b|m_1$, the equation $m_1 = qm + r$ yields that $a|r$ and $b|r$. Hence $r \in V$. We know $0 \leqslant r < m$. If r were not zero, then r would be a natural number in V smaller than the smallest natural number m in V, which is absurd. Thus $r = 0$, so $m_1 = qm$, and $m|m_1$. This shows that m satisfies (ii).

Now the uniqueness of m. Suppose m' satisfies the conditions (i), (ii), (iii), too. Then $a|m'$, $b|m'$ by (i), and so $m|m'$ by (ii). Also, $a|m$, $b|m$ by (i), and so $m'|m$ by (ii). Hence $m|m'$ and $m'|m$. By Lemma 5.2(12), we obtain

$|m| = |m'|$. From (iii), we have $m > 0$, $m' > 0$, which yields $m = m'$. Thus m is unique. □

5.20 Definition: Let $a,b \in \mathbb{Z}$, neither of them zero. The unique integer m in Theorem 5.19 is called the *least common multiple of a and b*.

The least common multiple of a and b will be denoted by $[a,b]$. From the proof of Theorem 5.19, we see that $[a,b]$ is indeed the smallest of the positive multiples of a and b ($[a,b]$ is the smallest number in V). From the fact that $a|m$ and $-a|m$ are equivalent, and likewise that $b|m$ and $-b|m$ are equivalent, it follows that the defining conditions (i), (ii), (iii) do not change when we replace a by $-a$ or b by $-b$. Therefore, $[a,b] = [-a,b] = [-a,-b] = [a,-b]$. In the same way, the conditions (i), (ii), (iii) in Theorem 5.19 are symmetric in a and b, and this gives $[a,b] = [b,a]$.

The greatest common divisor and the least common multiple of two integers will be connected in Lemma 5.22. We need a preliminary result.

5.21 Lemma: *Let a,b,m be integers and $a \neq 0$, $b \neq 0$. If $a|m$, $b|m$ and $(a,b) = 1$, then $ab|m$.*

Proof: This follows immediately from the fundamental theorem of arithmetic (Theorem 5.17), but we give another proof. Since $a|m$ and $b|m$, there are integers a_1,b_1 such that $a\,a_1 = m = b\,b_1$. Hence $a|b\,b_1$. Since $(a,b) = 1$, Theorem 5.12 yields $a|b_1$. So $b_1 = ac$ for some integer c and $m = b\,b_1 = bac = abc$, so $ab|m$, as claimed. □

5.22 Lemma: *Let a and b be integers, neither of them zero. Then we have $[a,b] = |ab|/(a,b)$.*

Proof: As neither $[a,b]$, nor (a,b), nor $|ab|$ changes when we replace a and b by their absolute values, we assume, without loss of generality, that $a > 0$, $b > 0$. We put $d = (a,b)$. We show that ab/d satisfies the three conditions (i), (ii), (iii) in Theorem 5.19. Let $a = a_1 d$, $b = b_1 d$, so that $(a_1,b_1) = 1$ by Lemma 5.11.

We have $a|ab_1$, so $a|a(b/d)$; and $b|a_1b$, so $b|(a/d)b$. Thus a divides ab/d and b divides ab/d. Hence ab/d satisfies (i). Clearly $ab/d > 0$, so ab/d satisfies (iii). We now show that ab/d satisfies (ii) as well; i.e., we show that ab/d divides m_1 whenever $a|m_1$ and $b|m_1$. Let $m_1 \in \mathbb{Z}$ and $a|m_1$, $b|m_1$. Then $d|m_1$ and in fact a_1 divides m_1/d and b_1 divides m_1/d. By Lemma 5.21 (with $a_1,b_1,m_1/d$ in place of a,b,m, respectively), we get a_1b_1 divides m_1/d, so $ab|m_1d$, so ab/d divides m_1. Thus ab/d satisfies (ii) and $ab/d = [a,b]$ □

It can be shown that $[[a,b],c] = [a,[b,c]]$ for any $a,b,c \in \mathbb{Z}$, provided a,b,c are all distinct from zero. The positive number $[[a,b],c]$ is called the *least common multiple of* a,b,c, and is denoted shortly by $[a,b,c]$. One proves easily that $[a,b,c]$ is the unique integer m such that

(i) $a|m$, $b|m$, $c|m$,
(ii) for all $m_1 \in \mathbb{Z}$, if $a|m_1$, $b|m_1$, $c|m_1$, then $m|m_1$,
(iii) $m > 0$.

Inductively, if the least common multiple of $n-1$ integers $a_1,a_2,\ldots,a_{n-1}$ has already been defined and denoted as $[a_1,a_2,\ldots,a_{n-1}]$, then the least common multiple $[a_1,a_2,\ldots,a_{n-1},a_n]$ of n integers $a_1,a_2,\ldots a_{n-1},a_n$ is defined to be $[[a_1,a_2,\ldots,a_{n-1}],a_n]$. One can show that their least common multiple $[a_1,a_2,\ldots,a_{n-1},a_n]$ is the unique integer m such that

(i) $a_1|m, a_2|m, \ldots, a_{n-1}|m, a_n|m$,
(ii) for all $m_1 \in \mathbb{Z}$, if $a_1|m_1, a_2|m_1, \ldots, a_{n-1}|m_1, a_n|m_1$, then $m|m_1$,
(iii) $m > 0$.

Exercises

1. Find (10897,16949) and express it in the form $10897x + 16949y$, where x and y are integers.

2. Assume $m,n \in \mathbb{N}$ and $m \neq n$. What is $(2^{2^m}+1, 2^{2^n}+1)$?

3. Let $a,b \in \mathbb{Z}$, neither of them equal to zero, and assume $(a,b) = 1$. Let x_0,y_0 be integers such that $ax_0 + by_0 = 1$. Prove that all integer pairs x,y satisfying $ax + by = 1$ are given by

$$x = x_0 + bt, \quad y = y_0 - at$$

as t runs through all integers.

4. Let a,b,c be integers, none of them equal to zero, and let $(a,b) = d$. Prove that there are integers x,y satisfying $ax + by = c$ if and only if $d|c$. Moreover, if $d|c$ and x_0,y_0 are integers such that $ax_0 + by_0 = c$, prove that all integer pairs x,y satisfying $ax + by = c$ are given by

$$x = x_0 + (b/d)t, \quad y = y_0 - (a/d)t$$

as t runs through all integers.

5. Prove the assertions in Remark 5.18.

6. Let m and n be two natural numbers and

$$m = p_1^{c_1}p_2^{c_2}\ldots p_r^{c_r}, \qquad n = p_1^{e_1}p_2^{e_2}\ldots p_r^{e_r},$$

where $p_1,p_2,\ldots,p_r$ are distinct prime numbers and $c_i \geqslant 0$, $e_i \geqslant 0$ for all $i = 1,2,\ldots,r$. Show that

$$[m,n] = p_1^{u_1}p_2^{u_2}\ldots p_r^{u_r}$$

with $u_i = \max\{c_i,e_i\}$ for all $i = 1,2,\ldots,r$. Here $\max\{x,y\}$ denotes the greater (maximum) of x and y when $x \neq y$ and denotes x when $x = y$.

7. Prove or disprove: $(a,[b,c]) = [(a,b),(a,c)]$ for all $a,b,c \in \mathbb{N}$.

8. Let $a,b \in \mathbb{Z}$, $b > 0$, and $a = qb + r$, with $q,r \in \mathbb{Z}$, $0 \leqslant r < b$. Prove directly that $(a,b) = (a,r)$.

9. Let $a,b \in \mathbb{Z}$ and $(a,b) = 1$. Show that $(a - b,a + b) = 1$ or 2.

10. Let a,m,n be natural numbers. Show that $(a^m - 1,a^n - 1) = a^{(m,n)} - 1$.

§6
Integers Modulo n

In Example 2.3(e), we have defined the congruence of two integers a,b with respect to a modulus $n \in \mathbb{N}$. Let us recall that $a \equiv b \pmod{n}$ means $n|a-b$. We have proved that congruence is an equivalence relation on $\mathbb{Z}$. The equivalence classes are called the *congruence classes* or *residue classes* (modulo n). The congruence class of $a \in \mathbb{Z}$ will be denoted by $\bar{a}$. Notice that there is ambiguity in this notation, for there is no reference to the modulus. Thus $\bar{1}$ represents the residue class of 1 with respect to the modulus 1, also with respect to the modulus 2, also with respect to the modulus 3, in fact with respect to any modulus. However, the modulus will be usually fixed throughout a particular discussion and $\bar{a}$ will represent the residue class of a with respect to that fixed modulus. The ambiguity is therefore harmless.

By the division algorithm (Theorem 5.3), any integer k can be written as $k = qn + r$, with $q,r \in \mathbb{Z}$, $0 \leqslant r < b$. So any integer k is congruent (mod n) to one of the numbers $0,1,2,\ldots,n-1$. Furthermore, no two distinct of the numbers $0,1,2,\ldots,n-1$ are congruent (mod n), for if $r_1, r_2 \in \{0,1,2,\ldots,n-1\}$ and $r_1 \equiv r_2 \pmod{n}$, then $n|r_1 - r_2$, so $n \leqslant |r_1 - r_2|$ by Lemma 5.2(11), and so $n \leqslant (n-1)-0$, which is impossible. Thus any integer is congruent to one of the numbers $0,1,2,\ldots,n-1$, and these numbers are pairwise incongruent. This means that $0,1,2,\ldots,n-1$ are the representatives of all the residue classes. Hence there are exactly n residue classes (mod n), namely

$$\begin{aligned}
\bar{0} &= \{x \in \mathbb{Z} : x \equiv 0 \pmod{n}\} &&= \{nz \in \mathbb{Z} : z \in \mathbb{Z}\} &&=: n\mathbb{Z} \\
\bar{1} &= \{x \in \mathbb{Z} : x \equiv 1 \pmod{n}\} &&= \{nz+1 \in \mathbb{Z} : z \in \mathbb{Z}\} &&=: n\mathbb{Z}+1 \\
\bar{2} &= \{x \in \mathbb{Z} : x \equiv 2 \pmod{n}\} &&= \{nz+2 \in \mathbb{Z} : z \in \mathbb{Z}\} &&=: n\mathbb{Z}+2
\end{aligned}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\overline{n-1} = \{x \in \mathbb{Z} : x \equiv n-1 \pmod{n}\} = \{nz+(n-1) \in \mathbb{Z} : z \in \mathbb{Z}\} =: n\mathbb{Z}+(n-1).$$

The set $\{\bar{0},\bar{1},\bar{2},\ldots,\overline{n-1}\}$ of residue classes (mod n) will be denoted by $\mathbb{Z}_n$. An element of $\mathbb{Z}_n$, thas is, a residue class (mod n) is called an *integer modulo n*, or an *integer mod n*. An integer mod n is not an integer, not

an element of $\mathbb{Z}$; it is a subset of $\mathbb{Z}$. An integer mod n is not an integer with a property "mod n". It is an object whose name consists of the three words "integer", "mod(ulo)", "n".

6.1 Lemma: *Let $n \in \mathbb{N}$, $a,a_1,b,b_1 \in \mathbb{Z}$. If $a \equiv a_1 \pmod n$ and $b \equiv b_1 \pmod n$, then $a + b \equiv a_1 + b_1 \pmod n$ and $ab \equiv a_1 b_1 \pmod n$.*

Proof: If $a \equiv a_1 \pmod n$ and $b \equiv b_1 \pmod n$, then $n \mid a - a_1$ and $n \mid b - b_1$. Hence $n \mid (a - a_1) + (b - b_1)$ by Lemma 5.2(5), which gives $n \mid (a + b) - (a_1 + b_1)$, so $a + b \equiv a_1 + b_1 \pmod n$. Also, $n \mid b(a - a_1) + a_1(b - b_1)$ by Lemma 5.2(7), which gives $n \mid ba - a_1 b_1$, so $ab \equiv a_1 b_1 \pmod n$. □

We want to define a kind of addition $\oplus$ and a kind of multiplication $\otimes$ on $\mathbb{Z}_n$. We put

$$\bar{a} \oplus \bar{b} = \overline{a + b} \qquad (*)$$

$$\bar{a} \otimes \bar{b} = \overline{ab} \qquad (**)$$

for all $\bar{a},\bar{b} \in \mathbb{Z}_n$ (for all $a,b \in \mathbb{Z}$). This is a very natural way of introducing addition and multiplication on $\mathbb{Z}_n$.

$(*)$ and $(**)$ seem quite innocent, but we must check that $\oplus$ and $\otimes$ are really binary operations on $\mathbb{Z}_n$. The reader might say at this point that $\oplus$ and $\otimes$ are clearly defined on $\mathbb{Z}_n$ and that there is nothing to check. But yes, there is. Let us remember that a binary operation on $\mathbb{Z}_n$ is a function from $\mathbb{Z}_n \times \mathbb{Z}_n$ into $\mathbb{Z}_n$ (Definition 3.18). As such, to each pair $(\bar{a},\bar{b})$ in $\mathbb{Z}_n \times \mathbb{Z}_n$, there must correspond a *single* element $\bar{a} \oplus \bar{b}$ and $\bar{a} \otimes \bar{b}$ if $\oplus$ and $\otimes$ are to be binary operations on $\mathbb{Z}_n$ (Definition 3.1) We must check that the rules $(*)$ and $(**)$ produce elements of $\mathbb{Z}_n$ that are uniquely determined by $\bar{a}$ and $\bar{b}$.

The rules $(*)$ and $(**)$ above convey the wrong impression that $\bar{a} \oplus \bar{b}$ and $\bar{a} \otimes \bar{b}$ are uniquely determined by $\bar{a}$ and $\bar{b}$. In order to penetrate into the matter, let us try to evaluate $X \oplus Y$, where $X,Y \in \mathbb{Z}_n$ are not given directly as the residue classes of integers $a,b \in \mathbb{Z}$. (We discuss $\oplus$; the discussion applies equally well to $\otimes$.) How do we find $X \oplus Y$? Since $X,Y \in \mathbb{Z}_n$, there are integers $a,b \in \mathbb{Z}$ with $\bar{a} = X, \bar{b} = Y$. Now add a and b in $\mathbb{Z}$ to get $a+b \in \mathbb{Z}$, then take the residue class of $a+b$. The result is $X \oplus Y$.

The result? The question is whether we have only *one* result to justify the article "the". We summarize telegrammatically. To find $X \oplus Y$,

1) choose $a \in \mathbb{Z}$ from X,
2) choose $b \in \mathbb{Z}$ from Y,
3) find $a + b$ in $\mathbb{Z}$,
4) take the residue class of $a + b$.

This sounds a perfectly good recipe for finding $X \oplus Y$, but notice that we use some auxiliary objects, namely a and b, to find $X \oplus Y$, which must be determined by X and Y alone. Indeed, the result $\overline{a + b}$ depends explicitly on the auxiliary objects a and b. We can use our recipe with different auxiliary objects. Let us do it. 1) I choose a from $X \subseteq \mathbb{Z}$ and you choose a_1 from X. 2) I choose b from $Y \subseteq \mathbb{Z}$ and you choose b_1 from Y. 3) I compute $a + b$ and you compute $a_1 + b_1$. In general, $a + b \neq a_1 + b_1$. Hence our recipe gives, generally speaking, distinct elements $a + b$ and $a_1 + b_1$. So far, both of us followed the same recipe. I cannot claim that my computation is correct and yours is false. Nor can you claim the contrary. Now we carry out the fourth step. I find the residue class of $a + b$ as $X \oplus Y$, and you find the residue class of $a_1 + b_1$ as $X \oplus Y$. Since $a + b \neq a_1 + b_1$ in $\mathbb{Z}$, it can very well happen that $\overline{a + b} \neq \overline{a_1 + b_1}$ in $\mathbb{Z}_n$. On the other hand, if $\oplus$ is to be a binary operation on $\mathbb{Z}_n$, we must have $\overline{a + b} = \overline{a_1 + b_1}$. This is the central issue. In order that $\oplus$ be a binary operation on $\mathbb{Z}_n$, there must work a mechanism which ensures $\overline{a + b} = \overline{a_1 + b_1}$ whenever $\overline{a} = \overline{a_1}$, $\overline{b_1} = \overline{b}$, even if $a + b \neq a_1 + b_1$. If there is such a mechanism, we say $\oplus$ is a well defined operation on $\mathbb{Z}_n$. This means $\oplus$ is really a genuine operation on $\mathbb{Z}_n$: $X \oplus Y$ is uniquely determined by X and Y alone. Any dependence of $X \oplus Y$ on auxiliary integers $a \in X$ and $b \in Y$ is only apparent. We will prove that $\oplus$ and $\otimes$ are well defined operations on $\mathbb{Z}_n$, but before that, we discuss more generally well definition of functions.

A function $f: A \to B$ is essentially a rule by which each element a of A is associated with a unique element of $f(a) = b$ of B. The important point is that the rule produces an element $f(a)$ that depends only on a. Sometimes we consider rules having the following form. To find $f(a)$,

1) do this and that
2) take an x related to a in such and such manner
3) do this and that to x
4) the result is $f(a)$.

A rule of this type uses an auxiliary object x. The result then depends on a and x. At least, it seems so. This is due to the ambiguity in the second step. This step states that we choose an x with such and such property, but there may be many objects $x,y,z, \ldots$ related to a in the prescribed manner. The auxiliary objects $x,y,z, \ldots$ will, in general, produce different results, so we should perhaps that the result is $f(a,x)$ (or $f(a,y), f(a,z), \ldots$). In order the above rule to be a function, it must produce the same result. Hence we must have $f(a,x) = f(a,y) = f(a,z) = \cdots$. The rule must be so constructed that the same result will obtain even if we use different auxiliary objects. If this be the case, the function is said to be *well defined*.

This terminology is somewhat unfortunate. It sounds as though there are two types of functions, well defined functions and not well defined functions (or badly defined functions). This is definitely not the case. A well defined function is simply a function. Badly defined functions do not exist. Being well defined is not a property, such as continuity, boundedness, differentiability, integrability etc. that a function might or might not possess. That a function $f: A \to B$ is well defined means: 1) the rule of evaluating $f(a)$ for $a \in A$ makes use of auxiliary, foreign objects, 2) there are many choices of these foreign objects, hence 3) we have reason to suspect that applying the rule with different choices may produce different results, which would imply that our rule does not determine $f(a)$ uniquely and f is not a function in the sense of Definition 3.1, but 4) our suspicion is not justified, for there is a mechanism, hidden under the rule, which ensures that same result will obtain even if we apply the rule with different auxiliary objects. The question as to whether a "function" is well defined arises only if that "function" uses objects not uniquely determined by the element a in its "domain" in order to evaluate $f(a)$. We wrote "function" in quotation marks, for such a thing may not be a function in the sense of Definition 3.1. Given such a "function", which we want to be a function in the sense of Definition 3.1, we check whether $f(a)$ is uniquely determined by a, that is, we check whether $f(a)$ is independent of the auxiliary objects that we use for evaluating $f(a)$. If this be the case, our supposed "function" f is indeed a function in the sense of Definition 3.1. We say then that f is well defined, or f is a well defined function. This means f is a function. In fact, it is more accurate to say that a function is defined instead of saying that a function is well defined.

6.2 Examples: (a) Let L be the set of all straight lines in the Euclidean plane, on which we have a cartesian coordinate system. We consider the "function" $s: L \to \mathbb{R} \cup \{\infty\}$, which assigns the slope of the line l to l. How do we find $s(l)$? As follows: 1) choose a point, say (x_1,y_1), on l; 2) choose another point, say (x_2,y_2), on l; 3) evaluate $x_2 - x_1$ and $y_2 - y_1$; 4) put $s(l) = (y_2 - y_1)/(x_2 - x_1)$ if $x_1 \neq x_2$ and $s(l) = \infty$ if $x_1 = x_2$. Clearly we can choose the points in many ways. For example, we might choose $(x_1',y_1') \neq (x_1,y_1)$ as the first point, $(x_2',y_2') \neq (x_2,y_2)$ as the second point. Then we have, in general, $x_2' - x_1' \neq x_2 - x_1$ and $y_2' - y_1' \neq y_2 - y_1$, so we might suspect that $(y_2' - y_1')/(x_2' - x_1') \neq (y_2 - y_1)/(x_2 - x_1)$. It is known from analytic geometry that these two quotients are equal, hence $s(l)$ depends only on l, and not on the points we choose. Thus s is a well defined function. Ultimately, this is due to the fact that there passes one and only one straight line through two distinct points. The next example shows that well definition breaks down if we modify the domain a little.

(b) Let C be the set of all curves in the Euclidean plane. We consider the "function" $s: C \to \mathbb{R} \cup \{\infty\}$, which assigns the "slope" of the curve c to c. How do we find $s(c)$? As follows: 1) choose a point, say (x_1,y_1), on l; 2) choose another point, say (x_2,y_2), on l; 3) evaluate $x_2 - x_1$ and $y_2 - y_1$; 4) put $s(l) = (y_2 - y_1)/(x_2 - x_1)$ if $x_1 \neq x_2$ and $s(l) = \infty$ if $x_1 = x_2$. This is the same rule as the rule in Example 6.2(a). Let us find the "slope" of the curve $y = x^2$. 1) Choose a point on this curve, for example (0,0). If you prefer, you might choose (-1,1). 2) Choose another point on this curve, for example (1,1). If you prefer, you might choose (3,9) of course. 3) Evaluate the differences of coordinates. We find 1 - 0 and 1 - 0. You find 3 -(-1) and 9 - 1. Hence 4) the slope is 1/1. You find it to be 8/4. So $s(c) = 1$ and $s(c) = 2$. This is nonsense. We see that different choices of the points on the curve (different choices of the auxiliary objects) give rise to different results. So the above rule is not a function. We do not say "s is not a well defined function". s is simply not a function at all. s is not defined.

(c) Let F be the set of all continuous functions on a closed interval $[a,b]$. We want to "define" an integral "function" $I: F \to \mathbb{R}$, which assignes the real number $\int_a^b f(x)dx$ to $f \in F$. So $I(f) = \int_a^b f(x)dx$. I is a "function" whose "domain" is a set of functions. How do we find $I(f)$? As follows. 1) Choose an indefinite integral of f, that is, choose a function F on $[a,b]$ such that $F'(x) = f(x)$ for all $x \in [a,b]$ (we take one-sided derivatives at a

and b). 2) Evaluate $F(a)$ and $F(b)$. 3) Put $I(f) = F(b) - F(a)$. There are many functions F with $F'(x) = f(x)$ for all $x \in [a,b]$. For two different choices F_1 and F_2, we have $F_1(b) \neq F_2(b)$ and $F_1(a) \neq F_2(a)$ in general. So we may suspect that $F_1(b) - F_1(a) \neq F_2(b) - F_2(a)$. In order to show that I is a well defined function, we must prove $F_1(b) - F_1(a) = F_2(b) - F_2(a)$ whenever F_1 and F_2 are functions on $[a,b]$ such that $F_1'(x) = f(x) = F_2'(x)$ for all $x \in [a,b]$. We know from the calculus that, when F_1 and F_2 have this property, there is a constant c such that $F_1(x) = F_2(x) + c$ for all $x \in [a,b]$. So $F_1(b) - F_1(a) = (F_2(b) + c) - (F_2(a) + c) = F_2(b) - F_2(a)$. Therefore, I is well defined.

After this lengthy digression, we return to the integers mod n and to the "operations" $\oplus$ and $\otimes$.

6.3 Lemma: *$\oplus$ and $\otimes$ are well defined operations on $\mathbb{Z}_n$.*

Proof: We are to prove $\bar{a} \oplus \bar{b} = \bar{a}' \oplus \bar{b}'$ and $\bar{a} \otimes \bar{b} = \bar{a}' \otimes \bar{b}'$ whenever $\bar{a} = \bar{a}'$ and $\bar{b} = \bar{b}'$ in $\mathbb{Z}_n$(different names for identical residue classes should not yield differerent results). This follows from Lemma 6.1. Indeed, if $\bar{a} = \bar{a}'$ and $\bar{b} = \bar{b}'$, then $a \equiv a' \pmod{n}$ and $b \equiv b' \pmod{n}$ by definition, so we obtain $a + b \equiv a' + b' \pmod{n}$ and $ab \equiv a'b' \pmod{n}$ by Lemma 6.1, hence $\overline{a + b} = \overline{a' + b'}$ and $\overline{ab} = \overline{a'b'}$, which gives $\bar{a} \oplus \bar{b} = \overline{a + b} = \overline{a' + b'} = \bar{a}' \oplus \bar{b}'$ and $\bar{a} \otimes \bar{b} = \overline{ab} = \overline{a'b'} = \bar{a}' \otimes \bar{b}'$. □

Having proved that $\oplus$ and $\otimes$ are well defined operations on $\mathbb{Z}_n$, we proceed to show that $\oplus$ and $\otimes$ possess many (but not all) properties of the usual addition and multiplication of integers. First we simplify our notation. From now on, we write $+$ and $\cdot$ instead of $\oplus$ and $\otimes$. In fact, we shall even drop $\cdot$ and use simply juxtaposition to denote a product of two integers mod n. Thus we will have $\bar{a} + \bar{b} = \overline{a + b}$ and $\bar{a} \cdot \bar{b} = \overline{ab}$ or simply $\bar{a}\,\bar{b} = \overline{ab}$. The reader should note that the same sign "+" is used to denote two very distinct operations: $\oplus$ in the old notation and the usual addition of integers. If anything, they are defined on distinct sets $\mathbb{Z}_n$ and $\mathbb{Z}$. The same remarks apply to multiplication.

6.4 Lemma: *For all* $\overline{a},\overline{b},\overline{c} \in \mathbb{Z}_n$, *the following hold.*

(1) $\overline{a} + \overline{b} \in \mathbb{Z}_n$;

(2) $(\overline{a} + \overline{b}) + \overline{c} = \overline{a} + (\overline{b} + \overline{c})$;

(3) $\overline{a} + \overline{0} = \overline{a}$;

(4) $\overline{a} + \overline{-a} = \overline{0}$;

(5) $\overline{a} + \overline{b} = \overline{b} + \overline{a}$;

(6) $\overline{a}\cdot\overline{b} \in \mathbb{Z}_n$;

(7) $(\overline{a}\cdot\overline{b})\cdot\overline{c} = \overline{a}\cdot(\overline{b}\cdot\overline{c})$;

(8) $\overline{a}\cdot\overline{1} = \overline{a}$;

(9) *if* $(a,n) = 1$, *then there is an* $\overline{x} \in \mathbb{Z}_n$ *such that* $\overline{a}\cdot\overline{x} = \overline{1}$;

(10) $\overline{a}\cdot\overline{b} = \overline{b}\cdot\overline{a}$;

(11) $\overline{a}\cdot(\overline{b} + \overline{c}) = \overline{a}\cdot\overline{b} + \overline{a}\cdot\overline{c}$ *and* $(\overline{b} + \overline{c})\cdot\overline{a} = \overline{b}\cdot\overline{a} + \overline{c}\cdot\overline{a}$;

(12) $\overline{a}\cdot\overline{0} = \overline{0}$.

Proof: (1) is obvious. (2) follows from the corresponding property of addition in $\mathbb{Z}$. We indeed have

$$\begin{aligned}(\overline{a} + \overline{b}) + \overline{c} &= \overline{(a + b)} + \overline{c} \\ &= \overline{(a + b) + c} \\ &= \overline{a + (b + c)} \\ &= \overline{a} + \overline{(b + c)} \\ &= \overline{a} + (\overline{b} + \overline{c}).\end{aligned}$$

The remaining assertions are proved in the same way by drawing bars over integers in the corresponding equations in $\mathbb{Z}$. We prove only (9), which is not as straightforward as the other claims. If $(a,n) = 1$, Then there are integers x,y with $ax - ny = 1$ (Lemma 5.10). Using (3) and (12), we get $\overline{1} = \overline{ax - ny} = \overline{ax} - \overline{ny} = \overline{a}\cdot\overline{x} - \overline{n}\cdot\overline{y} = \overline{a}\cdot\overline{x} - \overline{0}\cdot\overline{y} = \overline{a}\cdot\overline{x} - \overline{0} = \overline{a}\cdot\overline{x}$. □

Exercises

1. Determine whether the "function" $g: \mathbb{Z}_{13} \to \mathbb{N}$ is well-defined, if g is defined as follows.

(a) $g(\overline{a}) = (a,13)$;

(b) $g(\overline{a}) = (a,26)$;

(c) $g(\overline{a}) = (a,169)$;

(d) $g(\overline{a}) = (a^2,13)$;

(e) $g(\bar{a}) = (a^3, 169)$;
(f) $g(\bar{a}) = (a, 6)$;
(g) $g(\bar{a}) = (a^2, 65)$;

where $\bar{a} \in \mathbb{Z}_{13}$ and $a \in \mathbb{Z}$.

2) Let $f: \mathbb{Z}_{12} \times \mathbb{Z} \to \mathbb{Z}_{12}$ be such that $(\bar{a}, b) \mapsto \overline{a^2+ab+b^2}$. Is f well defined?

3) For an integer a, we denote by $\bar{a}$ the residue class of a (mod 12), by $\tilde{a}$ the residue class of a (mod 6), and by $\hat{a}$ the residue class of a (mod 5), so that $\bar{a} \in \mathbb{Z}_{12}$, $\tilde{a} \in \mathbb{Z}_6$ and $\hat{a} \in \mathbb{Z}_5$. Determine whether the following "functions" are well defined.

(a) $\mathbb{Z}_{12} \to \mathbb{Z}_6$, $\bar{a} \to \tilde{a}$;
(b) $\mathbb{Z}_6 \to \mathbb{Z}_{12}$, $\tilde{a} \to \bar{a}$;
(c) $\mathbb{Z}_{12} \to \mathbb{Z}_5$, $\bar{a} \to \hat{a}$;
(d) $\mathbb{Z}_5 \to \mathbb{Z}_6$, $\hat{a} \to \tilde{a}$;
(e) $\mathbb{Z}_5 \to \mathbb{Z}_6$, $\hat{a} \to \widetilde{a+1}$.

CHAPTER 2

Groups

§7
Basic Definitions

Before giving the formal definition of a group, we would rather present some concrete examples.

7.1 Examples: (a) Consider the addition of integers. From the numerous properties of this binary operation, we single out the following ones.

(i) $+$ is a binary operation on $\mathbb{Z}$, so, for any $a,b \in \mathbb{Z}$, we have $a+b \in \mathbb{Z}$.

(ii) For all $a, b, c \in \mathbb{Z}$, we have $(a+b)+c = a+(b+c)$.

(iii) There is an integer, namely $0 \in \mathbb{Z}$, which has the property

$$a+0=a \quad \text{for all } a \in \mathbb{Z}.$$

(iv) For all $a \in \mathbb{Z}$, there is an integer, namely $-a$, such that

$$a+(-a)=0.$$

(b) Consider the multiplication of positive real numbers. Let $\mathbb{R}^+$ be the set of positive real numbers. Here the multiplication enjoys properties analogous to the ones above.

(i) $\cdot$ is a binary operation on $\mathbb{R}^+$, so, for any $a,b \in \mathbb{R}^+$, we have $a \cdot b \in \mathbb{R}^+$.

(ii) For all $a, b, c \in \mathbb{R}^+$, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

(iii) There is a positive real number, namely $1 \in \mathbb{R}^+$, which has the property

$$a \cdot 1 = a \text{ for all } a \in \mathbb{R}^+.$$

(iv) For all $a \in \mathbb{R}^+$, there is a positive real number, namely $1/a$, such that

$$a \cdot \frac{1}{a} = 1.$$

(c) Let n be a natural number and consider the addition in $\mathbb{Z}_n$, which we introduced in §6.

(i) $+$ is a binary operation on $\mathbb{Z}_n$, so, for any $\overline{a}, \overline{b} \in \mathbb{Z}_n$, we have $\overline{a} + \overline{b} \in \mathbb{Z}_n$.

(ii) For all $\overline{a}, \overline{b}, \overline{c} \in \mathbb{Z}_n$, we have $(\overline{a} + \overline{b}) + \overline{c} = \overline{a} + (\overline{b} + \overline{c})$.

(iii) There is an integer mod n, namely $\overline{0} \in \mathbb{Z}_n$, which has the property

$$\overline{a} + \overline{0} = \overline{a} \text{ for all } \overline{a} \in \mathbb{Z}_n.$$

(iv) For all $\overline{a} \in \mathbb{Z}_n$, there is an integer mod n, namely $\overline{-a}$, such that

$$\overline{a} + (\overline{-a}) = \overline{0}.$$

(d) Let X be a nonempty set and let S_X be the set of all one-to-one mappings from X onto X. Consider the composition $\circ$ of mappings in S_X.

(i) $\circ$ is a binary operation on S_X, for if σ and τ are one-to-one mappings from X onto X, so is $\sigma \circ \tau$ by Theorem 3.13.

(ii) For all $\sigma, \tau, \mu \in S_X$, we have $(\sigma \circ \tau) \circ \mu = \sigma \circ (\tau \circ \mu)$ (Theorem 3.10).

(iii) There is a mapping in S_X, namely $\iota_X \in S_X$, such that

$$\sigma \circ \iota_X = \sigma \text{ for all } \sigma \in S_X \quad \text{(Example 3.9(a)).}$$

(iv) For all $\sigma \in S_X$, there is a mapping in S_X, namely σ^{-1}, such that

$$\sigma \circ \sigma^{-1} = \iota_X.$$

(See Theorem 3.14 and Theorem 3.16. That $\sigma^{-1} \in S_X$ follows from Theorem 3.17(1).)

These are examples of groups. In each case, we have a nonempty set and a binary operation on that set which enjoys some special properties. A group will be defined as a nonempty set and a binary operation on that set having the same properties as in the examples above. A group will thus consist of two parts: a set and a binary operation. Formally, a

group is an ordered pair whose components are the set and the operation in question.

7.2 Definition: An ordered pair $(G, \circ)$, where G is a nonempty set and $\circ$ is a binary operation on G, is called a *group* provided the following hold.

(i) $\circ$ is a (well defined) binary operation on G. Thus, for any $a, b \in G$, $a \circ b$ is a uniquely determined element of G.

(ii) For all $a, b, c \in G$, we have $(a \circ b) \circ c = a \circ (b \circ c)$.

(iii) There is an element e in G such that

$$a \circ e = a \text{ for all } a \in G$$

and which is furthermore such that

(iv) for all $a \in G$, there is an x with

$$a \circ x = e.$$

When $(G, \circ)$ is a group, we also say that G is (or builds, or forms) a group with respect to $\circ$ (or under $\circ$). Since a group is an ordered pair, two groups $(G, \circ)$ and $(H, *)$ are equal if and only if $G = H$ and the binary operation $\circ$ on G is equal to the binary operation $*$ on G (i.e., $\circ$ and $*$ are identical mappings from $G \times G$ into G). On one and the same set G, there may be distinct binary operations $\circ$ and $*$ under which G is a group. In this case, the groups $(G, \circ)$ and $(G, *)$ are distinct.

The four conditions (i)-(iv) of Definition 7.2 are known as the *group axioms*. The first axiom (i) is called the *closure axiom*. When (i) is true, we say G is *closed under* $\circ$.

A binary operation $\circ$ on a nonempty set G is said to be *associative* when (ii) holds. The associativity of $\circ$ enables us to write $a \circ b \circ c$ without ambiguity. Indeed, $a \circ b \circ c$ has first no meaning at all. We must write either $(a \circ b) \circ c$ or $a \circ (b \circ c)$ to denote a meaningful element in G. By associativity, we may and do make the convention that $a \circ b \circ c$ will mean $(a \circ b) \circ c = a \circ (b \circ c)$, for whether we read it as $(a \circ b) \circ c$ or $a \circ (b \circ c)$ does not make any difference. This would be wrong if $\circ$ were not associative. For instance, : (division) is not an associative operation on $\mathbb{Q}\backslash\{0\}$ and $(a:b):c \neq a:(b:c)$ unless $c = 1$ (here $a, b, c \in \mathbb{Q}\backslash\{0\}$). Thus $a:b:c$ is ambiguous.

An element e of a set G, on which there is a binary operation $\circ$, is called a *right identity element* or simply a *right identity* if $a \circ e = a$ for all a in

G. The third group axiom (iii) ensures that group G has a right identity element. We will show presently that group has precisely one identity element, but we have not proved it yet and we must be careful not to use the uniqueness of the right identity before we prove it. All we know at this stage is that a group has at least one right identity for which (iv) holds. As it is, there may be many right identities. In addition, there may be some right identities for which (iv) is true and also some for which (iv) is false. For the time being, these possibilities are not excluded.

They will be excluded in Lemma 7.3, where we will prove further that our unique right identity is also a left identity. A *left identity element* or a *left identity of* G, where G is a nonempty set with a binary operation $\circ$ on it, is by definition an element f of G such that $f \circ a = a$ for all $a \in G$. The group axioms say nothing about left identities. If $(G, \circ)$ is a group, we do not yet know if there is a left identity in G at all, nor do we know any relation between right and left identities. For the time being, there may be no or one or many left identities in G. If there is only one left identity, it may or may not be right identity. If there are many left identities, some or one or none of them may be right identities.

We mention all these possibilities so that the reader does not read in the axioms more than what they really say. The group axioms say nothing about left identities or about the uniqueness of the right identity.

The group axioms do say something about right inverses. If G is a nonempty set with a binary operation $\circ$ on it, and if e is a right identity in G, and $a \in G$, an element $x \in G$ is called a *right inverse* of a (with respect to e) when $a \circ x = e$. The group axioms state that, in case $(G, \circ)$ is a group, there is a right identity e in G with respect to which each element of G has at least one right inverse. Until we prove Lemma 7.3, there may be many right identities with this property. Also, some of the right identity elements may and some of the right identity elements may not have this property. Furthermore, some (or all) of the elements may have more than one right inverses with respect to some (or all) of the right identities. The group axioms make no uniqueness assertion about the right inverses.

Before we lose ourselves in chaos, we had better prove our lemma.

7.3 Lemma: *Let $(G, \circ)$ be a group and let e be a right identity element of G such that, for all $a \in G$, there exists a suitable x in G with $a \circ x = e$. The existence of e is assured by the group axioms* (iii) *and* (iv).

(1) *If $g \in G$ is such that $g \circ g = g$, then $g = e$.*

(2) *e is the unique right identity in G.*

(3) *A right inverse of an element in G is also a left inverse of the same element. In other words, if $a \circ x = e$, then $x \circ a = e$.*

(4) *e is a left identity in G. That is, $e \circ a = a$ for all $a \in G$.*

(5) *e is the unique left identity in G.*

(6) *Each element has a unique right inverse in G.*

(7) *Each element has a unique left inverse in G.*

(8) *The unique right inverse of any $a \in G$ is equal to the unique left inverse of a.*

Proof: (1) Let $g \in G$ be such that $g \circ g = g$. We choose a right inverse of g with respect to e. This is possible by the axiom (iv). Let us call it h. Thus $g \circ h = e$. Then

$$\begin{aligned} (g \circ g) \circ h &= g \circ h \\ g \circ (g \circ h) &= g \circ h && \text{(by associativity),} \\ g \circ e &= e && \text{(since } g \circ h = e\text{),} \\ g &= e && \text{(since } e \text{ is a right identity).} \end{aligned}$$

This proves part (1).

(2) The claim is that e is the unique right identity in G. This means: if $f \in G$ is a right identity, that is, if $a \circ f = a$ for all $a \in G$, then $f = e$. Suppose f is a right identity. Then $a \circ f = a$ for all $a \in G$. Writing f for a in particular, we see $f \circ f = f$. Hence $f = e$ by part (1).

(3) A right inverse x of an arbitrary element $a \in G$ is also a left inverse of a. This is what we are to prove. So we assume $a \circ x = e$ and try to derive $x \circ a = e$. We use part (1). If $a \circ x = e$, then

$$\begin{aligned} (x \circ a) \circ (x \circ a) &= [(x \circ a) \circ x] \circ a && \text{(by associativity)} \\ &= [x \circ (a \circ x)] \circ a && \text{(by associativity)} \\ &= [x \circ e] \circ a \\ &= x \circ a. \end{aligned}$$

So $g := (x \circ a)$ is such that $g \circ g = g$. By part (1), $g = e$. So $x \circ a = e$.

(4) We are to prove that e is a left identity. So we must show $e \circ a = a$ for all $a \in G$. Let $a \in G$ and let x be a right inverse of a. Then

$$a \circ x = e$$

$$
\begin{aligned}
a \circ x &= x \circ a && \text{(by part (3))} \\
(a \circ x) \circ a &= (x \circ a) \circ a \\
a \circ (x \circ a) &= (x \circ a) \circ a \\
a \circ e &= e \circ a \\
a &= e \circ a.
\end{aligned}
$$

Therefore, e is a left identity as well. This proves part (4).

(5) The claim is that e is the unique left identity in G. This means: if f is a left identity in G so that $f \circ a = a$ for all $a \in G$, then $f = e$. We know that the right identity e is a left identity (part (4)), and that e is the unique right identity (part (2)). So we conclude that e is the unique left identity. Is this correct? No, this is wrong. This would be correct if we knew that any left identity is also a right identity (and so the unique right identity by part (2)), which is not what part (4) states. For all we proved up to now, there may very well a unique right identity and many left identities (among them the right identity). We are to show in part (5) that this is impossible.
After so much fuss, now the correct proof, which is very short. Suppose $f \circ a = a$ for all $a \in G$. Write in particular f for a. Then $f \circ f = f$ and part (1) yields $f = e$.

(6) The claim is that each element $a \in G$ has a unique right inverse in G. We know that a has at least one right inverse, say x. We have $a \circ x = e$. We are to show: if $a \circ y = e$, then $y = x$ (here $y \in G$). Suppose then $a \circ x = e$ and $a \circ y = e$. We obtain

$$
\begin{aligned}
x \circ a &= e && \text{(by part (3))} \\
(x \circ a) \circ y &= e \circ y \\
x \circ (a \circ y) &= e \circ y \\
x \circ e &= e \circ y \\
x &= e \circ y \\
x &= y && \text{(by part (4)).}
\end{aligned}
$$

This proves part (6).

(7) and (8) Let $a \in G$ and let x be the unique right inverse of a. From part (3), we know that x is a left inverse of a, so that $x \circ a = e$. We must prove: if $x \circ a = e$ and $y \circ a = e$, then $y = x$. Suppose then $x \circ a = e$ and $y \circ a = e$. Then

$$
a \circ x = e
$$

$$
\begin{aligned}
y \circ (a \circ x) &= y \circ e \\
(y \circ a) \circ x &= y \\
e \circ x &= y \\
x &= y.
\end{aligned}
$$

This completes the proof. □

According to Lemma 7.3, a group $(G, \circ)$ has one and only one right identity, which is also the unique left identity. Therefore, we can refer to it as *the* identity of the group, without mentioning right or left. Similarly, since any $a \in G$ has a unique right inverse, which is also the unique left inverse of a, we may call it *the* inverse of a. The inverse of a is uniquely determined by a; for this reason, we introduce a notation displaying the fact that it depends on a alone. We write a^{-1} for the inverse of a (read: a inverse). Thus a^{-1} is the unique element of G such that $a \circ a^{-1} = a^{-1} \circ a = e$, where e is the identity of the group.

The group axioms, as presented in Definition 7.2, assert the existence of a right identity, and a right inverse of each element. We proved in Lemma 7.3 that a right identity is also a left identity and a right inverse of an element is also a left inverse of the same element. One could give an alternative definition of a group by so modifying the axioms that they assert the existence of a left identity, and a left inverse of each element. A lemma analogous to Lemma 7.3 would prove then that there is a unique left identity, which is also a unique right identity and that each element has a unique left inverse, which is also a unique right inverse of that element. Thus the existence of a right identity plus right inverses lead to the same algebraic structure (group) as the existence of a left identity plus left inverses.

However, existence of a right identity and the existence of left inverses do not always produce a group. For example, consider the set $\mathbb{Z} \times \mathbb{Z}$. For any (a,b), $(c,d) \in \mathbb{Z} \times \mathbb{Z}$, we put $(a,b) \Delta (c,d) = (a, b + d)$. Let us check if $(\mathbb{Z} \times \mathbb{Z}, \Delta)$ is a group.

(i) Δ is a binary operation on $\mathbb{Z} \times \mathbb{Z}$ since $a \in \mathbb{Z}$, $b + d \in \mathbb{Z}$ whenever $a,b,c,d \in \mathbb{Z}$. So $\mathbb{Z} \times \mathbb{Z}$ is closed under Δ.

(ii) Is Δ associative? For any (a,b), (c,d), $(e,f) \in \mathbb{Z} \times \mathbb{Z}$, we ask

$$
\begin{aligned}
[(a,b) \Delta (c,d)] \Delta (e,f) &\stackrel{?}{=} (a,b) \Delta [(c,d) \Delta (e,f)] \\
(a,b + d) \Delta (e,f) &\stackrel{?}{=} (a,b) \Delta (c,d + f)
\end{aligned}
$$

$$(a, (b + d) + f) \stackrel{?}{=} (a, b + (d + f))$$

Yes, this is true since + is an associative operation on $\mathbb{Z}$. Hence Δ is associative.

(iii) Is there an element in $\mathbb{Z} \times \mathbb{Z}$, (a_0,b_0) say, such that

$$(a,b) \Delta (a_0,b_0) = (a,b) \text{ for all } (a,b) \in \mathbb{Z} \times \mathbb{Z}?$$

Well, this is true if and only if $(a,b + b_0) = (a,b)$, which is equivalent to $b_0 = 0$. There is no condition on a_0. For example,

$$(a,b) \Delta (0,0) = (a,b + 0) = (a,b)$$
$$(a,b) \Delta (1,0) = (a,b + 0) = (a,b)$$

for all $(a,b) \in \mathbb{Z} \times \mathbb{Z}$, so (0,0) and (1,0) are right identities. In fact, any $(n,0) \in \mathbb{Z} \times \mathbb{Z}$ is a right identity.

From Lemma 7.3, we know that a group has one and only one right identity. so $\mathbb{Z} \times \mathbb{Z}$ is not a group under Δ. On the other hand, with respect to (0,0) for example (in fact, with respect to any right identity), each element (a,b) of $\mathbb{Z} \times \mathbb{Z}$ has a left inverse $(0,-b)$:

$$(0,-b) \Delta (a,b) = (0,-b + b) = (0,0)$$

(with respect to $(n,0)$, a left inverse of (a,b) is $(n,-b)$).

So $(\mathbb{Z} \times \mathbb{Z}, \Delta)$ is a system in which a right identity exists, plus a left inverse of each element; nevertheless, it fails to be a group. Likewise, fulfilling the existence of a left identity and right inverses is not enough for building a group.

We could define a group by including the claims of Lemma 7.3 directly into the definition. Then we would have

(iii)′ there is a unique $e \in G$ such that

$$a \circ e = e \circ a = a \text{ for all } a \in G$$

and

(iv)′ for all $a \in G$, there is a unique $a^{-1} \in G$ such that

$$a \circ a^{-1} = e = a^{-1} \circ a$$

in place of (iii) and (iv) of Definition 7.2. Some textbooks define groups in this way. This would save us from the trouble of proving Lemma 7.3. Why, then, did we not use this definition? Because we do not want to do unnecessary work. If we defined groups by (iii)′ and (iv)′ instead of (iii) and (iv), then, each time when we wanted to show that a set G builds a group under a binary operation $\circ$ on G, we had to check

1) that there is an $e \in G$ such that $a \circ e = a$ for all $a \in G$,
2) that this e is also such that $e \circ a = a$ for all $a \in G$,

3) that e is the unique element of G with these two properties,

4) that for each $a \in G$, there is an $a^{-1} \in G$ such that $a \circ a^{-1} = e$,

5) that $a^{-1} \circ a = e$ as well,

6) that this a^{-1} is the unique element of G with $a \circ a^{-1} = e = a^{-1} \circ a$,

which more than doubles our work. With our Definition 7.2, we need check only 1) and 4). The other items 2),3),5),6) follow from 1) and 4) automatically. We pay for our comfort by having to prove Lemma 7.3, but, once this is over, we have less work to do in order to see whether a given set G forms a group under a given operation $\circ$ on it, as in the following examples.

7.4 **Examples:** (a) For any two elements a,b of $\mathbb{Q}\backslash\{1\}$, we put $a \circ b = ab - a - b + 2$. We ask if $\mathbb{Q}\backslash\{1\}$ is a group under $\circ$. Let us check the group axioms.

(i) For all $a,b \in \mathbb{Q}\backslash\{1\}$, we observe $a \circ b = ab - a - b + 2 \in \mathbb{Q}$, but this is not enough. We must prove $a \circ b \neq 1$ also. Let $a,b \in \mathbb{Q}$, $a \neq 1 \neq b$. We suppose $a \circ b = 1$ and try to reach a contradiction. If $a \circ b = 1$, then

$$\begin{aligned} ab - a - b + 2 &= 1 \\ ab - a - b + 1 &= 0 \\ (a-1)(b-1) &= 0 \\ a - 1 = 0 \text{ or } b - 1 &= 0 \\ a = 1 \text{ or } b &= 1, \end{aligned}$$

a contradiction. So $a \circ b \in \mathbb{Q}\backslash\{1\}$ and $\circ$ is a binary operation on $\mathbb{Q}\backslash\{1\}$.

(ii) For all $a,b,c \in \mathbb{Q}\backslash\{1\}$, we ask if $(a \circ b) \circ c = a \circ (b \circ c)$.

$$\begin{aligned} (ab - a - b + 2) \circ c &\stackrel{?}{=} a \circ (bc - b - c + 2) \\ (ab-a-b+2)c-(ab-a-b+2)-c+2 &\stackrel{?}{=} a(bc-b-c+2)-a-(bc-b-c+2)+2 \\ abc-ac-bc+2c-ab+a+b-2-c+2 &\stackrel{?}{=} abc-ab-ac+2a-a-bc+b+c-2+2 \end{aligned}$$

The answer is "yes." So $\circ$ is associative.

(iii) We are looking for an $e \in \mathbb{Q}\backslash\{1\}$ such that $a \circ e = a$ for all $a \in \mathbb{Q}\backslash\{1\}$. Assuming such an e exists, we get

$$\begin{aligned} ae - a - e + 2 &= a \\ ae - a &= 2a - 2 \\ (a-1)e &= 2(a-1) \\ e &= 2 \qquad (\text{since } a - 1 \neq 0). \end{aligned}$$

We have not proved that $2 \in \mathbb{Q}\backslash\{1\}$ is a right identity element. We showed only that a right identity element, if it exists at all, has to be 2. Let us see if 2 is really a right identity. We observe

$$a \circ 2 = a2 - a - 2 + 2 = 2a - a = a$$

far all $a \in \mathbb{Q}\backslash\{1\}$. Since $2 \in \mathbb{Q}\backslash\{1\}$, 2 is indeed a right identity in $\mathbb{Q}\backslash\{1\}$.

(iv) For all $a \in \mathbb{Q}\backslash\{1\}$, we must find an $x \in \mathbb{Q}\backslash\{1\}$ such that $a \circ x = 2$. Well, this gives

$$\begin{aligned} ax - a - x + 2 &= 2 \\ ax - a - x + 1 &= 1 \\ (a - 1)(x - 1) &= 1 \\ x - 1 &= 1/(a - 1) \\ x &= a/(a - 1), \end{aligned}$$

which is meaningful since $a \neq 1$. We have not proved yet that $a/(a - 1)$ is a right inverse of a. We showed only that a right inverse of $a \in \mathbb{Q}\backslash\{1\}$, if it exists at all, has to be $a/(a - 1)$. We must now show that $a \circ a/(a - 1) = 2$ for all $a \in \mathbb{Q}\backslash\{1\}$ and also that $a/(a - 1) \in \mathbb{Q}\backslash\{1\}$. Good. We have

$$\begin{aligned} a \circ a/(a - 1) &= a(a/(a - 1)) - a - (a/(a - 1)) + 2 \\ &= (a - 1)(a/(a - 1)) - a + 2 \\ &= 2, \end{aligned}$$

and also $a/(a - 1) \neq 1$, for $a/(a - 1) \in \mathbb{Q}$ and $a/(a - 1) = 1$ would imply that $a = a - 1$, hence $0 = 1$, which is absurd.

Since all the group axioms hold, $\mathbb{Q}\backslash\{1\}$ is a group under $\circ$.

(b) Let us define an operation $*$ on $\mathbb{Z}$ by putting $a * b = a + b + 2$ for all $a,b \in \mathbb{Z}$. Does $\mathbb{Z}$ form a group under $*$?

(i) For any $a,b \in \mathbb{Z}$, $a * b = a + b + 2$ is an integer. So $\mathbb{Z}$ is closed under $*$.

(ii) For all $a,b,c \in \mathbb{Z}$, we ask if $(a * b) * c = a * (b * c)$. We have

$$\begin{aligned} (a * b) * c &= (a + b + 2) * c \\ &= (a + b + 2) + c + 2 \\ &= a + (b + 2 + c) + 2 \\ &= a + (b + c + 2) + 2 \\ &= a + (b * c) + 2 \\ &= a * (b * c). \end{aligned}$$

So $*$ is associative.

(iii) Is there an integer $e \in \mathbb{Z}$ such that $a * e = a$ for all $a \in \mathbb{Z}$? Well, this gives $a + e + 2 = a$ and $e = -2$. Let us check whether -2 is really a right identity element. We observe that $a * -2 = a + (-2) + 2 = a$ for all $a \in \mathbb{Z}$. So -2 is a right identity element.

(iv) Does each integer a have a right inverse in $\mathbb{Z}$? The condition $a * x = -2$ yields

$$a + x + 2 = -2$$

$$x = -4 - a \in \mathbb{Z}.$$

$-4 - a$ is indeed a right inverse of a since $a * (-4 - a) = a + (-4 - a) + 2 = -2$. Therefore $\mathbb{Z}$ is a group with respect to $*$.

(c) Let A be a nonempty set and let $\mathcal{A}$ be the set of all subsets of A. The elements of $\mathcal{A}$ are thus subsets of A. Consider the forming of symmetric differences (§1, Ex.7). $\mathcal{A}$ is a group under Δ:

(i) For all $S,T \in \mathcal{A}$, $S \Delta T$ is a subset of A, so $S \Delta T \in \mathcal{A}$ and $\mathcal{A}$ is closed under Δ.

(ii) Δ is associative (§1, Ex.8).

(iii) $\varnothing$ is a right identity (§1, Ex.8).

(iv) Each element S of $\mathcal{A}$ has a right inverse, namely S itself, as $S \Delta S = \varnothing$ for all $S \in \mathcal{A}$ (§1, Ex.8).

So $\mathcal{A}$ is a group under Δ.

We have seen many examples of groups. In some of the groups $(G, \circ)$, the underlying set G is infinite, in some finite. The number of elements of G, more precisely the cardinality of G, is called the *order* of the group $(G, \circ)$. We denote the order of $(G, \circ)$ by $|G|$. A group $(G, \circ)$ is called a *finite* group if $|G|$ is finite, and an *infinite* group if $|G|$ is infinite. One might distinguish between various infinite cardinalities, but we will not do so in this book. When the order of a group $(G, \circ)$ is infinite, we write $|G| = \infty$. The symbol ∞ will stand for all types of infinities.

A. Cayley (1821-1895) introduced a convenient device for investigating groups. Let $(G, \circ)$ be a finite group. We make a table that displays $a \circ b$ for each $a,b \in G$. We divide a square into $|G|^2$ parts by dividing the sides into $|G|$ parts. Each one of the rows will be indexed by an element of the group, usually written on the left of the row. Likewise, each one of the columns will be indexed by an element of the group, usually written above the column. Each element will index only one row and only one column. It is customary to use the same ordering of the elements to index the rows and columns. Also, the first row and the first column are customarily indexed by the identity element of the group. In the cell where the row of $a \in G$ and $b \in G$ meet, we write down $a \circ b$. This square is known as the *Cayley table* or the *operation table* (*multiplication* or *addition table*, as the case may be) *of the group* $(G, \circ)$.

As an illustration, we give the addition table of $(\mathbb{Z}_4,+)$ below. $(\mathbb{Z}_4,+)$ is a group by Example 7.1(c). We drop the bars for convenience.

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

We observe in this table that every element of $\mathbb{Z}_4$ appears once in each row and also in each column. This is a general property of groups: if $(G,\circ)$ is a group, then every element b of G appears once and only once in the row of any $a \in G$, say in the cell where the row of a and the column of $x \in G$ meet. A similar assertion holds for columns. This is the content of the next lemma.

7.5 Lemma: *Let $(G,\circ)$ be a group and $a,b \in G$.*
(1) *There is one and only one $x \in G$ such that $a \circ x = b$.*
(2) *There is one and only one $y \in G$ such that $y \circ a = b$.*

Proof: (1) We prove first that there can be at most one $x \in G$ such that $a \circ x = b$. Let $a \circ x = b = a \circ x_1$. We prove $x = x_1$. We have

$$\begin{aligned} a \circ x &= a \circ x_1 \\ a^{-1} \circ (a \circ x) &= a^{-1} \circ (a \circ x_1) \\ (a^{-1} \circ a) \circ x &= (a^{-1} \circ a) \circ x_1 \\ e \circ x &= e \circ x_1 \\ x &= x_1 \end{aligned}$$

by Lemma 7.3. So there can be at most one x with $a \circ x = b$.

The existence of at least one such x is easily seen when we put $x = a^{-1} \circ b$. Indeed, $a \circ (a^{-1} \circ b) = (a \circ a^{-1}) \circ b = e \circ b = b$.
So there is one and only one element x of G, namely $x = a^{-1} \circ b$, such that $a \circ x = b$. This proves (1).

The proof of (2) is similar and is left to the reader. □

We give an application of Lemma 7.5. We determine the Cayley table of groups of order 3. Let $(\{e,a,b\}, \circ)$ be a group of order 3, where e is the identity. The Cayley table of this group contains the information given in Figure 1. Now we fill the remaining four cells. What is $a \circ a$? The cell * cannot contain a, for a would otherwise appear more than once in the

$\circ$	e	a	b
e	e	a	b
a	a	*	
b	b		

Figure 1

$\circ$	e	a	b
e	e	a	b
a	a	b	
b	b		

Figure 2

$\circ$	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

Figure 3

second row (or column). So the cell * contains e or b. If it contained e, then the third entry in the second row had to be b and b would appear at least twice in the third column, contrary to Lemma 7.5. This leaves only the possibility $a \circ a = b$. Then we have the table in Figure 2. The remaining cells are necessarily filled in as in Figure 3.

We did not prove that Figure 3 is a Cayley table of a group of order 3. At this stage, we do not even know whether a group of order 3 exists. We proved: if there is a group of order 3 at all, then its Cayley table is the table of Figure 3. We now prove the existence of a group of order 3. We use Figure 3. Let $\{e,a,b\}$ be a set of 3 elements, and let the binary operation $\circ$ on this set be *defined* as in Figure 3. It is easy to check the group axioms (i),(iii),(iv). It remains to check associativity. We must verify $3.3.3 = 27$ equations $(x \circ y) \circ z = x \circ (y \circ z)$, where $x,y,z \in \{e,a,b\}$. An equation of this type is true when one of x,y,z is equal to e. So we are left with $2.2.2 = 8$ equations

$$(a \circ a) \circ a = a \circ (a \circ a) \qquad (b \circ a) \circ a = b \circ (a \circ a)$$
$$(a \circ a) \circ b = a \circ (a \circ b) \qquad (b \circ a) \circ b = b \circ (a \circ b)$$
$$(a \circ b) \circ a = a \circ (b \circ a) \qquad (b \circ b) \circ a = b \circ (b \circ a)$$
$$(a \circ b) \circ b = a \circ (b \circ b) \qquad (b \circ b) \circ b = b \circ (b \circ b)$$

and these are verified easily. Hence $(\{e,a,b\}, \circ)$ is a group. There is a group of order 3. Any two groups of order 3 have essentialy the same Cayley table, namely the table in Figure 3. This statement will be made precise in §20.

The Cayley tables of $(\mathbb{Z}_4,+)$ and $(\{e,a,b\}, \circ)$ are symmetric about the principal diagonal (that joins the upper-left and lower-right cells). What does this signify? The symmetry of the Cayley table of a group $(G, \circ)$ means that the cell where the i-th row and j-th column meet has the same entry as the cell where the j-th row and i-th column meet, and this for all $i,j = 1,2, \ldots, |G|$. Assuming the i-th row is the row of $a \in G$ and the j-th column is the column of $b \in G$ (and assuming we index the rows and columns by the elements of G in the same order), this means: $a \circ b = b \circ a$ for all $a,b \in G$. So the group is commutative in the following sense.

7.6 Definition: A group $(G, \circ)$ is called a *commutative* group or an *abelian* group, if, in addition to the group axioms (i)-(iv), a fifth axiom

(v) $a \circ b = b \circ a$ for all $a,b \in G$

holds.

A binary operation on a set G is called *commutative* when $a \circ b = b \circ a$ for all $a,b \in G$. So a commutative group is one where the operation is commutative. The term "abelian" is used in honor of N. H. Abel, a Norwegian mathematician (1802-1829).

We close this paragraph with some comments on the group axioms. The reader might ask why we should study the structures $(G, \circ)$ where $\circ$ satisfies the axioms (i),(ii),(iii),(iv). Why do we not study structures $(G, \circ)$ where $\circ$ satisfies the axioms (i),(iii),(iv),(v) or (i),(ii),(iii),(v)? What is the reason for preferring the axioms (i),(ii),(iii),(iv) to some other combination of (i),(ii),(iii),(iv),(v)? There is of course no reason why other combinations ought to be excluded from study. As a matter of fact, all combinations have a proper name and there are theories about them. However, they are very far from having the same importance as the combination (i),(ii), (iii),(iv).

A mathematical theory, if it deserves to be considered important, has to possess both generality and informative significance. Clearly, a theory whose axioms are too restrictive to hold in a variety of cases is bound to be insignificant for those who cannot fulfill them in their area of study, and the theory will have limited interest. An interesting theory is a general one. But generality costs content. When we wish that the axioms of a theory be fulfilled in diverse areas and in many contexts, we must also realize that the theory can only deal with what is common in these diverse areas, and this might be nil. There we have the danger that the theory will degenerate into a list of uninformative paraphrases of the axioms without substance. Imposing restrictions on the axioms diminishes the use and interest of a theory, and lifting restrictions tends to make the theory void. The balance between generality and content is very delicate. Group theory is one of the cases where this balance is attained successfully. Group theory has applications in literally every branch of mathematics, both pure and applied, as well as in theoretical physics and other sciences, and it is a theory full of deep, interesting, beautiful results. This is why the choice (i),(ii),(iii),(iv) is judicious. Other combinations of the axioms are not as fruitful as (i),(ii),(iii),(iv).

Exercises

1. Determine whether the following sets build groups with respect to the operations given. In each case, state which group axioms are satisfied.

(a) $\mathbb{R}$ under subtraction, multiplication and division.

(b) $\mathbb{R}\backslash\{0\}$, $\mathbb{Q}\backslash\{0\}$, $\mathbb{C}\backslash\{0\}$ under multiplication.

(c) $\{0,1\}$, $\{-1,1\}$ under multiplication.

(d) $\{z \in \mathbb{C}: |z| \leqslant 1\}$ under multiplication.

(e) $\{z \in \mathbb{C}: |z| = 1\}$ under multiplication.

(f) $5\mathbb{Z} = \{5z \in \mathbb{Z}: z \in \mathbb{Z}\}$ under multiplication and addition.

(g) $\{x\}$ under $\circ$, where $x \circ x = x$.

(h) $\{(t,u) \in \mathbb{Z} \times \mathbb{Z} : t^2 - 5u^2 = 4\}$ under $*$, where $*$ is defined by

$$(t_1,u_1) * (t_2,u_2) = \left(\frac{t_1t_2 + 5u_1u_2}{2}, \frac{t_1u_2 + t_2u_1}{2}\right)$$

for all $(t_1,u_1),(t_2,u_2)$ in this set.

(i) $\mathbb{Z}_6$ and $\mathbb{Z}_8$ under multiplication and addition.

(j) $\mathbb{Z}_7$ and $\mathbb{Z}_7\backslash\{\bar{0}\}$ under multiplication.

(k) $\{f,g\}$ under the composition of mappings, where $f: x \to x$ and $g: x \to 1/(1 - x)$ are functions from $\mathbb{R}\backslash\{1\}$ into $\mathbb{R}\backslash\{1\}$.

(l) $\{f,g,h\}$ under the composition of mappings, where $f: x \to x$ and $g: x \to 1/(1 - x)$ and $h: x \to (x - 1)/x$ are functions from $\mathbb{R}\backslash\{1,0\}$ into $\mathbb{R}\backslash\{1,0\}$.

(m) $\{f_{a,b}: a,b \in \mathbb{R}, a \neq 0\}$ under the composition of mappings, where $f_{a,b}$ is defined by $f_{a,b}(x) = ax + b$ as a function from $\mathbb{R}$ into $\mathbb{R}$.

2. For which $m \in \mathbb{N}$ is the set $\mathbb{Z}_m\backslash\{0\}$ a group under multiplication?

§8
Conventions and Some Computational Lemmas

In our study of groups, we are interested in how $a \circ b$ depends on a and b; not in the name or sign of the operation. For this reason, we suppress the operation sign altogether and use juxtaposition. Henceforward, we will write ab (and also occasionally $a \cdot b$) for $a \circ b$. We will refer to the operation as *multiplication*. Thus "multiplication" will be used in a broad sense. It can mean the usual multiplication of numbers, but also the composition of mappings, the taking of symmetric differences of two sets, or some rather artificial operation like those in Example 7.4. With this convention, there is no need to refer to the operation all the time when we discuss groups. So we call the *set* G a group, instead of the ordered pair $(G, \circ)$ (we keep in mind of course that there can be many groups on the same set). We say then that the group is written *multiplicatively* or that G is a *multiplicative* group. Conforming to this, we call ab the *product of a and b*. Also, we write 1 for the identity element of the group. Thus 1 is not necessarily the number one. It is perhaps the identity mapping, perhaps the empty set, perhaps some other object. What it is depends on the group we are investigating. But a warning: we will *not* write $\frac{1}{a}$ for the inverse a^{-1} of an element a in a group.

This is the multiplicative notation for groups. Sometimes, we shall use the *additive* notation, too, especially when the group is commutative. Then the operation is denoted by "+" and is called *addition*. Like "multiplication", "addition" is used in a general sense. We call $a + b$ the *sum of a and b*. When we have an *additive* group, the identity element of the group will be written as 0. So 0 is not necessarily the number zero. Also, we write $-a$ for the inverse of an element a in an additively written group. We call $-a$ the *opposite of a*.

8.1 Lemma: *Let G be a group and let $a,b,c \in G$.*
(1) *If $ab = ac$, then $b = c$* (left cancellation).
(2) *If $ba = ca$, then $b = c$* (right cancellation).

Proof: (1) If $ab = ac$, we multiply by a^{-1} on the left and get $a^{-1}(ab) = a^{-1}(ac)$. Using associativity, we obtain $(a^{-1}a)b = (a^{-1}a)c$. So $1b = 1c$. Since 1 is the identity element of G, we finally get $b = c$.

(2) The proof of (2) is similar and is left to the reader. □

We must be careful when we want to use Lemma 8.1 to make cancellation. If the group is not commutative, left multiplication by an element and right multiplication by the same element give in general different results. In the proof of Lemma 8.1, we multiplied by a^{-1} on the same side. We cannot conclude $b = c$ from $ab = ca$, for instance. Indeed, we have

$$ab = ca \implies a^{-1}(ab) = a^{-1}(ca) \implies (a^{-1}a)b = (a^{-1}c)a \implies b = a^{-1}ca$$

and this is all we can say. In general, $a^{-1}ca \neq c$, so $b \neq c$. You must always make sure that you cancel on the same side.

Cancellations are multiplications by inverse elements. We now evaluate the inverse of an inverse, and the inverse of a product.

8.2 Lemma: *Let G be a group and let $a,b \in G$. Then*
(1) $(a^{-1})^{-1} = a$,
(2) $(ab)^{-1} = b^{-1}a^{-1}$.

Proof: (1) $aa^{-1} = 1$ by the definition of a^{-1}. So a is a left inverse of a^{-1}. So a is the inverse of a^{-1}(Lemma 7.3).

(2) $(ab)(b^{-1}a^{-1}) = a(b(b^{-1}a^{-1})) = a((bb^{-1})a^{-1}) = a(1a^{-1}) = aa^{-1} = 1$, and so $b^{-1}a^{-1}$ is the inverse of ab. □

Therefore, the inverse of the inverse of an element is the element itself. Also, the inverse of a product is the product of the inverses, but in the *reverse* order. Do *not* write $(ab)^{-1} = a^{-1}b^{-1}$. This is wrong unless $a^{-1}b^{-1} = b^{-1}a^{-1}$, which is equivalent to $ab = ba$ (why?) and which is not true in general.

*

* *

We defined the product of *two* elements. The product of a and b is ab. We now want to define the product of n elements and prove that the usual exponentiation rules are valid. The rest of this paragraph is extremely dull. The reader may just glance at the assertions and skip the proofs if she (or he) wishes.

By the product of three elements a,b,c in a group G, we understand an element abc of G. Let us recall we agreed to denote by abc the element $(ab)c = a(bc)$. So the product of a,b,c in this order is evaluated by two successive multiplications. Either we evaluate ab first, then multiply it by c, or we evaluate bc first, then multiply a by it. In either way, we get the same result by associativity and this result is denoted by abc, without parentheses.

Now let us consider the product of four elements a,b,c,d. Their product in this order will be defined by three successive multiplications of two elements. This can be done in five distinct ways:

$$a(b(cd)),\ a((bc)d),\ (ab)(cd),\ ((ab)c)d,\ (a(bc))d,$$

but these five products are all equal by associativity. The first two products are equal since $b(cd) = (bc)d$. The last two products are equal since $(ab)c = a(bc)$. Further, we have $a(b(cd)) = (ab)(cd)$ [put $cd = e$, then $a(be) = (ab)e$] and $(ab)(cd) = ((ab)c)d$ [put $ab = f$, then $f(cd) = (fc)d$]. So the five products are equal. This renders it possible to drop the parentheses and write simply $abcd$. This is the product of a,b,c,d in the given order.

More generally, we want to define the product of n elements $a_1,a_2,\ldots,a_n$ in a group G ($n > 2$). The product of $a_1,a_2,\ldots,a_n$ will be defined by $n - 1$ successive multiplications of two elements. By inserting parentheses in all possible ways, we obtain many products (their exact number is $2.4..(4n - 6)/n!$), but associativity assures that these products are equal. Now we prove this. In view of some later applications, the following lemma is stated more generally than for groups.

8.3 Lemma: *Let G be a nonempty set and let there be defined an associative binary operation on G, denoted by juxtaposition. Let*

$a_1, a_2, \ldots, a_n \in G$. *Then the products of* $a_1, a_2, \ldots, a_n$ *are independent of the mode of putting parentheses. This means the following. We define*

$P_1(a_1) = \{a_1\}$

$P_2(a_1, a_2) = \{a_1 a_2\}$

$P_3(a_1, a_2, a_3) = \{(a_1 a_2)a_3, a_1(a_2 a_3)\}$

$= \{xy: x \in P_1(a_1), y \in P_2(a_2, a_3) \text{ or } x \in P_2(a_1, a_2), y \in P_1(a_3)\}$

$P_4(a_1, a_2, a_3, a_4) = \{a_1(a_2(a_3 a_4)), a_1((a_2 a_3)a_4), (a_1 a_2)(a_3 a_4), ((a_1 a_2)a_3)a_4, (a_1(a_2 a_3))a_4\}$

$= \{xy: x \in P_1(a_1), y \in P_3(a_2, a_3, a_4) \text{ or } x \in P_2(a_1, a_2), y \in P_2(a_3, a_4) \text{ or } x \in P_3(a_1, a_2, a_3), y \in P_1(a_4)\}$

.....

$P_k(a_1, a_2, \ldots, a_k) = \{xy: x \in P_i(a_1, a_2, \ldots, a_i), y \in P_{k-i}(a_{i+1}, \ldots, a_k) \text{ for some } i = 1, 2, \ldots, k\}$

for $k = 1, 2, \ldots, n$. *Thus* P_k *are subsets of* G *whose elements are the products of* $a_1, a_2, \ldots, a_k$, *reduced to* $k - 1$ *successive multiplications of two elements in* G.

Claim: For all $n \in \mathbb{N}$ *and for all* $a_1, a_2, \ldots, a_n \in G$, *the set* $P_n(a_1, a_2, \ldots, a_n)$ *contains one and only one element.*

Proof: The proof will be by induction on n (in the form 4.5). For $n = 1, 2$, it is evident that $P_1(a_1)$, $P_2(a_1, a_2)$ each have exactly one element. For $n = 3$, the claim is just the associativity of multiplication. For $n = 4$, the argument preceding the lemma proves the claim. Notice that we used only the associativity of multiplication there.

Suppose $n \geq 5$ and the lemma is proved for $1, 2, \ldots, n - 1$. Let $u, v \in P_n(a_1, a_2, \ldots, a_n)$. We are to prove $u = v$. By the definition of $P_n(a_1, a_2, \ldots, a_n)$, we have $u = xy$, $v = st$, where

$$x \in P_i(a_1, a_2, \ldots, a_i),\ y \in P_{n-i}(a_{i+1}, \ldots, a_n),\ i \in \mathbb{N},\ 1 \leq i \leq n-1,$$
$$s \in P_j(a_1, a_2, \ldots, a_j),\ t \in P_{n-j}(a_{j+1}, \ldots, a_n),\ j \in \mathbb{N},\ 1 \leq j \leq n-1.$$

We prove $u = v$ first under the assumption $i = j$. By induction, the set $P_i(a_1, a_2, \ldots, a_i)$ contains one and only one element. Hence $x = s$. Also,

applying the induction hypothesis to $n - i$, with the elements $a_{i+1},\ldots,a_n$, we conclude that $P_{n-i}(a_{i+1},\ldots,a_n)$ has one and only one element. This gives $y = t$. Then we get $u = xy = sy = st = v$. So the claim is proved in case $i = j$.

Now suppose $i \neq j$. Without losing generality, we assume $i < j$. We put $j = i + h$, with $h \in \mathbb{N}$. Now apply the induction hypothesis to j, with the elements $a_1,\ldots,a_j$. There is a unique element in $P_j(a_1,\ldots,a_j)$, which we called s. Also by induction, applied to i with the elements $a_1,\ldots,a_i$, there is a unique element in $P_i(a_1,a_2,\ldots,a_i)$, namely x. Again by induction, applied to h with the elements $a_{i+1},\ldots,a_j$ there is a unique element in $P_h(a_{i+1},\ldots,a_j)$, say b. By the definition of $P_j(a_1,\ldots,a_j)$, we have $xb \in P_j(a_1,\ldots,a_j)$, so $xb = s$.

We have $n - i = h + (n - j)$. By induction, applied to $n - i$ with the elements $a_{i+1},\ldots,a_n$, the set $P_{n-i}(a_{i+1},\ldots,a_n)$ has one and only one element, which we called y. Also by induction, applied to h with the elements $a_{i+1},\ldots,a_j$, there is a unique element in $P_h(a_{i+1},\ldots,a_j)$, namely b. Again by induction, applied to $n - j$ with the elements $a_{j+1},\ldots,a_n$, the set $P_{n-j}(a_{j+1},\ldots,a_n)$ has a unique element, namely t. By the definition of $P_{n-i}(a_{i+1},\ldots,a_{i+h},a_{j+1},\ldots,a_n)$, we have $bt \in P_{n-i}(a_{i+1},\ldots,a_n)$, so $bt = y$.

Thus $xb = s$ and $bt = y$. This gives $u = xy = x(bt) = (xb)t = st = v$. This completes the proof. □

8.4 Definition: The unique element in $P_n(a_1,a_2,\ldots,a_n)$ of Lemma 8.3 is called the *product of* $a_1,a_2,\ldots,a_n$ (in this order) and is denoted by $a_1a_2\ldots a_n$ or by $\prod_{i=1}^{n} a_i$.

So the product of n elements in a given order can be written without parentheses. This simplifies the notation enormously.
Using the notation of Definition 8.4, we can reformulate Lemma 8.3 as follows. If G is a nonempty set with an associative multiplication on it, and if $a_1, a_2, \ldots, a_n \in G$, then

$$a_1(a_2\ldots a_n) = (a_1a_2)(a_3\ldots a_n) = (a_1a_2a_3)(a_4\ldots a_n) = \cdots = (a_1a_2\ldots a_{n-1})a_n = a_1a_2\ldots a_n.$$

We write a^n for $a_1a_2\ldots a_n$ in case $a_1, a_2, \ldots, a_n$ are all equal to $a \in G$, $n \in \mathbb{N}$. In particular, $a^1 = a$. We have $a^n = a^{n-1}a = aa^{n-1}$. More generally, the above reformulation of Lemma 8.3 gives

$$a^m a^n = a^{m+n}, \text{ for all } a \in G \text{ and } m,n \in \mathbb{N}. \qquad (*)$$

In particular, $(a^m)^2 = a^m a^m = a^{m+m} = a^{2m} = a^{m2}$. We prove more generally $(a^m)^n = a^{mn}$ by induction on n. The case $n = 1$ is trivial and the case $n = 2$ has just been shown. Assume now $n \geqslant 3$ and $(a^m)^{n-1} = a^{m(n-1)}$ for all $a \in G$. We want to show $(a^m)^n = a^{mn}$ for all $a \in G$. We have $(a^m)^n = (a^m)^{1+(n-1)} = (a^m)^1(a^m)^{n-1} = a^m(a^m)^{n-1}$ by $(*)$, with $a^m, 1, n-1$ in place of a,m,n, respectively. Then we get $(a^m)^n = a^m a^{m(n-1)} = a^m a^{mn-m} = a^{m+(mn-m)}$ by $(*)$, with $a, m, mn-n$ in place of a,m,n, respectively. This gives $(a^m)^n = a^{mn}$. Thus we proved the

8.5 Lemma: *If there is an associative multiplication on a nonempty set G, denoted by juxtaposition, then*

$$a^m a^n = a^{m+n} \text{ and } (a^m)^n = a^{mn} \text{ for all } a \in G \text{ and } m,n \in \mathbb{N}. \qquad \square$$

Lemma 8.5 can be extended to arbitrary integral powers in the case of groups. We give the relevant definitions.

8.6 Definition: Let G be a group, $a \in G$, $m \in \mathbb{N}$. We put

$$a^0 = 1 = \text{identity of } G, \text{ and } a^{-m} = (a^m)^{-1} = \text{inverse of } a^m.$$

8.7 Lemma: *Let G be a group. Then*

(1) $a^m a^n = a^{m+n}$;

(2) $(a^{-1})^m = a^{-m}$;

(3) $(a^m)^n = a^{mn}$;

for all $a \in G$ and $m,n \in \mathbb{Z}$.

Proof: (1) We prove $a^m a^n = a^{m+n}$. If $m \geqslant 1$, $n \geqslant 1$, Lemma 8.5 yields the result. If $m = 0$, then $a^0 a^n = 1a^n = a^n = a^{0+n}$ for all $n \in \mathbb{Z}$; and if $n = 0$, then $a^m a^0 = a^m 1 = a^m = a^{m+0}$ for all $m \in \mathbb{Z}$. So we have

$$a^m a^n = a^{m+n} \text{ whenever } m,n \geqslant 0. \qquad (e)$$

We must prove this relation also when $m \geqslant 0$, $n \leqslant 0$; $m \leqslant 0$, $n \geqslant 0$; $m \leqslant 0$, $n \leqslant 0$. Changing our notation (replacing m,n by $|m|, |n|$) we must prove (i) $a^m a^{-n} = a^{m-n}$; (ii) $a^{-m}a^n = a^{-m+n}$; (iii) $a^{-m}a^{-n} = a^{-m+(-n)}$ for all $m,n \geqslant 0$.

(i) Let $m,n \geqslant 0$. If $m \geqslant n$, then $a^{m-n}a^n = a^m$ by (e). Multiplying by $(a^n)^{-1} = a^{-n}$ on the right, we get $a^{m-n} = a^m a^{-n}$ if $m \geqslant n$. Taking the inverses of both sides of this equation, we get, in case $m \geqslant n$, $a^n a^{-m} = [(a^n)^{-1}]^{-1}(a^m)^{-1} = [a^m(a^n)^{-1}]^{-1} = (a^m a^{-n})^{-1} = (a^{m-n})^{-1} = a^{-(m-n)} = a^{-m+n}$. Interchanging m and n, we get $a^m a^{-n} = a^{-n+m} = a^{m-n}$ in case $n \geqslant m$. So $a^m a^{-n} = a^{m-n}$, irrespective of whether $m \geqslant n$ or $n \geqslant m$.

(ii) Let $m,n \geqslant 0$. If $n \geqslant m$, then $a^m a^{-m+n} = a^n$ by (e). Multiplying by $(a^m)^{-1} = a^{-m}$ on the left, we get $a^{-m+n} = a^{-m}a^n$ if $n \geqslant m$. Taking the inverses of both sides of this equation, we get, in case $n \geqslant m$, $a^{-n}a^m = a^{-n}[(a^m)^{-1}]^{-1} = (a^n)^{-1}(a^{-m})^{-1} = (a^{-m}a^n)^{-1} = (a^{-m+n})^{-1} = (a^{n-m})^{-1} = a^{-(n-m)} = a^{-n+m}$. Interchanging n and m, we get $a^{-m}a^n = a^{-m+n}$ in case $m \geqslant n$. So $a^{-m}a^n = a^{-m+n}$, irrespective of whether $m \geqslant n$ or $n \geqslant m$.

(iii) Let $m,n \geqslant 0$. We have $a^m a^n = a^{m+n}$ by (e). Taking the inverses of both sides of this equation, we get $a^{-m}a^{-n} = (a^m)^{-1}(a^n)^{-1} = (a^n a^m)^{-1} = (a^{m+n})^{-1} = a^{-(m+n)} = a^{-m+(-n)}$ for all $m,n \geqslant 0$.

Thus $a^m a^n = a^{m+n}$ for all $a \in G$ and $m,n \in \mathbb{Z}$.

(2) We prove $(a^{-1})^m = a^{-m}$. This is true if $m = 1$, since $(a^{-1})^1 = a^{-1} = (a^1)^{-1}$. Suppose now $m \in \mathbb{N}$, $m \geqslant 2$ and $(a^{-1})^{m-1} = a^{-(m-1)}$. Then $(a^{-1})^m = (a^{-1})^{m-1}(a^{-1}) = a^{-(m-1)}a^{-1} = a^{-m+1}a^{-1} = a^{-m+1-1} = a^{-m}$. So $(a^{-1})^m = a^{-m}$ for all $m \in \mathbb{N}$ by induction. It is also true when $m = 0$, as $(a^{-1})^0 = 1 = a^0 = a^{-0}$. Now we must prove it for $m < 0$. With a slight change in notation, we are prove $(a^{-1})^{-m} = a^m$ for all $m \in \mathbb{N}$. We have indeed $(a^{-1})^{-m} = [(a^{-1})^m]^{-1} = (a^{-m})^{-1} = [(a^m)^{-1}]^{-1} = a^m$. The first equality in this chain follows from Definition 8.6, with a^{-1} in place of a, the second from the fact that $(a^{-1})^m = a^{-m}$ for all $m \in \mathbb{N}$, which we just proved and the third from Definition 8.6.

So $(a^{-1})^m = a^{-m}$ for all $m \in \mathbb{Z}$.

(3) We prove $(a^m)^n = a^{mn}$. If $m \geqslant 1$, $n \geqslant 1$, Lemma 8.5 yields the result. If $m = 0$, then $(a^0)^n = 1^n = 1 = a^0 = a^{0n}$ for all $n \in \mathbb{Z}$; and if $n = 0$, then $(a^m)^0 = 1 = a^0 = a^{m0}$ for all $m \in \mathbb{Z}$. So we have

$$(a^m)^n = a^{mn} \text{ whenever } m,n \geqslant 0. \qquad (e')$$

We must prove this relation also when $m \geqslant 0$, $n \leqslant 0$; $m \leqslant 0$, $n \geqslant 0$; $m \leqslant 0$, $n \leqslant 0$. Replacing m,n by $|m|, |n|$ we must prove (i) $(a^m)^{-n} = a^{m(-n)}$; (ii) $(a^{-m})^n = a^{(-m)n}$; (iii) $(a^{-m})^{-n} = a^{(-m)(-n)}$ for all $m,n \geqslant 0$.

Writing (e') with a^{-1} in place of a and using (2), we get $(a^m)^{-n} = [(a^m)^{-1}]^n = (a^{-m})^n = [(a^{-1})^m]^n = (a^{-1})^{mn} = a^{-(mn)} = a^{m(-n)}$. This proves (i). We also get $(a^{-m})^n = a^{-(mn)} = a^{-(m)n}$. This proves (ii). Finally, we have $(a^{-m})^{-n} = [(a^m)^{-1}]^{-n} = ([(a^m)^{-1}]^{-1})^n = (a^m)^n = a^{mn} = a^{(-m)(-n)}$. This proves (iii).
Thus $(a^m)^n = a^{mn}$ for all $m,n \in \mathbb{Z}$.

The proof is complete. □

8.8 Lemma: *Let G be a group and $a_1, a_2, \ldots, a_n \in G$. Then*

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

Proof: By induction on n. If $n = 2$, the assertion is true by Lemma 8.2(2). Suppose now $n \in \mathbb{N}$, $n \geqslant 3$ and $(a_1a_2\ldots a_{n-1})^{-1} = a_{n-1}^{-1}\ldots a_2^{-1}a_1^{-1}$. Then

$$\begin{aligned}(a_1a_2\ldots a_{n-1}a_n)^{-1} &= ((a_1a_2\ldots a_{n-1})a_n)^{-1}\\ &= a_n^{-1}(a_1a_2\ldots a_{n-1})^{-1}\\ &= a_n^{-1}(a_{n-1}^{-1}\ldots a_2^{-1}a_1^{-1})\\ &= a_n^{-1}a_{n-1}^{-1}\ldots a_2^{-1}a_1^{-1},\end{aligned}$$

as was to be proved. □

Lemma 8.8 gives an alternative proof of $(a^{-1})^m = (a^m)^{-1}$. When our group is commutative, we have additional results, for example $a^mb^n = b^na^m$.

8.9 Lemma: *Let G be a group and $a,b \in G$. If $ab = ba$, then*
(1) $ab^n = b^na$;
(2) $a^mb^n = b^na^m$;
for all $m,n \in \mathbb{Z}$.

Proof: (1) We prove $ab^n = b^na$. The case $n = 0$ is trivial. Also, $ab^1 = ab = ba = b^1a$ by hypothesis and the claim is true for $n = 1$. Suppose now $n \in \mathbb{N}$, $n \geqslant 2$ and the claim is proved for $n - 1$, so that $ab^{n-1} = b^{n-1}a$. Then $ab^n = a(b^{n-1}b) = (ab^{n-1})b = (b^{n-1}a)b = b^{n-1}(ab) = b^{n-1}(ba) = (b^{n-1}b)a = b^na$. By induction, $ab^n = b^na$ for all $n \in \mathbb{N}$.

We multiply this relation by b^{-n} on the left and on the right. This gives $b^{-n}a = ab^{-n}$ for $n \in \mathbb{N}$. So $ab^n = b^n a$ is true also when $n \leq -1$. So $ab^n = b^n a$ for all $n \in \mathbb{Z}$.

(2) We have $b^n a = ab^n$ by (1). We use this as a hypothesis and apply (1) with a,b,n replaced by b^n,a,m, respectively. Then we obtain $a^m b^n = b^n a^m$ for all $m,n \in \mathbb{Z}$. □

If G is not a group but merely a nonempty set with an associative multiplication on it, the proof remains valid for the case $m,n \in \mathbb{N}$; and also for the case $m = 0$ or $n = 0$, provided there is a unique identity e in G and we agree to write $a^0 = e$ for all $a \in G$:

8.10 Lemma: *Let G be a nonempty set with an associative multiplication on it. Let $a,b \in G$.*
(1) If $ab = ba$, then $a^m b^n = b^n a^m$ for all $m,n \in \mathbb{N}$.
(2) If, in addition, there is a unique $e \in G$ such that $ce = ec$ for all $c \in G$, and if we put $c^0 = e$ for all $c \in G$, then $a^m b^n = b^n a^m$ also when $m = 0$ or $n = 0$. □

8.11 Lemma: *Let G be a nonempty set with an associative multiplication on it. For any $m \in \mathbb{N}$ and for any $a_1,a_2,\ldots,a_m,b \in G$ such that*

$$a_i b = b a_i \text{ for all } i = 1,2,\ldots,m$$

there holds $(a_1 a_2 \ldots a_m)b = b(a_1 a_2 \ldots a_m)$.

Proof: By induction on m. The case $m = 1$ is included in the hypothesis. Suppose now $m \geq 2$ and the claim is true for $m - 1$. Then

$$\begin{aligned}(a_1 a_2 \ldots a_{m-1}a_m)b &= ((a_1 a_2 \ldots a_{m-1})a_m)b \\ &= (a_1 a_2 \ldots a_{m-1})(a_m b) \\ &= (a_1 a_2 \ldots a_{m-1})(b a_m) \\ &= ((a_1 a_2 \ldots a_{m-1})b)a_m \\ &= (b(a_1 a_2 \ldots a_{m-1}))a_m\end{aligned}$$

$$= b((a_1a_2\ldots a_{m-1})a_m)$$
$$= b(a_1a_2\ldots a_{m-1}a_m),$$

as was to be proved. □

Lemma 8.11 gives a new proof of Lemma 8.10 when we choose $a_1 = a_2 = \cdots = a_m = a$ and replace b by b^n.

We proved in Lemma 8.3 that the product of n elements in a group (or in a set with an associative multiplication on it) is independent of the mode of putting parentheses. When the elements commute, the product is also independent of the order of elements.

8.12 Lemma: *Let G be a nonempty set with an associative multiplication on it. For all $n \in \mathbb{N}$, for all $a_1, a_2, \ldots, a_n \in G$ such that*

$$a_ia_j = a_ja_i \text{ whenever } i,j = 1,2,\ldots,n,$$

there holds

$$a_{k_1}a_{k_2}\ldots a_{k_n} = a_1a_2\ldots a_n$$

for each arrangement $k_1, k_2, \ldots, k_n$ of $1,2,\ldots,n$ (i.e., for each $k_1, k_2, \ldots, k_n$ such that $\{k_1, k_2, \ldots, k_n\} = \{1,2,\ldots,n\}$).

Proof: By induction on n. The case $n = 1$ is trivial. Now assume $n \geqslant 2$ and the claim is proved for $n - 1$, for all pairwise commuting elements $b_1, b_2, \ldots, b_{n-1}$ of G, for all arrangements of $1,2,\ldots,n-1$. Let $a_1, a_2, \ldots, a_n$ be n arbitrary pairwise commuting elements of G and let $k_1, k_2, \ldots, k_n$ be an arbitrary arrangement of $1,2,\ldots,n$. Then $n = k_j$ for some $j \in \{1,2,\ldots,n\}$. We have

$$\begin{aligned}
a_{k_1}a_{k_2}\ldots a_{k_n} &= (a_{k_1}\ldots a_{k_{j-1}})a_{k_j}(a_{k_{j+1}}\ldots a_{k_n}) \\
&= (a_{k_1}\ldots a_{k_{j-1}})(a_{k_j}(a_{k_{j+1}}\ldots a_{k_n})) \\
&= (a_{k_1}\ldots a_{k_{j-1}})((a_{k_{j+1}}\ldots a_{k_n})a_{k_j}) \\
&= ((a_{k_1}\ldots a_{k_{j-1}})(a_{k_{j+1}}\ldots a_{k_n}))a_{k_j} \\
&= (a_{k_1}\ldots a_{k_{j-1}}a_{k_{j+1}}\ldots a_{k_n})a_n
\end{aligned}$$

and here $k_1,\ldots k_{j-1},k_{j+1},\ldots,k_n$ are simply the numbers $1,2,\ldots,n-1$ in some order. By the inductive hypothesis, applied to the elements $a_1,a_2,\ldots,a_{n-1}$

and the arrangement $k_1,\ldots k_{j-1},k_{j+1},\ldots,k_n$ of the numbers 1,2, $\ldots,n-1$, we have $a_{k_1}\ldots a_{k_{j-1}}a_{k_{j+1}}\ldots a_{k_n} = a_1a_2\ldots a_{n-1}$; therefore

$$\begin{aligned} a_{k_1}a_{k_2}\ldots a_{k_n} &= (a_{k_1}\ldots a_{k_{j-1}}a_{k_{j+1}}\ldots a_{k_n})a_n \\ &= (a_1a_2\ldots a_{n-1})a_n \\ &= a_1a_2\ldots a_{n-1}a_n \end{aligned}$$

and the induction argument goes through. In the chain of equations above, the term $(a_{k_1}\ldots a_{k_{j-1}})$ is absent if $j = 1$ and the term $(a_{k_{j+1}}\ldots a_{k_n})$ is absent if $j = n$. The argument remains valid in these cases. □

8.13 Lemma: *Let G be a commutative group and let $a_1,a_2,\ldots,a_n$ be arbitrary elements of G. Then*

$$a_{k_1}a_{k_2}\ldots a_{k_n} = a_1a_2\ldots a_n$$

for all arrangments $k_1,k_2,\ldots,k_n$ of the indices 1,2,$\ldots$,n.

Proof: This follows immediately from Lemma 8.12. □

8.14 Lemma: *Let G be a nonempty set with an associative multiplication on it and let $a,b \in G$.*
(1) *If $ab = ba$, then $(ab)^n = a^nb^n$ for all $n \in \mathbb{N}$.*
(2) *If, in addition, there is a unique $e \in G$ such that $ce = ec$ for all $c \in G$, and if we put $c^0 = e$ for all $c \in G$, then $(ab)^0 = a^0b^0$.*
(3) *If, in addition, G is a group, then $(ab)^n = a^nb^n$ for all $n \in \mathbb{Z}$.*

Proof: (1) The claim is trivially true when $n = 1$. Suppose now $n \geqslant 2$ and assume $(ab)^{n-1} = a^{n-1}b^{n-1}$. Then

$$\begin{aligned} (ab)^n = (ab)^{n-1}(ab) &= (a^{n-1}b^{n-1})(ab) \\ &= a^{n-1}(b^{n-1}a)b \\ &= a^{n-1}(ab^{n-1})b \quad \text{(by Lemma 8.10)} \\ &= (a^{n-1}a)(b^{n-1}b) \\ &= a^nb^n \end{aligned}$$

and the claim is true for n. So $(ab)^n = a^nb^n$ for all $n \in \mathbb{N}$.

(2) Writing e for c, we get $ee = e$. Thus $(ab)^0 = e = ee = a^0e = a^0b^0$.
(3) That $(ab)^n = a^nb^n$ is proved for $n \geqslant 0$. We are to prove it also when $n \leqslant -1$. Replacing n by $-n$, we are to prove that $(ab)^{-n} = a^{-n}b^{-n}$ for $n \in \mathbb{N}$. We note that $ab = ba$ implies $b^{-1}a^{-1} = (ab)^{-1} = (ba)^{-1} = a^{-1}b^{-1}$, so the

hypothesis of (1) is satisfied when we replace a by a^{-1} and b by b^{-1}. Using (1) with a^{-1},b^{-1} in place of a,b, respectively, we obtain

$$(ab)^{-n} = [(ab)^{-1}]^n = [(ba)^{-1}]^n = (a^{-1}b^{-1})^n = (a^{-1})^n(b^{-1})^n = a^{-n}b^{-n}$$

for $n \in \mathbb{N}$. Thus $(ab)^n = a^nb^n$ is valid also when $n \leqslant -1$. So $(ab)^n = a^nb^n$ for all $n \in \mathbb{Z}$. □

8.15 Lemma: *Let G be a commutative group. Then $(ab)^n = a^nb^n$ for all $a,b \in G$ and for all $n \in \mathbb{Z}$.*

Proof: This follows immediately from Lemma 8.14. □

So far, we dealt with multiplicative groups. For additive groups, there are some modifications. In the case of an additive group, the unique element in $P_n(a_1,a_2,\ldots,a_n)$ of Lemma 8.3 is called the *sum of* $a_1,a_2,\ldots,a_n$ and is denoted by $a_1+a_2+\cdots+a_n$ or by $\sum_{i=1}^{n} a_i$. We write na for $a_1+a_2+\cdots+a_n$ in case $n \in \mathbb{N}$ and $a_1,a_2,\ldots,a_n$ are all equal to $a \in G$. Also, we define $0a = 0$ (the first 0 is the integer 0, the second 0 is the identity element of G) and $(-m)a = -(ma)$ for $m \in \mathbb{N}$. Thus we defined na for all $n \in \mathbb{Z}, a \in G$.

8.16 Lemma: *Let G be an additively written commutative group. Then*
(1) $ma + na = (m + n)a$;
(2) $(-m)a = m(-a)$;
(3) $n(ma) = (nm)a$;
(4) $n(a + b) = na + nb$
for all $m,n \in \mathbb{Z}, a,b \in G$.

Proof: (1),(2),(3) follow from Lemma 8.7 and (4) from Lemma 8.15. Notice that commutativity is essential for (4). □

Exercises

1. Let G be a group such that $a^{2k+1} = 1$ for all $a \in G$. Prove that G is commutative.

2. Justify each step in the proof of Lemma 8.11.

3. Let G be a group and $a,b,c \in G$. Suppose $ab = ba$. Prove that $(a^m b^n c^r)^{-1} = c^{-r}a^{-m}b^{-n}$ for all $m,n,r \in \mathbb{Z}$, justifying each detail.

(4) Let G be a nonempty set with an associative multiplication on it and let $a_1, a_2, \ldots, a_n$ be pairwise commuting elements of G. Show that

$$(a_1 a_2 \ldots a_n)^m = a_1{}^m a_2{}^m \ldots a_n{}^m$$

for all $m \in \mathbb{N}$.

(5) Show that, if G is an additive commutative group, then

$$-(a_1 + a_2 + \cdots + a_n) = (-a_1) + (-a_2) + \cdots + (-a_n)$$

for all $a_1, a_2, \ldots, a_n$ in G.

§9
Subgroups

A group is a set with a binary operation on it which has some nice properties. Being a set, a group has subsets. Naturally, we are more interested in those subsets which reflect the algebraic structure of the group than in the other subsets. They help us understand the structure of the group. Foremost among them are the sets which are groups themselves. We give them a name.

9.1 Definition: Let G be a group. A nonempty subset H of G is called a *subgroup of* G if H itself is a group under the operation on G.

We write $H \leqslant G$ to express that H is a subgroup of G. Clearly, G is a subgroup of G, so $G \leqslant G$. If H is a subroup of G and a proper subset of G, i.e., if $H \leqslant G$ and $H \subset G$, we call H a *proper subgroup of* G. In this case, we write $H < G$. The notations $H \not\leqslant G$ and $H \not< G$ mean that H is not a subgroup respectively not a proper subgroup of G.

Given a group G and a nonempty subset H of G, we must check the group axioms for H in order to determine whether H is a subgroup of G. We now discuss each one of these axioms. It turns out that we can do without some of them.

First of all, there must be a binary operation on H. The operation on H is the operation on G. More precisely, the operation on H is the restriction of the operation on G to H. Hence, for $a,b \in H$, the element ab is computed as the product of a and b in G. In order to have a binary operation on H, given by $(a,b) \to ab$ as in G, it is necessary and sufficient that $ab \in H$ for all $a,b \in H$. Hence H must be closed under the multiplication on G. Then and only then is there a binary operation on H that is the restriction of the multiplication on G.

In the second place, we must check associativity. For all $a,b,c \in H$, we must show $(ab)c = a(bc)$. But we know that $(ab)c = a(bc)$ for all $a,b,c \in G$. Since $H \subseteq G$, we have all the more so $(ab)c = a(bc)$ for all $a,b,c \in H$. Indeed, if all the elements of G have a certain property, then all the elements of H will have the same property. Thus associativity holds in H automatically, so to speak. We do not have to check it.

In H, there must exist an identity, say $1_H \in H$ such that $a1_H = a$ for all $a \in H$. In particular, the identity 1_H of H has to be such that $1_H 1_H = 1_H$. Since $1_H \in H \subseteq G$, Lemma 7.3(1) yields $1_H = 1_G$ = identity element of G. So the identity element of G is also the identity element of H, *provided* it belongs to H. Then we do not have to look for an identity element of H, we must only check that the identity element of G does belong to H. We write 1 for the identity element of H, since it is the identity element of G.

Finally, for each $a \in H$, there must exist an $x \in H$ such that $ax = 1$. Reading this equation in G, we see $x = a^{-1}$ = the inverse of a in G. We know that the inverse of a exists. Where? The inverse of a exists in G. We must also check $a^{-1} \in H$. Thus we do not have to look for an inverse of a. We must only check that the inverse a^{-1} of a, which we know to be in G, is in fact an element of H.

Summarizing this discussion, we see that a nonempty subset H of a group G is a subgroup of G if and only if

(1) $ab \in H$ for all $a,b \in H$,

(2) $1 \in H$,

(3) $a^{-1} \in H$ for all $a \in H$.

Moreover, (2) follows from (1) and (3). Indeed, if $a \in H$ (remember that $H \neq \emptyset$), then $a^{-1} \in H$ by (3) and hence $aa^{-1} \in H$ by (1), which gives $1 \in H$. So (1),(2),(3) together is equivalent to (1),(3) together. We proved the following lemma.

9.2 Lemma (Subgroup criterion): *Let G be a group and let H be a nonempty subset of G. Then H is a subgroup of G if and only if*

(i) *for all $a,b \in H$, we have $ab \in H$ (H is closed under multiplication) and*

(ii) *for all* $a \in H$, *we have* $a^{-1} \in H$ (H is closed under the forming of inverses). □

So we can dispense with checking $1 \in H$ when we know $H \neq \emptyset$. On the other hand, when we do not know a priori that $H \neq \emptyset$, the easiest way to ascertain $H \neq \emptyset$ may be to check that $1 \in H$.

When our subset is finite, we can do even better.

9.3 Lemma: (1) *Let G be a group and let H be a nonempty finite subset of G. Then H is a subgroup of G if and only if H is closed under multiplication.*

(2) *Let G be a finite group and let H be a nonempty subset of G. Then H, is a subgroup of G if and only if H is closed under multiplication.*

Proof: (1) We prove that 9.2(ii) follows from 9.2(i) when H is finite, so that 9.2(i) and 9.2(ii) are together equivalent to 9.2(i), which is the claim. So, for all $a \in H$, we must show that $a^{-1} \in H$ under the assumption that H is finite and closed under multiplication.

If $a \in H$ and H is closed under multiplication, we have $aa = a^2 \in H$, $a^2a = a^3 \in H, \ldots$, in general $a^n \in H$ for all $n \in \mathbb{N}$. The infinitely many elements $a, a^2, a^3, \ldots, a^n, \ldots$ of H cannot be all distinct, because H is a finite set. Thus $a^m = a^k$ for some $m, k \in \mathbb{N}$, $m \neq k$. Without loss of generality, let us assume $m > k$. Then

$$a^{m-k-1}a = a^{m-k} = a^m a^{-k} = a^m(a^k)^{-1} = a^m(a^m)^{-1} = 1,$$

so that $a^{-1} = a^{m-k-1} \in H$. So H is closed under the forming of inverses.

(2) This follows from (1), since any subset of a finite set is finite. □

9.4 Examples: (a) For any group G, the subsets $\{1\}$ and G are subgroups of G. Here $\{1\}$ is called the *trivial subgroup of G*.

(b) If $K \leqslant H$ and $H \leqslant G$, then K is clearly a subgroup of G.

(c) Let $4\mathbb{Z} = \{4z \in \mathbb{Z} : z \in \mathbb{Z}\} = \{u \in \mathbb{Z} : 4|u\} \subseteq \mathbb{Z}$. Now $\mathbb{Z}$ is a group under addition (Example 7.1(a)), and $4\mathbb{Z}$ is closed under addition and under the forming of inverses by Lemma 5.2(5) and Lemma 5.2(1):

(i) if $x,y \in 4\mathbb{Z}$, then $4|x$ and $4|y$, then $4|x + y$, so $x + y \in 4\mathbb{Z}$,

(ii) if $x \in 4\mathbb{Z}$, then $4|x$, then $4|-x$, so $-x \in 4\mathbb{Z}$.

Hence $4\mathbb{Z} \leqslant \mathbb{Z}$.

(d) The additive group $\mathbb{Z}$ is a subgroup of the additive group $\mathbb{Q}$. Also, we have $\mathbb{Q} \leqslant \mathbb{R} \leqslant \mathbb{C}$, where the group operation is ordinary addition.

(e) Under multiplication, $\mathbb{Q}^+ := \{x \in \mathbb{Q} : x > 0\}$ is a subgroup of $\mathbb{Q}\setminus\{0\}$, since

(i) the product of two positive rational numbers is a positive rational number, and

(ii) the reciprocal, that is, the multiplicative inverse $1/a$ of any positive rational number a is a positive rational number.

($\mathbb{Q}\setminus\{0\}$ is a group under multiplication by §7,Ex.1(b).) Also, $\mathbb{Q}^+ \leqslant \mathbb{R}^+$ (see Example 7.1(b)) and $\mathbb{Q}\setminus\{0\} \leqslant \mathbb{R}\setminus\{0\}$. We have in fact $\mathbb{Q}^+ = (\mathbb{Q}\setminus\{0\}) \cap \mathbb{R}^+$.

(f) If H_1 and H_2 are subgroups of G, then $H_1 \cap H_2$ is a subgroup of G. Indeed, $H_1 \cap H_2 \neq \varnothing$ since $1 \in H_1$ and $1 \in H_2$. Also

(i) $a,b \in H_1 \cap H_2 \Rightarrow a,b \in H_1$ and $a,b \in H_2 \Rightarrow ab \in H_1$ and $ab \in H_2 \Rightarrow ab \in H_1 \cap H_2$,

(ii) $a \in H_1 \cap H_2 \Rightarrow a \in H_1$ and $a \in H_2 \Rightarrow a^{-1} \in H_1$ and $a^{-1} \in H_2$ $\Rightarrow a^{-1} \in H_1 \cap H_2$.

Thus $H_1 \cap H_2 \leqslant G$. More generally, if H_i are subgroups of G, where i runs through an index set I, then $\bigcap_{i\in I} H_i \leqslant G$. Indeed, $\bigcap_{i\in I} H_i \neq \varnothing$ since $1 \in H_i$ for all $i \in I$ and

(i) $a,b \in \bigcap_{i\in I} H_i \Rightarrow a,b \in H_i$ for all $i \in I \Rightarrow ab \in H_i$ for all $i \in I$ $\Rightarrow ab \in \bigcap_{i\in I} H_i$,

(ii) $a \in \bigcap_{i\in I} H_i \Rightarrow a \in H_i$ for all $i \in I \Rightarrow a^{-1} \in H_i$ for all $i \in I$ $\Rightarrow a^{-1} \in \bigcap_{i\in I} H_i$.

(g) Let $S_{[0,1]}$ be the set of all one-to-one mappings from $[0,1]$ into $[0,1]$, which is a group under the composition of mappings (Example 7.1(d)). Consider

$$T = \{\alpha \in S_{[0,1]} : 0\alpha = 0\}.$$

Then T is a subgroup of $S_{[0,1]}$, for T is not empty (why?) and

(i) $\alpha,\beta \in T \implies 0\alpha = 0$ and $0\beta = 0 \implies 0(\alpha\beta) = (0\alpha)\beta = 0\beta = 0$
$\implies \alpha\beta \in T$,

(ii) $\alpha \in T \implies 0\alpha = 0 \implies 0\alpha\alpha^{-1} = 0\alpha^{-1} \implies 0 = 0\alpha^{-1}$
$\implies 0 = 0\alpha^{-1} \implies \alpha^{-1} \in T$.

(h) Let $U = \{\bar{1},\bar{3},\bar{5},\bar{7}\} \subseteq \mathbb{Z}_8$ and consider the multiplication in $\mathbb{Z}_8$. We see

$$\begin{array}{llll} \bar{1}\,\bar{1} = \bar{1} & \bar{1}\,\bar{3} = \bar{3} & \bar{1}\,\bar{5} = \bar{5} & \bar{1}\,\bar{7} = \bar{7} \\ \bar{3}\,\bar{1} = \bar{3} & \bar{3}\,\bar{3} = \bar{1} & \bar{3}\,\bar{5} = \bar{7} & \bar{3}\,\bar{7} = \bar{5} \\ \bar{5}\,\bar{1} = \bar{5} & \bar{5}\,\bar{3} = \bar{7} & \bar{5}\,\bar{5} = \bar{1} & \bar{5}\,\bar{7} = \bar{3} \\ \bar{7}\,\bar{1} = \bar{7} & \bar{7}\,\bar{3} = \bar{5} & \bar{7}\,\bar{5} = \bar{3} & \bar{7}\,\bar{7} = \bar{1} \end{array}$$

so U is closed under multiplication. Since $\mathbb{Z}_8$ is a finite set, U is a subgroup of $\mathbb{Z}_8$ by Lemma 9.3. Right? No, this is wrong. This would be correct if $\mathbb{Z}_8$ were a group under multiplication, which it is not (for instance, $\bar{0}$ has no inverse by Lemma 6.4(12)). $\mathbb{Z}_8$ is a group under addition, but this is something else. When we want to use Lemma 9.2 or Lemma 9.3, we must make sure that the larger set is a group.

Nevertheless, U is a group under multiplication:

(i) U is closed under multiplication by the calculations above.

(ii) Multiplication on U is associative since it is in fact associative on $\mathbb{Z}_8$ (Lemma 6.4(7)).

(iii) $\bar{1} \in U$ and $\bar{a}\,\bar{1} = \bar{a}$ for all $\bar{a} \in U$. This follows from our calculations or from Lemma 6.4(8). So $\bar{1}$ is an identity element of U.

(iv) Each element of U has an inverse in U. This follows from the equations $\bar{1}\,\bar{1} = \bar{1}$, $\bar{3}\,\bar{3} = \bar{1}$, $\bar{5}\,\bar{5} = \bar{1}$, $\bar{7}\,\bar{7} = \bar{1}$ and from $\bar{1},\bar{3},\bar{5},\bar{7} \in U$.

So U is a group. Let us find its subgroups. Now we can use Lemma 9.3. This lemma shows that $\{\bar{1},\bar{3}\},\{\bar{1},\bar{5}\},\{\bar{1},\bar{7}\}$ are subgroups of U since they are closed under multiplication. The reader will easily see that these are the only nontrivial proper subgroups of U. Hence the subgroups of U have orders 1,2,4, which are all divisors of the order $|U| = 4$ of U.

(i) $E := \{1,-1,i,-i\} \subseteq \mathbb{C}\setminus\{0\}$ is a subgroup of the group $\mathbb{C}\setminus\{0\}$ of nonzero complex numbers under multiplication by Lemma 9.3 as it is closed under multiplication. The same lemma shows that $\{1,-1\}$ is a subgroup of

E. Also, E has no other nontrivial proper subgroup, for any subgroup of E that contains i or $-i$ must contain i^2,i^3,i^4 or $(-i)^2,(-i)^3,(-i)^4$ and thus must be E itself. So E has exactly three subgroups, one of order 1, one of order 2,one of order 4. Here, too, the orders of the subgroups are divisors of the order $|E| = 4$ of the group E.

(j) Lemma 9.3 may be false if the subset is not finite. For example, $\mathbb{Z}$ is a group under addition, $\mathbb{N}$ is a subset of $\mathbb{Z}$ and $\mathbb{N}$ is closed with respect to addition. Still, $\mathbb{N}$ is not a subgroup of $\mathbb{Z}$ since there is no additive identity in $\mathbb{N}$ ($0 \notin \mathbb{N}$).

Exercises

1. Let G be a group and let H be a nonempty subset of G. Show that H is a subgroup of G if and only if $ab^{-1} \in H$ for all $a,b \in H$.

2. Show that $n\mathbb{Z} := \{nz \in \mathbb{Z} : z \in \mathbb{Z}\} = \{u \in \mathbb{Z} : n|u\} \subseteq \mathbb{Z}$ is a subgroup of $\mathbb{Z}$ (under addition), where n is any natural number.

3. Let $M = \{\alpha \in S_{[0,1]} : 0\alpha = 0 \text{ or } 1\alpha = 1\}$. Is M a subgroup of $S_{[0,1]}$?

4. Let $L = \{\overline{1},\overline{2},\overline{4},\overline{5},\overline{7},\overline{8}\} \subseteq \mathbb{Z}_9$. Show that L is a group. Find all subgroups of L. Do the orders of the subgroups divide the order $|L| = 6$ of the group L?

5. Let G be a group and let $H \leqslant G, K \leqslant G$. Show that $H \cup K$ is not a subgroup of G unless $H \cup K = H$ or $H \cup K = K$. (The union of two subgroups is (generally) not a subgroup.)

6. Give an example of a group G and subgroups H,K,L of G such that $H \cup K \cup L \leqslant G$. (The union of three subgroups can be a subgroup.)

7. Let G be a group and let a be a fixed element of G. Determine whether the subsets

$$C = \{x \in G: ax = xa\} \text{ and } D = \{x \in G: ax = xa \text{ or } ax = xa^{-1}\}$$

of G are subgroups of G.

8. In Example 9.4(h), why cannot we use Lemma 6.4(9) to prove that the axiom (iv) holds?

§10
Lagrange's Theorem

The order of any subgroup of U in Example 9.4(h) divides the order of U. The same thing is true for the group E in Example 9.4(i). Likewise, the reader verified that the order of L in §9, Ex.4 is divisible by the order of any subgroup of L. These are special instances of a general theorem named after J. L. Lagrange (1736-1813), which asserts that the order of a subgroup divides the order of a group, provided, of course, the group has finite order so that we can meaningfully speak about divisibility. It is the first important theorem of group theory that we come across.

The proof of Lagrange's theorem requires the notion of cosets, which plays an important role in group theory.

10.1 Definition: Let G be a group, $H \leqslant G$ and $a \in G$. We put

$$Ha := \{ha \in G : h \in H\} \subseteq G$$

and call Ha a *right coset of H in G*. We put

$$aH := \{ah \in G : h \in H\} \subseteq G$$

and call aH a *left coset of H in G*.

Right and left cosets of H are subsets of G. When the group is written additively, we write $H + a = \{h + a \in G : h \in H\}$ and $a + H = \{a + h \in G : h \in H\}$ for the right and left cosets of H. A right coset is not necessarily a left coset and a left coset is not necessarily a right coset. However, when the group is commutative, the right and left cosets coincide, as is evident from the definition. During a particular discussion, we usually fix a subgroup H of a group G and consider its various (right or left) cosets. Then we refer to Ha as the *right coset of $a \in G$*, or as the *right coset of H determined by a*. We use similar expressions for aH.

Cosets are subsets of a group, so the equality of two cosets is defined by mutual inclusion. We ask when two cosets are equal. The next lemma gives an answer.

10.2 Lemma: *Let G be a group, $H \leqslant G$ and $a,b \in G$.*
(1) *The right coset $H1$ = the subgroup H = the left coset $1H$.*
(2) *$Ha = H$ if and only if $a \in H$; $aH = H$ if and only if $a \in H$.*
(3) *$Ha = Hb$ if and only if $a = hb$ for some $h \in H$; $aH = bH$ if and only if $a = bh$ for some $h \in H$.*
(4) *$Ha = Hb$ if and only if $a \in Hb$; $aH = bH$ if and only if $a \in bH$.*
(5) *$Ha = Hb$ if and only if $ab^{-1} \in H$; $aH = bH$ if and only if $a^{-1}b \in H$.*
(6) *$Ha = Hb$ if and only if $Hab^{-1} = H$; $aH = bH$ if and only if $a^{-1}bH = H$.*

Proof: We prove only the assertions for right cosets and leave the discussion of left cosets to the reader.

(1) From the definition of $H1$ and 1, we get

$$H1 = \{h1 \in G: h \in H\} = \{h \in G: h \in H\} = H.$$

(2) If $Ha = H$, then $a = 1a \in \{ha \in G: h \in H\} = Ha = H$, so $a \in H$. Conversely, if $a \in H$, then

$a \in H$ and $a^{-1} \in H$,

$ha \in H$ and $ha^{-1} \in H$ for any $h \in H$, since H is closed under multiplication,

$ha \in H$ and $h = (ha^{-1})a \in Ha$ for all $h \in H$,

$Ha \subseteq H$ and $H \subseteq Ha$,

so $Ha = H$.

(3) If $Ha = Hb$, then $a \in Ha = Hb$, so $a = hb$ for some $h \in H$. Conversely, assume $a = hb$, where $h \in H$. Then

$a = hb$ and $b = h^{-1}a$,

$h'a = h'hb \in Hb$ and $h'b = h'h^{-1}a \in Ha$ for all $h' \in H$,

$Ha \subseteq Hb$ and $Hb \subseteq Ha$,

so $Ha = Hb$.

(4) This is just a reformulation of (3).

(5) $Ha = Hb$ if and only if $a = hb$ for some $h \in H$, and there is a unique h with $a = hb$, namely $h = ab^{-1}$ (Lemma 7.5(2)); thus $a = hb$ for some $h \in H$ if and only if $ab^{-1} \in H$.

(6) $Ha = Hb$ if and only if $ab^{-1} \in H$ by (5), and $ab^{-1} \in H$ if and only if $Hab^{-1} = H$ by (2).

10.3 Lemma: *Let $H \leqslant G$. Then G is the union of the right cosets of H. The right cosets of H are mutually disjoint. Analogous statements hold for left cosets.*

Proof: As $Ha \subseteq G$ for any $a \in G$, we get $\bigcup_{a \in G} Ha \subseteq G$. Also, for any $g \in G$, we have $g \in Hg$, so $g \in \bigcup_{a \in G} Ha$, thus $G \subseteq \bigcup_{a \in G} Ha$. This proves $G = \bigcup_{a \in G} Ha$.

Now we prove that the right cosets of H are mutually disjoint. Assume $Ha \cap Hb \neq \emptyset$. We are to show $Ha = Hb$. Well, we take $c \in Ha \cap Hb$ if $Ha \cap Hb \neq \emptyset$. Then $c \in Ha$ and $c \in Hb$. So $Ha = Hc$ and $Hc = Hb$ by Lemma 10.2(4). We obtain $Ha = Hb$.

The left cosets are treated similarly. □

In the terminology of Theorem 2.5, right cosets of H form a partition of G. Theorem 2.5 tells us that the right cosets are the equivalence classes of a certain equivalence relation on G. By the proof of Theorem 2.5, we see that this equivalence relation $\sim$ is given by

$$\text{for all } a,b \in G\text{: } a \sim b \text{ if and only if } Ha = Hb,$$

which we can read as

$$\text{for all } a,b \in G\text{: } a \sim b \text{ if and only if } ab^{-1} \in H.$$

It may be worth while to obtain Lemma 10.3 from this relation $\sim$, instead of obtaining the relation $\sim$ from Lemma 10.3.

10.4 Definition: Let $H \leqslant G$ and $a,b \in G$. We write $a \equiv_r b \pmod{H}$ and say *a is right congruent to b modulo H* if $ab^{-1} \in H$. Similarly, we write $a \equiv_l b \pmod{H}$ and say *a is left congruent to b modulo H* if $a^{-1}b \in H$.

10.5 Lemma: *Let $H \leqslant G$. Right congruence modulo H and left congruence modulo H are equivalence relations on G.*

Proof: We give the proof for right congruence only. We check that it is reflexive, symmetric and transitive.

(i) For all $a \in G$, $a \equiv_r a \pmod{H}$, as this means $aa^{-1} = 1 \in H$. So right congruence is reflexive. Reflexivity of right congruence follows from the fact that $1 \in H$.

(ii) If $a \equiv_r b \pmod{H}$, then $ab^{-1} \in H$, then $(ab^{-1})^{-1} \in H$, hence $ba^{-1} \in H$ and $b \equiv_r a \pmod{H}$. So right congruence is symmetric. Symmetry of right congruence follows from the fact that H is closed under the forming of inverses.

(iii) If $a \equiv_r b \pmod{H}$ and $b \equiv_r c \pmod{H}$, then $ab^{-1} \in H$ and $bc^{-1} \in H$, then $(ab^{-1})(bc^{-1}) \in H$, hence $ac^{-1} \in H$ and $a \equiv_r c \pmod{H}$. So right congruence is transitive. Transitivity of right congruence follows from the fact that H is closed under multiplication.

Hence right congruence is an equivalence relation on G. □

According to Theorem 2.5, G is the disjoint union of right congruence classes. The right congruence class of $a \in G$ is the right coset of a:

$$\begin{aligned}
[a] &= \{x \in G : x \equiv_r a \pmod{H}\} \\
&= \{x \in G : xa^{-1} \in H\} \\
&= \{x \in G : xa^{-1} = h, \text{ where } h \in H\} \\
&= \{x \in G : x = ha, \text{ where } h \in H\} \\
&= \{ha \in G : h \in H\} \\
&= Ha.
\end{aligned}$$

This gives a new proof of Lemma 10.3.

10.6 Lemma: *Let $H \leqslant G$. There are as many distinct right cosets of H in G as there are distinct left cosets of H in G. More precisely, let $\mathcal{R}$ be the set of right cosets of H in G and let $\mathcal{L}$ be the set of left cosets of H in G. Then $\mathcal{R}$ and $\mathcal{L}$ have the same cardinality: $|\mathcal{R}| = |\mathcal{L}|$.*

Proof: We must find a one-to-one correspondence between $\mathcal{R}$ and $\mathcal{L}$. We put $\sigma: \mathcal{R} \to \mathcal{L}$

$$Ha \to a^{-1}H.$$

We show that σ is a one-to-one, onto mapping. First we prove it is a mapping. We have to do it. Indeed, how do we find $X\sigma$ if $X \in \mathcal{R}$? Well, we write $X = Ha$, that is, we choose an $a \in X$, then we find the inverse of this a, and "map" $X = Ha$ to the left coset $a^{-1}H$ of H determined by a^{-1}. So we must show that $X\sigma$ is independent of the element a we choose from X., i.e., that σ is a well defined function. We are to prove

$$Ha = Hb \implies (Ha)\sigma = (Hb)\sigma.$$

If $Ha = Hb$, then $ab^{-1} \in H$ by Lemma 10.2(5), then $(ab^{-1})^{-1} \in H$, so $ba^{-1} \in H$, so $a^{-1}H = b^{-1}H$ by Lemma 10.2(5), and $(Ha)\sigma = (Hb)\sigma$. Hence σ is indeed a well defined function.

σ is one-to-one since $(Ha)\sigma = (Hb)\sigma \implies a^{-1}H = b^{-1}H \implies (b^{-1})^{-1}a^{-1} \in H$ $\implies ba^{-1} \in H \implies (ba^{-1})^{-1} \in H \implies ab^{-1} \in H \implies Ha = Hb$, and σ is onto as well, since any $bH \in \mathcal{L}$ is the image of $Hb^{-1} \in \mathcal{R}$ under σ:

$$(Hb^{-1})\sigma = (b^{-1})^{-1}H = bH.$$

Hence $|\mathcal{R}| = |\mathcal{L}|$. □

10.7 Definition: Let G be a group and $H \leqslant G$. The (cardinal) number of distinct right cosets of H in G, which is also the (cardinal) number of distinct left cosets of H in G, is called the *index of H in G*, and is denoted by $|G:H|$.

So $|G:H|$ is a natural number or $|G:H| = \infty$. Notice that G is written before H in $|G:H|$, but when we read, "H" is pronounced before "G": index of H in G. Lemma 10.6 states essentially that we do not have to distinguish between "right" and "left" index.

Note that $|G:H| = 1$ means $H = H1$ is the only right coset of H in G, whence $a \in Ha = H$ for all $a \in G$ and so $G \subseteq H$. Thus $|G:H| = 1$ if and only if $H = G$.

We will be mostly interested in cases where $|G:H|$ is finite. This can happen even if $|G|$ is infinite. For instance, $4\mathbb{Z}$ is a subgroup of $\mathbb{Z}$ (under addition) by Example 9.4(c) and the left (right) cosets

$$0 + 4\mathbb{Z}, 1 + 4\mathbb{Z}, 2 + 4\mathbb{Z}, 3 + 4\mathbb{Z}$$

of $4\mathbb{Z}$ in $\mathbb{Z}$ are all the left cosets of $4\mathbb{Z}$ in $\mathbb{Z}$. Hence $|\mathbb{Z} : 4\mathbb{Z}| = 4$. Incidentally, we see that Definition 10.4 is a natural generalization of the congruence relation on $\mathbb{Z}$.

We need one more lemma for the proof of Lagrange's theorem.

10.8 Lemma: *Let G be a group and $H \leqslant G$. Any right coset of H and any left coset of H in G have the same (cardinal) number of elements as H. In fact, $|Ha| = |aH| = |H|$ for all $a \in G$.*

Proof: We prove the lemma for right cosets only. For any $a \in G$, we must find a one-to-one correspondence between H and Ha. What is more natural than the mapping

$$\varphi: H \to Ha$$
$$h \to ha$$

from H into Ha? Now φ is indeed a mapping H into Ha. It is one-to-one, for $h\varphi = h'\varphi$ $(h,h' \in H)$ implies $ha = h'a$, which gives $h = h'$ after cancelling a (Lemma 8.1(2)). Also, it is onto by the very definition of Ha. So we get $|Ha| = |H|$. □

10.9 Theorem (Lagrange's theorem): *If $H \leqslant G$, then $|G| = |G:H||H|$. In particular, if G is a finite group, then $|H| \mid |G|$.*

Proof: From Lemma 10.3, we know $G = \bigcup_{a \in G} Ha$ and that the Ha are mutually disjoint. Avoiding redundancies, we write

$$G = \bigcup_{Ha \in \mathcal{R}} Ha,$$

where $\mathcal{R}$ is the set of distinct right cosets of H in G. Since Ha are disjoint, we obtain

$$|G| = \sum_{Ha \in \mathcal{R}} |Ha|$$

when we count the elements. Since $|Ha| = |H|$ for all $Ha \in \mathcal{R}$ by Lemma 10.8, we get

$$|G| = \sum_{Ha \in \mathcal{R}} |Ha| = \sum_{Ha \in \mathcal{R}} |H| = |\mathcal{R}|\,|H| = |G:H||H|$$

as $|G:H| = |\mathcal{R}|$ by Definition 10.7. □

The basic idea of the preceding proof is simple. We have a disjoint union $G = \bigcup_{Ha \in \mathcal{R}} Ha$ and we count the elements. Then we get $|G| = \sum_{Ha \in \mathcal{R}} |Ha|$. In the sequel, we will prove some important results by a similar reasoning. We will have a disjoint union $S = \bigcup_{i \in I} T_i$ and, counting the elements, we will get $|S| = \sum_{i \in I} |T_i|$. See §§25,26.

Here is an application of Lagrange's theorem.

10.10 Theorem: *Let p be a positive prime number and G be a group of order p. Then G has no nontrivial proper subgroup.*

Proof: We are to show that $\{1\}$ and G are the only subgroups of G. Now if $H \leqslant G$, then $|H| \mid |G|$ by Lagrange's theorem, so $|H| \mid p$ and $|H| = 1$ or p. If $|H| = 1$, then necessarily $H = \{1\}$. If $|H| = p$, then $|H| = |G|$ and $H \leqslant G$ together yield $H = G$. □

If G is a finite group and $K \leqslant H \leqslant G$, Lagrange's theorem gives

$$|G:H| = \frac{|G|}{|H|}, \ |H:K| = \frac{|H|}{|K|}, \text{ so } |G:H||H:K| = \frac{|G|}{|H|}\frac{|H|}{|K|} = \frac{|G|}{|K|} = |G:K|.$$

We give another proof of this result which works also in the case of infinite groups and infinite indices.

10.11 Theorem: *If $K \leqslant H \leqslant G$, then $|G:H||H:K| = |G:K|$. In particular, if any two of $|G:H|, |G:K|, |H:K|$ is finite, then the third is finite, too.*

Proof: Let $\mathcal{R} = \{Ha_i : i \in I\}$ be the set of all distinct right cosets of H in G. We have

$$G = \bigcup_{i \in I} Ha_i, \text{ with } a_i \in G, Ha_i \neq Ha_{i_1} \text{ for } i \neq i_1, |I| = |G:H|. \tag{1}$$

Let $\mathcal{R}' = \{Kb_j : j \in J\}$ be the set of all distinct right cosets of K in H. Then

$$H = \bigcup_{j \in J} Kb_j, \text{ with } b_j \in H, Kb_j \neq Kb_{j_1} \text{ for } j \neq j_1, |J| = |H:K|. \tag{2}$$

We must prove $|G:K| = |I||J|$. Since $|I \times J| = |I||J|$, this will be accomplished if we can find a one-to-one correspondence between $I \times J$ and the set of right cosets of K in G. How we find this correspondence will be clear when we observe

$$\begin{aligned} Ha_i &= \{ha_i \in G: h \in H\} = \{ha_i \in G: h \in \bigcup_{j \in J} Kb_j\} \\ &= \{ha_i \in G: \text{there are } j \in J \text{ and } k \in K \text{ with } h = kb_j\} \\ &= \{kb_j a_i \in G: j \in J, k \in K\} \\ &= \bigcup_{j \in J} \{kb_j a_i \in G: k \in K\} = \bigcup_{j \in J} Kb_j a_i, \end{aligned}$$

so that $G = \bigcup_{i \in I} Ha_i = \bigcup_{i \in I} \bigcup_{j \in J} Kb_j a_i = \bigcup_{(i,j) \in I \times J} Kb_j a_i$. This suggests

$$(i,j) \to Kb_j a_i$$

as a mapping from $I \times J$ into the set of right cosets of K in G. Let us check if it works.

For each $(i,j) \in I \times J$, $b_j a_i$ is an element of G; hence $Kb_j a_i$ is a right coset of K in G. Thus the above correspondence is indeed a mapping from $I \times J$ into the set of right cosets of K in G.

It is onto, for if Kg ($g \in G$) is any right coset of K in G, then $g \in \bigcup_{(i,j) \in I \times J} Kb_j a_i$ by our observation, so, by the definition of union, there is $(i_0, j_0) \in I \times J$ with $g \in Kb_{j_0} a_{i_0}$. Then $Kg = Kb_{j_0} a_{i_0}$ and Kg is the image of $(i_0, j_0) \in I \times J$.

It is one-to-one: if $Kb_j a_i = Kb_{j_1} a_{i_1}$, then $b_j a_i = kb_{j_1} a_{i_1}$ for some $k \in K \subseteq H$ by Lemma 10.2(3), then $a_i = b_j^{-1} k b_{j_1} a_{i_1}$ with $b_j^{-1} k b_{j_1} \in H$, so $Ha_i = Ha_{i_1}$ by Lemma 10.2(3), so $i = i_1$ by (1). Thus $a_i = a_{i_1}$ and we get $Kb_j = Kb_j a_i a_i^{-1} = Kb_{j_1} a_{i_1} a_i^{-1} = Kb_{j_1} a_i a_i^{-1} = Kb_{j_1}$ by Lemma 10.2(6), which yields $j = j_1$ by (2). Hence $Kb_j a_i = Kb_{j_1} a_{i_1}$ implies $(i,j) = (i_1, j_1)$. The mapping is one-to-one.

Thus $|G:K| = |I \times J| = |I|\,|J| = |G:H||H:K|$. □

Exercises

1. Find the right cosets of all subgroups of U (Example 9.4(h)), of E (Example 9.4(i)) and of L (§9,Ex.4).

2. Let T be the subgroup of $S_{[0,1]}$ that we discussed in Example 9.4(g). Show that $\{\sigma \in S_{[0,1]}: 0\sigma = 1\}$ is a right coset of T in $S_{[0,1]}$. Is it a left coset of T? Would your answer be different if we wrote the functions on the left? What is $|S_{[0,1]}:T|$?

3. Find all cosets of $n\mathbb{Z}$ in $\mathbb{Z}$. What is $|\mathbb{Z} : n\mathbb{Z}|$?

4. Why do we not use the "mapping" $\mathfrak{R} \to \mathfrak{L}$ $Ha \to aH$ in the proof of Lemma 10.6?

§11
Cyclic Groups

Let G be a group and $a \in G$. Consider the set $\{a^n \in G: n \in \mathbb{Z}\}$ of all integral powers of a. We designate this subset of G shortly by $\langle a \rangle$. It is not empty and is in fact a subgroup of G:

(i) if $a^m, a^n \in \langle a \rangle$, then $a^m a^n = a^{m+n} \in \langle a \rangle$, as $m + n \in \mathbb{Z}$ when $m, n \in \mathbb{Z}$,

(ii) if $a^m \in \langle a \rangle$, then $(a^m)^{-1} = a^{-m} \in \langle a \rangle$, as $-m \in \mathbb{Z}$ when $m \in \mathbb{Z}$.

11.1 Definition: Let G be a group and $a \in G$. Then $\langle a \rangle = \{a^n \in G: n \in \mathbb{Z}\}$ is called the *cyclic subgroup of G generated by a*. If it happens that $\langle a \rangle = G$, then G is called a *cyclic* group and a is called a *generator of G*.

Any cyclic group is abelian. Indeed, if G is a cyclic group, generated by a, then any two elements a^m, a^n $(m, n \in \mathbb{Z})$ of G commute:

$$a^m a^n = a^{m+n} = a^{n+m} = a^n a^m.$$

The converse is false. There are abelian groups which are not cyclic. For example, the group U of Example 9.4(h) is abelian but not cyclic since the cyclic subgroups generated by $\overline{1}, \overline{3}, \overline{5}, \overline{7}$ are all proper subgroups of U.

11.2 Examples: (a) Consider the subgroup $\langle i \rangle$ of $\mathbb{C}\backslash\{0\}$ under multiplication. We have

$$i^0 = 1,\ i^1 = i,\ i^2 = -1,\ a^3 = -i$$

and other powers of i do not give rise to other complex numbers. To see this, let $n \in \mathbb{Z}$ and divide n by 4 to get $n = 4q + r$, $0 \leqslant r \leqslant 3$, $q, r \in \mathbb{Z}$. Then

$$i^n = i^{4q+r} = i^{4q} i^r = (i^4)^q i^r = 1^4 i^r = i^r \subseteq \{1, i, -1, -i\}.$$

Hence $\langle i \rangle = \{1, i, -1, -i\}$ is a cyclic group of order 4.

(b) In §9, Ex.4, the reader proved that $L = \{\overline{1}, \overline{2}, \overline{4}, \overline{5}, \overline{7}, \overline{8}\}$ is a group under multiplication (mod 9). Let us find the cyclic subgroup of L generated by $\overline{2}$. We have

$$\bar{2}^0 = \bar{1},\ \bar{2}^1 = \bar{2},\ \bar{2}^2 = \bar{4},\ \bar{2}^3 = \bar{8},\ \bar{2}^4 = \bar{7},\ \bar{2}^5 = \bar{5},$$
$$L = \{\bar{1},\bar{2},\bar{4},\bar{5},\bar{7},\bar{8}\} = \{\bar{2}^n \in L : n = 0,1,2,3,4,5\} \subseteq \{\bar{2}^n \in L : n \in \mathbb{Z}\} = \langle\bar{2}\rangle,$$

thus $L = \langle\bar{2}\rangle$. So L is a cyclic group and $\bar{2}$ is a generator of L. We see

$$\langle\bar{4}\rangle = \{\bar{1},\bar{4},\bar{7}\},\ \langle\bar{8}\rangle = \{\bar{1},\bar{8},\},\ \langle\bar{7}\rangle = \{\bar{1},\bar{7},\bar{4}\}$$

are proper subgroups of L. In particular, $\bar{4},\bar{7},\bar{8}$ are not generators of L. On the other hand,

$$\langle\bar{5}\rangle = \{\bar{1},\bar{5},\bar{7},\bar{8},\bar{4},\bar{2}\} = L$$

and $\bar{5}$ is another generator of L.

A cyclic group has many generators. The number of generators of a cyclic group will be determined later in this paragraph.

11.3 Definition: Let G be a group and $a \in G$. The order $|\langle a\rangle|$ of the cyclic subgroup of G generated by a is called the *order of* a and is denoted by $o(a)$.

Thus $o(a)$ is either a natural number or ∞. Of course, if G is a finite group, then every element a of G will have finite order, in fact $o(a)\,|\,|G|$ by Lagrange's theorem. An infinite group, on the other hand, has in general, elements of finite order as well as elements of infinite order.

11.4 Lemma: *Let G be a group and $a \in G$. Then $o(a)$ is finite if and only if there is a natural number n with $a^n = 1$. If this is the case, then $o(a)$ is the smallest natural number s such that $a^s = 1$.*

Proof: We put $A = \{n \in \mathbb{N} : a^n = 1\}$. The claim is that $o(a)$ is finite if and only if A is not empty. First we suppose $o(a)$ is finite and prove that A is not empty. If $o(a)$ is finite, then $\langle a\rangle$ is a finite subgroup of G and the infinitely many elements

$$a^1, a^2, a^3, a^4, \ldots$$

of $\langle a\rangle$ cannot be all distinct. So $a^k = a^m$ for some $k,m \in \mathbb{N}$ with $k \neq m$. Assuming $k < m$ without loss of generality, we obtain $a^{m-k} = a^m a^{-k} = a^m(a^k)^{-1} = a^m(a^m)^{-1} = 1$, so $m - k \in A$ and $A \neq \varnothing$.

Suppose now there are natural numbers n with $a^n = 1$, that is, suppose that $A \neq \varnothing$. We prove that $o(a)$ is finite, and is in fact the smallest natural number in A. To this end, let s be the smallest natural number in A. We show first $s \leq o(a)$ and then $o(a) \leq s$.

Consider the s elements $a^0, a^1, a^2, \ldots, a^{s-1}$ of $\langle a \rangle$. These are all distinct, for if

$$a^i = a^j,\ i \neq j,\ 0 \leq i,j \leq s-1$$

say with $i < j$, then

$$a^{j-i} = 1,\ j - i \leq (s-1) - 0,\ j - i \in \mathbb{N},$$
$$j - i \in A,\quad j - i \leq s - 1,$$

contradicting that s is the smallest natural number in A. So there are at least s distinct elements in $\langle a \rangle$. This gives $s \leq |\langle a \rangle| = o(a)$.

Next we show that there are at most s distinct elements in $\langle a \rangle$. If $a^h \in \langle a \rangle$, where $h \in \mathbb{Z}$, we divide h by s to get

$$h = qs + r,\qquad q,r \in \mathbb{Z},\qquad 0 \leq r \leq s-1,$$
$$a^h = a^{qs+r} = a^{sq}a^r = (a^s)^q a^r = 1^q a^r = a^r,$$

so

$$\langle a \rangle \subseteq \{a^0, a^1, a^2, \ldots, a^{s-1}\},$$
$$|\langle a \rangle| \leq |\{a^0, a^1, a^2, \ldots, a^{s-1}\}|,$$
$$o(a) \leq s$$

since the elements $a^0, a^1, a^2, \ldots, a^{s-1}$ are all distinct.

From $s \leq o(a)$ and $o(a) \leq s$, we get $o(a) = s$. □

11.5 Lemma: *Let G be a group and $a \in G$. Then $o(a) = \infty$ if and only if powers of a with distinct exponents are distinct, i.e., if and only if $a^m \neq a^k$ whenever $m \neq k$ $(m,k \in \mathbb{Z})$.*

Proof: If $a^m \neq a^k$ whenever $m \neq k$, then the infinitely many elements

$$\ldots, a^{-3}, a^{-2}, a^{-1}, a^0, a^1, a^2, a^3, \ldots$$

of $\langle a \rangle$ are all distinct. So $\langle a \rangle$ is an infinite group and $o(a) = \infty$.

Suppose now the condition in the lemma does not hold. Then there are $m,k \in \mathbb{Z}$ with $a^m = a^k$, $m \neq k$. Assume $m > k$ without loss of generality. Then $m - k \in \mathbb{N}$ and $a^{m-k} = 1$. There is a natural number n, namely n

$= m - k$, with $a^n = 1$. Then $o(a)$ is finite by Lemma 11.4. Hence $o(a) = \infty$ implies that $a^m \neq a^k$ whenever $m \neq k$ $(m,k \in \mathbb{Z})$. □

11.6 Lemma: *Let G be a group and let $a \in G$ be of finite order. Let $n \in \mathbb{Z}$. Then $a^n = 1$ if and only if $o(a) | n$.*

Proof: We put $s = o(a)$. If $s|n$, then $n = sq$ for some $q \in \mathbb{Z}$, hence $a^n = a^{sq} = (a^s)^q = 1^q = 1$ since $a^s = 1$ by Lemma 11.4. Conversely, suppose $a^n = 1$. We divide n by s and get

$$n = qs + r, \qquad q,r \in \mathbb{Z}, \qquad 0 \leq r \leq s - 1,$$

$$1 = a^n = a^{qs+r} = a^{sq}a^r = (a^s)^q a^r = 1^q a^r = a^r.$$

If $r \neq 0$, then r would be a natural number smaller than s with $a^r = 1$, contradicting Lemma 11.4. So $r = 0$, $n = qs$ and $s|n$. □

11.7 Lemma: *If G is a finite group, then $a^{|G|} = 1$ for all $a \in G$.*

Proof: For any $a \in G$, $o(a) = |\langle a \rangle|$ divides $|G|$ by Lagrange's theorem. So $a^{|G|} = 1$ by Lemma 11.6. □

Next we show that subgroups of cyclic groups are also cyclic.

11.8 Theorem: *Let G be a cyclic group and let $H \leq G$. Then H is cyclic. More informatively, let $G = \langle a \rangle$. Then $\{1\} = \langle 1 \rangle$ and if $H \neq \{1\}$, then $H = \langle a^t \rangle$, where t is the smallest natural number in the set $\{n \in \mathbb{N}: a^n \in H\}$.*

Proof: The subgroup $\{1\}$ of $G = \langle a \rangle$ is clearly the cyclic subgroup of G generated by 1, hence $\{1\} = \langle 1 \rangle$ is cyclic. Suppose now $\{1\} \neq H \leq G$. We prove that H is cyclic, and in fact $H = \langle a^t \rangle$ as stated in the theorem. Since $H \neq \{1\}$ by assumption, there is a nonidentity element in H, say $a^{m'} \in H$, with $m \in \mathbb{Z}\backslash\{0\}$. Then $a^{-m} \in H$ since H is closed under the forming of inverses. So $a^m, a^{-m} \in H$, $m \neq 0$. So there is a natural number n such that $a^n \in H$, for instance $n = |m|$. Thus the set $\{n \in \mathbb{N}: a^n \in H\}$ is not empty.

From the natural numbers in this set, we choose the smallest one and call it t.

Now $a^t \in H$. Also $a^{-t} = (a^t)^{-1} \in H$. Since H is closed under multiplication, we obtain $a^{kt} = (a^t)^k = a^t a^t \ldots a^t \in H$ and $a^{-kt} = (a^{-t})^k = a^{-t}a^{-t}\ldots a^{-t} \in H$ for all $k \in \mathbb{N}$. Since $a^{0t} = 1 \in H$, we see $a^{mt} = a^{tm} \in H$ for all $m \in \mathbb{Z}$. Thus we have $\langle a^t \rangle = \{a^{tm} \in G : m \in \mathbb{Z}\} \subseteq H$.

Assume next $b \in H$, where $b \in G = \langle a \rangle$. We write $b = a^n$ with a suitable n in $\mathbb{Z}$ and divide n by t. This gives

$$n = tq + r, \quad q,r \in \mathbb{Z}, \qquad 0 \leq r \leq t - 1,$$
$$a^r = a^{n-tq} = a^n(a^{-t})^q \in H,$$

since $a^n, a^{-t} \in H$. If $r \neq 0$, then r would be a natural number smaller than t such that $a^r \in H$, contradicting the definition of t. So $r = 0$, $n = tq$, $t|n$ and $b = a^n = a^{tq} \in \langle a^t \rangle$. This holds for all $b \in H$. Hence $H \subseteq \langle a^t \rangle$.

From $\langle a^t \rangle \subseteq H$ and $H \subseteq \langle a^t \rangle$, we get $H = \langle a^t \rangle$, as claimed. □

11.9 Lemma: *Let G be a group and $a \in G$. Let $k \in \mathbb{Z}$, $k \neq 0$.*
(1) If $o(a) = \infty$, then $o(a^k) = \infty$.
(2) If $o(a) = n \in \mathbb{N}$, then $o(a^k) = n/(n,k)$.

Proof: (1) Suppose $o(a) = \infty$. If $o(a^k)$ were finite, say $o(a^k) = m \in \mathbb{N}$, then $(a^k)^m = 1$, so $a^{km} = 1 = a^0$, although km and 0 are distinct integers, contrary to Lemma 11.5. So $o(a) = \infty$ implies $o(a^k) = \infty$.

(2) Now let us suppose $o(a) = n \in \mathbb{N}$. Then $\langle a^k \rangle \leq \langle a \rangle$ and so $o(a^k)$ is finite. By Lemma 11.4,

$$\begin{aligned} o(a^k) &= \text{smallest natural number } s \text{ such that } (a^k)^s = 1 \\ &= \text{smallest natural number } s \text{ such that } a^{ks} = 1 \\ &= \text{smallest natural number } s \text{ such that } n|ks \qquad \text{(Lemma 11.6)} \\ &= \text{smallest natural number } s \text{ such that } \frac{n}{(n,k)} \Big| \frac{k}{(n,k)} s \\ &= \text{smallest natural number } s \text{ such that } \frac{n}{(n,k)} \Big| s \qquad \text{(Lemma 5.11 and Theorem 5.12)} \\ &= \frac{n}{(n,k)}. \end{aligned}$$

□

From Lemma 11.9(1), we infer that any nontrivial subgroup of an infinite cyclic group is infinite. Using Lemma 11.9(2), we can find the number of generators of a finite cyclic group. Let $G = \langle a \rangle$ be a cyclic group of order $n \in \mathbb{N}$. Which elements are the generators of G? Any element a^k generates a subgroup $\langle a^k \rangle$ of $\langle a \rangle$ and a^k is a generator of $\langle a \rangle$ if and only if $\langle a^k \rangle = \langle a \rangle$. We know $\langle a^k \rangle \leqslant \langle a \rangle$, so, since $|\langle a \rangle| = n$ is finite, a^k is a generator of $\langle a \rangle$ if and only if $|\langle a^k \rangle| = |\langle a \rangle|$. Thus a^k is a generator of $\langle a \rangle$ if and only if $o(a^k) = o(a)$, that is, if and only if $n = n/(n,k)$, and so if and only if $(n,k) = 1$. There are n distinct elements $a^0, a^1, a^2, \ldots, a^{n-1}$ in $\langle a \rangle$, and among these,

$$\{a^k : (n,k) = 1, 0 \leqslant k \leqslant n-1\} = \{a^k : (n,k) = 1, 1 \leqslant k \leqslant n\}$$

is the set of generators of $\langle a \rangle$. Hence the number of generators of $\langle a \rangle$ is the number of positive integers smaller than (or equal to) n and relatively prime to n. This number is traditionally denoted by $\varphi(n)$. For example

$$\varphi(1) = 1, \quad \varphi(2) = 1, \quad \varphi(3) = 2, \quad \varphi(4) = 2, \quad \varphi(5) = 4,$$
$$\varphi(6) = 2, \quad \varphi(7) = 6, \quad \varphi(8) = 4, \quad \varphi(9) = 6, \quad \varphi(10) = 4.$$

The function $\varphi: \mathbb{N} \to \mathbb{N}$ is known as Euler's *phi function* or Euler's *totient function* (L. Euler, a Swiss mathematician (1707-1783)).

Lagrange's theorem asserts that $m \mid |G|$ when there is a subgroup H of order $|H| = m$ (provided G is a finite group). The converse of Lagrange's theorem is false: if G is a finite group and $m \mid |G|$, then it is not necessarily true that G has a subgroup of order m (see §16, Ex.7). However, for cyclic groups, the converse of Lagrange's theorem is true.

11.10 Lemma: *Let $G = \langle a \rangle$ be a cyclic group of order $|G| = n$. For any positive divisor m of n, there is a unique subgroup H of order $|H| = m$, namely $\langle a^{n/m} \rangle$.*

Proof: $o(a) = n$ by hypothesis. We write $n = mk$. Consider the subgroup $\langle a^k \rangle$ of $\langle a \rangle$. We observe $|\langle a^k \rangle| = o(a^k) = n/(n,k) = mk/(mk,k) = mk/k = m$, so $\langle a^k \rangle$ is a subgroup of order m.

We now show that $\langle a^k \rangle$ is the unique subgroup of G of order m. Let L be a subgroup of order m. We want to prove $L = \langle a^k \rangle$. Since $|L| = |\langle a^k \rangle| = m$ is finite, it will suffice to prove that $L \leqslant \langle a^k \rangle$. This is certainly true if $L = \{1\}$, that is, if $m = 1$. When $m \neq 1$, we have, by Theorem 11.8, $L = \langle a^t \rangle$, where t is the smallest natural number such that $a^t \in L$. In order to show $\langle a^t \rangle = L \leqslant \langle a^k \rangle$, we need only prove $a^t \in \langle a^k \rangle$, i.e., we need only prove $k|t$. This is easy: since $o(a^t) = |\langle a^k \rangle| = |L| = m$, we get $(a^t)^m = 1$ by Lemma 11.6, so $a^{tm} = 1$, so $n|tm$ by Lemma 11.6 again, which gives $km|tm$, hence $k|t$. □

Lemma 11.10 implies that a finite cyclic group G has, for any positive divisor k of $|G|$, a unique subgroup of index k. This reformulation of Lemma 11.10 extends immediately to infinite cyclic groups.

11.11 Lemma: *Let $G = \langle a \rangle$ be a cyclic group of infinite order. For any $m \in \mathbb{N}$, there is a unique subgroup H of G of index $|G:H| = m$, namely $H = \langle a^m \rangle$. Any nontrivial subgroup of G has finite index in G.*

Proof: We have $G = \langle a \rangle$, $o(a) = \infty$. The elements of G are the symbols a^k, where k runs through the set of integers. By Lemma 11.5, $a^k \neq a^j$ for $k \neq j$. Two symbols are multiplied by adding the exponents: $a^k \cdot a^j = a^{k+j}$. Also, a^0 is the identity and $(a^k)^{-1}$ is the symbol a^{-k}. Essentially, we have the group of integers under addition, but the integers are written as exponents.

First we prove that a nontrivial subgroup of G has finite index in G. Let $L \leqslant G = \langle a \rangle$, $L \neq \{1\}$. From Theorem 11.8, we know $L = \langle a^t \rangle$, where t is the smallest natural number such that $a^t \in L$. Any element a^n of $G = \langle a \rangle$ can be written as a^{tq+r}, with some uniquely determined integers q, r, where $0 \leqslant r \leqslant t - 1$. Thus any element a^n of G belongs to one and only one of the subsets

$$\{a^{tq} : q \in \mathbb{Z}\}, \{a^{tq+1} : q \in \mathbb{Z}\}, \{a^{tq+2} : q \in \mathbb{Z}\}, \ldots, \{a^{tq+(t-1)} : q \in \mathbb{Z}\},$$

which are just the right cosets

$$\begin{array}{ccccc} \langle a^t \rangle a^0, & \langle a^t \rangle a^1, & \langle a^t \rangle a^2, & \ldots, & \langle a^t \rangle a^{t-1} \\ La^0, & La^1, & La^2, & \ldots, & La^{t-1} \end{array}$$

of L. The uniqueness of q and r implies that these cosets are distinct. Alternatively, one can show that these cosets are distinct by noting that $La^i = La^j$ $(0 \leq i,j \leq t-1)$ implies, when $i \neq j$, say when $i < j$, that $L = La^{j-i}$ and thus (Lemma 10.2(2)) $a^{j-i} \in L$, where $0 < j - i \leq t - 1$, contrary to the definition of t as the smallest natural number such that $a^t \in L$. So there are exactly t distinct right cosets of L in G and $|G:L| = t$ is finite.

We proved in fact that $|G:\langle a^t\rangle| = t$ when $t \in \mathbb{N}$. Thus, for any $m \in \mathbb{N}$, there is a subgroup of G of index m, namely $\langle a^m\rangle$. We proceed to show that $\langle a^m\rangle$ is the unique subgroup of G of index m. Assume $K \leq G$ with $|G:K| = m \in \mathbb{N}$. We are to show $K = \langle a^m\rangle$. Now $K = \langle a^k\rangle$, where k is the smallest natural number such that $a^k \in K$ (as $|G:K|$ is finite, $K \neq \{1\}$). So $m = |G:K| = |G:\langle a^k\rangle| = k$ and $a^m = a^k$, which yields $K = \langle a^k\rangle = \langle a^m\rangle$. Therefore $\langle a^m\rangle$ is the unique subgroup of G of index m. □

We learned the structure of cyclic groups quite well, but we had only a few examples. We have not seen any cyclic group of order 5 or 7. For all we know about cyclic groups up to now, it is feasible that there is no cyclic group of order 5 or 7. We show next that there is a cyclic group of any order. Incidentally, this shows that there are groups of all orders.

11.12 Theorem: *There is a cyclic group of infinite order. Also, for any $n \in \mathbb{N}$, there is a cyclic group of order n.*

Proof: We give examples of cyclic groups in additive notation. In this notation, $\langle a\rangle$ is the group $\{na : n \in \mathbb{Z}\}$, the group operation being $na + ma = (n + m)a$, the additive counterpart of the rule $a^n a^m = a^{n+m}$.

$\mathbb{Z}$ (under addition) is a cyclic group of infinite order as $\mathbb{Z} = \{m1 : m \in \mathbb{Z}\} = \langle 1\rangle$ is generated by $1 \in \mathbb{Z}$.

$\mathbb{Z}_n$ (under addition) is a cyclic group of order n as $\mathbb{Z}_n = \{m\bar{1} : m \in \mathbb{Z}\} = \langle \bar{1}\rangle$ is generated by $\bar{1} \in \mathbb{Z}_n$. □

11.13 Theorem: *Let p be a prime number. If G is a group of order p, then G is cyclic.*

Proof: Since p is prime, $|G| = p \neq 1$ and so G does not consist of the identity element only. Let a be any element of G distinct from the identity. Then $\{1\} \neq \langle a \rangle \leq G$ and $|\langle a \rangle|$ is a positive divisor of $|G| = p$ by Lagrange's theorem. Since $\langle a \rangle \neq \{1\}$, we have $|\langle a \rangle| \neq 1$, and so $|\langle a \rangle| = p = |G|$. This forces $G = \langle a \rangle$. Thus G is a cyclic group. (In fact, any nonidentity element of G is a generator of G.) □

Exercises

1. Let G be a group and let a be an element of finite order n in G. Show that, for all $m,k \in \mathbb{Z}$, the equality $a^m = a^k$ holds if and only if $m \equiv k \pmod{n}$.

2. Find all subgroups of a cyclic group of order 8, of a cyclic group of order 10, and of a cyclic group of order 12.

3. Let G be a group, $a \in G$ and $o(a) = 36$. What are the orders of a^2, a^3, a^4, a^7, a^{12}, a^{15}, a^{17}?

4. Let G be a group and $a \in G$. Let $n,k \in \mathbb{N}$ and let $m = [n,k]$ be the least common multiple of n and k. Prove that $\langle a^n \rangle \cap \langle a^k \rangle = \langle a^m \rangle$.

5. Let G be a group and $a \in G$ with $o(a) = n_1 n_2 \in \mathbb{N}$, where n_1,n_2 are relatively prime natural numbers. Show that there are uniquely determined elements a_1,a_2 of G such that

$$a_1 a_2 = a = a_2 a_1$$

and

$$o(a_1) = n_1,\ o(a_2) = n_2.$$

6. Let G be a group and $a,b \in G$. Assume that $o(a) \in \mathbb{N}$, $o(b) \in \mathbb{N}$ and that $o(a),o(b)$ are relatively prime. Prove: if $ab = ba$, then $o(ab) = o(a)o(b)$. Prove also that $o(ab) = o(a)o(b)$ is not necessarily true when the hypothesis $ab = ba$ is omitted.

7. Show that, if $p,n \in \mathbb{N}$ and p is prime, then $\varphi(p^n) = p^n - p^{n-1}$.

§12
Group of Units Modulo n

Let n be a natural number and consider $\mathbb{Z}_n$. We defined two operations on this set, namely addition and multiplication (Lemma 6.3). With respect to addition, $\mathbb{Z}_n$ forms a group. What about multiplication? With respect to multiplication, $\mathbb{Z}_n$ is not a group unless $n = 1$. This can be easily seen from the fact that $\bar{0}$ has no multiplicative inverse in $\mathbb{Z}_n$ (Lemma 6.4(12); note that $\bar{0} \neq \bar{1}$ when $n \neq 1$). However, as in Example 9.4(h), a suitable subset of $\mathbb{Z}_n$ is a group under multiplication.

12.1 Lemma: *Let $n \in \mathbb{N}$ and $a,b \in \mathbb{Z}$. If $\bar{a} = \bar{b}$ in $\mathbb{Z}_n$, then $(a,n) = (b,n)$.*

Proof: If $\bar{a} = \bar{b}$ in $\mathbb{Z}_n$, then $a \equiv b \pmod{n}$, so $n|b - a$, so $nk = b - a$ for some $k \in \mathbb{Z}$. We put $d_1 = (a,n)$ and $d_2 = (b,n)$. We have $d_1|n$ and $d_1|a$, thus $d_1|nk + a$, thus $d_1|b$. From $d_1|n$ and $d_1|b$, we get $d_1|(b,n)$, so $d_1|d_2$. Likewise we obtain $d_2|d_1$. So $|d_1| = |d_2|$ by Lemma 5.2(12) and, since d_1,d_2 are positive, we have $d_1 = d_2$. □

The preceding lemma tells that the mapping $\mathbb{Z}_n \to \mathbb{N}$ is well defined. The
$\bar{a} \to (a,n)$
claim of the lemma is not self-evident and requires proof. Compare it to the apparrently similar but *wrong* assertion that $\bar{a} = \bar{b}$ implies $(a,n^2) = (b,n^2)$. By Lemma 12.1, the following definition is meaningful.

12.2 Definition: Let $n \in \mathbb{N}$ and $\bar{a} \in \mathbb{Z}_n$, where $a \in \mathbb{Z}$. If $(a,n) = 1$, then $\bar{a}$ is called a *unit in* $\mathbb{Z}_n$. The set of all units in $\mathbb{Z}_n$ will be denoted by $\mathbb{Z}_n^*$.

The reader will observe that U in Example 9.4(h) is exactly $\mathbb{Z}_8^*$. We see $\mathbb{Z}_7^* = \{\bar{1},\bar{2},\bar{3},\bar{4},\bar{5},\bar{6}\}$. More generally, $\mathbb{Z}_p^* = \{\bar{1},\bar{2}, \ldots, \overline{p-1}\}$ for any prime number p. So $|\mathbb{Z}_p^*| = p - 1$. When $n > 1$, $\mathbb{Z}_n^*$ consists of the residue classes of the numbers among $1,2,3, \ldots, n-1, n$ that are relatively prime to n. By

the definition of Euler's phi function, we conclude $|\mathbb{Z}_n^\times| = \varphi(n)$. So $\varphi(12) = 4$ and in fact $\mathbb{Z}_{12}^\times = \{\overline{1},\overline{5},\overline{7},\overline{11}\}$. Also, $\varphi(15) = 8$ and $\mathbb{Z}_{15}^\times = \{\overline{1},\overline{2},\overline{4},\overline{7},\overline{8},\overline{11},\overline{13},\overline{14}\}$.

12.3 Lemma: *Let $n \in \mathbb{N}$ and $a,b \in \mathbb{Z}$. If $(a,n) = (b,n) = 1$, then $(ab,n) = 1$.*

Proof: This follows from the fundamental theorem of arithmetic (Theorem 5.17), but we give another proof. We put $d = (ab,n)$ and assume, by way of contradiction, that $d > 1$. Then $p|d$ for some prime number p (Theorem 5.13). So

$p\|ab$	and	$p\|n$	
$p\|a$ or $p\|b$	and	$p\|n$	(Euclid's lemma)
$p\|a$ and $p\|n$	or	$p\|b$ and $p\|n$	
$p\|(a,n)$	or	$p\|(b,n)$,	

contrary to the hypothesis $(a,n) = 1 = (b,n)$. So $(ab,n) = d = 1$. □

12.4 Theorem: *For any $n \in \mathbb{N}$, $\mathbb{Z}_n^\times$ is a group under multiplication.*

Proof: (cf. Example 9.4(h).) We check the group axioms.

(i) Is $\mathbb{Z}_n^\times$ closed under multiplication? Let $\overline{a},\overline{b} \in \mathbb{Z}_n^\times$, so that a,b are integers with $(a,n) = 1 = (b,n)$. We ask whether $\overline{ab} \in \mathbb{Z}_n^\times$, i.e., which is equivalent to asking whether $(ab,n) = 1$. By Lemma 12.3, ab is indeed relatively prime to n and so $\mathbb{Z}_n^\times$ is closed under multiplication.

(ii) Multiplication in $\mathbb{Z}_n^\times$ is associative since it is in fact associative in $\mathbb{Z}_n$ (Lemma 6.4(7)).

(iii) $\overline{1} \in \mathbb{Z}_n^\times$ as $(1,n) = 1$ and $\overline{a}\,\overline{1} = \overline{a1} = \overline{a}$ for all $\overline{a} \in \mathbb{Z}_n^\times$. Hence $\overline{1}$ is an identity element of $\mathbb{Z}_n^\times$.

(iv) Each element in $\mathbb{Z}_n^\times$ has an inverse in $\mathbb{Z}_n^\times$. This follows from Lemma 6.4(9). Let us recall its proof. If $\overline{a} \in \mathbb{Z}_n^\times$, with $a \in \mathbb{Z}$ and $(a,n) = 1$, then there are integers x,y such that $ax + ny = 1$. From this we get $\overline{ax} = \overline{1}$, so $\overline{x}$ is an inverse of $\overline{a}$. Yes, but this is not enough. We must further show that $\overline{x} \in \mathbb{Z}_n^\times$, or equivalently that $(x,n) = 1$. This follows from

the equation $ax + ny = 1$, since $d = (x,n)$ implies $d|x$, $d|n$, so $d|\ ax + ny$, so $d|1$, so $d = 1$.

Hence $\mathbb{Z}_n^*$ is a group under multiplication. □

$\mathbb{Z}_n^*$ is a finite group of order $\varphi(n)$. Using Lemma 11.7, we obtain $\bar{a}^{\varphi(n)} = \bar{1}$ for all $\bar{a} \in \mathbb{Z}_n^*$. Writing this in congruence notation, we get an important theorem of number theory due to L. Euler.

12.5 Theorem (Euler's theorem): *Let $n \in \mathbb{N}$. For all integers that are relatively prime to n, we have*

$$a^{\varphi(n)} \equiv 1 \pmod{n}.$$ □

The case when n is a prime number had already been observed by Pierre de Fermat (1601-1665). The result is known as Fermat's theorem or as Fermat's little theorem.

12.6 Theorem (Fermat's theorem): *If p is a positive prime number then*

$$a^{p-1} \equiv 1 \pmod{p}$$

for all integers a that are relatively prime to p (i.e., for all integers a such that $p \nmid a$. □

Multiplying both sides of the congruence $a^{p-1} \equiv 1 \pmod{p}$ by a, we get $a^p \equiv a \pmod{p}$. The latter congruence is true also without the hypothesis $(a,p) = 1$, since both a^p and a are congruent to 0 (mod p) when $(a,p) \neq 1$. This is also knows as Fermat's (little) theorem.

12.7 Theorem (Fermat's theorem): *If p is a prime number, then*

$$a^p \equiv a \pmod{p}$$

for all integers a. □

Exercises

1. Prove that $\mathbb{Z}_n^\times$ is an abelian group under multiplication.

2. Construct the multiplication tables of $\mathbb{Z}_n^\times$ for $n = 2,4,6,10,12$.

3. What are the orders of $\bar{2}$ in $\mathbb{Z}_3^\times$, $\bar{2}$ in $\mathbb{Z}_5^\times$, $\bar{3}$ in $\mathbb{Z}_7^\times$, $\bar{2}$ in $\mathbb{Z}_{11}^\times$, $\bar{2}$ in $\mathbb{Z}_{13}^\times$, $\bar{3}$ in $\mathbb{Z}_{17}^\times$, $\bar{2}$ in $\mathbb{Z}_{19}^\times$, $\bar{5}$ in $\mathbb{Z}_{23}^\times$? What do you guess?

4. Show that $\mathbb{Z}_3^\times$, $\mathbb{Z}_{3^2}^\times$, $\mathbb{Z}_{3^3}^\times$, $\mathbb{Z}_{3^4}^\times$ are cyclic.

5. Assume p is prime, $\mathbb{Z}_p^\times$ is cyclic, and $m \in \mathbb{N}$, $m \geqslant 2$. Prove that $\mathbb{Z}_{p^m}^\times$ is cyclic by establishing that, if $\bar{a}$ in $\mathbb{Z}_p^\times$ is a generator of $\mathbb{Z}_p^\times$, then either $\bar{a}$ or $\overline{a+p}$ in $\mathbb{Z}_{p^m}^\times$ is a generator of $\mathbb{Z}_{p^m}^\times$.

6. Find the order of $\bar{3}$ in $\mathbb{Z}_8^\times$, in $\mathbb{Z}_{16}^\times$, in $\mathbb{Z}_{32}^\times$, in $\mathbb{Z}_{64}^\times$.

7. Prove or disprove: if $a \in \mathbb{Z}$ and $a \equiv 5 \pmod 8$, then the order of $\bar{a}$ in $\mathbb{Z}_{2^m}^\times$ is 2^{m-2} for all $m \geqslant 3$.

8. Show that $\mathbb{Z}_{pq}^\times$ is not cyclic if p and q are positive odd prime numbers. (Hint: What is $\varphi(pq)$ and what is $a^{(p-1)(q-1)/2}$ congruent to (mod pq) if a is an integer relatively prime to pq?)

§13
Groups of Isometries

For any nonempty set X, the set S_X of all one-to-one mappings from X onto X is a group under the composition of mappings (Example 7.1(d)). In particular, if X happens to be the Euclidean plane E, then E is the set of all points in the plane and S_E is a group. We note that E is not merely an ordinary set of points. An important feature of E is that there is a measure of distance between the points of E. Among the mappings in S_E, we examine those functions which preserve the distance between any two points. Clearly, such functions will be more important than other ones in S_E, since such mappings respect an important structure of the Euclidean plane E.

We choose an arbitrary but fixed cartesian coordinate system on E. Each point P in E will then be represented by the ordered pair (x,y) of its coordinates. We will not distinguish between the point P and the ordered pair (x,y). So we write $(x,y)\alpha$ in place of $P\alpha$, where $\alpha \in S_E$. The distance between two points P,Q in E is given by $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$, if P and Q have coordinates $(x_1,y_1),(x_2,y_2)$, respectively. This distance will be denoted by $d(P,Q)$ or by $d((x_1,y_1),(x_2,y_2))$.

13.1 Definition: A mapping $\alpha \in S_E$ is called an *isometry (of E)* if

$$d(P\alpha,Q\alpha) = d(P,Q)$$

for any two points P,Q in E.

This word is derived from "isos" and "metron", meaning "equal" and "measure" in Greek. The set of all isometries of E will be denoted by *Isom* E. Since the identity mapping $\iota_E: E \to E$ is evidently an isometry, *Isom* E is a nonempty subset of S_E. In fact, *Isom* $E \leqslant S_E$.

13.2 Theorem: *Isom E is a subgroup of* S_E.

Proof: We must show that the product of two isometries and the inverse of an isometry are isometries (Lemma 9.2).

(i) Let $\alpha,\beta \in Isom\ E$. Then, for any two points P,Q in E

$$\begin{aligned} d(P\alpha\beta,Q\alpha\beta) &= d((P\alpha)\beta,(Q\alpha)\beta) \\ &= d(P\alpha,Q\alpha) && \text{(since } \beta \text{ is an isometry)} \\ &= d(P,Q) && \text{(since } \alpha \text{ is an isometry),} \end{aligned}$$

so $\alpha\beta \in Isom\ E$. Hence $Isom\ E$ is closed under multiplication.

(ii) Let $\alpha \in Isom\ E$ and let P,Q be any two points in E. Since $\alpha \in S_E$, there are uniquely determined points P',Q' in E such that $P'\alpha = P$, $Q'\alpha = Q$. Thus $P' = P\alpha^{-1}$, $Q' = Q\alpha^{-1}$. Then

$$\begin{aligned} d(P',Q') &= d(P'\alpha,Q'\alpha) && \text{(since } \alpha \text{ is an isometry)} \\ d(P\alpha^{-1},Q\alpha^{-1}) &= d(P,Q) \\ \alpha^{-1} &\in Isom\ E. \end{aligned}$$

Hence $Isom\ E \leqslant S_E$. □

We examine some special types of isometries, namely translations, rotations and reflections.

Loosely speaking, a translation shifts every point of E by the same amount in the same direction. In more detail, a translation is a mapping which "moves" any point (x,y) in E by a units in the direction of the x-axis and by b units in the direction of the y-axis (the directions being reversed when a or b is negative). See Figure 1. The formal definition is as follows.

13.3 Definition: A mapping $E \to E$ is called a *translation* if there are two real numbers a,b such that

$$(x,y) \to (x + a, y + b)$$

for all points (x,y) in E under this mapping.

The translation $(x,y) \to (x + a, y + b)$ will be denoted by $\tau_{a,b}$.

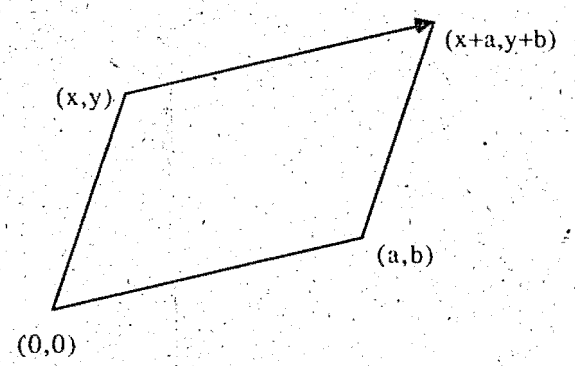


Figure 1

13.4 Lemma: *Let $\tau_{a,b}$ and $\tau_{c,d}$ be arbitrary translations.*

(1) $\tau_{a,b}\tau_{c,d} = \tau_{a+c,b+d}$.

(2) $\tau_{0,0} = \iota_E = \iota$.

(3) $\tau_{-a,-b}\tau_{a,b} = \iota = \tau_{a,b}\tau_{-a,-b}$.

Proof: (1) We have
$$\begin{aligned}(x,y)(\tau_{a,b}\tau_{c,d}) &= ((x,y)\tau_{a,b})\tau_{c,d}\\ &= (x + a, y + b)\tau_{c,d}\\ &= ((x + a) + c, (y + b) + d)\\ &= (x + (a + c), y + (b + d))\\ &= (x,y)\tau_{a+c,b+d}\end{aligned}$$
for all $(x,y) \in E$. Thus $\tau_{a,b}\tau_{c,d} = \tau_{a+c,b+d}$.

(2) We have $(x,y)\tau_{0,0} = (x + 0, y + 0) = (x,y) = (x,y)\iota$ for all $(x,y) \in E$. Thus $\tau_{0,0} = \iota$.

(3) From (1) and (2) we get $\tau_{-a,-b}\tau_{a,b} = \tau_{(-a)+a,(-b)+b} = \tau_{0,0} = \iota$ and likewise $\tau_{a,b}\tau_{-a,-b} = \tau_{a+(-a),b+(-b)} = \tau_{0,0} = \iota$. □

13.5 Lemma: *Any translation is an isometry.*

Proof: First of all, we must show that any translation belongs to S_E. Let $\tau_{a,b}$ be an arbitrary translation ($a,b \in \mathbb{R}$). There is a mapping $\psi: E \to E$ such that $\tau_{a,b}\psi = \iota = \psi\tau_{a,b}$, namely $\psi = \tau_{-a,-b}$ by Lemma 13.4(3). Thus $\tau_{a,b}$ is one-to-one and onto by Theorem 3.17(2). Hence $\tau_{a,b} \in S_E$.

Next we show $d((x_1,y_1),(x_2,y_2)) = d((x_1,y_1)\tau_{a,b}, (x_2,y_2)\tau_{a,b})$ for any two points $(x_1,y_1),(x_2,y_2)$ in E. We have

$$\begin{aligned} d((x_1,y_1)\tau_{a,b}, (x_2,y_2)\tau_{a,b}) &= d((x_1 + a, y_1 + b), (x_2 + a, y_2 + b)) \\ &= \sqrt{[(x_1+a) - (x_2+a)]^2 + [(y_1+a) - (y_2+a)]^2} \\ &= \sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \\ &= d((x_1,y_1),(x_2,y_2)) \end{aligned}$$

and so $\tau_{a,b} \in Isom\ E$. □

13.6 Theorem: *The set T of all translations is a subgroup of Isom E.*

Proof: Let $T = \{\tau_{a,b}: a,b \in \mathbb{R}\}$ be the set of all translations. T is a subset of *Isom* E by Lemma 13.5. From Lemma 13.4(2), $\iota = \tau_{0,0} \in T$, so $T \neq \emptyset$. Now we use our subgroup criterion (Lemma 9.2).

(i) The product of two translations $\tau_{a,b}$ and $\tau_{c,d}$ is a translation $\tau_{a+c,b+d} \in T$ by Lemma 13.4(1). So T is closed under multiplication.

(ii) The inverse of any translation $\tau_{a,b} \in T$ is also a translation $\tau_{-a,-b} \in T$ by Lemma 13.4(3). So T is closed under taking inverses.

Thus T is a subgroup of *Isom* E. □

Next we investigate rotations. By a rotation about a point C through an angle φ, we want to understand a mapping from E into E which sends the point C to C and whose effect on any point $P \neq C$ is as follows: we turn the line segment CP about the point C through the angle φ into a new line segment, say to CQ; the point P will be sent to the point Q (see Figure 2). We recall that positive values of φ measure counterclockwise angles and negative values of φ measure clockwise angles.

Rotations are most easily described in a polar coordinate system. We choose the center of rotation, the point C, as the pole. The initial ray is chosen arbitrarily. The point P with polar coordinates (r,θ) is then sent to the point Q whose polar coordinates are $(r,\theta + \varphi)$. If C is the origin and

the initial ray is the positive x-axis of a cartesian coordinate system, then the polar and cartesian coordinates of a point P are connected by

$$x = r\cos\theta \qquad y = r\sin\theta.$$

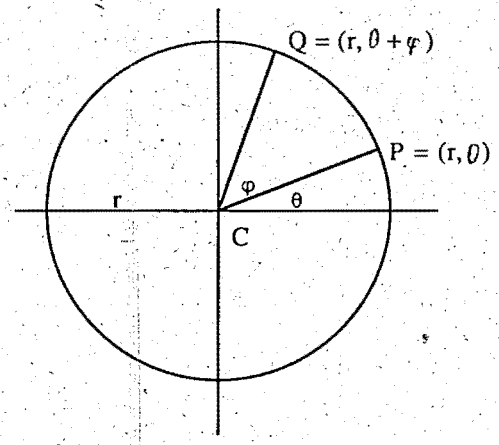


Figure 2

In our fixed cartesian coordinate system, the image of any point $P = (x,y)$ can be found as follows. If P has polar coordinates (r,θ), then its image will have polar coordinates $(r,\theta + \varphi)$, so the cartesian coordinates x',y' of $Q := (r,\theta + \varphi)$ are

$$\begin{aligned} x' = r\cos(\theta + \varphi) &= r(\cos\theta\cos\varphi - \sin\theta\sin\varphi) \\ &= (r\cos\theta)\cos\varphi - (r\sin\theta)\sin\varphi \\ &= x\cos\varphi - y\sin\varphi, \\ y' = r\sin(\theta + \varphi) &= r(\sin\theta\cos\varphi + \cos\theta\sin\varphi) \\ &= (r\sin\theta)\cos\varphi + (r\cos\theta)\sin\varphi \\ &= y\cos\varphi + x\sin\varphi \\ &= x\sin\varphi + y\cos\varphi. \end{aligned}$$

This suggests the following formal definition.

13.7 Definition: A mapping $E \to E$ is called a *rotation about the origin through an angle* φ if there is a real number φ such that

$$(x,y) \to (x\cos\varphi - y\sin\varphi, x\sin\varphi + y\cos\varphi)$$

for all points (x,y) in E under this mapping.

The rotation $(x,y) \to (x\cos\varphi - y\sin\varphi, x\sin\varphi + y\cos\varphi)$ will be denoted by ρ_φ. We have an analogue of Lemma 13.4 for rotations.

13.8 Lemma: *Let ρ_φ and ρ_ψ be arbitrary rotations about the origin.*
(1) $\rho_\varphi\rho_\psi = \rho_{\varphi+\psi}$.
(2) $\rho_0 = \iota_E = \iota$.
(3) $\rho_{-\varphi}\rho_\varphi = \iota = \rho_\varphi\rho_{-\varphi}$.

Proof: (1) We have $(x,y)(\rho_\varphi\rho_\psi) = ((x,y)\rho_\varphi)\rho_\psi$
$= (x\cos\varphi - y\sin\varphi, x\sin\varphi + y\cos\varphi)\rho_\psi$
$= ((x\cos\varphi - y\sin\varphi)\cos\psi - (x\sin\varphi + y\cos\varphi)\sin\psi\ ,\ (x\cos\varphi - y\sin\varphi)\sin\psi + (x\sin\varphi + y\cos\varphi)\cos\psi)$
$= (x(\cos\varphi\cos\psi - \sin\varphi\sin\psi) - y(\sin\varphi\cos\psi + \cos\varphi\sin\psi),$
$\qquad x(\cos\varphi\sin\psi + \sin\varphi\cos\psi) + y(-\sin\varphi\sin\psi + \cos\varphi\cos\psi))$
$= (x\cos(\varphi+\psi) - y\sin(\varphi+\psi), x\sin(\varphi+\psi) + y\cos(\varphi+\psi))$
$= (x,y)\rho_{\varphi+\psi}$
for all $(x,y) \in E$. Thus $\rho_\varphi\rho_\psi = \rho_{\varphi+\psi}$.

(2) We have $(x,y)\rho_0 = (x\cos 0 - y\sin 0, x\sin 0 + y\cos 0) = (x - 0, 0 + y)$
$= (x,y) = (x,y)\iota$
for all $(x,y) \in E$. Thus $\rho_0 = \iota$.

(3) From (1) and (2) we get $\rho_{-\varphi}\rho_\varphi = \rho_{(-\varphi)+\varphi} = \rho_0 = \iota$ and likewise $\rho_\varphi\rho_{-\varphi} = \rho_{\varphi+(-\varphi)} = \rho_0 = \iota$. □

Lemma 13.8 was to be expected. When we carry out a rotation through an angle φ and then a rotation through an angle ψ, we have in effect a rotation through an angle $\varphi + \psi$. This is what Lemma 13.8(1) states. Also, when we carry out a rotation through an angle φ and then a rotation through the same angle in the reverse direction, the final result will be: no net motion at all. This is what Lemma 13.8(3) states.

13.9 Lemma: *Any rotation about the origin is an isometry.*

Proof: First of all, we must show that any rotation about the origin belongs to S_E. Let ρ_φ be an arbitrary rotation about the origin ($\varphi \in \mathbb{R}$). There is a mapping $\psi: E \to E$ such that $\rho_\varphi\psi = \iota = \psi\rho_\varphi$, namely $\psi = \rho_{-\varphi}$ by Lemma 13.8(3). Thus ρ_φ is one-to-one and onto by Theorem 3.17(2). Hence $\rho_\varphi \in S_E$.

Now we prove that ρ_φ preserves distance. For any two points $(x,y),(u,v)$ in E, we have $d^2((x,y)\rho_\varphi,(u,v)\rho_\varphi)$

$= d^2((x\cos\varphi - y\sin\varphi, x\sin\varphi + y\cos\varphi),(u\cos\varphi - v\sin\varphi, u\sin\varphi + v\cos\varphi))$

$= [(x-u)\cos\varphi - (y-v)\sin\varphi]^2 + [(x-u)\sin\varphi + (y-v)\cos\varphi]^2$

$= (x-u)^2\cos^2\varphi - 2(x-u)(y-v)\cos\varphi\sin\varphi + (y-v)^2\sin^2\varphi$

$+(x-u)^2\sin^2\varphi + 2(x-u)(y-v)\cos\varphi\sin\varphi + (y-v)^2\cos^2\varphi$

$= (x-u)^2(\cos^2\varphi + \sin^2\varphi) + (y-v)^2(\sin^2\varphi + \cos^2\varphi)$

$= (x-u)^2 + (y-v)^2$

$= d^2((x,y),(u,v))$,

hence $d((x,y)\rho_\varphi,(u,v)\rho_\varphi) = d((x,y),(u,v))$. So ρ_φ is an isometry. □

13.10 Theorem: *The set R of all rotations about the origin is a subgroup of Isom E.*

Proof: Let $R = \{\rho_\varphi : \varphi \in \mathbb{R}\}$ be the set of all rotations about the origin. R is a subset of *Isom* E by Lemma 13.9. By Lemma 13.8(2), $\iota = \rho_0 \in R$, so $R \neq \varnothing$. Now we use our subgroup criterion (Lemma 9.2).

(i) The product of two rotations ρ_φ and ρ_ψ about the origin is a rotation $\rho_{\varphi+\psi} \in R$ about the origin by Lemma 13.8(1). So R is closed under multiplication.

(ii) The inverse of any rotation $\rho_\varphi \in R$ about the origin is also a rotation $\rho_{-\varphi} \in R$ about the origin by Lemma 13.8(3). So R is closed under taking inverses.

Thus R is a subgroup of *Isom* E. □

So far, we have been dealing with rotations about the origin. What about rotations about an arbitrary point C, whose coordinates are (a,b), say. A rotation about C through an angle φ will map a point P with coordinates $(x+a, y+b)$ to a point Q with coordinates $(x'+a, y'+b)$, where (x',y') is the point to which (x,y) is mapped under a rotation about the origin through an angle φ. So the image of $(x,y)\tau_{a,b}$ will be $(x,y)\rho_\varphi\tau_{a,b}$. See Figure 3. This suggests the following formal definition.

13.11 Definition: Let $C = (a,b)$ be a point in E. The mapping $(\tau_{a,b})^{-1}\rho_\varphi\tau_{a,b}: E \to E$ is called a *rotation about C through an angle φ*.

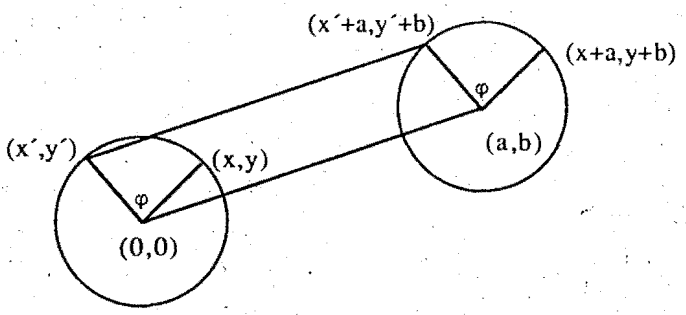


Figure 3

We put $(\tau_{a,b})^{-1}R\tau_{a,b} := \{(\tau_{a,b})^{-1}\rho_\varphi\tau_{a,b}: \rho_\varphi \in R\}$. This is the set of all rotations about the point (a,b). It is a subgroup of *Isom* E. The proof of this statement is left to the reader.

Now we examine reflections. The cartesian equations of a reflection are very cumbersome. For this reason, we give a coordinate-free definition of reflections. We need some notation. Let P,Q be distinct points in the plane E. In what follows, $\overrightarrow{PQ}$ will denote the line through P and Q, and $\overline{PQ}$ will denote the line segment between P and Q. So $\overline{PQ}$ is the set of points R in E such that $d(P,R) + d(R,Q) = d(P,Q)$.

The geometric idea of a reflection is that there is a line m and that each point P is mapped to its "mirror image" Q on the other side of m. So $\overline{PQ}$ is perpendicular to m and $d(P,R) = d(R,Q)$, where R is the point of intersection of m and $\overline{PQ}$. See Figure 4.

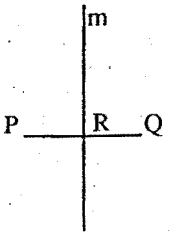


Figure 4

13.12 Definition: Let m be a straight line in E and let $\sigma_m : E \to E$ be the mapping defined by

$P\sigma_m = P$ if P is on m

and

$P\sigma_m = Q$ if P is not on m and if m is the perpendicular bisector of $\overline{PQ}$.

σ_m is called the *reflection in the line m.*

The perpendicular bisector of $\overline{PQ}$ is the line that is perpendicular to $\overline{PQ}$ and that intersects $\overline{PQ}$ at a point R such that $d(P,R) = d(R,Q)$. It is also the locus of all points in E which are equidistant from P and Q. So it is the set $\{R \in E: d(P,R) = d(Q,R)\}$. We will make use of this description of the perpendicular bisector in the sequel without explicit mention.

13.13 Lemma: *Let σ_m be the reflection in a line m. Then $\sigma_m \neq \iota = \sigma_m^2$.*

Proof: $P\sigma_m \neq P$ if P is not on the line m and so $\sigma_m \neq \iota$. Now we prove that $\sigma_m^2 = \iota$. We have $P\sigma_m^2 = P(\sigma_m\sigma_m) = (P\sigma_m)\sigma_m = P\sigma_m = P$ when P is a point on m by definition. It remains to show $P\sigma_m^2 = P$ also when P is not on m. Let P be a point not on m and let $Q = P\sigma_m$, $P_1 = Q\sigma_m$. Then Q is not on m and m is the perpendicular bisector of $\overline{PQ}$ as well as of $\overline{QP_1}$. So $\overline{PQ}$ and $\overline{QP_1}$ are parallel lines. Since they have a point Q in common, they are identical lines. Let R be the point at which m and $\overline{PQ}$ intersect. So $P_1 \neq Q$ and P_1 is that point on $\overline{PQ}$ for which $d(Q,R) = d(P_1,R)$. Since P is on $\overline{PQ}$ and $d(Q,R) = d(P,R)$, we obtain $P = P_1$, as was to be proved. □

13.14 Lemma: *Any reflection in a line is an isometry.*

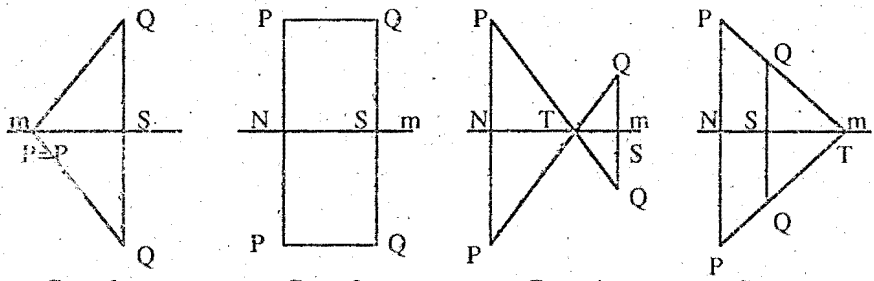
Proof: Let m be a line, σ_m the reflection in m and let P,Q be arbitrary points in the plane. We put $P_1 = P\sigma_m$, $Q_1 = Q\sigma_m$. We are to show $d(P,Q) = d(P_1,Q_1)$. We distinguish several cases.

Case 1. Assume both P and Q are on m. Then $P_1 = P$ and $Q_1 = Q$. So $d(P,Q) = d(P_1,Q_1)$.

Case 2. Assume one of the points is on m, the other is not. We suppose, without loss of generality, that P is on m and Q is not on m. Let $Q\,Q_1$ intersect m at S. Then $d(Q,S) = d(S,Q_1)$, $d(P,S) = d(S,P_1)$ since $P = P_1$ and

the angles $\angle PSQ$ and $\angle P_1Q_1S$ are both right angles. By the side-angle-side condition, the triangles $\triangle QPS$ and $\triangle Q_1P_1S$ are congruent. So the corresponding sides $\overline{PQ}$ and $\overline{P_1Q_1}$ have equal length. This means $d(P,Q) = d(P_1,Q_1)$.

From now on, assume that neither P nor Q is on m. Let m intersect $\overline{PP_1}$ at N and $\overline{QQ_1}$ at S.



Case 2 Case 3 Case 4 Case 4

Figure 5

Case 3. Assume that $\overline{PQ}$ is parallel to m. Then the quadrilaterals $\square NPQS$ and $\square NP_1Q_1S$ are rectangles. So $\square PP_1Q_1Q$ is a rectangle and the sides $\overline{PQ}$ and $\overline{P_1Q_1}$ have equal length. This means $d(P,Q) = d(P_1,Q_1)$.

Case 4. Assume that $\overline{PQ}$ is not parallel to m. Then $\overline{PQ}$ intersects m at a point T. As in case 2, $\triangle PTN$ and $\triangle P_1TN$ are congruent, and $\triangle QST$ and $\triangle Q_1ST$ are congruent, so $d(P,T) = d(P_1,T)$ and $d(Q,T) = d(Q_1,T)$. Also, $\angle STQ_1 = \angle STQ = \angle NTP = \angle NTP_1$, which shows that P_1,T,Q_1 lie on a straight line. Then we obtain

$$\begin{aligned} d(P_1,Q_1) &= d(P_1,T) \mp d(T,Q_1) \\ &= d(P_1,T) \mp d(Q_1,T) \\ &= d(P,T) \mp d(Q,T) \\ &= d(P,T) \mp d(T,Q) \\ &= d(P,Q), \end{aligned}$$

where the upper or lower sign is to be taken according as whether P,Q are on the same or on the opposite sides of m. □

13.15 Theorem: *Let m be a line in E. Then $\{\iota,\sigma_m\}$ is a subgroup of Isom E.*

Proof: $\{\iota,\sigma_m\}$ is a finite nonempty subset of *Isom E* by Lemma 13.14. It is closed under multiplication by Lemma 13.13. So it is a subgroup of *Isom E* by Lemma 9.3(1). □

Translations, rotations and reflections are isometries. Thus the products of any number of these mappings, carried out in any order, will be isometries, too. We show in the rest of this paragraph that all isometries are obtained in this way. We need some lemmas.

13.16 Lemma: *Let P,Q,R be arbitrary points in E.*
(1) *There is a unique translation that maps P to Q.*
(2) *If $d(P,Q) = d(P,R)$, there is a rotation about P that maps Q to R.*

Proof: (1) When $P = (a,b)$ and $Q = (c,d)$, say, then $\tau_{m,n}$ maps P to Q if and only if $(a + m, b + n) = (c,d)$, i.e., if and only if $m = c - a$, $n = d - b$. So $\tau_{c-a,d-b}$ is the unique translation that maps P to Q.

(2) We draw the circle whose center is at P and whose radius is equal to $d(P,Q)$. This circle passes through R by hypothesis. Let φ be the angle which the circular arc $\overset{\frown}{QR}$ subtends at the center P. Then a rotation about P through an angle φ maps Q to R. □

The next lemma states that an isometry is completely determined by its effect on three points not lying on a line.

13.17 Lemma: *Let P,Q,R be three distinct points in E that do not lie on a straight line. Let α,β be isometries such that $P\alpha = P\beta$, $Q\alpha = Q\beta$, $R\alpha = R\beta$. Then $\alpha = \beta$.*

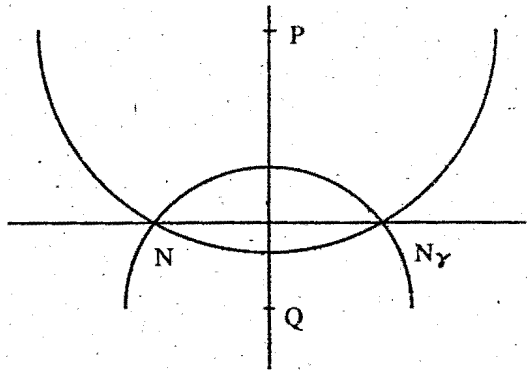


Figure 6

Proof: We put $\alpha\beta^{-1} = \gamma$. We suppose $\gamma \neq \iota$ and try to reach a contradiction. If $\gamma \neq \iota$, then there is a point N in E such that $N \neq N\gamma$. Since $P\alpha = P\beta$ by hypothesis, $P\gamma = P$ and so $P \neq N$. Similarly $Q \neq N$ and $R \neq N$. Now γ is an isometry, so $d(P,N\gamma) = d(P\gamma,N\gamma) = d(P,N)$ and likewise $d(Q,N\gamma) = d(Q,N)$ and $d(R,N\gamma) = d(R,N)$. So the circle with center at P and radius $d(P,N)$ and the circle with center at Q and radius $d(Q,N)$ intersect at the points N and $N\gamma$. Then $\overline{PQ}$ is the perpendicular bisector of $\overline{NN\gamma}$. Here we used $N \neq N\gamma$. But $d(R,N\gamma) = d(R,N)$ and R lies therefore on the perpendicular bisector of $\overline{NN\gamma}$, i.e., R lies on $\overline{PQ}$, contrary to the hypothesis that P,Q,R do not lie on a straight line. Hence necessarily $\gamma = \iota$ and $\alpha = \beta$. □

13.18 Theorem: *Let P,Q,R be three distinct points in E that do not lie on a straight line and let P',Q',R' be three distinct points in E. Assume that $d(P,Q) = d(P',Q')$, $d(P,R) = d(P',R')$, $d(Q,R) = d(Q',R')$. Then there is a translation τ, a rotation ρ (about an appropriate point and through a suitable angle) and a reflection σ such that*

$$P' = P\beta,\ Q' = Q\beta,\ R' = R\beta,$$

where β denotes $\tau\rho$ or $\tau\rho\sigma$.

Proof: By Lemma 13.16(1), there is a translation τ that maps P to P'. We put $Q_1 = Q\tau$ and $R_1 = R\tau$. Since τ is an isometry, $d(P,Q) = d(P\tau,Q\tau) = d(P',Q_1)$, so $d(P',Q_1) = d(P',Q')$. By Lemma 13.16(2), there is a rotation about P' that maps Q_1 to Q'. Let us denote this rotation by ρ. Then $P'\rho = P'$. Put $R_1\rho = R_2$.

Here it may happen that $R_2 = R'$. Putting $\beta = \tau\rho$ in this case, we have $P' = P\beta$, $Q' = Q\beta$, $R' = R\beta$, as claimed.

Suppose now $R_2 \neq R'$. From $d(P',R') = d(P,R) = d(P\tau\rho, R\tau\rho) = d(P',R_2)$ and $d(Q',R') = d(Q,R) = d(Q\tau\rho, R\tau\rho) = d(Q',R_2)$, we deduce that both P' and Q' lie on the perpendicular bisector of $\overline{R'R_2}$. Denoting the reflection in the line $\overleftrightarrow{P'Q'}$ by σ, we get $P'\sigma = P'$, $Q'\sigma = Q'$ and $R_2\sigma = R'$. Putting $\beta = \tau\rho\sigma$ in this case, we have $P' = P\beta$, $Q' = Q\beta$, $R' = R\beta$, as claimed.

The proof is summarized schematically below.

$$\begin{array}{ccccccc} & \tau & & \rho & & \sigma & \\ P & \to & P' & \to & P' & \to & P' \\ Q & \to & Q_1 & \to & Q' & \to & Q' \\ R & \to & R_1 & \to & R_2 & \to & R' \end{array}$$ □

13.19 Theorem: *Every isometry can be written as a product of translations, rotations and reflections. In fact, if α is an arbitrary isometry, then there is a translation τ, a rotation ρ and a reflection σ such that*

$$\alpha = \tau\rho \text{ or } \tau\rho\sigma.$$

Proof: Let α be an arbitrary isometry. Choose any three distinct points P,Q,R in E not lying on a straight line. Then $d(P,Q) = d(P\alpha, Q\alpha)$, $d(P,R) = d(P\alpha, R\alpha)$, $d(Q,R) = d(Q\alpha, R\alpha)$. So the hypotheses of Theorem 13.18 are satisfied with $P' = P\alpha$, $Q' = Q\alpha$, $R' = R\alpha$ and there is a translation τ, a rotation ρ and a reflection σ such that

$$P\alpha = P\beta, Q\alpha = Q\beta, R\alpha = R\beta,$$

where $\beta = \tau\rho$ or $\tau\rho\sigma$. By Lemma 13.17, $\alpha = \beta$. Thus $\alpha = \tau\rho$ or $\tau\rho\sigma$. □

Exercises

1. Let m be the line in E whose cartesian equation is $ax + by + c = 0$. Show that the reflection σ_m in the line m is given by

$$(u,v)\sigma_m = \left(u - \frac{2a}{a^2 + b^2}(au + bv + c),\ v - \frac{2b}{a^2 + b^2}(au + bv + c)\right).$$

2. Let m and n be two distinct lines intersecting at a point P. Show that $\sigma_m\sigma_n$ is a rotation about P. Through which angle?

3. Let m and n be parallel lines. Show that $\sigma_m \sigma_n$ is a translation.

4. Prove that every rotation and every translation can be written as a product of two reflections.

5. Prove that every isometry can be written as a product of reflections.

6. A *halfturn* $\sigma_P = \sigma_{(a,b)}$ about a point $P = (a,b)$ is defined as the mapping given by

$$(x,y) \rightarrow (2a - x, 2b - y)$$

for all points (x,y) in E. Show that any halfturn is an isometry of order two. Prove that the product of three halfturns is a halfturn.

7. Prove that a halfturn σ_P is the product of any two reflections in lines intersecting perpendicularly at P.

8. Prove that a product of two halfturns is a translation.

9. Show that the set of all translations and halfturns is a subgroup of *Isom E*.

10. Prove that the product of four reflections can be written as a product of two reflections.

11. Show that $\rho_{2\pi/n}$ generates a cyclic subgroup or order n of *Isom E*.

12. Prove that every nonidentity translation is of infinite order and that ρ_φ is of finite order if and only if φ is a rational multiple of π.

§14
Dihedral Groups

In this paragraph, we examine the symmetry groups of regular polygons.

Let F be any nonempty subset of the Euclidean plane E. Here F might be a set with a single point, a line, a geometric figure or an arbitrary subset of E. Let $\alpha \in S_E$. We put

$$F\alpha = \{x\alpha : x \in F\} = \{y \in E : y = x\alpha \text{ for some } x \in F\}.$$

$F\alpha$ is called the *image of* F *under* α. Clearly,

$$F1 = \{x1 : x \in F\} = \{x : x \in F\} = F$$

and we have
$$\begin{aligned} F(\alpha\beta) &= \{x(\alpha\beta) : x \in F\} = \{(x\alpha)\beta : x \in F\} \\ &= \{(x\alpha)\beta : x\alpha \in F\alpha\} = \{y\beta : y \in F\alpha\} = (F\alpha)\beta \end{aligned}$$

for all $\alpha, \beta \in S_E$. We record this as a lemma.

14.1 Lemma: *Let F be a nonempty subset of E and let $\alpha, \beta \in S_E$. Then*

$$F1 = F$$

and
$$F(\alpha\beta) = (F\alpha)\beta. \qquad \square$$

Let P be a point in E and $\alpha \in S_E$. We say α *fixes* P if $P\alpha = P$. We also say P is a *fixed point of* α in this case. Let $\emptyset \neq F \subseteq E$. We say α *fixes* F (as a set) if $F\alpha = F$. This means of course $F\alpha \subseteq F$ and $F \subseteq F\alpha$, so $P\alpha \in F$ for all $P \in F$ and also, for every $Q \in F$, there is a $P \in F$ such that $Q = P\alpha$. The reader should not confuse this with α fixing F pointwise. We say that α *fixes* F *pointwise* if $P\alpha = P$ for all $P \in F$, i.e., if α fixes every point of F. As an example, let A be the x-axis $\{(x,0) : x \in \mathbb{R}\}$. The translation $\tau_{1,0}$ fixes A as a set, but not pointwise. On the other hand, the reflection $\sigma : (x,y) \to (x,-y)$ in the x-axis fixes A pointwise.

This terminology is meaningful for all elements of S_E, but we consider only isometries in this paragraph.

14.2 Definition: Let F be a nonempty subset of E and let $\alpha \in Isom\ E$. If $F\alpha = F$, then α is called a *symmetry of* F.

So a symmetry of F is an isometry that fixes F as a set. A symmetry of F is *not* a property of F. It is a mapping.

14.3 Examples: (a) Let $F = \{(0,0)\}$ be the subset of E consisting of the origin only. Any rotation ρ_φ about the origin is a symmetry of F, since any rotation about the origin is an isometry and fixes the origin (or, equivalently, fixes F).

(b) Let $F = \{(x,y) \in E: y = mx\}$ be the line whose cartesian equation is $y = mx$ (where $m \in \mathbb{R}$). Then the translation $\tau_{1,m}$ is a symmetry of F since

$$\begin{aligned} F\tau_{1,m} &= \{f\tau_{1,m} \in E: f \in F\} \\ &= \{(x,y)\tau_{1,m} \in E: y = mx\} \\ &= \{(x+1,y+m) \in E: y = mx\} \\ &= \{(x+1,(x+1)m) \in E: x \in \mathbb{R}\} \\ &= \{(u,v) \in E: v = mu\} \\ &= F. \end{aligned}$$

Similarly, all translations of the form $\tau_{a,am}$ is a is a symmetry of F. We note that such translations form a group. In fact, the symmetries of any nonempty subset of E form a group.

14.4 Theorem: *Let F be a nonempty subset of the Euclidean plane E and let Sym F be the set of all symmetries of F, so that*

$$Sym\ F := \{\alpha \in Isom\ E: F\alpha = F\}.$$

Then Sym F is a subgroup of Isom E.

Proof: We have $F\iota = F$ by Lemma 14.1, so $\iota \in Sym\ F$ and $Sym\ F$ is not empty. Now we use Lemma 9.2.

(i) If $\alpha,\beta \in Sym\ F$, then $F\alpha = F$ and $F\beta = F$, so $F(\alpha\beta) = (F\alpha)\beta = F\beta = F$ by Lemma 14.1. Thus $\alpha\beta \in Sym\ F$.

(ii) If $\alpha \in Sym\ F$, then $F\alpha = F$, so $F(\alpha^{-1}) = (F\alpha)\alpha^{-1} = F(\alpha\alpha^{-1}) = F\iota = F$ by Lemma 14.1. Thus $\alpha^{-1} \in Sym\ F$.

It follows that $Sym\ F \leqslant Isom\ E$. □

14.5 Definition: Let F be a nonempty subset of E. Then

$$Sym\ F = \{\alpha \in Isom\ E: F\alpha = F\}$$

is called the *symmetry group of* F.

We now study the symmetry groups of regular polygons. For our purposes, it will be convenient to define regular polygons as follows. Let K be a circle and let $P_1,P_2,\ldots,P_n$ be n points on this circle K such that each one of the arcs $\overset{\frown}{P_1P_2},\overset{\frown}{P_2P_3},\ldots,\overset{\frown}{P_{n-1}P_n}$ subtends an angle of $2\pi/n$ radians at the center of K (where $n \geqslant 3$). So the points $P_1,P_2,\ldots,P_n$ divide the circle K into n circular arcs of equal length. The union of the line segments $\overline{P_1P_2},\overline{P_2P_3},\ldots,\overline{P_{n-1}P_n},\overline{P_nP_1}$ is called a *regular n-gon*. The circle K is called *the circumscribing circle* of this regular n-gon. This is justified since a regular n-gon has a unique circumscribing circle. The center of the circumscribing circle is called the *center* of the regular n-gon and the points $P_1,P_2,\ldots,P_n$ are called the *vertices* of the regular n-gon.

Let F be a regular n-gon. We want to determine $Sym\ F$. It is geometrically evident that any α in $Sym\ F$ maps a vertex to a vertex and fixes the center of F. We use this fact without proof. A proof is outlined in the exercises at the end of this paragraph. Let $P_1,P_2,\ldots,P_n$ be the vertices and let C be the center of F. We assume the notation so chosen that $P_1,P_2,\ldots,P_n$ are consecutive vertices as we trace the regular n-gon counterclockwise. In the following discussion, P_{n+1} will stand for P_1, P_{n+2} for P_2, in general P_{n+k} for P_k. In other words, the indices will be read modulo n.

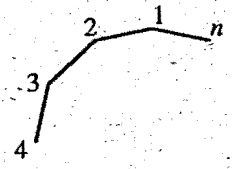


Figure 1

Now let $\alpha \in Sym\ F$. Then α is completely determined by its effect on three distinct points not on a straight line (Lemma 13.17). For example, α is determined by $C\alpha, P_1\alpha, P_2\alpha$. We have already remarked that $C\alpha = C$. Also $P_1\alpha = P_k$ for some $k \in \{1,2, \ldots, n\}$. What about P_2? Since α is an isometry, $P_2\alpha$ will be a vertex whose distance from P_k is equal to the distance between P_1 and P_2. Thus $P_2\alpha$ will be adjacent to P_k: it is either P_{k-1} or P_{k+1}. We see that there are n choices for $P_1\alpha$ and, once the choice for $P_1\alpha$ has been made, there are two choices for $P_2\alpha$. Hence there are at most $n.2 = 2n$ isometries in $Sym\ F$. We exhibit $2n$ symmetries of F and this will prove $|Sym\ F| = 2n$.

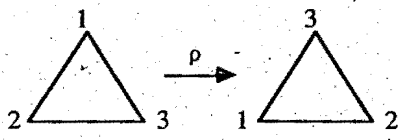


Figure 2

First we examine the special case $n = 3$, when F is an equilateral triangle. Consider a rotation about the center of F through an angle of $2\pi/3$ radians, which we denote by ρ. Under ρ, the vertices P_1, P_2, P_3 take the places of $P_2, P_3, P_4 = P_1$ respectively. It is seen from Figure 3 that ρ^2 maps P_1, P_2, P_3 respectively to P_3, P_2, P_1 and that ρ^3 fixes P_1, P_2, P_3, which implies $\rho^3 = \iota$. We found three symmetries of F, namely ι, ρ, ρ^2. Since $\rho^3 = \iota \neq \rho$, we see that $<\rho>$ is a cyclic subgroup of order 3 of $Sym\ F$.

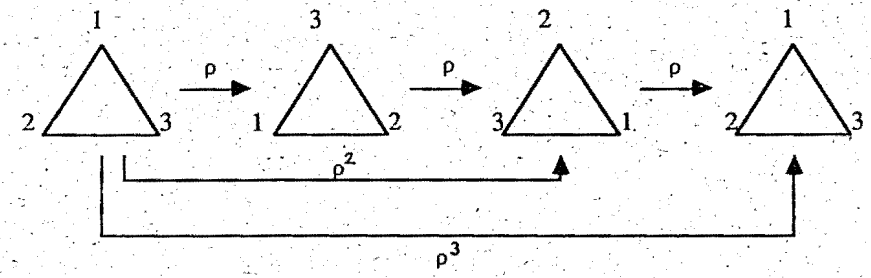


Figure 3

Now consider the reflection in the perpendicular bisector of $\overline{P_2P_3}$, which passes through P_1. We designate this reflection as σ. Under σ, the vertex P_1 remains fixed and the vertices P_2 and P_3 exchange their places. We know $\sigma^2 = \iota$ (Lemma 13.13). From Figure 4, we read off $\sigma\rho^{-1} = \sigma\rho^2 = \rho\sigma$. Using ρ and σ, we obtain two new symmetries of F, namely $\rho\sigma$ and $\rho^2\sigma$. The reader may check that $\rho\sigma$ is the reflection in the perpendicular bisector of $\overline{P_1P_3}$ and that $\rho^2\sigma$ is the reflection in the perpendicular bisector of $\overline{P_1P_2}$. From the geometric meaning of these mappings, or from their effect on P_1, P_2, P_3, we infer that $\iota, \rho, \rho^2, \sigma, \rho\sigma, \rho^2\sigma$ are distinct. Thus they form the symmetry group of F: $Sym\ F = \{\iota, \rho, \rho^2, \sigma, \rho\sigma, \rho^2\sigma\}$. In particular, $|Sym\ F|$ is equal to 6.

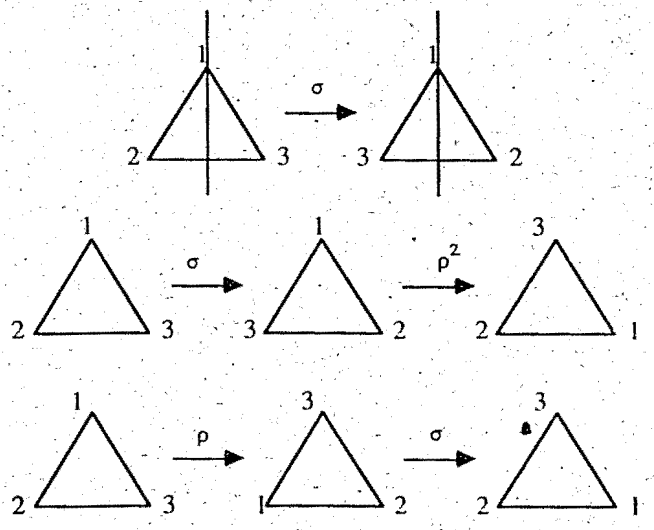


Figure 4

The discussion of a general regular polygon follows much the same lines. Consider a rotation about the center of F through an angle of $2\pi/n$ radians, which we denote by ρ. Under ρ, the vertices $P_1,P_2, \ldots ,P_n$ are mapped respectively to $P_2,P_3, \ldots ,P_n,P_1$. It is seen that ρ^k maps $P_1,P_2, \ldots ,P_n$ respectively to $P_{k+1},P_{k+2}, \ldots ,P_{k+n}$, where k is any integer. Thus $\rho^k = \iota$ if and only if $P_{k+i} = P_i$, that is, if and only if $k + i \equiv i \pmod{n}$ for all i, so if and only if $n|k$, from which we obtain $o(\rho) = n$. In this way, we found n symmetries of F, namely $\iota,\rho,\rho^2,\ldots,\rho^n$. Here $<\rho>$ is a cyclic subgroup of order n of *Sym F*.

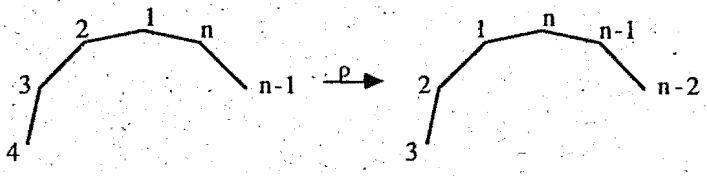


Figure 5

Now consider the reflection σ in the angular bisector of the angle $\angle P_nP_1P_2$. The bisector of this angle passes through $P_{(n/2)+1}$ if n is even and through the midpoint of $\overline{P_{(n+1)/2}P_{(n+3)/2}}$ if n is odd. One reads off from Figure 6 that $P_k\sigma = P_{n+2-k}$ for $k = 1,2, \ldots ,n$.

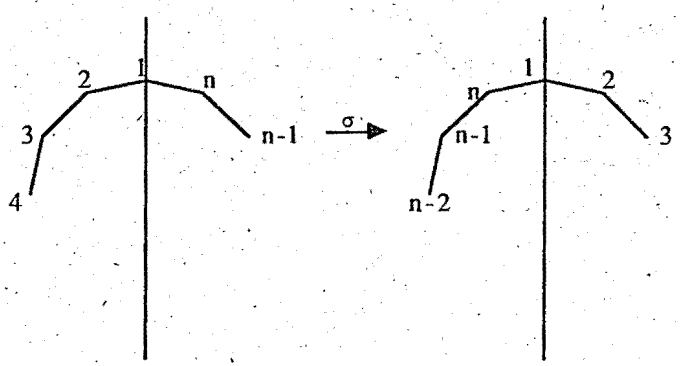


Figure 6

Thus we have, for any $j = 1,2, \ldots ,n$,

$$P_j\sigma\rho = P_{n+2-j}\rho = P_{(n+2-j)+1} = P_{n-j+3},$$

$$P_j\rho^{-1}\sigma = P_{j-1}\sigma = P_{(n+2)-(j-1)} = P_{n-j+3},$$

so $\sigma\rho = \rho^{-1}\sigma$, as can be seen from Figure 7 too.

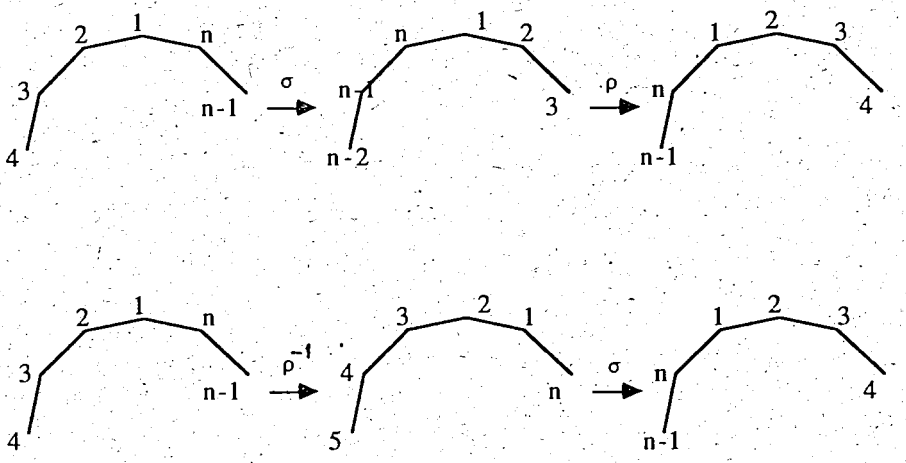


Figure 7

Using ρ and σ, we obtain $n - 1$ new symmetries $\rho\sigma, \rho^2\sigma, \ldots, \rho^{n-1}\sigma$ of F. These are reflections in certain lines. The reader may verify this assertion in the case of squares, regular pentagons, regular hexagons and regular heptagons. In particular, we have $(\rho^m\sigma)^2 = 1$ for any $m = 0,1, \ldots, n - 1$. This follows also from the lemma below.

14.6 Lemma: *Let G be a group and let $\rho,\sigma \in G$ be such that $\sigma^2 = 1$ and $\sigma\rho = \rho^{-1}\sigma$. Then $\sigma\rho^n = \rho^{-n}\sigma$ for all $n \in \mathbb{Z}$.*

Proof: The claim is certainly true when $n = 0$, and also when $n = 1$ by hypothesis. We prove $\sigma\rho^n = \rho^{-n}\sigma$ for all $n \in \mathbb{N}$ by induction on n. Suppose we proved it for $n = k \in \mathbb{N}$, so that $\sigma\rho^k = \rho^{-k}\sigma$, then it is true for $n = k + 1$, since $\sigma\rho^{k+1} = \sigma(\rho^k\rho) = (\sigma\rho^k)\rho = (\rho^{-k}\sigma)\rho = \rho^{-k}(\sigma\rho) = \rho^{-k}(\rho^{-1}\sigma) = (\rho^{-k}\rho^{-1})\sigma = \rho^{-(k+1)}\sigma$. This shows $\sigma\rho^n = \rho^{-n}\sigma$ for all $n \geqslant 0$. We must further show this when $n < 0$, or, equivalently, that $\sigma\rho^{-n} = \rho^n\sigma$ for all $n \in \mathbb{N}$. This will follow from what we proved above, with ρ^{-1},σ in place of ρ,σ. Observe that $\sigma^2 = 1$, $\sigma\rho = \rho^{-1}\sigma$ implies

$$\begin{aligned} \sigma\sigma\rho &= \sigma\rho^{-1}\sigma \\ \rho &= \sigma\rho^{-1}\sigma \\ \rho\sigma &= \sigma\rho^{-1}\sigma\sigma \end{aligned}$$

$$\rho\sigma = \sigma\rho^{-1}$$

and, taking inverses, $\sigma\rho^{-1} = (\rho^{-1})^{-1}\sigma$,

so the hypothesis is valid with ρ^{-1}, σ in place of ρ, σ. Then we get

$$\sigma(\rho^{-1})^n = (\rho^{-1})^{-n}\sigma \qquad \text{for all } n \in \mathbb{N},$$

and thus $\sigma\rho^{-n} = \rho^n\sigma$ for all $n \in \mathbb{N}$. This completes the proof. □

We found $2n$ symmetries of F: $\iota, \rho, \rho^2, \ldots, \rho^{n-1}, \sigma, \rho\sigma, \rho^2\sigma, \ldots, \rho^{n-1}\sigma$. These are distinct (why?) From $|Sym\ F| \leqslant 2n$, we get $|Sym\ F| = 2n$ and

$$Sym\ F = \{\iota, \rho, \rho^2, \ldots, \rho^{n-1}, \sigma, \rho\sigma, \rho^2\sigma, \ldots, \rho^{n-1}\sigma\}.$$

Every element in *Sym F* can be written as a product of a suitable power of ρ by suitable power of σ, which remark we summarize by saying that ρ *and* σ *generate Sym F*. We also say *Sym F is generated by* ρ *and* σ, and that ρ *and* σ *are generators of Sym F*. The notation $\langle\rho,\sigma\rangle$ denotes a group with two generators. Including the relations $\rho^n = \iota$, $\sigma^2 = \iota$ and $\sigma\rho = \rho^{-1}\sigma$, we write

$$Sym\ F = \langle\rho,\sigma : \rho^n = \iota,\ \sigma^2 = \iota,\ \sigma\rho = \rho^{-1}\sigma\rangle.$$

14.7 Definition: Let G be a group having elements a, b such that

$$o(a) = n,\ o(b) = 2,\ ba = a^{-1}b,$$

$$G = \{a^k b^r : k = 0,1,\ldots,n-1,\ r = 0,1\},$$

where $n \geqslant 2$. Then G is called a *dihedral group of order* n.

It is easily seen from $o(a) = n$ and $o(b) = 2$ that the elements of G displayed in Definition 14.7 are indeed distinct. Using Lemma 14.6, products in G can be brought to the form $a^k b^r$. Hence G is really of order n. In a dihedral group of order 8, for example, we have

$$a^2bab^3a^5a^7b^{-1} = a^2baba^{12}b = a^2.bab.a^4b = a^2.a^{-1}.a^4b = a^5b$$

with the foregoing notation. (The exponent of a changes sign when b "passes through" a to the other side.)

We see that symmetry groups of regular polygons are dihedral groups. We write D_{2n} for a dihedral group of order $2n$. (Warning: some authors write D_n for a dihedral group of order $2n$.) Henceforward, we write D_{2n} instead of *Sym F* (*F* being a regular polygon with n sides). The ambiguity in "D_{2n}" (whether it designates an arbitrary dihedral group or the particular dihedral group *Sym F*) is harmless.

Some people use Definition 14.7 only when $n \geqslant 3$. They do not consider D_4 as a dihedral group. This is consistent with the fact that D_4 is not the symmetry group of any regular polygon (see however Ex.10). But then they have to formulate the following theorem of Leonardo da Vinci (yes, of Leonardo da Vinci (1452-1519)) less beautifully.

14.8 Theorem: *A finite subgroup of Isom E is either a cyclic group or a dihedral group.*

This theorem will not be used in the sequel and its proof is left to the reader.

Exercises

1. Let α be an isometry and F_1, F_2 nonempty subsets of E. Show that $(F_1 \cup F_2)\alpha = F_1\alpha \cup F_2\alpha$ and $(F_1 \cap F_2)\alpha = F_1\alpha \cap F_2\alpha$. Generalize to arbitrary unions and intersections.

2. Let α be an isometry and P, Q two distinct points in E. Show that $(\overleftrightarrow{PQ})\alpha = \overleftrightarrow{P\alpha Q\alpha}$ and $(\overline{PQ})\alpha = \overline{P\alpha Q\alpha}$. (Hint: $\overline{PQ} = \{R \in E : d(P,R) + d(R,Q) = d(P,Q)\}$.)

3. Let α be an isometry and R the midpoint of $\overline{PQ}$. Show that $R\alpha$ is the midpoint of $\overline{P\alpha Q\alpha}$.

4. Let α be an isometry and $\triangle PQR$ a triangle (i.e., the union of the segments $\overline{PQ}, \overline{QR}, \overline{PR}$). Show that $(\triangle PQR)\alpha = \triangle P\alpha Q\alpha R\alpha$. (By the side-side-side condition, the triangles $\triangle PQR$ and $\triangle P\alpha Q\alpha R\alpha$ are congruent, hence

$\angle PQR$ and $\angle P\alpha Q\alpha R\alpha$ are equal: isometries preserve angles. In particular, isometries preserve perpendicularity and so also parallelity.)

5. Let $P_1, P_2, \ldots, P_n$ be the vertices of a regular n-gon. Prove that the center C of the regular n-gon is uniquely determined. (For instance, if n happens to be even, C is the midpoint of $P_iP_{i+(n/2)}$. As the radius of a circumscribing circle is equal to $d(P_i,C)$, this shows that a circumscribing circle is completely determined by the vertices. Hence there is a unique circumscribing circle of a regular n-gon.)

6. Let α be an isometry and M a regular n-gon. let C be the center of the regular n-gon. Prove the following assertions.

(a) $M\alpha$ is a regular n-gon with center $C\alpha$.

(b) If α is a symmetry of M, then $C\alpha = C$.

(c) If α is a symmetry of M, then $\{P_1,P_2, \ldots,P_n\}\alpha = \{P_1,P_2, \ldots,P_n\}$.

(Under a symmetry of M, a vertex is mapped to a vertex. Hint: A point P on M is a vertex if and only if $d(P,C)$ = radius of the circumscribing circle.)

7. Let m be a real number. Prove that $\{\tau_{a,am}: \alpha \in \mathbb{R}\}$ is a subgroup of *Isom* E without appealing to Theorem 14.4.

8. Let P be a point and m a line. Find all isometries that fix P, all isometries that fix m and all isometries that fix m pointwise. Show directly that these three sets are subgroup of *Isom* E.

9. Let F be a nonempty subset of E. Is $\{\alpha \in Isom\ E: F\alpha \subseteq F\}$ necessarily a subgroup of *Isom* E?

10. Find the symmetry group of a rectangle that is not a square.

11. With the notation of Definition 14.7, what is $|D_{2n}:\langle a\rangle|$?

12. Prove Theorem 14.8.

13. Let $\tau: \mathbb{R} \to \mathbb{R}$, $\sigma: \mathbb{R} \to \mathbb{R}$. Prove the following assertions.

$x \to x + 1$ $\quad x \to -x$

(a) $\tau,\sigma \in S_{\mathbb{R}}$.

(b) $|x - y| = |x\tau - y\tau| = |x\sigma - y\sigma|$. (So τ and σ preserve distance in $\mathbb{R}$. For this reason, they are said to be *isometries of* $\mathbb{R}$.)

(c) $o(\tau) = \infty$ and $o(\sigma) = 2$.

(d) $\sigma\tau = \tau^{-1}\sigma$. (Thus τ,σ satisfy the conditions an a,b in Definition 14.7, except n is replaced by ∞ here. A *dihedral group of infinite order* is a group D_∞ having element a,b such that $o(a) = \infty$, $o(b) = 2$, $ba = a^{-1}b$ and

$$G = \{a^k b^r : k \in \mathbb{Z}, r = 0,1\}.$$

The group generated by τ,σ is an example of a dihedral group of infinite order.)

14. Prove that a group generated by two distinct elements a,b such that $o(a) = 2 = o(b)$ is a dihedral group (of finite or infinite order).

15. Let n be any natural number or ∞. Find a group G and $a,c \in G$ such that $o(a) = 2 = o(c)$ and $o(ac) = n$. (So $o(ac)$ cannot be determined from $o(a)$ and $o(c)$ alone.)

§15
Symmetric Groups

For any nonempty set X, the set S_X of all one-to-one mappings from X onto X is a group under the composition of functions (Example 7.1(d)). In particular, choosing X to be the set $\{1,2,\ldots,n\}$ of the first n natural numbers, we get a group $S_{\{1,2,\ldots,n\}}$. We abbreviate this group as S_n.

15.1 Definition: Let $n \in \mathbb{N}$. The group of all one-to-one mappings from $\{1,2,\ldots,n\}$ onto $\{1,2,\ldots,n\}$ is called the *symmetric group (on n letters)* and is written S_n. The elements of S_n are called *permutations (of* $1,2,\ldots,n$).

The reader should not confuse the symmetric group with the symmetry group of a figure in the Euclidean plane.

15.2 Theorem: S_n is a group of order $n!$.

Proof: Let $\pi \in S_n$ be a permutation of $1,2,\ldots,n$. Then 1π is one of the numbers $1,2,\ldots,n$. Since π is one-to-one, $1\pi \neq 2\pi$, so 2π is one of the remaining $n-1$ numbers among $1,2,\ldots,n$ after 1π has been determined. Since π is one-to-one, $1\pi \neq 3\pi$ and $2\pi \neq 3\pi$, so 3π is one of the remaining $n-2$ numbers among $1,2,\ldots,n$ after 1π and 2π have been determined. Proceeding in this way, we see that, for any $k = 1,2,\ldots,n$, the number $k\pi$ must be one of the numbers among $1,2,\ldots,n$ which are distinct from $1\pi,2\pi,\ldots,(k-1)\pi$. Hence there are n choices for 1π; and $n-1$ choices for 2π; ...; and $n-(k-1)$ choices for $k\pi$; ...; and $n-(n-1)$ choices for $n\pi$; and all these choices give a permutation of $1,2,\ldots,n$. Therefore there are

$$n.(n-1).(n-2).\ldots.2.1 = n!$$

permutations in S_n. □

We introduce a notation for permutations. Let $n \in \mathbb{N}$ and $\pi \in S_n$. Then π is a mapping $\pi: \{1,2, \ldots,n\} \to \{1,2, \ldots,n\}$ and can be exhibited by associating any number in $\{1,2, \ldots,n\}$ with its image by an arrow. Thus $\pi \in S_n$ for which $1\pi = 3$, $2\pi = 1$, $3\pi = 2$, $4\pi = 5$, $5\pi = 4$ can be displayed as

$$\begin{aligned} 1 &\to 3 \\ 2 &\to 1 \\ 3 &\to 2 \\ 4 &\to 5 \\ 5 &\to 4, \end{aligned}$$

or, in order to save space, as

$$\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 3 & 1 & 2 & 5 & 4 \end{array}.$$

We simplify this notation further by deleting the arrows and enclosing the two rows of numbers in parentheses. Thus we arrive at

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

for our π. In general, we write any $\sigma \in S_n$ as

$$\begin{pmatrix} \ldots & a & \ldots \\ \ldots & a\sigma & \ldots \end{pmatrix}$$

In this notation, there are two rows of n elements and n columns of two elements. The rows consist of the numbers $1,2, \ldots,n$. The image under σ of any $a \in \{1,2, \ldots,n\}$ is written just below a in the second row. This notation is due to A. Cauchy (1789-1857).

The order of the columns is immaterial in this notation. For example

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

are all equal permutations in S_6.

The identity permutation $\iota \in S_n$ maps any a to a, so the rows will be identical. Thus

$$\iota = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 1 & 2 & 3 & \dots & n \end{pmatrix}.$$

The inverse of any $\sigma \in S_n$ is found easily. By definition, σ^{-1} is the function (permutation) that maps $a\sigma$ to a, for all $a \in \{1,2, \dots,n\}$. Let σ be $\begin{pmatrix} \dots & a & \dots \\ \dots & a\sigma & \dots \end{pmatrix}$.

Then, under σ^{-1}, any element in the second row is mapped to the number just above it. σ^{-1} is therefore obtained by interchanging the rows of σ. For instance, in S_7, we have

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & 6 & 3 & 5 & 4 & 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 7 & 6 & 3 & 5 & 4 & 1 & 2 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{pmatrix},$$ which may also be written as $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 6 & 7 & 3 & 5 & 4 & 2 & 1 \end{pmatrix}$. Two permutations in S_n, say π and σ, are multiplied as follows. We have $\pi = \begin{pmatrix} \dots & a & \dots \\ \dots & a\pi & \dots \end{pmatrix}$ and $\sigma = \begin{pmatrix} \dots & b & \dots \\ \dots & b\sigma & \dots \end{pmatrix}$. What is $\pi\sigma$? By definition, $\pi\sigma$ is the permutation that maps a to $(a\pi)\sigma$, for all a in $\{1,2,\dots,n\}$. To evaluate $(a\pi)\sigma$, we locate a in the first row of π, then read the number below it, which is $a\pi$, and locate this $a\pi$ in the first row of σ. The number below it is $(a\pi)\sigma$. We do this for $a = 1,2, \dots,n$ and in each case, write the number we obtain below a. Enclosing this configuration in parentheses, we get $\pi\sigma$ in double row notation. Here is an example.

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

In the first permutation, below 1, we see 5 and in the second permutation, below 5, we see 4. So, in the product, below 1, we write 4. Then, in the first permutation, below 2, we see 3 and in the second permutation, below 3, we see 5. So, in the product, below 2, we write 5:

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

The remaining entries are found by the same method and we get

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

The product of three or more permutations is evaluated in the same way:

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

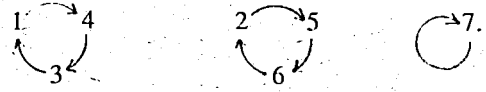
We now introduce a more efficient notation for permutations. The permutation $\sigma=\begin{pmatrix}1&2&3&4&5&6&7\\4&5&1&3&6&2&7\end{pmatrix}$ in S_7 is a mapping given explicitly as

$$\begin{aligned}1&\to 4\\2&\to 5\\3&\to 1\\4&\to 3\\5&\to 6\\6&\to 2\\7&\to 7.\end{aligned}$$

A more compact description of σ can be given as

$$1\to 4\to 3\to 1;\qquad 2\to 5\to 6\to 2;\qquad 7\to 7,$$

or as



We drop the arrows and enclose the numbers in a "cycle" within parentheses, in the order indicated by the arrows in a "cycle". Thus we get

$$(143)(256)(7)$$

after juxtaposing the parentheses. The meaning of this symbolism is as follows. Each number a is mapped, under σ, to the number that follows it in the parenthetical expression ("cycle") which contains a. If a happens to be the last entry in a "cycle", then the first number in that "cycle" is considered to follow a. For example,

$$(15234)(6897)\ \in S_9$$

is the permutation by which 1 is mapped to 5, 5 to 2, 2 to 3, 3 to 4, 4 to 1, 6 to 8, 8 to 9, 9 to 7, 7 to 6. Thus

$$(15234)(6897) = \begin{pmatrix} 1 & 5 & 2 & 3 & 4 & 6 & 8 & 9 & 7 \\ 5 & 2 & 3 & 4 & 1 & 8 & 9 & 7 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 3 & 4 & 1 & 2 & 8 & 6 & 9 & 7 \end{pmatrix}.$$

Here (15234) can also be written as (23415) or as (34152), (41523), (52341). Similar remarks are valid for (6897).

An arbitrary permutation $\pi \in S_n$ is written as follows. We open a parenthesis and write down an arbitrary number $a \in \{1,2, \ldots,n\}$. If $a\pi = a$, we close the parenthesis and obtain the expression (a). If $a\pi = b \neq a$, we write b after a. Now we have $(ab$. Here $b\pi \neq b$, because $b\pi \neq a\pi$ (π is one-to-one). If $b\pi = a$, we close the parenthesis and obtain the expression (ab). If $b\pi = c \neq b$, we write c after b. Now we have $(abc$. Here $c\pi \neq b,c$, because π is one-to-one. We evaluate $c\pi$. If $c\pi = a$, close the parenthesis and obtain the expression (abc). If $c\pi = d \neq a$, we repeat our procedure, each time writing down the image of a number after that number. Since we have n numbers at our disposal, we meet, after at most n steps, one of the numbers for a second time. If this happens when we have the expression

$$(abc\ldots g$$

where $a,b,c,\ldots,g$ are all distinct, but $g\pi$ is one of them, we conclude that $g\pi \neq b,c,\ldots,g$, since $b = a\pi$, $c = b\pi$, ... and π is one-to-one. Hence $g\pi = a$. We close the parenthesis and obtain the expression $(abc\ldots g)$.

If $a,b,c,\ldots,g$ exhaust all the numbers $1,2, \ldots,n$, we are done. Otherwise, we select an arbitrary number from $1,2, \ldots,n$ that is distinct from $a,b,c,\ldots,g$. Let us call it h. We open a new parenthesis starting with h and repeat our procedure. After finitely many steps, we get an expression of the form

$$(abc\ldots g)(h\ldots k)\ldots(t\ldots x),$$

where $\{a,b,c, \ldots,g,h, \ldots,k, \ldots,t, \ldots x\} = \{1,2, \ldots,n\}$. We call each one of the expressions $(abc\ldots g),(h\ldots k), \ldots,(t\ldots x)$ a "cycle".

We will presently give a rigorous definition of a cycle and prove that every permutation can be written as a product of disjoint cycles. But let us consider some examples first.

Let us write $\begin{pmatrix} 1&2&3&4&5&6&7&8&9 \\ 2&6&7&1&5&4&9&8&3 \end{pmatrix}$ in cycle notation. This is done in the following steps, which are carried out mentally at once in practice.

(1	(12	(126	(1264
(1264)	(1264)(3	(1264)(37	(1264)(379
(1264)(379)	(1264)(379)(5	(1264)(379)(5)	
(1264)(379)(5)(8	(1264)(379)(5)(8)		

In this notation, the order of the cycles is not important. We can write the permutation above also as

$$(5)(379)(8)(1264) \quad \text{or as} \quad (379)(5)(1264)(8).$$

Besides, one can start a cycle with any number in that cycle. Our permutation can thus be written as

$$(5)(793)(8)(6412) \quad \text{or as} \quad (937)(5)(2641)(8).$$

The identity permutation is given by $(1)(2)(3)\ldots(n)$. For obvious reasons, we prefer to write ι instead of $(1)(2)(3)\ldots(n)$ for the identity permutation.

The inverse of a permutation is found easily. Let $\sigma \in S_n$, $a,b \in \{1,2,\ldots,n\}$. In the cycles of σ, the number $a\sigma$ follows a. By definition, $b\sigma^{-1}$ is that number a for which $a\sigma = b$. Hence $b\sigma^{-1}$ is the number which is followed by b. Stated otherwise, $b\sigma^{-1}$ is the number that comes just before b in the cycles of σ. So σ^{-1} consists of the same cycles, but the entries being written in the *reverse* order. For example,

$$[(12)(357)(64)]^{-1} = (21)(753)(46), \quad [(326)(15)(4)]^{-1} = (623)(51)(4).$$

Two permutations in S_n, say π and σ, are multiplied as follows. We have $\pi = (\ldots a a\pi \ldots)$ and $\sigma = (\ldots a a\sigma \ldots)$. What is $\pi\sigma$? By definition, $\pi\sigma$ is the permutation that maps a to $(a\pi)\sigma$, for all $a \in \{1,2,\ldots,n\}$. To evaluate $(a\pi)\sigma$, we locate a in the cycle of π containing a, then read the number that follows it, which is $a\pi$, and locate this $a\pi$ in the cycle of σ. The number that follows it is $(a\pi)\sigma$. Opening a parenthesis with an arbitrary number a, we find $(a\pi)\sigma = a(\pi\sigma)$ in this way, and write it after a. So we get an expression $(ab$, say. We find $b(\pi\sigma) = c$. We write $(abc$. We repeat this process until we get a. Then we close our cycle. At this step, we have

$(abc\ldots g)$, say. If there are numbers among $1,2,\ldots,n$ not used up in this cycle, we select an arbitrary one of them and obtain a second cycle starting with that number. We continue in this fashion until all the numbers $1,2,\ldots,n$ are used up.

Let us compute the product $(1256)(347).(157)(24)(3)(6)$ in S_7. We start with the smallest number 1, for example. We write (1 . Now 1 is followed by 2 in the first permutation and 2 is followed by 4 in the second permutation. Thus we get (14 . Now 4 is followed by 7 in the first permutation and 7 is followed by 1 in the second permutation. We close our first cycle. We have (14). We open a new cycle with 2, for example. Now 2 is followed by 5 in the first permutation and 5 is followed by 7 in the second permutation. We have (14)(27 . Continuing in this way, we find $(1256)(347).(157)(24)(3)(6) = (14)(273)(56)$. Another example: $(152)(3476).(1724)(563) = (1654273)$.

We make a convention. Whenever there appears a cycle consisting of a single number, we suppress it. Hence, whenever a number j does not appear in the cycles of a permutation σ, we understand $j\sigma = j$. With this convention, we write shortly

$(123)(47)$ for $(123)(47)(5)(6)$ in S_7,
$(245)(3876)$ for $(245)(3876)(1)$ in S_8.

This convention simplifies multiplication: if a number does not appear in the cycles of one or more of the factors, it is mapped to itself by the permutations in question. For example,

$$(123).(12) = (23)$$
$$(254).(12)(34) = (25341).$$

The way we multiply permutations, either in double row or in cycle notation, reflects the fact that we write functions to the right of the elements. If we had written functions on the left, then $\pi\sigma$ would mean: first σ, then π. A product would be evaluated in double row notation by reading the permutations from right to left. In the cycle notation, we would be reading the cycles from right to left, but the numbers in the cycles from left to right. Writing our functions on the right, we avoid backward or inconsistent reading. We read everything in the correct order.

The alert reader will have noticed that the same symbol in cycle notation stands for many different permutations. Thus (123)(45) stands

for (123)(45) in S_5, for (123)(45)(6) in S_6, for (123)(45)(6)(7) in S_7, etc. So an isolated symbol (123)(45) is ambiguous. Also, our thumb rule for finding inverses in the cycle notation works only when the cycles are disjoint. It is time that we discuss these points rigorously.

15.3 Definition: Let $\sigma \in S_n$ and $m \in \{1,2, \ldots,n\}$. When $m\sigma = m$, we say that m *is fixed by* σ or that σ *fixes* m. When $m\sigma \neq m$, then m is said to be *moved by* σ or σ is said to *move* m.

15.4 Definition: Let $\pi,\sigma \in S_n$. If the set of numbers moved by π and the set of numbers moved by σ are disjoint, then π and σ are called *disjoint permutations*. We also say π is *disjoint from* σ in this case.

15.5 Lemma: *Let* $\alpha,\beta \in S_n$ *and* $k \in \{1,2, \ldots,n\}$. *Assume* α,β *are disjoint.*
(1) *If* k *is moved by* α, *then* $k\alpha$ *is also moved by* α.
(2) *If* k *is moved by* α *and fixed by* β, *then* $k\alpha$ *is fixed by* β.

Proof: (1) If k were fixed by α, so that $(k\alpha)\alpha = k\alpha$, we would apply α^{-1} to both sides of this equation and get $k\alpha = (k\alpha)\alpha\alpha^{-1} = k\alpha\alpha^{-1} = k$, contrary to the hypothesis that k is moved by α. So $k\alpha$ is moved by α.

(2) $k\alpha$ is moved by α according to part (1). If $k\alpha$ were moved by β, then $k\alpha$ would be moved both by α and by β, contrary to the hypothesis that α and β are disjoint permutations. Thus $k\alpha$ is fixed by β. □

We can now prove that disjoint permutations always commute.

15.6 Theorem: *If* $\sigma,\tau \in S_n$ are disjoint permutations, then $\sigma\tau = \tau\sigma$.

Proof: We must show $m(\sigma\tau) = m(\tau\sigma)$ for all $m \in \{1,2, \ldots,n\}$. Since σ and τ are disjoint, for each $m \in \{1,2, \ldots,n\}$, there are three possibilities:

I. m is moved by σ, fixed by τ.
II. m is fixed by σ, moved by τ.
III. m is fixed by σ, fixed by τ.

In case I, $m\sigma$ fixed by τ by Lemma 15.5(2) (with m,σ,τ in place of k,α,β), hence $(m\sigma)\tau = m\sigma$ and

$$m(\sigma\tau) = (m\sigma)\tau = m\sigma,$$
$$m(\tau\sigma) = (m\tau)\sigma = m\sigma \qquad \text{(as } m \text{ is fixed by } \tau\text{)},$$

so $m(\sigma\tau) = m(\tau\sigma)$.

In case II, $m\tau$ fixed by σ by Lemma 15.5(2) (with m,τ,σ in place of k,α,β), hence $(m\tau)\sigma = m\tau$ and

$$m(\sigma\tau) = (m\sigma)\tau = m\tau \qquad \text{(as } m \text{ is fixed by } \sigma\text{)},$$
$$m(\tau\sigma) = (m\tau)\sigma = m\tau,$$

so $m(\sigma\tau) = m(\tau\sigma)$.

In case III, we have

$$m(\sigma\tau) = (m\sigma)\tau = m\tau = m,$$
$$m(\tau\sigma) = (m\tau)\sigma = m\sigma = m,$$

so $m(\sigma\tau) = m(\tau\sigma)$.

In all three cases, we have $m(\sigma\tau) = m(\tau\sigma)$. Since this holds for all m in the set $\{1,2,\ldots,n\}$, we conclude $\sigma\tau = \tau\sigma$. □

In order to prepare our way for a formal definition of cycle, let us examine the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 4 & 1 & 2 & 3 & 6 & 7 & 5 & 8 & 10 & 9 \end{pmatrix}$$

in S_{10}. Informally, we write this as (1432)(567)(8)(9,10) and call (1432), (567), (8), (9,10) "cycles" (we use a comma to avoid confusion when we have a number with more than one digits). The idea is to consider (1432) etc. as a permutation by itself. Then

$$(1432)(567)(8)(9,10)$$

is a product of four permutations. We observe that {1432}, {567}, {8}, {9,10} are pairwise disjoint subsets of {1,2,3,4,5,6,7,8,9,10} and yield a partition of {1,2,3,4,5,6,7,8,9,10}. So there is an equivalence relation on {1,2,3,4,5,6,7,8,9,10} with these subsets as equivalence classes (Theorem 2.5). We want to find this equivalence relation.

15.7 Lemma: *Let π be a permutation in S_n. We put, for $a,b \in \{1,2,\ldots,n\}$,*

$$a \overset{\pi}{\sim} b$$

if and only if there is an integer $k \in \mathbb{Z}$ such that $a\pi^k = b$. Then $\overset{\pi}{\sim}$ is an equivalence relation on $\{1,2,\ldots,n\}$.

Proof: (i) For all $a \in \{1,2,\ldots,n\}$, we have $a\pi^0 = a$, with $0 \in \mathbb{Z}$. So $a \overset{\pi}{\sim} a$ for all a and $\overset{\pi}{\sim}$ is reflexive.

(ii) If $a \overset{\pi}{\sim} b$, then $a\pi^k = b$ for some $k \in \mathbb{Z}$, so $b\pi^{-k} = a$ with $-k \in \mathbb{Z}$ and therefore $b \overset{\pi}{\sim} a$. So $\overset{\pi}{\sim}$ is symmetric.

(iii) If $a \overset{\pi}{\sim} b$ and $b \overset{\pi}{\sim} c$, then $a\pi^k = b$ and $b\pi^m = c$ for some $k,m \in \mathbb{Z}$, then $a\pi^{k+m} = a\pi^k\pi^m = b\pi^m = c$, with $k + m \in \mathbb{Z}$ and therefore $a \overset{\pi}{\sim} c$. So $\overset{\pi}{\sim}$ is transitive. □

The reader will check easily that {1432}, {567}, {8}, {9,10} are the equivalence classes of $\overset{\pi}{\sim}$ in {1,2,3,4,5,6,7,8,9,10} if π denotes the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 4 & 1 & 2 & 3 & 6 & 7 & 5 & 8 & 10 & 9 \end{pmatrix}$$

we treated above. So the equivalence relation of Lemma 15.7 seems promising.

15.8 Lemma: *Let $\pi \in S_n$ and let $A \subseteq \{1,2,\ldots,n\}$ be an equivalence class under the equivalence relation of Lemma 15.7. We define π_A by*

$$b\pi_A = \begin{cases} b\pi \text{ if } b \in A \\ b \text{ if } b \notin A \end{cases}$$

for $b \in \{1,2,\ldots,n\}$. Then π_A is a permutation in S_n and, whenever x and y are moved by π_A, there is an integer k such that $x\pi_A^k = y$.

Proof: By definition of $\overset{\pi}{\sim}$, there holds $x \overset{\pi}{\sim} x\pi$ for all $x \in \{1,2,\ldots,n\}$ and $x\pi$ belongs to the equivalence class of x. So the equivalence class of x and the equivalence class of $x\pi$ are identical. Hence $x\pi \in A$ if and only if $x \in A$.

Using this remark, we prove, for any $n \in \mathbb{N}$, that $x\pi^n = x\pi_A^n$ for all $x \in A$. This is true when $n = 1$. If it is true for $n = k - 1$, we have, for any $x \in A$,

$$\begin{aligned} x\pi_A^k &= (x\pi_A)\pi_A^{k-1} \\ &= (x\pi)\pi_A^{k-1} \\ &= (x\pi)\pi^{k-1} \\ &= x\pi^k, \end{aligned}$$

and it is true for $n = k$. Thus it is true for all $n \in \mathbb{N}$.

In particular, it is true for $m = o(\pi)$ and

$$x\pi_A^m = \begin{cases} x\pi^m \text{ if } x \in A \\ x \quad \text{if } x \notin A \end{cases} = x,$$

hence $\pi_A^{m-1}\pi_A = \iota = \pi_A\pi_A^{m-1}$. By Theorem 3.17(2), π_A is one-to-one and onto. Thus $\pi_A \in S_n$.

Finally, if x and y are moved by π_A, then necessarily $x,y \in A$ in view of the definition of π_A, so there is an integer $k \in \mathbb{Z}$ with $x\pi^k = y$ and thus $x(\pi_A)^k = y$ by what we proved above (since $x \in A$). This completes the proof. □

15.9 Theorem: *Let $\pi \in S_n$ and let $A_1, A_2, \ldots, A_h$ be the equivalence classes of $\{1,2,\ldots,n\}$ under the equivalence relation $\overset{\pi}{\sim}$ in Lemma 15.7. Let $\pi_{A_1}, \pi_{A_2}, \ldots, \pi_{A_h}$ be the associated permutations as in Lemma 15.8.*

(1) *$\pi_{A_1}, \pi_{A_2}, \ldots, \pi_{A_h}$ are pairwise commuting permutations in S_n.*
(2) $\pi = \pi_{A_1}\pi_{A_2}\ldots\pi_{A_h}$.

Proof: (1) The equivalence classes $A_1, A_2, \ldots, A_h$ are pairwise disjoint sets. Now π_{A_i} either moves no number at all (this happens if and only if A_i has exactly one element), or moves only the numbers in A_i. Therefore, the numbers moved by π_{A_i} and π_{A_j} make up disjoint sets whenever $i \neq j$. So the permutations $\pi_{A_1}, \pi_{A_2}, \ldots, \pi_{A_h}$ are pairwise disjoint permutations (Definition 15.4) and they commute by Theorem 15.6.

(2) We have $\pi_{A_1}\pi_{A_2}\ldots\pi_{A_h} = \pi_{A_{1'}}\pi_{A_{2'}}\ldots\pi_{A_{h'}}$ for any arrangement $1',2',\ldots,h'$ of the numbers $1,2,\ldots,h$. (Lemma 8.12). We want to show $b\pi = b\pi_{A_1}\pi_{A_2}\ldots\pi_{A_h}$ for all $b \in \{1,2,\ldots,n\} = A_1 \cup A_2 \cup \ldots \cup A_h$. So let b be in $\{1,2,\ldots,n\}$. Renumbering $A_1,A_2,\ldots,A_h$ if need be, we may assume, without loss of generality that $b \in A_h$. Then $b \notin A_1$, $b \notin A_2,\ldots,b \notin A_{h-1}$ and thus $b\pi_{A_1} = b\pi_{A_2} = \cdots = b\pi_{A_{h-1}} = b$ by the definition of these functions. Thus $b\pi_{A_1}\pi_{A_2}\ldots\pi_{A_h} = b_{A_h}$ and the proof will be complete when we show $b\pi = b_{A_h}$. But this follows immediately from the definition of π_{A_h} since $b \in A_h$. □

In our example, the associated permutations are

$$(1432)(5)(6)(7)(8)(9)(10) = (1432)$$
$$(567)(1)(2)(3)(4)(8)(9)(10) = (567)$$
$$(8)(1)(2)(3)(4)(5)(6)(7)(9)(10) = (8)\ (= \iota)$$
$$(9,10)(1)(2)(3)(4)(5)(6)(7)(8) = (9,10).$$

In view of this, we define cycles as the associated permutations. Cycles will be distinguished from other permutations by the property stated in Lemma 15.8.

15.10 Definition: A permutation $\pi \in S_n$ is called a *cycle* if, for all x,y in $\{1,2,\ldots,n\}$ that are moved by π, there is an integer k such that $x\pi^k = y$.

The identity permutation is vacuously a cycle. Lemma 15.8 states that π_A is a cycle when A is an equivalence class under the equivalence relation in Lemma 15.7. Since the cycles are disjoint, we may reformulate Theorem 15.9 as follows.

15.9 Theorem: *Every permutation π in S_n can be written as a product of disjoint cycles. These cycles are completely determined by π, and they commute in pairs.* □

Let σ be a cycle in S_n distinct from ι. Let $a_1, a_2, \ldots, a_m$ be the numbers moved by σ. Since σ is one-to-one, $a_1\sigma, a_2\sigma, \ldots, a_m\sigma$ are all distinct and we may assume the numbering so chosen that

$$a_1\sigma = a_2,\ a_2\sigma = a_3,\ \ldots,\ a_{m-1}\sigma = a_m,\ a_m\sigma = a_1.$$

In this case, we write $(a_1a_2\ldots a_m)$ for σ. Then m is called the *length* of the cycle $(a_1a_2\ldots a_m)$ and $(a_1a_2\ldots a_m) = \sigma$ is called an *m-cycle*. The identity permutation is called a *1-cycle*.

With this notation, we have

$$a_1\sigma = a_2 \neq a_1,\ a_1\sigma^2 = a_3 \neq a_1,\ \ldots,\ a_1\sigma^{m-1} = a_m \neq a_1$$

and so $\sigma \neq \iota$, $\sigma^2 \neq \iota, \ldots, \sigma^{m-1} \neq \iota$. On the other hand,

$$a_1\sigma^m = a_1 \quad \text{and} \quad a_k\sigma^m = a_1\sigma^{k-1}\sigma^m = a_1\sigma^m\sigma^{k-1} = a_1\sigma^{k-1} = a_k$$

for all $k = 1,2, \ldots, n$. So σ^m fixes $a_1, a_2, \ldots, a_m$. But σ fixes the numbers among $1,2, \ldots, n$ which are distinct from $a_1, a_2, \ldots, a_m$, and then σ^m fixes them, too. Hence $b\sigma^m = b$ for all $b \in \{1,2, \ldots, n\}$. Thus m is the smallest natural number such that $\sigma^m = \iota$. Using Lemma 11.4, we obtain the following Theorem, which is also true when $m = 1$.

15.11 Theorem: *The order of a cycle is its length. In other words, if* $\sigma = (a_1a_2\ldots a_m)$, *then* $o(\sigma) = m$. □

15.12 Remarks: (1) The inverse of a cycle $\sigma = (a_1a_2\ldots a_m) \in S_n$, for which $a_1\sigma = a_2, a_2\sigma = a_3, \ldots, a_{m-1}\sigma = a_m, a_m\sigma = a_1$ and which fixes any other number in $\{1,2, \ldots, n\}$ (if any) is by definition the mapping π whose effect on $a_1, a_2, \ldots, a_m$ is given by $a_m\pi = a_{m-1}, \ldots, a_3\pi = a_2, a_2\pi = a_1, a_1\pi = a_m$ and which fixes the other numbers (if any). Thus $\sigma^{-1} = \pi$ is the cycle

$$(a_ma_{m-1}\ldots a_2a_1).$$

(2) Let $\pi \in S_n$ be written as $\pi = \pi_{A_1}\pi_{A_2}\ldots\pi_{A_k}$ with the notation of Theorem 15.9. A cycle π_{A_i} is the identity if there is only one number in A_i. Then the cycle π_{A_i} may be deleted from the product.

(3) If $\pi = \pi_{A_1}\pi_{A_2}\ldots\pi_{A_k}$ is the representation of π as a product of disjoint cycles, then $\pi = \pi_{A_k}\ldots\pi_{A_2}\pi_{A_1}$ and so $\pi^{-1} = \pi_{A_1}^{-1}\pi_{A_2}^{-1}\ldots\pi_{A_k}^{-1}$. But this is true only

when A_i are disjoint. In any case, it is safer to reverse the order of the cycles as well as the ordering of the numbers in each cycle when we want to find the inverse of a product of cycles, as this is valid also in the case the cycles are not pairwise disjoint and is a more consistent procedure: you reverse everything. For example,

$$[(15)(243)(687)]^{-1} = (786)(342)(51).$$

(4) The ambiguity in cycle notation is harmless, as it will be either clear from the context which symmetric group we are working in, or the results will be independent of the symmetric group.

In the rest of this paragraph, we determine the order of a permutation written as a product of disjoint cycles. We start with a general lemma.

15.13 Lemma: *Let G be a group and $a,b \in G$. Suppose $ab = ba$ and assume that $o(a)$ and $o(b)$ are finite. Suppose further that $\langle a\rangle \cap \langle b\rangle = \{1\}$. Then $o(ab)$ is finite. In fact, $o(ab)$ is the least common multiple of $o(a)$ and $o(b)$: we have $o(ab) = [o(a),o(b)]$.*

Proof: First we show that $(ab)^k = 1$ if and only if $o(a)|k$ and $o(b)|k$ (where $k \in \mathbb{Z}$). Indeed, if $o(a)|k$ and $o(b)|k$, then $a^k = 1$ and $b^k = 1$ (Lemma 11.6) and so $(ab)^k = a^k b^k = 1.1 = 1$ (Lemma 8.14(3); here we use $ab = ba$). Conversely, if $(ab)^k = 1$, then $a^k b^k = 1$, so $a^k = b^{-k} \in \langle a\rangle \cap \langle b\rangle = \{1\}$. So we have $a^k = 1 = b^{-k}$, and $a^k = 1 = b^k$, and thus $o(a)|k$ and $o(b)|k$. Therefore $(ab)^k = 1$ if and only if $o(a)|k$ and $o(b)|k$.

Then, by Lemma 11.4,

$$\begin{aligned} o(ab) &= \text{smallest number in } \{k \in \mathbb{N}: (ab)^k = 1\}, \\ &\qquad \text{provided this set is not empty,} \\ &= \text{smallest number in } \{k \in \mathbb{N}: o(a)|k \text{ and } o(b)|k\}, \\ &\qquad \text{provided this set is not empty,} \\ &= \text{the least common multiple of } o(a) \text{ and } o(b), \text{ as} \\ &\qquad \text{the set is not empty,} \\ &= [o(a),o(b)] \end{aligned}$$

Generally speaking, we cannot determine the order of a and b from $o(a)$ and $o(b)$ alone. $o(ab)$ depends also on the role the elements a,b play in the group. (See §14, Ex.15.) Lemma 15.13 is one of the rare situations where $o(ab)$ is determined in terms of $o(a)$ and $o(b)$.

Lemma 15.13 will be used to find the order of a product of disjoint permutations. We need the following result.

15.14 Lemma: (1) *If* σ_1 *and* τ, *as well as* σ_2 *and* τ *are disjoint permutations in* S_n, *then* $\sigma_1\sigma_2$ *and* τ *are disjoint.*

(2) *If* $\sigma_1,\sigma_2,\ldots,\sigma_m$ *are disjoint from* τ, *then* $\sigma_1\sigma_2\ldots\sigma_m$ *and* τ *are disjoint.*

(3) *If* σ *and* τ *are disjoint, then* σ^{-1} *and* τ *are disjoint.*

(4) *If* σ *and* τ *are disjoint, then* σ^m *and* τ *are disjoint for all* $m \in \mathbb{Z}$.

(5) *If* σ *and* τ *are disjoint, then* σ^m *and* τ^r *are disjoint for all* $m,r \in \mathbb{Z}$.

Proof: (1) By hypothesis, any $k \in \{1,2,\ldots,n\}$ that is moved by τ is fixed by σ_1 and σ_2. So $k\tau \neq k$ implies $k\sigma_1 = k$ and $k\sigma_2 = k$. So $k\tau \neq k$ implies $k(\sigma_1\sigma_2) = (k\sigma_1)\sigma_2 = k\sigma_2 = k$ and $\sigma_1\sigma_2$ fixes every number that τ moves. Hence $\sigma_1\sigma_2$ and τ are disjoint. (The argument is valid also when $\tau = \iota$.)

(2) This follows from (1) by induction on m. The details are left to the reader.

(3) Let $k \in \{1,2,\ldots,n\}$ be moved by τ. We wish to show that k is fixed by σ^{-1}. Since σ and τ are disjoint, k is fixed by σ. So $k\sigma = k$. Applying σ^{-1} to both sides, we get $(k\sigma)\sigma^{-1} = k\sigma^{-1}$, hence $k = k\sigma^{-1}$ and k is fixed by σ^{-1}. Therefore σ^{-1} and τ are disjoint.

(4) Let $m \in \mathbb{N}$. Choosing $\sigma_1,\sigma_2,\ldots,\sigma_m$ all equal to σ in (2), we deduce that σ^m and τ are disjoint. Now applying (3) with σ^m,τ in place of σ,τ, we get that $\sigma^{-m} = (\sigma^m)^{-1}$ is disjoint from τ, for any $m \in \mathbb{N}$. As $\tau^0 = \iota$ is trivially disjoint from τ, we conclude that σ^m and τ are disjoint for all $m \in \mathbb{Z}$.

(5) When σ and τ are disjoint and $m,r \in \mathbb{Z}$, then σ^m and τ are disjoint by (4), and using (4) with r,τ,σ^m respectively in place of m,σ,τ, we deduce that τ^r and σ^m are disjoint. Hence σ^m and τ^r are disjoint for all $m,r \in \mathbb{Z}$. □

15.15 Theorem: *Let σ and τ be disjoint permutations in S_n. Then*

$$o(\sigma\tau) = [o(\sigma),o(\tau)].$$

Proof: We use Lemma 15.13. Since σ and τ are disjoint, $\sigma\tau = \tau\sigma$ by Theorem 15.6. Also, $o(\sigma)$ and $o(\tau)$ are finite since S_n is a finite group by Theorem 15.2. We must also show that $\langle\sigma\rangle \cap \langle\tau\rangle = \{\iota\}$. When we do this, the hypotheses of Lemma 15.13 will be satisfied and it will yield $o(\sigma\tau) = [o(\sigma),o(\tau)]$. So we show $\langle\sigma\rangle \cap \langle\tau\rangle \leqslant \{\iota\}$.

Suppose $\langle\sigma\rangle \cap \langle\tau\rangle \not\leqslant \{\iota\}$. Then there is an $\alpha \in \langle\sigma\rangle \cap \langle\tau\rangle$ with $\alpha \neq \iota$ and $\alpha = \sigma^m = \tau^r$ for some integers m,r. Since $\alpha \neq \iota$, there is a $j \in \{1,2,\ldots,n\}$ such that $j\alpha \neq j$. So j is moved by σ^m and also by τ^r. On the other hand, σ^m and τ^r are disjoint by Lemma 15.14(5) and there cannot be any number in $\{1,2,\ldots,n\}$ which is moved both by σ^m and by τ^r. This is a contradiction. Thus $\langle\sigma\rangle \cap \langle\tau\rangle \leqslant \{\iota\}$. As remarked above, this completes the proof. □

15.16 Theorem: *Let $\sigma_1,\sigma_2,\ldots,\sigma_m$ be pairwise disjoint permutations in S_n. Then $o(\sigma_1\sigma_2\ldots\sigma_m) = [o(\sigma_1),o(\sigma_2),\ldots,o(\sigma_m)]$.*

Proof: By induction on m. The case $m = 2$ is treated in Theorem 15.15. The inductive step is left to the reader. □

15.17 Theorem: *The order of $\sigma \in S_n$ is the least common multiple of the lengths of the disjoint cycles in the representation of σ as a product of disjoint cycles.*

Proof: The disjoint cycles are pairwise disjoint and the order of a cycle is its length (Theorem 15.11). The claim follows now immediately from Theorem 15.16. □

For instance, (134)(275698) has order 6, (124)(3756) has order 12 and (34)(79)(12586) has order 10.

Exercises

1. Evaluate $\begin{pmatrix}1&2&3&4&5\\5&3&1&2&4\end{pmatrix}\begin{pmatrix}1&2&3&4&5\\2&1&3&5&4\end{pmatrix}$, $\begin{pmatrix}1&2&3&4&5&6\\6&1&2&3&4&5\end{pmatrix}\begin{pmatrix}1&2&3&4&5&6\\2&4&6&5&3&1\end{pmatrix}$ and

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

2. Evaluate (1253)(24315), (1542)(376)(1754) and $(1243)(345)(265)(1452)(135)^{-1}(3246)$.

3. Write the permutations in Ex. 1 in cycle notation. Carry out the multiplication in cycle notation and compare the results.

4. Write the permutations in Ex. 2 in double row notation. Carry out the multiplication in double row notation and compare the results.

5. Write all elements in S_1, S_2, S_3, S_4.

6. Construct multiplication tables of S_1, S_2, S_3, S_4.

7. Find the orders of all elements in S_3 and S_4.

8. Show that $V_4 := \{\iota,(12)(34),(13)(24),(14)(23)\}$ is a subgroup of S_4.

9. Show that $D := \{\iota,(13),(24),(12)(34),(13)(24),(14)(23),(1234),(1432)\}$ is a subgroup of S_4. Prove that it is a dihedral group in the sense of Definition 14.7.

10. Find all subgroups of S_3 and S_4.

11. Let $H \leqslant S_n$. For $a,b \in G$, put $a \overset{H}{\sim} b$ if and only if there is a $\sigma \in H$ such that $a\sigma = b$. Show that $\overset{H}{\sim}$ is an equivalence relation on $\{1,2,\ldots,n\}$. (Lemma 15.7 is a special case when $H = <\pi>$.)

12. Let $a_1, a_2, \ldots, a_m$ be pairwise commuting elements of finite order in a group G such that $\langle a_i \rangle \cap \langle a_j \rangle = \{1\}$ whenever $i \neq j$. Show that $o(a_1 a_2 \ldots a_m) = [o(a_1), o(a_2), \ldots, o(a_m)]$. This gives an alternative proof of Theorem 15.16.

13. For $\sigma \in S_4$, we put $\sigma V_4 := \{\sigma\pi : \pi \in V_4\}$ and $V_4\sigma := \{\pi\sigma : \pi \in V_4\}$ (Ex. 8). Find σV_4 and $V_4\sigma$ when $\sigma = \iota$, $\sigma = (12)$, $\sigma = (123)$, $\sigma = (12)(34)$, $\sigma = (1234)$.

14. For $H \subseteq S_4$, $\sigma \in S_4$, we put $\sigma H := \{\sigma\pi : \pi \in H\}$ and $H\sigma := \{\pi\sigma : \pi \in H\}$. Thus σH and $H\sigma$ are subsets of S_4.

Let $H_1 = \{\iota, (13), (24), (12)(34)\}$. Check whether $(12)H_1 = H_1(12)$, $(13)H_1 = H_1(13)$, $(123)H_1 = H_1(123)$, $(12)(34)H_1 = H_1(12)(34)$, $(1234)H_1 = H_1(1234)$.

Let $H_2 = \{\iota, (12), (34), (12)(34)\}$. Check whether $(12)H_2 = H_2(12)$, $(13)H_2 = H_2(13)$, $(123)H_2 = H_2(123)$, $(12)(34)H_2 = H_2(12)(34)$, $(1234)H_2 = H_2(1234)$.

Compare to Ex. 13.

15. Show that, for any $\sigma \in S_n$, there holds $\sigma^{-1}(123)\sigma = (abc)$ with suitable a, b, c. How are a, b, c related to σ? (Work out some specific examples.) Generalize your conclusion to $\sigma^{-1}\pi\sigma$.

§16
Alternating Groups

In this paragraph, we examine an important subgroup of S_n, called the alternating group on n letters. We begin with a definition that will play an important role throughout this paragraph.

16.1 Definition: A cycle of length 2 in S_n (where $n \geqslant 2$) is called a *transposition.*

A transposition is therefore a permutation of the form (ab) and has order 2 (Theorem 15.11). We remark that $(ab) = (ba)$.

16.2 Theorem: *Any permutation in S_n (where $n \geqslant 2$) can be written as a product of transpositions.*

Proof: Since any permutation in S_n can be written as a product of (disjoint) cycles (Theorem 15.9), it suffices to prove that any cycle can be written as a product of transpositions. This follows from $(abc\ldots e) = (ab)(ac)\ldots(ae)$ for cycles of length > 1. Also $\iota = (12)(12)$ is a product of transpositions. This completes the proof. □

There is no uniqueness claim in Theorem 16.2. A permutation can be written as a product of different transpositions. For instance,

$$(12345) = (12)(13)(14)(15) = (45)(41)(42)(43).$$

is written as a product of different transpositions. Nor is the number of transpositions is unique. The permutation (132546) can be written as a product of five or nine transpositions:

$$(132546) = (13)(12)(15)(14)(16) = (24)(12)(14)(23)(46)(14)(16)(45)(16).$$

In fact, we can attach a product of two transpositions $(ab)(ab) = \iota$ at will and increase the number of transpositions by 2. Hence a product of n transpositions can be written also as a product of $n + 2, n + 4, n + 6, \ldots$ transpositions. We note that this does not change the *parity* of the number of transpositions. The parity of the number of transpositions is unique. If a permutation can be written as a product of an odd (even) number of transpositions, then, in any representation of this permutation as a product of transpositions, the number of transpositions is odd (even). A permutation cannot be written as a product of an odd number of transpositions and also as a product of an even number of transpositions. We proceed to prove this assertion. We need the notion of inversions of a permutation.

Let $\sigma \in S_n$. We write σ in double row notation, where, in the first row, the numbers $1,2, \ldots,n$ are in their natural order:

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ 1\sigma & 2\sigma & \ldots & n\sigma \end{pmatrix}.$$

Corresponding to the correct inequalities

$$\begin{array}{llll} 1 < 2 & 1 < 3 & \ldots\ldots & 1 < n \\ & 2 < 3 & \ldots\ldots & 2 < n \\ & & \ldots\ldots\ldots\ldots & \\ & & & n - 1 < n \end{array}$$

among the numbers in the first row, we obtain the inequalities

$$\begin{array}{llll} 1\sigma < 2\sigma & 1\sigma < 3\sigma & \ldots\ldots & 1\sigma < n\sigma \\ & 2\sigma < 3\sigma & \ldots\ldots & 2\sigma < n\sigma \\ & & \ldots\ldots\ldots\ldots & \\ & & & (n - 1)\sigma < n\sigma \end{array}$$

among the numbers in the second row when we replace each k by $k\sigma$ $(k = 1,2, \ldots,n)$. These inequalities will be referred to as the inequalities of σ. In general, some of the inequalities of σ will be correct, some will be wrong (if $\sigma \neq \iota$, there will be a wrong inequality of σ). A wrong inequality $i\sigma < j\sigma$ of σ means: $i < j$ but $i\sigma > j\sigma$ i.e., the natural order of i and j is inverted in the second row (that is, the larger one precedes the smaller

one). We call each wrong inequality of σ an *inversion of* σ. For example, $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 5 & 6 & 3 & 1 & 4 \end{pmatrix}$ has the inequalities

$$\begin{array}{lllll} 2<5 & 2<6 & 2<3 & 2<1 & 2<4 \\ & 5<6 & 5<3 & 5<1 & 5<4 \\ & & 6<3 & 6<1 & 6<4 \\ & & & 3<1 & 3<4 \\ & & & & 1<4, \end{array}$$

eight of which are wrong, namely $2 < 1$, $5 < 3$, $5 < 1$, $5 < 4$, $6 < 3$, $6 < 1$, $6 < 4$, $3 < 1$. Hence there are eight inversions of σ.

The main work of this paragraph is done in the next lemma.

16.3 Lemma: *Let $n \geqslant 2$, $\sigma \in S_n$ and let (ik) be a transposition in S_n. If σ has an odd number of inversions, then $(ik)\sigma$ has an even number of inversions. If σ has an even number of inversions, then $(ik)\sigma$ has an odd number of inversions.*

Proof: Since $(ik) = (ki)$, we assume, without loss of generality, that $i < k$. We have

$$\sigma = \begin{pmatrix} 1 & \dots & i & \dots & k & \dots & n \\ 1\sigma & \dots & i\sigma & \dots & k\sigma & \dots & n\sigma \end{pmatrix}, \quad (ik)\sigma = \begin{pmatrix} 1 & \dots & i & \dots & k & \dots & n \\ 1\sigma & \dots & k\sigma & \dots & i\sigma & \dots & n\sigma \end{pmatrix}.$$

Tne second rows of σ and $(ik)\sigma$ are identical, aside from the locations of $i\sigma$ and $k\sigma$. Here σ gives rise to the inequalities

1.	$h\sigma < i\sigma$,	$h\sigma < k\sigma$	where $h \in \{1, \dots, i-1\} =: \mathrm{H}$,
	$i\sigma < j\sigma$,		where $j \in \{i+1, \dots, k-1\} =: \mathrm{J}$,
	$i\sigma < k\sigma$,		
2.	$i\sigma < m\sigma$,		where $m \in \{k+1, \dots, n\} =: \mathrm{M}$,
	$j\sigma < k\sigma$,		where $j \in \mathrm{J}$,
3.	$k\sigma < m\sigma$,		where $m \in \mathrm{M}$,

and to certain other inequalities that do not involve $i\sigma$ or $k\sigma$. And $(ik)\sigma$ gives rise to the inequalities

1.	$h\sigma < k\sigma$,	$h\sigma < i\sigma$	where $h \in \mathrm{H}$,
	$k\sigma < j\sigma$,		where $j \in \mathrm{J}$,
	$k\sigma < i\sigma$,		

3. $k\sigma < m\sigma$, where $m \in M$,
$j\sigma < i\sigma$, where $j \in J$,
2. $i\sigma < m\sigma$, where $m \in M$,

and to certain other inequalities that do not involve $i\sigma$ or $k\sigma$.

In the cases $i = 1$, $k = i + 1$, $k = n$, there holds respectively $H = \emptyset$, $J = \emptyset$, $M = \emptyset$ and the corresponding inequalities should be deleted. This does not impair the argument below.

We are to show that the number of inversions of σ and the number of inversions of $(ik)\sigma$ differ by an odd number.

The inequalities of σ and of $(ik)\sigma$ that do not involve $i\sigma$ or $k\sigma$ are identical. Also, the inequalities 1., 2., 3. of σ and $(ik)\sigma$ are the same (or absent). So only the inequalities

I. $i\sigma < j\sigma$, $i\sigma < k\sigma$, $j\sigma < k\sigma$ (where $j \in J$) of σ

and

II. $k\sigma < j\sigma$, $k\sigma < i\sigma$, $j\sigma < i\sigma$ (where $j \in J$) of $(ik)\sigma$

are different. We must prove that the number of wrong inequalities in I and II differ by an odd number.

Since one of $i\sigma < k\sigma$, $k\sigma < i\sigma$ is correct and the other is wrong, we must prove only that the number of wrong inequalities in

A. $i\sigma < j\sigma$, $j\sigma < k\sigma$ (where $j \in J$)

and in

B. $k\sigma < j\sigma$, $j\sigma < i\sigma$ (where $j \in J$)

differ by an even number.

Suppose there are s wrong inequalities $i\sigma < j\sigma$ and t wrong inequalities $j\sigma < k\sigma$ in A, where $|J| \geqslant s \geqslant 0$ and $|J| \geqslant t \geqslant 0$ (including the case $J = \emptyset$, $|J| = 0$). Then there are $s + t$ wrong inequalities and there are $(|J| - s) + (|J| - t) = 2|J| - (s + t)$ correct inequalities in A. Since B consists of the negations of the inequalities in A, there are $2|J| - (s + t)$ wrong inequalities in B. So

(no. of wrong inequalities in A) - (no. of wrong inequalities in B)
$= (s + t) - (2|J| - (s + t)) = 2(s + t - |J|) =$ an even number.

This completes the proof. □

16.4 Definition: Let $n \in \mathbb{N}$ and let $\sigma \in S_n$. If σ has an odd number of inversions, then σ is called an *odd permutation.* If σ has an even number of inversions, then σ is called an *even permutation.*

As the number of inversions of a permutation is uniquely determined, it is clear that a permutation cannot be both odd and even. With this terminology, Lemma 16.3 reads as follows.

16.3 Lemma: *Let $n \geqslant 2$ and $\sigma \in S_n$. Let (ik) be a transposition in S_n. If σ is odd, then $(ik)\sigma$ is even. If σ is even, then $(ik)\sigma$ is odd.* □

Applying Lemma 16.3 r times, we have

16.5 Lemma: *Let $n \geqslant 2$, $\sigma \in S_n$ and let $\tau_1, \tau_2, \ldots, \tau_r$ be transpositions in S_n. If r is odd, then σ and $\tau_1\tau_2 \ldots \tau_r\sigma$ have the opposite "parity" (i.e., one of them is odd, the other is even). If r is even, then σ and $\tau_1\tau_2 \ldots \tau_r\sigma$ have the same "parity".* □

16.6 Theorem: *Let $n \geqslant 2$, $\pi \in S_n$. Then π is an odd (even) permutation if and only if π can be written as a product of an odd (even) number of transpositions. In particular, π cannot be written as a product of an odd number of transpositions and also as a product of an even number of transpositions.*

Proof: We use Lemma 16.5 with $\sigma = \iota$. Let π be written as a product of transpositions, say $\pi = \tau_1\tau_2 \ldots \tau_r$. Lemma 16.5 tells us that $\pi = \tau_1\tau_2 \ldots \tau_r\iota$ and ι have opposite or same "parities" according as whether r is odd or even. Since ι has 0 inversions, ι is an even permutation. So $\pi = \tau_1\tau_2 \ldots \tau_r$ is an odd permutation or an even permutation according as whether r is an odd number or an even number. The other assertion follows from the remark made after Definition 16.4. □

We describe the "parity" of a product.

16.7 Theorem: *Let $n \geqslant 2$. The product of two permutations in S_n has the "parity" given by the following law.*

$$(odd)(odd) = (even) \qquad (odd)(even) = (odd)$$
$$(even)(odd) = (odd) \qquad (even)(even) = (even).$$

Proof: Let $\sigma, \pi \in S_n$. We want to find the "parity" of $\sigma\pi$. Let $\sigma = \tau_1\tau_2 \ldots \tau_s$ and $\pi = \tau_1'\tau_2' \ldots \tau_p'$, where $\tau_1, \tau_2, \ldots, \tau_s, \tau_1', \tau_2', \ldots, \tau_p'$ are transpositions (Theorem 16.2). Then $\sigma\pi = \tau_1\tau_2 \ldots \tau_s\tau_1'\tau_2' \ldots \tau_p'$ is a product of $s + p$ transpositions.

If σ is an odd permutation and π is an odd permutation, then s is an odd number and p is an odd number (Theorem 16.6), so $s + p$ is an even number, so $\sigma\pi$ is an even permutation (Theorem 16.6). Thus (odd)(odd) = (even). The other cases are proved similarly. □

The assertion of Theorem 16.7 resembles the rule for finding the sign of a product of two real numbers: the product of a negative number by a negative number is positive, etc. In order to exploit this analogy, we introduce a new term.

16.8 Definition: Let $n \in \mathbb{N}$ and $\sigma \in S_n$. The *sign of* σ is the integer 1 or -1. We write $\varepsilon(\sigma)$ for the sign of σ, and define it as follows.

$$\varepsilon(\sigma) = \begin{cases} 1 \text{ if } \sigma \text{ is an even permutation} \\ -1 \text{ if } \sigma \text{ is an odd permutation.} \end{cases}$$

With this definition, the content of Theorem 16.7 can be expressed more succintly.

16.7 Theorem: *For any σ, π in S_n, there holds* $\varepsilon(\sigma\pi) = \varepsilon(\sigma)\varepsilon(\pi)$. □

16.9 Theorem: *Let $n \geq 2$. The number of odd permutations in S_n is equal to the number of even permutations in S_n. This number is $n!/2$.*

Proof: We must find a one-to-one correspondence between the set of odd permutations and the set of even permutations in S_n. Now

$$\begin{aligned} T: \{\sigma \in S_n: \varepsilon(\sigma) = -1\} &\to \{\sigma \in S_n: \varepsilon(\sigma) = 1\} \\ \sigma &\to (12)\sigma \end{aligned}$$

is a one-to-one mapping (by Lemma 8.1(1)) from the set of odd permutations in S_n into the set of even permutations in S_n (by Lemma 16.3), which is in fact onto, since any even permutation π is the image, under T, of the odd permutation $(12)\pi$ (Lemma 16.3). So T is a one-to-one correspondence between these sets and they contain equal number of elements, say k elements. Since these sets are disjoint, and their union is S_n, there are $2k$ elements in S_n, whose order is $n!$ by Theorem 15.2. Hence $k = n!/2$. □

Theorem 16.7 asserts that the set of even permutations in S_n is closed under multiplication. So it is a subgroup of S_n by Lemma 9.3(2).

16.10 Definition: The subgroup of even permutations in S_n ($n \geq 2$) is called the *alternating group (on n letters)* and is written as A_n.

16.11 Theorem: *For $n \geq 2$, A_n is a group of order $n!/2$.*

Proof: Theorem 16.9. □

Exercises

1. Find the sign of (13524) and of (153462).

2. Show that a cycle of length m is odd (even) if and only if m is even (odd).

3. Prove that $\varepsilon(\sigma_1\sigma_2\ldots\sigma_t) = \varepsilon(\sigma_1)\varepsilon(\sigma_2)\ldots\varepsilon(\sigma_t)$ for all permutations $\sigma_1, \sigma_2, \ldots \sigma_t$ in S_n.

4. Find the sign of (143)(1245)(243) and of (1435)(25643) without evaluating these products.

5. Write all elements in A_2, A_3, A_4.

6. Construct multiplication tables of A_2, A_3, A_4.

7. Find all subgroups of A_4. Does A_4 have a subgroup of order 6?

8. Verify Lemma 16.3 by going through the argument in its proof in the specific cases below.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 1 & 5 & 7 & 2 & 4 & 6 \end{pmatrix}, \quad (ik) = (12),\ (14),\ (23),\ (26),\ (27),\ (67).$$

§17
Groups of Matrices

In this paragraph, we examine some groups whose elements are matrices. The reader probably knows matrices (whose entries are real or complex numbers), but this is not a prerequisite for understanding this paragraph. We give an elementary account of the theory of matrices as far as needed here. Matrix theory will be taken systematically in Chapter 4, §43.

We allow the entries to be elements of any field. Fields will be formally introduced in Chapter 3, §29 (Definition 29.13). Until then, we shall be content with the following definition.

17.1 Temporary Definition: A *field* is one of the sets $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{Z}_p$, where p is a prime number.

After having learned about fields in Chapter 3, the reader may check that the theory in this paragraph carries over to the more general situation where the term "field" is used in the sense of Definition 29.13.

We note that K is a commutative group under addition, whose identity element we shall denote by 0 (so that 0 is the number 0 in case K is one of $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and it is the residue class $\overline{0} = 0 + p\mathbb{Z}$ in case K is $\mathbb{Z}_p$ for some prime number p), and that $K\backslash\{0\}$ is a group under multiplication. This will be used many times in this paragraph.

17.2 Definition: Let K be a field. A *matrix over* K is an array

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

of four elements a,b,c,d of K, arranged in two rows and two columns, and enclosed within parentheses. (The plural of "matrix" is "matrices".)

Thus $\begin{pmatrix} 1 & 2 \\ -4 & 0 \end{pmatrix}$ is matrix over $\mathbb{Q}$ (and also over $\mathbb{R}$ and $\mathbb{C}$), $\begin{pmatrix} \pi & \sqrt{2} \\ 5 & -7 \end{pmatrix}$ is a matrix over $\mathbb{R}$ (and also over $\mathbb{C}$). In addition, $\begin{pmatrix} \bar{2} & \bar{3} \\ \bar{5} & \bar{4} \end{pmatrix}$ is a matrix over $\mathbb{Z}_7$, when bars mean residue classes modulo 7.

The set of all matrices over a field K will be denotded by $Mat_2(K)$. The subscript 2 signifies that there are 2 rows and 2 columns in a matrix (in the sense of Definition 17.2).

If K is a field and A,B are matrices from $Mat_2(K)$, we say A *is equal to* B provided the corrsponding entries in A and B are equal. More exactly,

$$A: \begin{pmatrix} a & b \\ c & d \end{pmatrix} \text{ is equal to } B: \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$$

if and only if $a = a'$, $b = b'$, $c = c'$, $d = d'$. In this case, we write $A = B$. A single matrix equation is equivalent to four equations between the elements of the underlying field. It is clear that matrix equality is an equivalence relation on $Mat_2(K)$. In particular, it is legitimate to say that A and B are equal when A is equal to B.

In this definition of matrix equality, the location of the entries are taken into account. Thus $\begin{pmatrix} 5 & 1 \\ 0 & 2 \end{pmatrix}$ and $\begin{pmatrix} 2 & 5 \\ 1 & 0 \end{pmatrix}$ are different matrices, although they are made up of the same numbers.

We introduce two binary operations on $Mat_2(K)$, addition and multiplication. Addition is defined in the most obvious way.

17.3 Definition: Let K be a field. For any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ in $Mat_2(K)$, we define the *sum of* A *and* B as the matrix

$$\begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}.$$

The sum of A and B will be denoted by $A + B$. Taking sums in $Mat_2(K)$ will be called *addition* (of matrices).

Addition of matrices is essentially the addition in the underlying field, carried out four times. Not surprisingly, many properties of addition in the field are reflected in matrix addition. For example, just like a field is a group under addition, matrices over a field form a group under addition, too.

17.4 Theorem: *Let K be a field. Then $Mat_2(K)$ is a commutative group under addition.*

Proof: We check the group axioms

(i) For any matrices $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ in $Mat_2(K)$, we have

$a + e, b + f, c + g, d + h \in K$ since $a,b,c,d,e,f,g,h \in K$ and K is closed under addition. Hence

$$A + B = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix} \in K$$

and $Mat_2(K)$ is closed under (matrix) addition.

(ii) Associativity of addition in $Mat_2(K)$ follows from associativity of addition in K. Indeed, for any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$, $C = \begin{pmatrix} k & m \\ n & p \end{pmatrix}$ in $Mat_2(K)$, we have

$$(A + B) + C = \left[\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix}\right] + \begin{pmatrix} k & m \\ n & p \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix} + \begin{pmatrix} k & m \\ n & p \end{pmatrix}$$

$$= \begin{pmatrix} (a+e)+k & (b+f)+m \\ (c+g)+n & (d+h)+p \end{pmatrix}$$

$$= \begin{pmatrix} a+(e+k) & b+(f+m) \\ c+(g+n) & d+(h+p) \end{pmatrix}$$

$$= \begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e+k & f+m \\ g+n & h+p \end{pmatrix}$$

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

(iii) What can be the identity element? Well, probably the matrix $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, where 0 denotes the zero element of the field K (for

instance, when K is $\mathbb{Z}_p$ for some prime number p, 0 is the residue class $\overline{0}$ $= p\mathbb{Z}$). Indeed, we have, for any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$,

$$A + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} a+0 & b+0 \\ c+0 & d+0 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A$$

and $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ is a right identity of $Mat_2(K)$. The matrix $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ will be called the *zero matrix (over K)* and will be designated by the symbol 0. This should not be confused with the zero element of the underlying field K.

(iv) Any matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$ has a right inverse (opposite) $-A$ in $Mat_2(K)$, namely $\begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix}$ (since $-a, -b, -c, -d \in K$):

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix} = \begin{pmatrix} a+(-a) & b+(-b) \\ c+(-c) & d+(-d) \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0.$$

Thus $Mat_2(K)$ is a group under addition. We finally check commutativity.

(v) Commutativity of addition in $Mat_2(K)$ follows from commutativity of addition in K. Indeed, for any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ in $Mat_2(K)$, we have

$$A + B = \begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix} = \begin{pmatrix} e+a & f+b \\ g+c & h+d \end{pmatrix} = \begin{pmatrix} e & f \\ g & h \end{pmatrix} + \begin{pmatrix} a & b \\ c & d \end{pmatrix} = B + A.$$

So $Mat_2(K)$ is a commutative group under addition. □

The additive group $Mat_2(K)$ is somewhat dull. It is just four copies of the additive group K. More interesting matrix groups arise when the operation is multiplication. We introduce this operation now.

17.5 Definition: Let K be a field. For any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ in $Mat_2(K)$, we define the *product of A and B* as the matrix

$$\begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}.$$

The product of A and B will be denoted by $A \cdot B$ or simply by AB. Taking products in $Mat_2(K)$ will be called *multiplication* (of matrices).

This definition looks bizarre. One would expect the product of A and B, with the notation of Definition 17.5, to be $\begin{pmatrix} ae & bf \\ cg & dh \end{pmatrix}$. Some motivation for Definition 17.5 can be gained as follows. With each matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ (over $\mathbb{R}$, say), there is associated a coordinate transformation

$$\begin{aligned} x &= ax' + by' \\ y &= cx' + dy' \end{aligned}$$

of the Euclidean plane. Carrying out the transformations associated with $\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ successively, we obtain

$$\begin{aligned} x &= ax' + by' & x' &= ex'' + fy'' \\ y &= cx' + dy' & y' &= gx'' + hy'', \end{aligned}$$

which gives

$$\begin{aligned} x &= a(ex'' + fy'') + b(gx'' + hy'') = (ae+bg)x'' + (af+bh)y'' \\ y &= c(ex'' + fy'') + d(gx'' + hy'') = (ce+dg)x'' + (cf+dh)y'', \end{aligned}$$

so the product of the matrices is the one which is associated with the successive application of the transformation.

If matrix multiplication is new to you, you are urged to write down matrices over $\mathbb{R}$ and multiply them in order to acquire dexterity in performing this operation.

We collect some basic properties of matrix multiplication in the next theorem. Let us recall that $K\backslash\{0\}$ is a group under multiplication. The identity element of this group will be denoted by 1. Thus 1 is the number 1 when K is one of $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, and the residue class $\overline{1} = 1 + p\mathbb{Z}$ when $K = \mathbb{Z}_p$ for some prime number p.

17.6 Theorem: *Let K be a field, whose zero element is 0 and whose identity element is 1.*

(1) *$Mat_2(K)$ is closed under matrix multiplication.*

(2) *$(AB)C = A(BC)$ for all $A,B,C \in Mat_2(K)$.*

(3) *Let $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then $AI = IA = A$ for all $A \in Mat_2(K)$.*

(4) *$A(B + C) = AB + AC$ and $(B + C)A = (BA + CA)$ for all $A,B,C \in Mat_2(K)$.*

Proof: Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$, $C = \begin{pmatrix} k & m \\ n & p \end{pmatrix}$ be arbitrary elements of $Mat_2(K)$.

(1) Since a field is closed under addition and multiplication, $ae + bg$, $af + bh$, $ce + dg$, $cf + dh \in K$ whenever $a,b,c,d,e,f,g,h \in K$. So $AB \in Mat_2(K)$ for all $A,B \in Mat_2(K)$ and $Mat_2(K)$ is closed under multiplication.

(2) This is routine calculation. We evaluate $(AB)C$ and $A(BC)$:

$$(AB)C = \left[\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix}\right]\begin{pmatrix} k & m \\ n & p \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}\begin{pmatrix} k & m \\ n & p \end{pmatrix}$$

$$= \begin{pmatrix} (ae+bg)k + (af+bh)n & (ae+bg)m + (af+bh)p \\ (ce+dg)k + (cf+dh)n & (ce+dg)m + (cf+dh)p \end{pmatrix}$$

$$= \begin{pmatrix} aek+bgk+afn+bhn & aem+bgm+afp+bhp \\ cek+dgk+cfn+dhn & cem+dgm+cfp+dhp \end{pmatrix}, \qquad \text{(i)}$$

$$A(BC) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\left[\begin{pmatrix} e & f \\ g & h \end{pmatrix}\begin{pmatrix} k & m \\ n & p \end{pmatrix}\right] = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} ek+fn & em+fp \\ gk+hn & gm+hp \end{pmatrix}$$

$$= \begin{pmatrix} a(ek+fn)+b(gk+hn) & a(em+fp)+b(gm+hp) \\ c(ek+fn)+d(gk+hn) & c(em+fp)+d(gm+hp) \end{pmatrix}$$

$$= \begin{pmatrix} aek+afn+bgk+bhn & aem+afp+bgm+bhp \\ cek+cfn+dgk+dhn & cem+cfp+dgm+dhp \end{pmatrix}. \qquad \text{(ii)}$$

Since addition is commutative in K, the matrices (i) and (ii) are equal. Hence $(AB)C = A(BC)$ for all $A,B,C \in Mat_2(K)$.

(3) We compute $AI = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a1+b0 & a0+b1 \\ c1+d0 & c0+d1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A,$

$$IA = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 1a+0c & 1b+0d \\ 0a+1c & 0b+1d \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = A,$$

as claimed.

(4) We have $A(B + C) = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\left[\begin{pmatrix} e & f \\ g & h \end{pmatrix} + \begin{pmatrix} k & m \\ n & p \end{pmatrix}\right]$

$$= \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e+k & f+m \\ g+n & h+p \end{pmatrix}$$

$$= \begin{pmatrix} a(e+k)+b(g+n) & a(f+m)+b(h+p) \\ c(e+k)+d(g+n) & c(f+m)+d(h+p) \end{pmatrix}$$

$$= \begin{pmatrix} ae+ak+bg+bn & af+am+bh+bp \\ ce+ck+dg+dn & cf+cm+dh+dp \end{pmatrix}$$

$$= \begin{pmatrix} ae+bg+ak+bn & af+bh+am+bp \\ ce+dg+ck+dn & cf+dh+cm+dp \end{pmatrix}$$

$$= \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix} + \begin{pmatrix} ak+bn & am+bp \\ ck+dn & cm+dp \end{pmatrix}$$

$$= \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix} + \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} k & m \\ n & p \end{pmatrix}$$

$$= AB + AC.$$

The proof of $(B + C)A = (BA + CA)$ follows similar lines and is left to the reader. □

Theorem 17.6 seems promising. Three of the group axioms are satisfied, with I as the identity. It remains to investigate whether every matrix over a field has a right inverse.

Suppose K is a field and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$. Then A has a right inverse $X = \begin{pmatrix} x & y \\ z & u \end{pmatrix}$ in $Mat_2(K)$ if and only if $AX = I$, which is equivalent to

$$\begin{array}{ll} (1) \quad ax + bz = 1, & (2) \quad ay + bu = 0, \\ (3) \quad cx + dz = 0, & (4) \quad cy + du = 1. \end{array}$$

We multiply the equation (1) by d, (3) by $-b$ and add them side by side. Using associativity of addition in K, distributivity of multiplication over addition, and *commutativity* of multiplication in K, we get

$$(ad - bc)x = d.$$

We multiply (2) by d, (4) by $-b$ and add them. We multiply (1) by $-c$, (3) by a and add them. We multiply (2) by $-c$, (4) by a and add them. We get

$$(ad - bc)y = -b, \qquad (ad - bc)z = -c, \qquad (ad - bc)u = a.$$

We emphasize again that commutativity of multiplication in K is used crucially to derive these equations.

The element $ad - bc$ appears in each one of these equations. In view of its importance, we give it a name.

17.7 Definition: Let K be a field and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$. Then the element $ad - bc$ in K is called the *determinant of* A, written as $det(A)$ or as $det\ A$.

We have shown: if K is a field and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$, and if $X = \begin{pmatrix} x & y \\ z & u \end{pmatrix}$ in $Mat_2(K)$ is a right inverse of A, then

$$\begin{array}{ll} (det\ A)x = d & (det\ A)y = -b \\ (det\ A)z = -c & (det\ A)u = a. \end{array} \qquad (D)$$

These equations impose certain conditions on a matrix having a right inverse. We cannot expect that every matrix has a right inverse. Those having a right inverse are characterized very simply as the matrices with a nonzero determinant.

17.8 Theorem: *Let K be a field and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$. Then A has a right inverse if and only if $det\ A \neq 0$. If this is the case, then there is a unique right inverse of A, namely the matrix*

$$\begin{pmatrix} (det\ A)^{-1}d & -(det\ A)^{-1}b \\ -(det\ A)^{-1}c & (det\ A)^{-1}a \end{pmatrix}$$

where $(det\ A)^{-1}$ is the inverse of $det\ A \in K\backslash\{0\}$ in the multiplicative group $K\backslash\{0\}$.

Proof: First we assume $det\ A = 0$ and show that A has no right inverse. Indeed, if $det\ A = 0$ and A had a right inverse, then the equations (D) would become

$$\begin{aligned} d &= 0 & b &= 0 \\ c &= 0 & a &= 0, \end{aligned}$$

and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ would be the zero matrix $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. The existence of a right inverse $X = \begin{pmatrix} x & y \\ z & u \end{pmatrix}$ would yield

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I = AX = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x & y \\ z & u \end{pmatrix} = \begin{pmatrix} 0x+0z & 0y+0u \\ 0x+0z & 0y+0u \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

hence $1 = 0$ in K, a contradiction. Thus A has no right inverse if $det\ A = 0$.

Now let us assume $det\ A \neq 0$ and show that A has a unique right inverse. Since $det\ A \in K\backslash\{0\}$ and $K\backslash\{0\}$ is a group under multiplication, $det\ A$ has an inverse in $K\backslash\{0\}$, which we denote by $(det\ A)^{-1}$. This is the nonzero element of the field K such that $(det\ A)^{-1}(det\ A) = (det\ A)(det\ A)^{-1} = 1 =$ the identity element of $K\backslash\{0\}$. So we can solve for x,y,z,u in (D) by multiplying the equations in (D) by $(det\ A)^{-1}$. We get

$$\begin{aligned} x &= (det\ A)^{-1}d, & y &= -(det\ A)^{-1}b, \\ z &= -(det\ A)^{-1}c, & u &= (det\ A)^{-1}a. \end{aligned}$$

Thus, if A has a right inverse at all, this right inverse must be the matrix written in the enunciation of the theorem (in particular, A has a unique right inverse). It is easy to check that this matrix is indeed a right inverse of A:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} (det\ A)^{-1}d & -(det\ A)^{-1}b \\ -(det\ A)^{-1}c & (det\ A)^{-1}a \end{pmatrix}$$

$$= \begin{pmatrix} (det\ A)^{-1}(ad-bc) & (det\ A)^{-1}(-ab+ba) \\ -(det\ A)^{-1}(cd-dc) & (det\ A)^{-1}(-cb+da) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I.$$

Hence A does have a unique right inverse and it is the matrix given in this theorem. □

We will prove presently that the matrices with right inverses form a group under multiplication. From Lemma 7.3, it will then follow that the unique right inverse of a matrix with a nonzero determinant is also the unique left inverse of the same matrix. We shall refer to is as its inverse. The rule for finding the inverse of $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is simple: interchange a and d, then put a minus sign in front of b and c, and multiply each entry by $(det\ A)^{-1}$ [i.e., divide each entry by $det\ A$]. For example, the inverse of $\begin{pmatrix} 5 & 2 \\ 1 & 2 \end{pmatrix} \in Mat_2(\mathbb{Q})$ is $\begin{pmatrix} \frac{1}{8}2 & -\frac{1}{8}2 \\ -\frac{1}{8}1 & \frac{1}{8}5 \end{pmatrix} = \begin{pmatrix} 1/4 & -1/4 \\ -1/8 & 5/8 \end{pmatrix}$ and that of

$\begin{pmatrix} \overline{5} & \overline{2} \\ \overline{1} & \overline{4} \end{pmatrix} \in Mat_2(\mathbb{Z}_7)$ is $\begin{pmatrix} \overline{2} & \overline{4} \\ -\overline{2} & \overline{1} \end{pmatrix} \begin{pmatrix} -\overline{2} & \overline{2} \\ \overline{2} & \overline{5} \end{pmatrix} = \begin{pmatrix} \overline{1} & \overline{3} \\ \overline{5} & \overline{3} \end{pmatrix}$ since the determinant is equal to $\overline{18} = \overline{4}$ and $\overline{4}^{-1} = \overline{2}$.

17.9 Theorem: *Let K be a field.*
(1) *$det\ (AB) = (det\ A)(det\ B)$ for all $A,B \in Mat_2(K)$.*
(2) *$det\ I = 1\ (\in K)$.*
(3) *If $AX = I$, then $det\ X = (det\ A)^{-1}$.*

Proof: (1) We use the notation of Definition 17.5. We get

$$\begin{aligned} det\ (AB) &= (ae + bg)(cf + dh) - (af + bh)(ce + dg) \\ &= aecf + aedh + bgcf + bgdh - afce - afdg - bhce - bhdg \\ &= aedh - afdg + bgcf - bhce \\ &= ad(eh - fg) - bc(eh - fg) \\ &= (ad - bc)(eh - fg) \\ &= (det\ A)(det\ B). \end{aligned}$$

(2) $det\ I = 1.1 - 0.0 = 1 - 0 = 1$.

(3) This follows from (1) and (2): if $AX = I$, then $1 = det\ I = det(AX) = (det\ A)(det\ X)$, so $det\ X = (det\ A)^{-1}$. □

The formula $det\ AB = (det\ A)(det\ B)$ is known as the multiplication rule of determinants. Loosely speaking, the determinant of a product is the product of the determinants. By induction on n, it is extended to n factors: $det\ (A_1A_2...A_n) = (det\ A_1)(det\ A_2)...(det\ A_n)$.

We finally have a group of matrices under multiplication.

17.10 Theorem: *Let K be a field. Then*

$$\{A \in Mat_2(K): det\ A \neq 0\}$$

is a group under matrix multiplication.

Proof: We check the group axioms. Let us call our set G for brevity.

(i) For $A, B \in G$, we have $det\ A \neq 0 \neq det\ B$. In the field K, product of nonzero elements is nonzero ($K\backslash\{0\}$ is a group, and closed under multiplication). So $det\ AB = (det\ A)(det\ B) \neq 0$ by Theorem 17.9(1) and consequently $AB \in G$. Thus G is closed under multiplication.

(ii) Associativity of multiplication in G follows from Theorem 17.6(2).

(iii) I is a right identity element of G, for $det\ I = 1 \neq 0$ by Theorem 17.9(2), so $I \in G$; and $AI = A$ for all $A \in G$ by Theorem 17.6(3).

(iv) Any $A \in G$ has a right inverse in G. Indeed, if $A \in G$, then $det\ A \neq 0$, so A has a right inverse X in $Mat_2(K)$. As $det\ X = (det\ A)^{-1} \neq 0$ (Theorem 17.9(3)), we see $X \in G$. Thus A has a right inverse in G.

Therefore, G is a group. □

17.11 Definition: Let K be a field. The group of Theorem 17.10 is called the *general linear group (of degree 2) over K*, and is written as $GL(2,K)$.

Since $GL(2,K)$ is a group, the unique right inverse of any matrix A in $GL(2,K)$ is also the unique left inverse of that matrix (Lemma 7.3). It will be called the *inverse of A*, and will be written as A^{-1}, in conformity with the usual terminology and notation. The matrix I will be called the *identity matrix*. Elements of $GL(2,K)$ are called *invertible* matrices or *regular* matrices. Matrices whose determinants are zero are called *singular*.

The next theorem furnishes another matrix group.

17.12 Theorem: *Let K be a field. Then*

$$\{A \in Mat_2(K): det\ A = 1\}$$

is a group under matrix multiplication.

Proof: Let us call this set S for brevity. As $1 \neq 0$ in K, we get $S \subseteq GL(2,K)$. We use the subgroup criterion (Lemma 9.2) to check that S is a subgroup of $GL(2,K)$.

(i) For $A,B \in S$, we have $det\ A = 1 = det\ B$, therefore $det\ AB = (det\ A)(det\ B) = 1.1 = 1$ by Theorem 17.9(1) and consequently $AB \in S$. Thus S is closed under multiplication.

(ii) For any $A \in S$, we have $det\ A = 1$, so $det\ (A^{-1}) = (det\ A)^{-1} = 1^{-1} = 1$ by Theorem 17.9(3) and $A^{-1} \in S$. Thus S is closed under the forming of inverses.

Therefore, S is a subgroup of $GL(2,K)$. □

17.13 Definition: Let K be a field. The group of Theorem 17.12 is called the *special linear group (of degree 2) over* K, and is written as $SL(2,K)$.

We close this paragraph with a group that plays an important role in number theory and in complex analysis.

17.14 Theorem: *The set*

$$\left\{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(\mathbb{Q}): a,b,c,d \in \mathbb{Z},\ ad - bc = 1\right\}$$

is a group under matrix multiplication.

Proof: Let us call this set H for brevity. Clearly $\varnothing \neq H \subseteq SL(2,\mathbb{Q})$. We check that H is a subgroup of $SL(2,\mathbb{Q})$.

(i) Suppose $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$ are elements of H. Then $AB = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$. Here the entries of AB, namely $ae+bg$, $ce+dg$, $af+bh$, $cf+dh$ are integers, because a,b,c,d,e,f,g,h are integers. Also, $det\ A = 1 = det\ B$, therefore $det\ AB = (det\ A)(det\ B) = 1.1 = 1$ by Theorem 17.9(1) and $AB \in H$. Thus H is closed under multiplication.

(ii) Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in H$. Then $det\ A = 1$ and so $A^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$ by Theorem 17.8. The entries d, $-b$, $-c$, a of A^{-1} are integers, because a, b, c, d are integers. Also, we have $det\ A = 1$, so $det\ (A^{-1}) = (det\ A)^{-1} = 1^{-1} = 1$ by Theorem 17.9(3) (or $det\ (A^{-1}) = da - (-b)(-c) = ad - bc = 1$). So $A^{-1} \in H$ and H is closed under the forming of inverses.

Therefore, H is a subgroup of $SL(2,\mathbb{Q})$. □

17.15 Definition: The group of Theorem 17.14 is called the *special linear group (of degree 2) over* $\mathbb{Z}$, or the *modular group*, and is written as $SL(2,\mathbb{Z})$ or as Γ.

Exercises

1. Let K be a field. Show that $GL(2,K)$ is not an abelian group.

2. Find all elements of $GL(2,\mathbb{Z}_2)$. What is the order of $GL(2,\mathbb{Z}_2)$?

3. Write down the multiplication table of $GL(2,\mathbb{Z}_2)$. Compare it (eventually after reordering the rows and columns) with the multiplication table of S_3.

4. Find all elements of $SL(2,\mathbb{Z}_3)$. What is the order of $SL(2,\mathbb{Z}_3)$?

5. Write down the multiplication table of $SL(2,\mathbb{Z}_3)$

6. Let K be a field and let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$. When $a = 0 = b$, we have $det\ A = 0$. In case $(a,b) \neq (0,0)$, prove that $det\ A = 0$ if and only if there is an element k in K such that $c = ka$, $d = kb$. Use this result and show that $|GL(2,\mathbb{Z}_p)| = (p^2 - 1)(p^2 - p)$.

7. Determine how many elements in $GL(2,\mathbb{Z}_p)$ have the same determinant. Find the order of $SL(2,\mathbb{Z}_p)$.

8. Show that $\left\{ \begin{pmatrix} 1 & 0 \\ a & b \end{pmatrix} \in Mat_2(K): b \neq 0 \right\}$ is a subgroup of $GL(2,K)$.

9. Prove that $\left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \in Mat_2(K): ad \neq 0 \right\}$ is a group under multiplication. Its elements are called *triangular* matrices.

10. Let K be a field. For any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$, we define the *trace of* A to be the element $a + d$ of K (sum of the entries in the upper-left lower-right diagonal). Show that the trace of AB is equal to the trace of BA for all $A,B \in Mat_2(K)$.

11. Let K be a field. For any $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$, we define the *transpose of* A to be the matrix $\begin{pmatrix} a & c \\ b & d \end{pmatrix} \in Mat_2(K)$, which is written A^t. Show that $det\ A^t = det\ A$ and $(AB)^t = B^tA^t$ for all $A,B \in Mat_2(K)$.

12. Let $m \geqslant 2$ and put $Mat_2(\mathbb{Z}_m) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z}_m \right\}$. Show that the theory in the text, until Theorem 17.8, remains valid for the elements of $Mat_2(\mathbb{Z}_m)$, which are called *matrices over* $\mathbb{Z}_m$.

In place of Theorem 17.8, prove that $A \in Mat_2(\mathbb{Z}_m)$ has a unique right inverse if and only if $det\ A \in \mathbb{Z}_m^\times$.

Put $GL(2,\mathbb{Z}_m) = \{A \in Mat_2(\mathbb{Z}_m) : det\ A \in \mathbb{Z}_m^\times\}$. Show that $GL(2,\mathbb{Z}_m)$ is a group under multiplication.

Prove that Theorem 17.12 remains true if "K" is replaced by "$\mathbb{Z}_m$".

13. Develope a theory of matrices over $\mathbb{Z}$ by modifying the theory of matrices over $\mathbb{Z}$. How do you define $GL(2,\mathbb{Z})$?

14. Let $H = \left\{ \begin{pmatrix} a & \bar{b} \\ b & \bar{a} \end{pmatrix} : a,b \in \mathbb{C} \right\} \subseteq Mat_2(\mathbb{C})$, where $\bar{x}$ is the complex conjugate of $x \in \mathbb{C}$. Prove that H is closed under addition and multiplication. Show that $H\backslash\{0\}$ is a group under multiplication.

15. If K is a field and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Mat_2(K)$, we write $-A = \begin{pmatrix} -a & -b \\ -c & -d \end{pmatrix}$. Let

$$1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ i = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},\ j = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},\ k = \begin{pmatrix} 0 & -i \\ -i & 0 \end{pmatrix} \in Mat_2(\mathbb{C}).$$

Thus 1 is the identity matrix over $\mathbb{C}$. Show that $ij = k$, $jk = i$, $ki = j$. Prove that $\{1,-1,i,-i,j,-j,k,-k\}$ is a group under multiplication, called a *quaternion* group of order 8 and is denoted as Q_8. Show that Q_8 has exactly one element of order 2. Find all subgroups of Q_8.

§18
Factor Groups

In this paragraph, we learn a way of constructing new groups from a given one. This construction is a generalization of obtaining the additive group $\mathbb{Z}_n$ from the additive group $\mathbb{Z}$. We recall that the elements of $\mathbb{Z}_n$ are certain subsets of $\mathbb{Z}$, namely the cosets of the subgroup $\{nz: z \in \mathbb{Z}\}$ in $\mathbb{Z}$ (cf. §10, Ex. 3). Addition in $\mathbb{Z}_n$ is induced from addition in $\mathbb{Z}$ (see §6). We want to do the same thing with an arbitrary group G. We start with a group G and a subgroup H of G. On the set of cosets of H in G, we wish to define a binary operation which reflects the operation on G and which makes the set of cosets into a group.

Two questions present themselves immediately. First, we have a set $\mathcal{R}$ of right cosets of H in G and a set $\mathcal{L}$ of left cosets of H in G. If G is an abelian group, the right cosets and the left cosets coincide. However, in general, the right cosets of H in G are different from the left cosets of H in G. Thus we have two different sets of cosets: $\mathcal{R} \neq \mathcal{L}$. Do we want to make $\mathcal{R}$ into a group or $\mathcal{L}$? Is it possible to make both $\mathcal{R}$ and $\mathcal{L}$ into groups? If so, how are these groups related? If not, why not?

Another question is about the operation. The central issue in §6, where we introduced the operations on $\mathbb{Z}_n$, was whether these operations were well defined. Once we knew that addition in $\mathbb{Z}_n$ is a well defined operation, it was straightforward to prove that $\mathbb{Z}_n$ is a group. Not surprisingly, we have the same problem here. The main point of the following discussion is to show that we have a well defined operation (Theorem 18.4). Once we know it, it is easy to show that our set of cosets is a group (Theorem 18.7).

It turns out that these questions are intimately connected and they will be resolved simultaneously.

18.1 Suggestion: *Let G be a group, H a subgroup of G, and let $\mathcal{R}$ be the set $\{Ha: a \in G\}$ of all right cosets of H in G. We suggest that we define a binary operation on $\mathcal{R}$, to be denoted by $\cdot$ or by juxtaposition, according to the "rule"*

$$Ha.Hb = Hab$$

for all $a,b \in G$.

This is the most natural way of defining a binary operation on $\mathfrak{R}$. Now we have to ask whether this is a well defined operation on $\mathfrak{R}$, for the "rule" of evalutating a product $Ha.Hb$ makes use of the elements a,b of G, which can be chosen in many ways. The "rule" says, in order to evaluate the product $X.Y$ of X and Y in $\mathfrak{R}$, that we (1) take an $a \in X$ so that $X = Ha$; (2) that we take an $b \in Y$ so that $Y = Hb$; (3) that we evaluate ab in G; (4) that we find the right coset $Hab \in \mathfrak{R}$ of this ab. The right coset Hab is supposed to be the product $X.Y$. We must make sure that we get the same right coset at the end, even if we choose different elements from the right cosets X and Y. We investigate when this "rule" yields a well defined operation on $\mathfrak{R}$.

The operation suggested in 18.1 is well defined if and only if the implication

$$\text{for all } a,a_1,b,b_1 \in G, \quad Ha = Ha_1 \text{ and } Hb = Hb_1 \implies Hab = Ha_1b_1$$

is valid. Using Lemma 10.2, we write this in the equivalent form

$$\text{for all } a,a_1,b,b_1 \in G, h,h_1 \in H, \ a_1 = ha \text{ and } b_1 = h_1b \implies a_1b_1 \in Hab$$

which simplifies to

$$\text{for all } a,a_1,b,b_1 \in G, h,h_1 \in H, \qquad ahh_1b \in Hab.$$

Using Lemma 10.2 again, we can write this as

$$\text{for all } a \in G, h_1 \in H, \quad ah_1 \in Ha$$

or as

$$\text{for all } a \in G, \qquad aH \subseteq Ha. \tag{o}$$

Thus the operation sugested in 18.1 is well defined if and only if H is a subgroup of G such that $aH \subseteq Ha$ for all $a \in G$. This is not true for every G and for every subgroup H of G. After we give other descriptions of such subgroups, we will see some examples.

18.2 Lemma: *Let $H \leqslant G$. For $a \in G$, let $a^{-1}Ha$ be the set*

$$\{a^{-1}ha \in G: h \in H\} = \{b \in G: aba^{-1} \in H\}.$$

The following are equivalent.

(1) $a^{-1}ha \in H$ *for all* $a \in G, h \in H$.
(2) $a^{-1}Ha \subseteq H$ *for all* $a \in G$.
(3) $a^{-1}Ha = H$ *for all* $a \in G$.
(4) $Ha = aH$ *for all* $a \in G$.
(5) $aH \subseteq Ha$ *for all* $a \in G$.

Proof: (1) $\Rightarrow$ (2) This follows from the definition of the set $a^{-1}Ha$.

(2) $\Rightarrow$ (3) Suppose $a^{-1}Ha \subseteq H$ for all $a \in G$. Then, for any $a \in G$, it is true that $(a^{-1})^{-1}Ha \subseteq H$. Hence, for any $h \in H, a \in G$, we have $aha^{-1} \in H$, so $h = a^{-1}(aha^{-1})a \in a^{-1}Ha$. Since this holds for all $h \in H$, we obtain $H \subseteq a^{-1}Ha$, for all $a \in G$. Together with the hypothesis $a^{-1}Ha \subseteq H$ for all $a \in G$, this yields $a^{-1}Ha = H$ for all $a \in G$.

(3) $\Rightarrow$ (4) If $a^{-1}Ha = H$, then

$$\begin{aligned} Ha &= \{ha \in G: h \in H\} \\ &= \{a(a^{-1}ha) \in G: h \in H\} \\ &= \{ax \in G: x \in a^{-1}Ha\} \\ &= \{ax \in G: x \in H\} \\ &= aH. \end{aligned}$$

(4) $\Rightarrow$ (5) This is trivial.

(5) $\Rightarrow$ (1) Suppose $aH \subseteq Ha$ for all $a \in G$. Then $a^{-1}H \subseteq Ha^{-1}$ for all $a \in G$. Keeping a fixed, we see $a^{-1}h \in Ha^{-1}$ for all $h \in H$. Thus, for all $h \in H$, there is an $h_1 \in H$ such that $a^{-1}h = h_1a^{-1}$. So $a^{-1}ha = h_1 \in H$. So $a^{-1}ha \in H$ for all h in H, and this holds for all $a \in G$. □

18.3 Definition: Let $H \leqslant G$. If H satisfies one (and hence all) of the conditions in Lemma 18.2, then H is called a *normal subgroup of* G, or *normal in* G.

We employ the symbol $H \trianglelefteq G$ to denote that H is a normal subgroup of G. Also, $H \not\trianglelefteq G$ means that H is not a normal subgroup of G. If H is a proper and normal subgroup of G, we write $H \triangleleft G$. Finally, $H \not\triangleleft G$ means that H is not a proper normal subgroup of G.

18.4 Theorem: *The operation suggested in* 18.1 *is well defined if and only if* $H \trianglelefteq G$.

Proof: This follows from (o), Lemma 18.2(5) and Definition 18.3. □

By Lemma 18.2(4), any right coset of H in G is a left coset of H in G if and only if $H \trianglelefteq G$. So the set $\mathfrak{R}$ of right cosets of H is equal to the set $\mathfrak{L}$ of left cosets of H if and only if $H \trianglelefteq G$. Theorem 18.4 shows that we have a well defined operation on $\mathfrak{R}$ if and only if $\mathfrak{R} = \mathfrak{L}$. This answers our two questions. We do not have to bother about the distinction between $\mathfrak{R}$ and $\mathfrak{L}$: if (and only if) the operation is well defined, there is no distinction between $\mathfrak{R}$ and $\mathfrak{L}$. See also Ex. 1 at the end of this paragraph.

18.5 Examples: (a) For any group G, it is clear that $G \trianglelefteq G$. Also, $\{1\} \trianglelefteq G$, since $a^{-1}1a \in \{1\}$ for all $a \in G$. We make a convention here. The trivial subgroup $\{1\}$ will henceforward be written simply as 1. It will be clear from the context whether 1 stands for the identity element or for the trivial subgroup. Thus $1 \trianglelefteq G$ and $G \trianglelefteq G$.

(b) Any subgroup of an abelian group is normal in that group. Indeed, if G is abelian and $H \leqslant G$, then $hg = gh$ for all $h \in H, g \in G$, hence $Hg = gH$ for all $g \in G$. Thus $H \trianglelefteq G$ by Lemma 18.2(4).

In the abelian group case, $Hg = gH$ is satisfied trivially, for $hg = gh$ for all $h \in H, g \in G$. You should notice, however, $Hg = gH$ does *not* mean that g commutes with every element of H. This is an equation between certain sets, so is equivalent to the inclusions $Hg \subseteq gH$ and $gH \subseteq Hg$. The first inclusion means

$$\text{for all } h \in H, \text{ there is } h_1 \in H \text{ such that } hg = gh_1.$$

Here $h_1 \neq h$ in general and therefore $hg = gh_1 \neq gh$. The second inclusion has a similar meaning.

$Hg = gH$ means that, when we multiply the elements of H by g on the right and on the left, we get the same *collection* of elements. It does not mean that, when we multiply any element of H by g on the right and on the left, we get the same product.

Many beginners misunderstand this point. Be careful not to read more than set equality in $Hg = gH$. Compare this with an isometry fixing a subset F of the Euclidean plane E and one fixing F pointwise (§14).

(c) Consider the subgroup $A_3 = \{\iota,(123),(132)\}$ of S_3. There are $|S_3:A_3| = |S_3|/|A_3| = 6/3 = 2$ right cosets and 2 left cosets of A_3 in S_3. These are

$$A_3 \text{ and } A_3(12) = \{(12),(23),(13)\}$$
$$A_3 \text{ and } (12)A_3 = \{(12),(13),(23)\}$$

and so any right coset of A_3 in S_3 is also a left coset of A_3 in S_3. Thus $A_3 \trianglelefteq S_3$.

(d) The result in Example 18.5(c) can be generalized. Let $H \leqslant G$ of index $|G:H| = 2$. Then there are two right cosets of H in G and two left cosets of H in G. Let H and X be the right cosets, H and Y the left cosets. From the disjoint unions

$$G = H \cup X \quad \text{and} \quad G = H \cup Y,$$

we read off

$$X = G \setminus H = Y,$$

so the right cosets H,X of H in G coincide with the left cosets H,X of H in G. Hence $H \trianglelefteq G$: if H has index two in G, then H is normal in G.

(e) Consider the subgroup $H := \{\iota,(12)\}$ of S_3. Now $|S_3:H| = 6/2 = 3$. The three right cosets of H and the three left cosets of H are

$$H = \{\iota,(12)\} \qquad H = \{\iota,(12)\}$$
$$H(13) = \{(13),(123)\} \qquad (13)H = \{(13),(132)\}$$
$$H(23) = \{(23),(132)\} \qquad (23)H = \{(23),(123)\}$$

and the right coset $\{(13),(123)\}$ is not a left coset. So $H \not\trianglelefteq S_3$. In the same way, $\{\iota,(13)\}$ and $\{\iota,(23)\}$ are not normal subgroups of S_3.

(f) Let $H = \{\iota,(12),(34),(12)(34)\}$. It is easy to see that $H \leqslant S_4$. Is H normal in S_4? We compare the right and left cosets of H in S_4. Aside from H, we see that the right coset

$$H(13)(24) = \{(13)(24),(1423),(3241),(14)(23)\}$$

is a left coset:

$$(13)(24)H = \{(13)(24),(1324),(1423),(14)(23)\}$$

since $(3241) = (1324)$. This is of course not enough to conclude $H \trianglelefteq S_4$. We must examine the other cosets also. We see

$$H(13) = \{(13),(123),(341),(1234)\}$$
$$(13)H = \{(13),(132)$$

and we stop here. This shows $H(13) \neq (13)H$. Hence $H \not\trianglelefteq S_4$.

(g) Let $V_4 = \{\iota,(12)(34),(13)(24),(14)(23)\}$. It is easily seen that $V_4 \leqslant S_4$. The subgroup V_4 is known as *Klein's four group* (after the German mathematician Felix Klein (1849-1925); Vierergruppe, whence V_4). The cosets of V_4 in S_4 are

$$\begin{array}{llll} & V_4 & & V_4 \\ V_4(12) &= \{(12),(34),(1324),(1423)\}, & (12)V_4 &= \{(12),(34),(1423),(1324)\} \\ V_4(13) &= \{(13),(1234),(24),(1432)\}, & (13)V_4 &= \{(13),(1432),(24),(1234)\} \\ V_4(23) &= \{(23),(1342),(1243),(14)\}, & (23)V_4 &= \{(23),(2431),(2134),(14)\} \\ V_4(123) &= \{(123),(134),(243),(142)\}, & (123)V_4 &= \{(123),(243),(142),(134)\} \\ V_4(132) &= \{(132),(234),(124),(143)\}, & (132)V_4 &= \{(132),(143),(234),(124)\} \end{array}$$

and since each right coset is a left coset, $V_4 \trianglelefteq S_4$. For a more conceptual proof of this result, see Ex. 5 at the end of this paragraph.

(h) Consider $K = \{\iota,(12),(13),(23),(123),(132)\} \leqslant S_4$. Is K normal in S_4? We observe $(14)^{-1}K(14) = (14)K(14) = \{\iota,(42),(43),(23),(423),(432)\} \neq K$ and so $K \not\trianglelefteq S_4$.

(i) Normality is not an intrinsic property of a subgroup. It is meaningles to speak about normality of a subgroup H itself. It is only meaningful to speak about normality of H in a group G. We have to specify the group G as well as the subgroup H when we speak about normality. It is possible that $H \trianglelefteq G_1$ and $H \not\trianglelefteq G_2$ for two groups G_1, G_2 containing H. Here is an example. Take

$$G_1 = D_3 = \langle \rho,\sigma : \rho^4 = 1, \sigma^2 = 1, \sigma^{-1}\rho\sigma = \rho^{-1} \rangle$$

$$G_2 = \langle \rho^2, \sigma \rangle = \{1, \rho^2, \sigma, \rho^2\sigma\} \leqslant G_1$$
$$H = \langle \sigma \rangle = \{1, \sigma\}.$$

Then $H \leqslant G_1$ and $H \leqslant G_2$. Now $|G_2 : H| = 2$, so $H \trianglelefteq G_2$ by Example 18.5(d) above. However

$$\rho^{-1}H\rho = \{1, \rho^{-1}\sigma\rho\} = \{1, \rho^{-1}\rho^{-1}\sigma\} = \{1, \rho^2\sigma\} \neq H$$

and thus $H \not\trianglelefteq G_1$.

Incidentally, $G_2 \trianglelefteq G_1$ since $|G_1 : G_2| = 2$. This shows that normality is not a transitive relation. It is possible that $H \trianglelefteq G_2, G_2 \trianglelefteq G_1$, yet $H \not\trianglelefteq G_1$.

(j) For any field K, we have $SL(2,K) \trianglelefteq GL(2,K)$. Indeed, if $S \in SL(2,K)$, then $det\ S = 1$ and, for any $G \in GL(2,K)$,

$$\begin{aligned} det\ (G^{-1}SG) &= det\ (G^{-1}.SG) = det\ G^{-1}.det\ (SG) = (det\ G)^{-1}.(det\ S)(det\ G) \\ &= (det\ G)^{-1}1(det\ G) = 1, \\ G^{-1}SG &\in SL(2,K) \text{ for all } S \in SL(2,K), G \in GL(2,K), \end{aligned}$$

and so $SL(2,K) \trianglelefteq GL(2,K)$ by Lemma 18.2(1).

(k) If $H \trianglelefteq G$ and $K \trianglelefteq G$, then $H \cap K \trianglelefteq G$. More generally, if $H_i \trianglelefteq G$ (where $i \in I$, an index set), then $\bigcap_{i \in I} H_i \trianglelefteq G$. We show this. Put $H = \bigcap_{i \in I} H_i$ for brevity. From $H_i \leqslant G$, it follows that $H \leqslant G$ (Example 9.4(f)). Also, for any $h \in H$ and $g \in G$,

$$\begin{aligned} h &\in H_i \text{ for all } i \in I, \\ g^{-1}hg &\in H_i \text{ for all } i \in I, \\ g^{-1}hg &\in H \end{aligned}$$

and $H \trianglelefteq G$ by Lemma 18.2(1).

(l) If $H \trianglelefteq G$ and $K \leqslant G$, then $H \cap K \trianglelefteq K$. Indeed, let $h \in H \cap K$, $k \in K$. Then $k^{-1}hk \in H$ since $h \in H$ and H is normal in G. Also, $k^{-1}hk \in K$ because $h \in K$ and K is closed under multiplication. Thus $k^{-1}hk \in H \cap K$ for all $h \in H \cap K$ and for all $k \in K$. Thus $H \cap K \trianglelefteq K$ by Lemma 18.2(1).

18.6 Definition: When $H \trianglelefteq G$, the set of all right cosets of H in G, which is also the set of all left cosets of H in G by Lemma 18.2(4), will be denoted by G/H, read *G by H*, or *G modulo H*, or *G mod H*.

Most authors do not insist on the condition $H \trianglelefteq G$ when they write G/H. They write G/H for the set $\mathfrak{R}$ of right cosets of H in G (or for the set of left cosets, especially when they write functions on the left) and employ some other symbol for the the set of left cosets (or for the the set of right cosets). Throughout this book, whenever we write G/H, it will be tacitly supposed that $H \trianglelefteq G$. The notation G/H is meaningless if $H \not\trianglelefteq G$ and will not be used in this case.

18.7 Theorem: *Let $H \trianglelefteq G$. Then G/H is a group under the operation suggested in* 18.1, *by which*

$$Ha.Hb = Hab \quad \text{for all } Ha, Hb \in G/H.$$

Proof: We check the group axioms.

(i) The operation on G/H is well defined by Theorem 18.4 and the product of two right cosets is again a right coset. So G/H is closed under this operation.

(ii) For all $Ha,Hb,Hc \in G/H$, we have $(Ha.Hb)Hc = Hab.Hc = H(ab.c) = H(a.bc) = Ha.Hbc = Ha(Hb.Hc)$ since $ab.c = a.bc$ for all $a,b,c \in G$. The operation is therefore associative.

(iii) $H = H1 \in G/H$ is a right identity element of since
$Ha.H1 = Ha1 = Ha$ for all $Ha \in G/H$.

(iv) Any $Ha \in G/H$ has a right inverse in G/H, namely Ha^{-1}:
$Ha.Ha^{-1} = Ha.a^{-1} = H1 = H$ = identity element of G/H.

Therefore G/H is a group. □

18.8 Definition: Let $H \trianglelefteq G$. The group G/H of Theorem 18.7 is called the *factor group of G with respect to H*, or the *factor group G by H*, or the *factor group G mod(ulo) H*. Instead of the term "factor group", the term "quotient group" is also used. The group operation is called *multiplication (of cosets)*.

Please notice that G/H is *not* a subgroup of G. The elements of G/H are subsets of G, not elements of G.

Since the multiplication on G/H is based on the multiplication on G, we expect that some properties of G are inherited by G/H. Here are some properties that are taken over by the factor groups.

18.9 Lemma: *Let $H \trianglelefteq G$.*
(1) $|G/H| = |G:H|$. *In particular, if G is finite, so is G/H and $|G/H| = |G|/|H|$.*
(2) *If G is abelian, so is G/H.*
(3) *If G is cyclic, so is G/H.*

Proof: (1) The elements of G/H are the cosets of H in G and there are $|G:H|$ cosets of H in G by Definition 10.7. So the order of G/H is the index of H in G. The second assertion follows from Lagrange's theorem.

(2) If G is abelian, then $ab = ba$ for all $a,b \in G$ and $Ha.Hb = Hab = Hba = Hb.Ha$ for all $Ha, Hb \in G/H$. Thus G/H is abelian, too.

(3) Assume that G is cyclic, say $G = \langle g \rangle$. Then any element x of G is of the form g^n, where $n \in \mathbb{Z}$. Hence any coset of H in G is of the form $Hx = Hg^n = (Hg)^n$. This shows $G/H = \langle Hg \rangle$. □

The converses of the claims in Lemma 18.9 are false. The factor group G/H can be finite (abelian, cyclic) without G being finite (abelian, cyclic).

We close this paragraph with some examples of factor groups.

18.10 Examples: **(a)** Let G be a group and $H = 1 = \{1\}$. Then $H \trianglelefteq G$ (Example 18.5(a)). The cosets of $H = 1$ the subsets of G having only one element:

$$Ha = \{1\}a = \{a\} \quad \text{for all } a \in G$$

and multiplication in $G/H = G/1$ is given by

$$\{a\}\{b\} = \{ab\}.$$

The factor group $G/1$ is governed by the same operation as G. Thus $G/1$ is almost the same group as G. The only difference is that the elements of G are enclosed within braces in $G/1$.

(b) Let $\mathbb{Z}$ be the additive group of integers and let $n\mathbb{Z} = \{nz \in \mathbb{Z} : z \in \mathbb{Z}\}$ be the subgroup of $\mathbb{Z}$ consisting of integers divisible by n. Since $\mathbb{Z}$ is abelian, $n\mathbb{Z} \trianglelefteq \mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}$ consists of the n *cosets*

$$n\mathbb{Z}, n\mathbb{Z} + 1, n\mathbb{Z} + 2, \ldots, n\mathbb{Z} + n - 1$$

which are usually abbreviated as

$$\overline{0}, \overline{1}, \overline{2}, \ldots, \overline{n-1}$$

(see §6; we write the cosets additively of course). Thus $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$ as sets.

In the factor group $\mathbb{Z}/n\mathbb{Z}$, the operation is given by

$$(n\mathbb{Z} + a) + (n\mathbb{Z} + b) = n\mathbb{Z} + (a + b) \qquad \text{for all } a,b \in \mathbb{Z}$$

which can be written shortly as

$$\overline{a} + \overline{b} = \overline{a + b} \qquad \text{for all } \overline{a},\overline{b} \in \mathbb{Z}/n\mathbb{Z}.$$

This is the definition of addition in $\mathbb{Z}_n$. So the operation in $\mathbb{Z}/n\mathbb{Z}$ coincides with the operation on $\mathbb{Z}_n$ that we learned in §6. Hence $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}_n$ as (additive) groups.

We understand the real reason why addition on $\mathbb{Z}_n$, as defined in §6, is a well defined operation. It is well defined only because $n\mathbb{Z} \trianglelefteq \mathbb{Z}$.

(c) Let $G = C_{12} = \langle g : g^{12} = 1 \rangle$ be a cyclic group of order 12 and let $H = \langle g^3 \rangle = \{1, g^3, g^6, g^9\} \leqslant G$. Since G is abelian, $H \trianglelefteq G$ and G/H consists of the cosets

$$H = \{1, g^3, g^6, g^9\},\ Hg = \{g, g^4, g^7, g^{10}\},\ Hg^2 = \{g^2, g^5, g^8, g^{11}\}.$$

The multiplication table of G/H is given below.

	H	Hg	Hg^2
H	H	Hg	Hg^2
Hg	Hg	Hg^2	H
Hg^2	Hg^2	H	Hg

(d) We know $V_4 \trianglelefteq S_4$ (Example 18.5(g)). The elements of S_4/V_4 are $V_4, V_4(12), V_4(13), V_4(23), V_4(123), V_4(132)$ and the multiplication table of S_4/V_4 is

.	V_4	$V_4(12)$	$V_4(13)$	$V_4(23)$	$V_4(123)$	$V_4(132)$
V_4	V_4	$V_4(12)$	$V_4(13)$	$V_4(23)$	$V_4(123)$	$V_4(132)$
$V_4(12)$	$V_4(12)$	V_4	$V_4(123)$	$V_4(132)$	$V_4(13)$	$V_4(23)$
$V_4(13)$	$V_4(13)$	$V_4(132)$	V_4	$V_4(123)$	$V_4(23)$	$V_4(12)$
$V_4(23)$	$V_4(23)$	$V_4(123)$	$V_4(132)$	V_4	$V_4(12)$	$V_4(13)$
$V_4(123)$	$V_4(123)$	$V_4(23)$	$V_4(12)$	$V_4(13)$	$V_4(132)$	V_4
$V_4(132)$	$V_4(132)$	$V_4(13)$	$V_4(23)$	$V_4(12)$	V_4	$V_4(123)$

This is almost identical with with the multiplication table of S_3:

.	ι	(12)	(13)	(23)	(123)	(132)
ι	ι	(12)	(13)	(23)	(123)	(132)
(12)	(12)	ι	(123)	(132)	(13)	(23)
(13)	(13)	(132)	ι	(123)	(23)	(12)
(23)	(23)	(123)	(132)	ι	(12)	(13)
(123)	(123)	(23)	(12)	(13)	(132)	ι
(132)	(132)	(13)	(23)	(12)	ι	(123)

Thus S_4/V_4 is almost the same group as S_3. They are not the same groups, of course, for the underlying sets are different. Nevertheless, it is clear

from the tables above that the operations on S_4/V_4 and on S_3 are closely related. This will be made more precise in §20.

Exercises

1. Let $H \leqslant G$ and let $\mathfrak{L}$ be the set of all left cosets of H in G. We suggest that we define a binary operation on $\mathfrak{L}$, according to the "rule"

$$aH.bH = abH$$

for all $a,b \in G$. Show that this operation is well defined if and only if $H \trianglelefteq G$.

2. Let $H \leqslant G$. Prove that $H \trianglelefteq G$ if and only if $Ha \subseteq aH$ for all $a \in G$.

3. Prove that, if $H \leqslant G$, $a \in G$ and Ha is a left coset of H in G, then $Ha = aH$.

4. Find a group G, a subgroup H of G, and an element a of G such that $a^{-1}Ha \subseteq H$ but $a^{-1}Ha \neq H$. Why does this not contradict Lemma 18.2?

5. Let $\{a,b,c,d\} = \{1,2,3,4\}$. Show that, for any $\sigma \in S_4$,

$$(ab)(cd)\sigma = \sigma(a\sigma,b\sigma)(c\sigma,d\sigma)$$

and thus $\sigma^{-1}\alpha\sigma \in V_4$ for all $\alpha \in V_4$. This proves $V_4 \trianglelefteq S_4$. Compare with §15, Ex. 15.

6. Find all normal subgroups of S_4 (cf. §15, Ex.10).

7. Find all normal subgroups of $SL(2,\mathbb{Z}_3)$.

8. Determine whether the following are normal subgroups in the groups indicated.

$\{g \in GL(2,\mathbb{R})$: $det\ g \geqslant 5\}$ in $GL(2,\mathbb{R})$
$\{g \in GL(2,\mathbb{R})$: $det\ g \geqslant 0\}$ in $GL(2,\mathbb{R})$
$\{g \in GL(2,\mathbb{R})$: $det\ g > 0\}$ in $GL(2,\mathbb{R})$
$\{g \in GL(2,\mathbb{C})$: $det\ g = 1\}$ in $GL(2,\mathbb{C})$
$\{g \in GL(2,\mathbb{C})$: $(det\ g)^{18} = 1\}$ in $GL(2,\mathbb{C})$
$\{g \in GL(2,\mathbb{Z}_{11})$: $det\ g = \overline{1}$ or $\overline{3}$ or $\overline{4}$ or $\overline{5}$ or $\overline{9}\}$ in $GL(2,\mathbb{Z}_{11})$.

9. Let K be a field. Then $K\backslash\{0\}$ is a group under multiplication. Suppose U is a subgroup of $K\backslash\{0\}$. Prove that $\{g \in GL(2,K)$: $det\ g \in U\}$ is a subgroup of $GL(2,K)$.

10. Let $n \in \mathbb{N}$ and put

$$\Gamma_n = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z}): \begin{matrix} a \equiv 1, b \equiv 0 \\ c \equiv 0, d \equiv 1 \end{matrix} \pmod{n} \right\}.$$

Determine if $\Gamma_n \trianglelefteq SL(2,\mathbb{Z})$.

11. Let $H \trianglelefteq G$ and let $Ha \in G/H$. Show that $o(Ha) = n$ (n is a natural number) if and only if n is the smallest natural number such that $x^n \in H$.

12. Show by counterexamples that the converses of the claims in Lemma 18.9 are false.

§19
Product Sets in Groups

In the preceding paragraph, we introduced a multiplication on the set of right cosets of a subgroup H of a given group G. This involved selecting elements from the cosets to be multiplied. Selecting elements from the cosets is an artificial step in this coset multiplication. We showed in Theorem 18.4 that the resulting coset is independent of the elements chosen when (and only when) H is a normal subgroup of G. However, this does not get rid of the inherent artificiality of the coset multiplication we studied in §18. A more natural multiplication would treat the elements of cosets on equal standing, rather than distinguishing (selecting) one of them (as in Suggestion 18.1) and then showing (as in Theorem 18.4) that no injustice to the remaining elements has been commited. We introduce in this paragraph a natural multiplication of cosets, and in fact more generally of arbitrary nonempty subsets in a group. The new multiplication will coincide with the one of Suggestion 18.1.

19.1 Definition: Let G be a group. For any nonempty subsets X,Y of G, the *product set* XY is defined to be

$$XY = \{xy \in G: x \in X, y \in Y\}.$$

When X has only one element, say when $X = \{x\}$, we write xY instead of $\{x\}Y$. Likewise, we write Xy instead of $X\{y\}$. This is consistent with the definition of cosets (Definition 10.1).

This multiplication is associative.

19.2 Lemma: *Let G be a group. For any nonempty subsets X,Y,Z of G, there holds $(XY)Z = X(YZ)$.*

Proof: This follows from the associativity of multiplication in G:

$$(XY)Z = \{uz \in G: u \in XY, z \in Z\}$$

$$= \{(xy)z \in G: x \in X, y \in Y, z \in Z\}$$
$$= \{x(yz) \in G: x \in X, y \in Y, z \in Z\}$$
$$= \{xv \in G: x \in X, v \in YZ\}$$
$$= X(YZ). \qquad \square$$

Using Lemma 8.3, we may and do drop the parentheses in any product set involving more than two subsets. For example, we write $XYZUV$ for $(XY)(Z(UV))$.

19.3 Examples: (a) Let $H \leqslant G$. As we have remarked earlier, $H\{x\} = Hx = \{hx: h \in H\}$ is the right coset of H in G containing $x \in G$. Analogously, $\{x\}H = xH$ is the left coset of H that contains x.

(b) Let G be a group, $H \leqslant G$ and $x \in G$. Then $x^{-1}Hx = \{x^{-1}hx: h \in H\}$ (see Lemma 18.2) is the product of the sets $\{x^{-1}\}, H, \{x\}$.

(c) Let G be a group and let X be a nonempty subset of G. Then XX consists of all products x_1x_2, where x_1 and x_2 run through X independently. Notice that $XX \neq \{x^2 \in G: x \in X\}$ in general. X is a multiplicatively closed subset of G if and only if $XX \subseteq X$. In particular, $HH \subseteq H$ for any subgroup H of G.

(d) Let G be a group and let X, Y be nonempty subsets of G. It follows from Definition 19.1 that

$$XY = \bigcup_{y \in Y} Xy = \bigcup_{x \in X} xY.$$

(e) Let $X = \{\iota, (12)\}$, $Y = \{\iota, (13)\}$. Now X and Y are subsets of S_3. Then $XY = \{\iota\iota, \iota(13), (12)\iota, (12)(13)\} = \{\iota, (13), (12), (123)\}$. Notice that X, Y are subgroups of S_3, but XY is not. So the product of two subgroups is not necessarily a subgroup.

(f) Let $X = \{\iota, (13)\}$ and $V_4 = \{\iota, (12)(34), (13)(24), (14)(23)\}$. Then $X \leqslant S_4$ and $V_4 \trianglelefteq S_4$ (Example 18.10(d)). Here XV_4
$= \{\iota\iota, \iota(12)(34), \iota(13)(24), \iota(14)(23), (13)\iota, (13)(12)(34), (13)(13)(24), (13)(14)(23)\}$
$= \{\iota, (12)(34), (13)(24), (14)(23), (13), (1432), (24), (1234)\}$
is easily seen to be closed under multiplication, hence XV_4 is a subgroup of S_4 (Lemma 9.3(2)), but not a normal subgroup of S_4, for $(13) \in XV_4$ but

$(12)^{-1}(13)(12) = (23) \notin XV_4$ (Lemma 18.2(1)). We see that the product of two subgroups is not necessarily a normal subgroup even if one of the factors is a normal subgroup.

In Example 19.3(f) above, it is easy to see $XV_4 = V_4X$. This is the basic reason why XV_4 turns out to be a subgroup of S_4. The next lemma describes the situation.

19.4 Lemma: *Let $H \leqslant G$ and $K \leqslant G$.*
(1) *$HK \leqslant G$ if and only if $HK = KH$.*
(2) *If $H \trianglelefteq G$ or $H \trianglelefteq G$, then $HK \leqslant G$.*
(3) *If $H \trianglelefteq G$ and $H \trianglelefteq G$, then $HK \trianglelefteq G$.*

Proof: Before we present the proof, it will be worthwhile to discuss the equation $HK = KH$. What does it mean? Well, HK and KH are subsets of G and equality of them is equivalent to the inclusions

$$HK \subseteq KH \text{ and } KH \subseteq HK.$$

The first inclusion means, for any $h \in H$ and $k \in K$, the element hk of$G$ belongs to KH, so that there are $k_1 \in K$ and $h_1 \in H$ such that $hk = k_1h_1$. Similarly, the second inclusion means, for any $k \in K$ and $h \in H$, there are $h_2 \in H$ and $k_2 \in K$ such that $kh = h_2k_2$.

$HK = KH$ does *not* mean that $hk = kh$ for all $h \in H, k \in K$. Of course, if $hk = kh$ for all $h \in H, k \in K$, then trivially $HK = KH$. However, it does not follow from $HK = KH$ that $hk = kh$ for all $h \in H, k \in K$. From $HK = KH$, it follows only that, for any $h \in H, k \in K$, there are $k_1 \in K, h_1 \in H$ and $h_2 \in H, k_2 \in K$ such that $hk = k_1h_1$ and $kh = h_2k_2$.

Now the proof.

(1) We are to show: (a) if $HK = KH$, then $HK \leqslant G$; and (b) if $HK \leqslant G$, then $HK = KH$.

(a) Suppose first $HK = KH$. We prove that HK is closed under multiplication and the forming of inverses (Lemma 9.2).

(i) If $HK = KH$, then $HK.HK = H.KH.K = H.HK.K = HH.KK \subseteq HK$ and so HK is closed under multiplication (see Example 19.3(c)). If you are not satisfied with this demonstration, here is another. Let $x,y \in HK$, say $x = hk$, $y = h_1k_1$ with $h,h_1 \in H$ and $k,k_1 \in K$. We wish to show $xy \in HK$. Now $xy = hk.h_1k_1 = h.kh_1.k_1$ and $kh_1 \in KH = HK$ by hypothesis, so $kh_1 = h_2k_2$ for some $h_2 \in H$, $k_2 \in K$. So $xy = h.kh_1.k_1 = h.h_2k_2.k_1 = hh_2.k_2k_1 \in HK$ since $hh_2 \in H$ and $k_2k_1 \in K$ as $H \leqslant G$ and $K \leqslant G$. Thus HK is closed under multiplication.

(ii) Let $x \in HK$, say $x = hk$ with $h \in H$, $k \in K$. We are to show that $x^{-1} \in HK$. We have $x^{-1} = (hk)^{-1} = k^{-1}h^{-1} \in KH = HK$, because $k^{-1} \in K$ and $h^{-1} \in H$ as K and H are subgroups of G. So HK is closed under the forming of inverses.

This proves that $HK \leqslant G$ whenever $H \leqslant G, K \leqslant G$ and $HK = KH$.

(b) Now suppose $H \leqslant G, K \leqslant G$ and $HK \leqslant G$. We want to show $HK = KH$, that is, $HK \subseteq KH$ and $KH \subseteq HK$. These inclusions follow from the fact that HK is closed under taking inverses. Indeed, if $x \in HK$, then $x^{-1} \in HK$, say $x^{-1} = hk$ with $h \in H, k \in K$. Then $x = (hk)^{-1} = k^{-1}h^{-1} \in KH$. So $HK \subseteq KH$. The other inclusion is proved in the same way.

This proves that $H \leqslant G, K \leqslant G$ and $HK \leqslant G$ implies $HK = KH$.

The proof of (1) is complete.

(2) We suppose $H \trianglelefteq G, K \leqslant G$ and prove that $HK \leqslant G$. According to part (1), it suffices to show $HK = KH$. First we prove $HK \subseteq KH$. Let $h \in H, k \in K$. Then $k^{-1}hk \in H$ since $H \trianglelefteq G$ (Lemma 18.2(1)) and $hk = k.k^{-1}hk \in KH$. This proves $HK \subseteq KH$. Now we prove $KH \subseteq HK$. For any $h \in H, k \in K$, we have $khk^{-1} \in H$ since $H \trianglelefteq G$ and thus $kh = khk^{-1}.k \in HK$ and $KH \subseteq HK$. Therefore $HK = KH$ and $HK \leqslant G$.

The proof of $HK \leqslant G$ under the hypotheses $H \leqslant G, K \trianglelefteq G$ follows similar lines and is left to the reader.

(3) We now assume $H \trianglelefteq G, K \trianglelefteq G$. From part (2), we get $HK \leqslant G$. We are to show $HK \trianglelefteq G$. To do that, we prove $g^{-1}xg \in HK$ for all $g \in G, x \in HK$ (Lemma 18.2(1)). For any $x \in HK$, there are $h \in H, k \in K$ with $x = hk$ and $g^{-1}xg = g^{-1}hkg = g^{-1}hg.g^{-1}kg \in HK$ since $g^{-1}hg \in H$ and $g^{-1}kg \in K$ as $H \trianglelefteq G$ and $K \trianglelefteq G$. Hence $HK \trianglelefteq G$.

This completes the proof. □

In Lemma 19.4(3), it would not be enough to prove that $g^{-1}xg \in HK$ for all $g \in G, x \in HK$. It is necessary to show $HK \leqslant G$ also. Generally speaking, "$A \trianglelefteq B$" summarizes two conditions on A and B: that A is a subgroup of B and that A is normal in B. We must check both of them whenever we want to show $A \trianglelefteq B$.

We turn our attention to the product of two right cosets. The product of two right cosets, as in Definition 19.1, is a subset of the group under discussion. When is it a right coset? The next lemma gives the answer.

19.5 Lemma: *Let $H \leqslant G$. The product of arbitrary right cosets of H in G, according to Definition 19.1, is always a right coset of H in G if and only if $H \trianglelefteq G$.*

Proof: The product of Ha and Hb (where $a,b \in G$) is

$$HaHb = \{hah_1b \in G: h,h_1 \in H\}$$

and $ab = 1a1b \in HaHb$. Thus $HaHb$ is a right coset of H in G if and only if it is the right coset of H in G to which ab belongs:

$$H \text{ is the right coset of } H \text{ in } G \iff HaHb = Hab.$$

We show that $HaHb = Hab$ for all $a,b \in G$ if and only if $H \trianglelefteq G$.

If $H \trianglelefteq G$, then $HaHb = Hab$ for all $a,b \in G$. Indeed, if $H \trianglelefteq G$, then $aH = Ha$ for all $a \in G$ (Lemma 18.2(4)), and, for any $a,b \in G$, we have

$$HaHb = H.aH.b = H.Ha.b = HH.ab = Hab.$$

Here we use $HH = H$, which follows from $HH \subseteq H$ (Example 19.3(c)) and $H = 1H \subseteq HH$.

Conversely, assume $HaHb = Hab$ for all $a,b \in G$. Then

$$HaH = (HaH)(bb^{-1}) = (HaHb)b^{-1} = (Hab)b^{-1} = (Ha)bb^{-1} = Ha$$

$$HaH = Ha$$

$$aH = 1aH \subseteq HaH = Ha$$

and so $aH \subseteq Ha$ for all $a \in G$. From Lemma 18.2(5), we obtain $H \trianglelefteq G$ □

The product of any two right cosets of $H \leqslant G$, as in Suggestion 18.1, is always a right coset of H, provided this multiplication is well defined, and it is well defined if and only if $H \trianglelefteq G$. On the other hand, the product of any two right cosets of $H \leqslant G$, as in Definition 19.1, is always a definite subset of G, but this subset is a right coset of H if and only if $H \trianglelefteq G$. The relation $HaHb = Hab$ in the proof of Lemma 19.5 shows that these two multiplications are identical when $H \trianglelefteq G$.

We know that HK is not necessarily a subgroup of G even if $H \leqslant G, K \leqslant G$. It is a subset of G. We now determine the number of elements in it.

19.6 Lemma: *Let $H,K \leqslant G$ and assume that H and K are finite. Then HK is a finite subset of G, whose cardinality is given by*

$$|HK| = \frac{|H|\,|K|}{|H \cap K|}.$$

Proof: We list all products hk, where h and k run through H and K, respectively. In this way, we get $|H|\,|K|$ elements of G. These are the elements of HK. Naïvely, we expect $|HK|$ to be equal to $|H|\,|K|$, but there may be repetitions in our list: the same element of HK may be written more than once. We have to keep account of repetitions. We show that each of the $|H|\,|K|$ products hk appears exactly n times in our list, where $n := |H \cap K|$. Thus there are $|H|\,|K|/n$ distinct elements in the list and $|HK| = |H|\,|K|/n$. In other words, the mapping

$$\begin{aligned} \varphi: H \times K &\to HK \\ (h,k) &\to hk \end{aligned}$$

is an n-to-one mapping. By this, we understand that exactly n elements in the domain $H \times K$ have the same image under φ.

To prove that φ is an n-to-one mapping, let us investigate when we have $(h_1,k_1)\varphi = (h_2,k_2)\varphi$. Well, $(h_1,k_1)\varphi = (h_2,k_2)\varphi$ if and only if $h_1k_1 = h_2k_2$ and therefore if and only if $h_1^{-1}h_2 = k_1k_2^{-1} = s$ belongs to $H \cap K$. Thus (h_1,k_1) and (h_2,k_2) have the same image under φ if and only if $h_2 = h_1s$ and

$k_2 = s^{-1}k_1$ for some $s \in H \cap K$. Denoting by $1 = s_1, s_2, \ldots, s_n$ the $n = |H \cap K|$ elements of $H \cap K$, we conclude that the n ordered pairs

$$(h_1,k_1),(h_1s_2,s_2^{-1}k_1),(h_1s_3,s_3^{-1}k_1), \ldots,(h_1s_n,s_n^{-1}k_1)$$

and only these ordered pairs have the image h_1k_1 under φ. This proves that φ is indeed n-to-one, and consequently

$$|HK| = \frac{|H|\,|K|}{|H \cap K|}.$$

Exercises

1. Let X,Y be arbitrary nonempty subsets of a group G and let g be an arbitrary element of G. Prove the following equivalences.

$$\begin{aligned} &X \subseteq Y \iff gX \subseteq gY \iff g^{-1}Xg \subseteq g^{-1}Yg; \\ &X = Y \iff gX = gY \iff Xg = Yg \iff g^{-1}Xg = g^{-1}Yg; \\ &X = gY \iff g^{-1}X = Y. \end{aligned}$$

2. Let H,K be subgroups of a group G. Assume that G is finite, $|H| \geqslant \sqrt{|G|}$ and $|K| \geqslant \sqrt{|G|}$. Prove that $H \cap K \neq 1$.

3. Let A,B,C be subgroups of a group G, with $A \leqslant C$. Prove that

$$A(B \cap C) = AB \cap C.$$

4. Let H,K be subgroups of a group G and let $g \in G$. Prove that

$$|HgK| = \frac{|H|\,|K|}{|g^{-1}Hg \cap K|}.$$

(A subset of the form HgK is called a *double coset*.)

§20
Group Homomorphisms

In Example 18.10(d), we have observed that the groups S_4/V_4 and S_3 have almost the same multiplication table. They have the same structure. In this paragraph, we study groups with the same structure.

20.1 Definition: Let G and G_1 be groups and let $\varphi: G \to G_1$ be a mapping from G into G_1. If

$$(ab)\varphi = a\varphi.b\varphi \quad \text{for all } a,b \in G,$$

then φ is called a *(group) homomorphism.*

The equation $(ab)\varphi = a\varphi.b\varphi$ is paraphrased by saying that φ preserves multiplication or that φ preserves products. Loosely speaking, a homomorphism is a mapping under which the image of a product is the product of the images.

Here "products" might refer to different operations. For $a,b \in G$, the product $ab \in G$ is clearly the result of the binary operation of the group G, whereas $a\varphi, b\varphi \in G_1$ and $a\varphi.b\varphi \in G_1$ is the result of the binary operation of the group G_1. This is implicit in the equation $(ab)\varphi = a\varphi.b\varphi$ which does not make any sense unless ab is the product of a,b in G and $a\varphi.b\varphi$ is the product of $a\varphi, b\varphi$ in G_1.

More precisely, if $\circ$ is the binary operation on G and if $*$ is the binary operation on G_1, then $\varphi: G \to G_1$ is a homomorphism provided

$$(a \circ b)\varphi = a\varphi * b\varphi \text{ for all } a,b \in G.$$

20.2 Examples: (a) One homomorphism is very well known to the reader. It is the logarithm function

$$log: \mathbb{R}^+ \to \mathbb{R}$$

from the group $\mathbb{R}^+$ of positive real numbers (under multiplication) into the group $\mathbb{R}$ of all real numbers (under addition). The homomorphism property of the logarithm function is the well known identity

$$log\ ab = log\ a + log\ b$$

that holds for all $a,b \in \mathbb{R}^+$.

(b) The determinant mapping

$$det: GL(2,\mathbb{Q}) \to \mathbb{Q}\backslash\{0\}$$

is a homomorphism from $GL(2,\mathbb{Q})$ into the group of nonzero rational numbers under multiplication, for

$$det\ AB = (det\ A)(det\ B)$$

for all $A,B \in GL(2,\mathbb{Q})$ by Theorem 17.9(1). The same thing is true for the mapping $det: GL(2,K) \to K\backslash\{0\}$, where K is any arbitrary field.

(c) The sign mapping

$$\varepsilon: S_n \to \{1,-1\}$$

is a homomorphism from S_n into the multiplicative group $\{1,-1\}$ since

$$\varepsilon(\sigma\pi) = \varepsilon(\sigma)\varepsilon(\pi)$$

for all $\sigma,\pi \in S_n$ by Theorem 16.7.

(d) The absolute value function

$$\varphi: \mathbb{R}\backslash\{0\} \to \mathbb{R}^+$$
$$a \to |a|$$

is a homomorphism from the group of all nonzero real numbers (under multiplication) into the group of positive real numbers (under multiplication) since

$$(ab)\varphi = |ab| = |a||b| = a\varphi b\varphi$$

for all $a,b \in \mathbb{R}\backslash\{0\}$.

(e) The signum function

$$\text{sgn}: \mathbb{R}\backslash\{0\} \to \{1,-1\}$$
$$x \to \begin{cases} 1 \text{ if } x \text{ is positive} \\ -1 \text{ if } x \text{ is negative} \end{cases}$$

is a homomorphism from the group of nonzero real numbers into the group $\{1,-1\}$.

(f) Let G be a group. Then the identity mapping

$$\iota: G \to G$$

is a homomorphism from G into G since

$$(ab)\iota = ab = a\iota b\iota$$

for all $a,b \in G$. More generally, let H be a subgroup of G and let

$$\mu: H \to G$$
$$h \to h$$

be the inclusion mapping (Example 3.2(a)).Then

$$(ab)\mu = ab = a\mu b\mu$$

for all $a,b \in H$. Hence μ is a homomorphism. Both ι and μ are one-to-one homomorphisms.

(g) Let $\varphi: G \to G_1$ be a group homomorphism and let $H \leqslant G$. Then the restriction

$$\varphi_H: H \to G_1$$

of φ to H (Example 3.2(i)) is a homomorphism from H into G_1 since

$$(ab)\varphi_H = (ab)\varphi = (a)\varphi(b)\varphi = (a)\varphi_H(b)\varphi_H$$

for all $a,b \in H$.

20.3 Lemma: *Let $\varphi: G \to G_1$ be a homomorphism of groups.*

(1) $1\varphi = 1$.

(2) $(a^{-1})\varphi = (a\varphi)^{-1}$ *for all* $a \in G$.

(3) $(a_1a_2 \ldots a_n)\varphi = (a_1\varphi)(a_2\varphi)\ldots(a_n\varphi)$ *for all* $a_1,a_2, \ldots,a_n \in G, n \in \mathbb{N}, n \geqslant 2$.

(4) $(a^n)\varphi = (a\varphi)^n$ *for all* $a \in G, n \in \mathbb{Z}$.

(5) *If* $o(a\varphi) = \infty$, *then* $o(a) = \infty$. *If* $o(a) = n \in \mathbb{N}$, *then* $o(a\varphi)$ *divides* n; *in particular,* $o(a\varphi) \leqslant o(a)$.

Proof: (1) Here we use the same symbol "1" with two different meanings. In "1φ", 1 is the identity element of the group G. On the right hand side, 1 is the identity element of the group G_1. A more accurate way of writing the claim is

$$(1_G)\varphi = 1_{G_1},$$

where 1_G is the identity element of G and 1_{G_1} is that of G_1. For the homomorphisms in Examples 20.2(a)-(e), the assertion means

$$\log 1 = 0;\ \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1;\ \iota \text{ is an even permutation};\ |1| = 1;\ 1 \text{ is positive}$$

respectively.

The proof is easy. We have $1\varphi = (1.1)\varphi = 1\varphi.1\varphi$, hence 1φ is the identity of G_1 by Lemma 7.3(1). One can also use $a\varphi 1\varphi = (a1)\varphi = a\varphi$ with some $a \in G$ to conclude $1\varphi = 1$.

(2) For any $a \in G$, we have $a\varphi(a^{-1}\varphi) = (aa^{-1})\varphi = 1\varphi = 1 = (a\varphi)(a\varphi)^{-1}$, hence $a^{-1}\varphi = (a\varphi)^{-1}$.

(3) We make induction on n. The case $n = 2$ is covered by the very definition of a homomorphism. Supposing the claim to be true for $n = k$, i.e., supposing $(a_1a_2\ldots a_k)\varphi = (a_1\varphi)(a_2\varphi)\ldots(a_k\varphi)$ for all $a_1, a_2, \ldots, a_k \in G$, we get

$$\begin{aligned}(a_1a_2\ldots a_ka_{k+1})\varphi &= ((a_1a_2\ldots a_k)a_{k+1})\varphi \\ &= (a_1a_2\ldots a_k)\varphi(a_{k+1})\varphi \\ &= (a_1\varphi)(a_2\varphi)\ldots(a_k\varphi)(a_{k+1})\varphi\end{aligned}$$

and the claim is true for $n = k + 1$. Hence it is true for all $n \in \mathbb{N}$, $n \geqslant 2$.

(4) We prove $(a^n)\varphi = (a\varphi)^n$ for all $a \in G$, $n \in \mathbb{Z}$. If $n > 0$, this follows from (3) when we take $a_1 = a_2 = \cdots = a_n = a$. If $n < 0$, this follows from (3) when we take $a_1 = a_2 = \cdots = a_n = a^{-1}$. If $n = 0$, the claim is proved in (1).

(5) Suppose $o(a\varphi) = \infty$. If $o(a)$ were a natural number m, then we would obtain $a^m = 1$, so $(a\varphi)^m = (a^m)\varphi = 1\varphi = 1$ and $a\varphi$ would be of finite order by Lemma 11.4, a contradiction. Thus $o(a\varphi) = \infty$ implies $o(a) = \infty$.

Suppose $o(a) = n \in \mathbb{N}$. Then $a^n = 1$ and $(a\varphi)^n = (a^n)\varphi = 1\varphi = 1$, so $o(a\varphi) \mid n$ by Lemma 11.4 and Lemma 11.6. □

Next we show that composition of homomorphisms is also a homomorphism.

20.4 Theorem: *Let* $\varphi: G \to G_1$ *and* $\psi: G_1 \to G_2$ *be group homomorphisms. Then the composition mapping*

$$\varphi\psi: G \to G_2$$

is a homomorphism from G *into* G_2.

Proof: We are to show that $(ab)\varphi\psi = (a)\varphi\psi.(b)\varphi\psi$ for all $a,b \in G$. This follows immediately:

$$\begin{aligned}(ab)\varphi\psi &= ((ab)\varphi)\psi && \text{(definition of } \varphi\psi)\\ &= (a\varphi.b\varphi)\psi && (\varphi \text{ is a homomorphism})\\ &= (a\varphi)\psi.(b\varphi)\psi && (\psi \text{ is a homomorphism})\\ &= (a)\varphi\psi.(b)\varphi\psi && \text{(definition of } \varphi\psi)\end{aligned}$$

for all $a,b \in G$. Hence $\varphi\psi$ is indeed a homomorphism. □

20.5 Definition: Let $\varphi: G \to G_1$ be a group homomorphism. The set

$$\{a\varphi \in G_1: a \in G\} = \{b \in G_1: b = a\varphi \text{ for some } a \in G\}$$

of all images (under φ) of the elements of G is called the *image of* φ and is denoted by $Im\ \varphi$ or by $G\varphi$. The set

$$\{a \in G: a\varphi = 1\}$$

of all elements of the domain G that are mapped to the identity of the range group G_1 is called the *kernel of* φ and is written as $Ker\ \varphi$.

Thus $Im\ \varphi \subseteq G_1$ and $Ker\ \varphi \subseteq G$. It is immediate from the definition of $Im\ \varphi$ that $Im\ \varphi \neq \varnothing$, for $G \neq \varnothing$. Also, $1 = 1_G \in Ker\ \varphi$ by Lemma 20.3(1), so $Ker\ \varphi \neq \varnothing$. We prove now that $Im\ \varphi$ is a subgroup of G_1 and that $Ker\ \varphi$ is a subgroup of G. In fact, $Ker\ \varphi$ is a normal subgroup of G. This is a very important fact.

20.6 Theorem: *Let $\varphi: G \to G_1$ be a group homomorphism. Then*

$$Im\ \varphi \leqslant G_1 \quad \text{and} \quad Ker\ \varphi \trianglelefteq G.$$

Proof: First we prove $Im\ \varphi \leqslant G_1$. We know $Im\ \varphi \neq \varnothing$. We use our subgroup criterion (Lemma 9.2).

(i) Let $x,y \in Im\ \varphi$. We are to show $xy \in Im\ \varphi$. Now $x,y \in Im\ \varphi$ means $x = a\varphi$, $y = b\varphi$ for some $a,b \in G$. Then $xy = (a\varphi)(b\varphi) = (ab)\varphi$ is the

image (under φ) of an element in G, namely of $ab \in G$. So $xy \in Im\,\varphi$ and $Im\,\varphi$ is closed under multiplication.

(ii) Let $x \in Im\,\varphi$. We are to show $x^{-1} \in Im\,\varphi$. Now $x \in Im\,\varphi$ means $x = a\varphi$ for some $a \in G$. Then $x^{-1} = (a\varphi)^{-1} = a^{-1}\varphi$ is the image (under φ) of an element in G, namely of $a^{-1} \in G$. So $x^{-1} \in Im\,\varphi$ and $Im\,\varphi$ is closed under taking inverses.

Thus $Im\,\varphi \leqslant G_1$.

Now we prove $Ker\,\varphi \trianglelefteq G$. First $Ker\,\varphi \leqslant G$. We know $Ker\,\varphi \neq \emptyset$. Compare the following with the proof of Theorem 17.12.

(i) For any $a,b \in Ker\,\varphi$, we have $a\varphi = 1 = b\varphi$, so $(ab)\varphi = (a\varphi)(b\varphi) = 1.1 = 1$ and $ab \in Ker\,\varphi$. Thus $Ker\,\varphi$ is closed under multiplication.

(ii) For any $a \in Ker\,\varphi$, we have $a\varphi = 1$, so $a^{-1}\varphi = (a\varphi)^{-1} = 1^{-1} = 1$ and $a^{-1} \in Ker\,\varphi$. Thus $Ker\,\varphi$ is closed under taking inverses.

Therefore $Ker\,\varphi \leqslant G$. Now we prove that $Ker\,\varphi$ is a normal subgroup of G. Compare the following with Example 18.5(j).

We must show that $g^{-1}kg \in Ker\,\varphi$ for any $g \in G$, $k \in Ker\,\varphi$ (Lemma 18.2(1)). This is easy: if $k \in Ker\,\varphi$, then $k\varphi = 1$ and, for any $g \in G$,

$$(g^{-1}kg)\varphi = (g^{-1}\varphi)(k\varphi)(g\varphi) = (g\varphi)^{-1}.1.g\varphi = 1,$$

so $g^{-1}kg \in Ker\,\varphi$. Thus $Ker\,\varphi \trianglelefteq G$. □

The elements of a group which have the same image under a homomorphism make up a coset of the kernel of that homomorphism.

20.7 Lemma: *Let $\varphi: G \to G_1$ be a group homomorphism. For any $a,b \in G$, there holds $a\varphi = b\varphi$ if and only if $(Ker\,\varphi)a = (Ker\,\varphi)b$.*

Proof: Let $a,b \in G$. Then $a\varphi = b\varphi$ if and only if

$$\begin{aligned}
(a\varphi)(b\varphi)^{-1} &= 1, && \text{so if and only if}\\
(a\varphi)(b^{-1}\varphi) &= 1, && \text{so if and only if}\\
(ab^{-1})\varphi &= 1, && \text{so if and only if}\\
ab^{-1} &\in Ker\,\varphi, && \text{so if and only if}
\end{aligned}$$

$$(Ker\ \varphi)a = (Ker\ \varphi)b$$

by Lemma 10.2(5). □

Since $Ker\ \varphi \trianglelefteq G$ by Theorem 20.6, we also have $a(Ker\ \varphi) = \{b \in G: b\varphi = a\varphi\}$. Alternatively, one may prove a lemma analogous to Lemma 20.7, stating that a and b have the same image under φ if and only if the left cosets $a(Ker\ \varphi)$ and $b(Ker\ \varphi)$ are equal, and combine it Lemma 20.7 to get $a(Ker\ \varphi) = (Ker\ \varphi)a$, thereby proving $Ker\ \varphi \trianglelefteq G$ anew.

It follows from Lemma 20.7 that φ is a one-to-one homomorphism if and only if $Ker\ \varphi$ has only one element. We give a direct proof of this.

20.8 Theorem: *Let $\varphi: G \to G_1$ be a group homomorphism. Then φ is one-to-one if and only if $Ker\ \varphi = 1$.*

Proof: Here 1 is the trivial subgroup of G (Example 18.5(a)). We prove φ is not one-to-one if and only if $Ker\ \varphi \neq 1$.

If φ is not one-to-one, then there are $a,b \in G$ with $a\varphi = b\varphi$ and $a \neq b$. We obtain then $1 = a\varphi.(a\varphi)^{-1} = a\varphi.(b\varphi)^{-1} = a\varphi.b^{-1}\varphi = (ab^{-1})\varphi$, with $ab^{-1} \neq 1$. Thus $1 \neq ab^{-1} \in Ker\ \varphi$ and $Ker\ \varphi \neq 1$.

Conversely, if $Ker\ \varphi \neq 1$, then there is an $a \in Ker\ \varphi$ with $a \neq 1$. Then we have $a\varphi = 1 = 1\varphi$ and $a \neq 1$. So φ is not one-to-one. □

We can determine whether a homomorphism is one-to-one by examining its kernel. A homomorphism φ is one-to-one if and only if $Ker\ \varphi = 1$. Also, we can determine whether a homomorphism is onto by examining its image. A homomorphism φ is onto if and only if $Im\ \varphi$ is the whole range. Homomorphisms which are both one-to-one and onto will have a name.

20.9 Definition: A group homomorphism $\varphi: G \to G_1$ is called an *isomorphism* if it is one-to-one and onto. If there is an isomorphism from

G onto G_1, we say G *is isomorphic to* G_1, and write $G \cong G_1$. If G is not isomorphic to G_1, we write $G \ncong G_1$.

20.10 Examples: (a) The logarithm function is well known to be a one-to-one function onto the set of real numbers. Thus

$$log: \mathbb{R}^+ \to \mathbb{R}$$

is an isomorphism.

(b) For any group G, the identity mapping

$$\iota: G \to G$$

is an isomorphism.

(c) Let G be a group. Then

$$\begin{aligned} \varphi: G &\to G/1 \\ g &\to \{g\} \end{aligned}$$

is an isomorphism from G onto $G/1$ (see Example 18.10(a)). Thus $G \cong G/1$.

(d) The mapping

$$\begin{aligned} \varphi: S_3 &\to S_4/V_4 \\ \sigma &\to V_4\sigma \end{aligned}$$

(where, on the right hand side, σ is the permutation in S_4 that fixes 4 and maps 1,2,3 as $\sigma \in S_3$ does) is an homomorphism. This is evident from the tables in Example 18.10(d). Also, φ is clearly one-to-one and onto. So φ is an isomorphism and $S_3 \cong S_4/V_4$.

An isomorphism, being one-to-one and onto, has an inverse mapping. It is natural to ask if the inverse of an isomorphism is an isomorphism. Also, is it true that composition of two isomorphisms is an isomorphism?

20.11 Lemma: *Let* $\varphi: G \to G_1$ *and* $\psi: G_1 \to G_2$ *be group isomorphisms.*
(1) *The composition* $\varphi\psi: G \to G_2$ *is an isomorphism from* G *onto* G_2.
(2) *The inverse* $\varphi^{-1}: G_1 \to G$ *of* φ *is an isomorphism from* G_1 *onto* G.

Proof: (1) The composition $\varphi\psi$ is a homomorphism by Theorem 20.4. It is one-to-one and onto by Theorem 3.13. So $\varphi\psi$ is an isomorphism.

(2) For any $x,y \in G_1$, we must show $(xy)\varphi^{-1} = x\varphi^{-1}.y\varphi^{-1}$. Since φ is onto, there are $a,b \in G$ such that $a\varphi = x$ and $b\varphi = y$. Now a and b are unique with this property, for φ is one-to-one, and $a = x\varphi^{-1}$, $b = y\varphi^{-1}$. This is the definition of the inverse mapping. Since φ is a homomorphism, we have

$$(ab)\varphi = a\varphi.b\varphi = xy$$

Hence, by definition of φ^{-1}, we get $ab = (xy)\varphi^{-1}$. Thus

$$(xy)\varphi^{-1} = ab = x\varphi^{-1}.y\varphi^{-1}$$

and this holds for all $x,y \in G_1$. So $\varphi^{-1}:G_1 \to G$ is a homomorphism. As it is one-to-one and onto by Theorem 3.17(1), φ^{-1} is an isomorphism. □

From Example 20.10(b) and Lemma 20.11, we see that

$$G \cong G$$
$$\text{if } G \cong G_1, \text{ then } G_1 \cong G$$
$$\text{if } G \cong G_1 \text{ and } G_1 \cong G_2, \text{ then } G \cong G_2$$

for any groups G,G_1,G_2. We are tempted to say that $\cong$ is an equivalence relation on the set of all groups. It is true indeed that $\cong$ is an equivalence relation, but we must avoid the phrase "the set of all groups". This phrase leads to logical difficulties. For more information about this point, the reader is referred to the appendix.

Since $G \cong G_1$ implies $G_1 \cong G$, it is legitimate to say G and G_1 are isomorphic when G is isomorphic to G_1.

We are not interested in the nature of the elements in a group. The essential thing is the algebraic structure of the group. If $G \cong G_1$, then any algebraic property of G is immediately carried over to G_1. For this reason, we do not distinguish between isomorphic groups. For example, any two cyclic groups of the same order are easily seen to be isomorphic. By abuse of language, we call any cyclic group of order $n \in \mathbb{N}$ *the* cyclic group of order n, and write C_n for it. Likewise, any two dihedral groups of order $2n$ are isomorphic, and we speak of *the* dihedral group of order $2n$, and write D_{2n} for it.

We saw in Theorem 20.6 that the kernel of any homomorphism is a normal subgroup of the domain. We show now conversely that any

normal subgroup of G is the kernel of some homomorphism from G. Into which group? Since we are given only a normal subgroup of G, the only range group that we can construct out of G and its normal subgroup is the factor group with respect to that normal subgroup.

20.12 Theorem: *Let $N \trianglelefteq G$. Then the mapping*

$$\begin{aligned} \nu: G &\to G/N \\ a &\to Na \end{aligned}$$

is a homomorphism. It is onto G/N and $Ker\ \nu = N$.

Proof: ν is a homomorphism, for $(ab)\nu = N(ab) = Na.Nb = a\nu.b\nu$ for all a,b in G, by the very definition of multiplication in G/N. Obviously, any element Na of G/N is the image of $a \in G$ under ν, so ν is onto. Finally

$$\begin{aligned} Ker\ \nu &= \{a \in G: a\nu = \text{identity of } G/N\} \\ &= \{a \in G: a\nu = N1\} \\ &= \{a \in G: Na = N\} \\ &= \{a \in G: a \in N\} \qquad \text{(Lemma 10.2(2))} \\ &= N. \end{aligned}$$

□

20.13 Definition: Let $N \trianglelefteq G$. The mapping $\nu: G \to G/N$ is called the
$a \to Na$
natural (or *canonical*) *homomorphism from G onto G/N.*

20.14 Theorem: *Let $N \trianglelefteq G$. Then there is a homomorphism $\varphi: G \to G_1$ with $Ker\ \varphi = N$.*

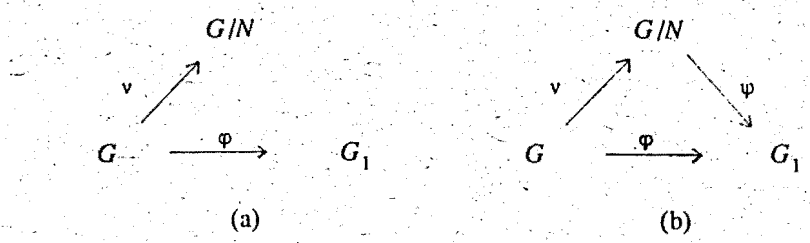
Proof: N is the kernel of the natural homomorphism $\nu: G \to G/N$ by Theorem 20.12. □

The coincidence of kernels with normal subgroups shows that normal subgroups, factor groups and homomorphisms are closely related. Theorem 20.12 describes the relation between $N \trianglelefteq G$, G/N and the natural homomorphism $\nu: G \to G/N$. We prove next that any homomorphism φ is

connected in the same way to $Ker\ \varphi$ and $G/Ker\ \varphi$ as ν is connected to N and G/N. This is done by showing that φ is essentially a natural homomorphism.

20.15 Theorem (Fundamental theorem on homomorphisms): *Let $\varphi: G \to G_1$ be a homomorphism of groups. Let $N = Ker\ \varphi$, which is normal in G by Theorem 20.6, and let $\nu: G \to G/N$ be the associated natural homomorphism.*
Then there is a one-to-one homomorphism $\psi: G/N \to G_1$ such that $\nu\psi = \varphi$.

[This theorem may be summarized in a diagram. The hypothesis is the diagram (a) below. The claim is that there is a one-to-one homomorphism y such that both paths from G to G_1 ($\nu\psi$ and φ) in diagram (b) have the same effect.



The equation $\nu\psi = \varphi$ can be regarded as a factorization of φ. Since the path $\nu\psi$ passes through $G/Ker\ \varphi$, we say φ *factors through* $G/Ker\ \varphi$.]

Proof: We must find a suitable $\psi: G/N \to G_1$. We want $\varphi = \nu\psi$, so that $a\varphi = a(\nu\psi) = (a\nu)\psi = (Na)\psi$. So we define

$$\psi: G/N \to G_1$$
$$Na \to a\varphi$$

In order to find the image of any coset of N under ψ, we have to choose an element a from that coset, which can be done, generally speaking, in many ways. So we have to make sure that ψ is a well defined function. Thus we have to show

$$\text{for all } a,b \in G, \quad Na = Nb \implies (Na)\psi = (Nb)\psi.$$

From the definition of $N = Ker\ \varphi$ and of ψ, we see that this implication is equivalent to

$$\text{for all } a,b \in G, \quad (Ker\ \varphi)a = (Ker\ \varphi)b \implies a\varphi = b\varphi,$$

and this is true by Lemma 20.7. Thus ψ is indeed a well defined mapping.

ψ is a homomorphism. This is verified easily:

$$\begin{aligned} (Na.Nb)\psi &\stackrel{?}{=} (Na)\psi.(Nb)\psi \quad \text{for all } a,b \in G \\ (Nab)\psi &\stackrel{?}{=} (Na)\psi.(Nb)\psi \quad \text{for all } a,b \in G \\ (ab)\varphi &\stackrel{?}{=} a\varphi.b\varphi \quad \text{for all } a,b \in G \end{aligned}$$

Since φ is a homomorphism, the last line is true. Hence ψ is a homomorphism.

ψ is one-to-one. To prove this, we need only show $Ker\ \psi = \{N\}$ (see Theorem 20.8; N is the identity of G/N). We observe

$$\begin{aligned} Ker\ \psi &= \{Na \in G/N: (Na)\psi = 1 = 1_{G_1}\} \\ &= \{Na \in G/N: a\varphi = 1\} \\ &= \{Na \in G/N: a \in Ker\ \varphi\} \\ &= \{Na \in G/N: a \in N\} \\ &= \{N\} \end{aligned}$$

by Lemma 10.2(2) and ψ is one-to-one.

From the definition of ψ, we have $a(\nu\psi) = (a\nu)\psi = (Na)\psi = a\varphi$ for all $a \in G$, so $\nu\psi = \varphi$.

This completes the proof. □

20.16 Theorem: *Let $\varphi: G \to G_1$ be a group homomorphism. Then*

$$G/Ker\ \varphi \cong Im\ \varphi.$$

In more detail: there is an isomorphism $\varphi': G/Ker\ \varphi \to Im\ \varphi$ such that $\nu\varphi'\mu = \varphi$, where $\nu: G \to G/Ker\ \varphi$ is the natural homomorphism and $\mu: Im\ \varphi \to G_1$ is the inclusion homomorphism (Example 20.2(f)). This means that the diagram

$$\begin{array}{ccc} G & \xrightarrow{\varphi} & G_1 \\ \nu\downarrow & & \uparrow\mu \\ G/Ker\,\varphi & & Im\,\varphi \end{array}$$

can be so completed with a homomorphism φ' *that both paths* $\nu\varphi'\mu$ *and* φ

$$\begin{array}{ccc} G & \xrightarrow{\varphi} & G_1 \\ \nu\downarrow & & \uparrow\mu \\ G/Ker\,\varphi & \xrightarrow{\varphi'} & Im\,\varphi \end{array}$$

have the same effect.

Proof: We use the homomorphism $\psi: G/N \to G_1$ of Theorem 20.15, where $N = Ker\,\varphi$. Obviously,

$$Im\,\psi = \{(Na)\psi: Na \in G/N\} = \{a\varphi: a \in G\} = Im\,\varphi$$

and since ψ is one-to-one, ψ is an isomorphism from $G/Ker\,\varphi$ onto $Im\,\psi = Im\,\varphi$. We observe

$$a(\nu\psi\mu) = (a\nu)(\psi\mu) = (Na)(\psi\mu) = ((Na)\psi)\mu = (a\varphi)\mu = a\varphi$$

for all $a \in G$, as μ maps any element of $Im\,\varphi$ to itself. Hence $\nu\psi\mu = \varphi$. The theorem follows when we write φ' in place of ψ. □

According to Theorem 20.16, any homomorphism $\varphi: G \to G_1$ is factored into three homomorphisms ν,φ',μ:

$$G \xrightarrow{\nu} G/Ker\,\varphi \xrightarrow{\varphi'} Im\,\varphi \xrightarrow{\mu} G_1$$

where (a) ν is onto $G/Ker\,\varphi$; (b) φ' is one-to-one and onto $Im\,\varphi$ and (c) μ is one-to-one. So $\nu\varphi'$ is onto and $\varphi'\mu$ is one-to-one (Theorem 3.11). Hence, if φ fails to be onto, it is only due to the fact that μ is not onto. Also, if φ fails to be one-to-one, it is only due to the fact that ν is not one-to-one. We see that any homomorphism φ is essentially an isomorphism φ', "diluted" by a natural homomorphism which (eventualy) accounts for its failure to be one-to-one and by an inclusion mapping which (eventualy) accounts for its failure to be onto. In fact, φ is one-to-one if and only if the associated natural homomorphism $\nu: G \to G/Ker\,\varphi$ is one-to-one and φ is onto if and only if the associated inclusion mapping $\mu: Im\,\varphi \to G_1$ is onto.

20.17 Examples: (a) Let $\langle a\rangle = \{a^n : n \in \mathbb{Z}\}$ be a cyclic group. The mapping

$$\begin{aligned} \psi: \mathbb{Z} &\to \langle a\rangle \\ n &\to a^n \end{aligned}$$

is a homomorphism from the additive group $\mathbb{Z}$ into $\langle a\rangle$, because

$$(m+n)\psi = a^{m+n} = a^m a^n = m\psi . n\psi$$

for all $m,n \in \mathbb{Z}$. From Theorem 20.16, we obtain

$$\mathbb{Z}/Ker\ \psi \cong Im\ \psi .$$

The homomorphism ψ is onto by definition of $\langle a\rangle$, hence $Im\ \psi = \langle a\rangle$ and

$$\mathbb{Z}/Ker\ \psi \cong \langle a\rangle .$$

We see that any cyclic group is isomorphic to a factor group of $\mathbb{Z}$. In order to get more information, we distinguish two cases, where $\langle a\rangle$ has finite or infinite order.

First suppose that $\langle a\rangle$ has finite order $k \in \mathbb{N}$. Then $o(a) = k$ and

$$\begin{aligned} Ker\ \psi &= \{n \in \mathbb{Z} : n\psi = 1 = a^0\} \\ &= \{n \in \mathbb{Z} : a^n = 1\} \\ &= \{n \in \mathbb{Z} : o(a) \mid n\} \qquad \text{(Lemma 11.6)} \\ &= k\mathbb{Z}. \end{aligned}$$

Thus $\mathbb{Z}/k\mathbb{Z} \cong \langle a\rangle$. We see that any cyclic group of order k is isomorphic to $\mathbb{Z}/k\mathbb{Z}$. Consequently, any two cyclic groups of order k are isomorphic to each other. For this reason, we speak of *the* cyclic group of order k.

In the second case, suppose $\langle a\rangle$ has infinite order. Then

$$\begin{aligned} Ker\ \psi &= \{n \in \mathbb{Z} : n\psi = 1 = a^0\} \\ &= \{n \in \mathbb{Z} : a^n = 1\} \\ &= \{0\} \qquad \text{(Lemma 11.5)} \end{aligned}$$

and so $\mathbb{Z}/\{0\} \cong \langle a\rangle$. From $\mathbb{Z}/\{0\} \cong \mathbb{Z}$ (Example 20.10(c)), we infer $\mathbb{Z} \cong \langle a\rangle$. We see that any infinite cyclic group is isomorphic to $\mathbb{Z}$. Consequently, any two cyclic groups of infinite order are isomorphic. For this reason, we speak of *the* infinite cyclic group.

(b) The determinant homomorphism (Example 20.2(b))

$$det: GL(2,\mathbb{Q}) \to \mathbb{Q}\backslash\{0\}$$

is onto $\mathbb{Q}\backslash\{0\}$, because any $a \in \mathbb{Q}\backslash\{0\}$ is the determinant of $\begin{pmatrix} a & 0 \\ 0 & 1 \end{pmatrix}$ in $GL(2,\mathbb{Q})$. Thus $Im\ det = \mathbb{Q}\backslash\{0\}$. Also

$$ker\ det = \{A \in GL(2,\mathbb{Q}): det\ A = 1\} = SL(2,\mathbb{Q}).$$

From Theorem 20.16, we obtain

$$GL(2,\mathbb{Q})/SL(2,\mathbb{Q}) \cong \mathbb{Q}\backslash\{0\}.$$

In the same way, $GL(2,K)/SL(2,K) \cong K\backslash\{0\}$

for any field K.

(c) The sign homomorphism

$$\varepsilon: S_n \to \{1,-1\} = C_2$$

is onto C_2 when $n \geqslant 2$, because $\varepsilon(\iota) = 1$ and $\varepsilon((12)) = -1$. Hence $Im\ \varepsilon = C_2$. As $Ker\ \varepsilon = \{\sigma \in S_n: \varepsilon(\sigma) = 1\} = A_n$ by definition, the relation

$$S_n/Ker\ \varepsilon \cong Im\ \varepsilon$$

yields $$S_n/A_n \cong C_2.$$

(d) Consider the absolute value homomorphism

$$\begin{aligned} \varphi: \mathbb{R}\backslash\{0\} &\to \mathbb{R}^+ \\ a &\to |a| \end{aligned}$$

Here $Im\ \varphi = \{|a|: a \in \mathbb{R}\backslash\{0\}\} = \mathbb{R}^+$ and $Ker\ \varphi = \{a \in \mathbb{R}\backslash\{0\}: |a| = 1\} = \{1,-1\} = C_2$. Thus

$$S_n/Ker\ \varphi \cong Im\ \varphi$$

gives $$(\mathbb{R}\backslash\{0\})/C_2 \cong \mathbb{R}^+.$$

(e) The mapping

$$\begin{aligned} \varphi: \mathbb{R} &\to \mathbb{C}\backslash\{0\} \\ x &\to e^{2\pi x i} \end{aligned}$$

is a homomorphism from the additive group $\mathbb{R}$ into $\mathbb{C}\backslash\{0\}$:

$$(x + y)\varphi = e^{2\pi(x+y)i} = e^{2\pi xi}e^{2\pi yi} = x\varphi . y\varphi$$

for all $x,y \in \mathbb{R}$. We have $\mathbb{R}/Ker\ \varphi \cong Im\ \varphi$. The reader may verify that

$$Im\ \varphi = \{z \in \mathbb{C}: |z| = 1\}.$$

As for the kernel,
$$\begin{aligned} Ker\ \varphi &= \{x \in \mathbb{R}: e^{2\pi xi} = 1\} \\ &= \{x \in \mathbb{R}: \cos 2\pi x + i\sin 2\pi x = 1\} \\ &= \{x \in \mathbb{R}: \cos 2\pi x = 1, \sin 2\pi x = 0\} \\ &= \mathbb{Z}. \end{aligned}$$

Thus

$$\mathbb{R}/\mathbb{Z} \cong \{z \in \mathbb{C}: |z| = 1\},$$

where $\mathbb{R}/\mathbb{Z}$ is an additive, the right hand side is a multiplicative group.

Exercises

1. Show that the mapping $exp: \mathbb{R} \to \mathbb{R}^+$ $x \to e^x$ is an isomorphism.

2. Determine whether the mapping $x \to log(log\ x)$ is a homomorphism.

3. Find an isomorphism from $\mathbb{Q}\backslash\{0\}$ under multiplication onto the group of Example 7.4(a).

4. Find an isomorphism from $\mathbb{Z}$ under addition onto the group of Example 7.4(b).

5. Let $\varphi_i: G_i \to G_{i+1}$ be homomorphisms of groups, where $i = 1,2, \ldots,n$. Show that $\varphi_1\varphi_2 \ldots \varphi_n$ is a homomorphism from G_1 onto G_{n+1}. Prove a corresponding result for isomorphisms.

6. Let $\varphi: G \to G_1$ be an isomorphism. Prove that $o(a) = o(a\varphi)$ for all $a \in G$.

7. Let $\varphi: G \to G_1$ be a homomorphism. Show that $a\varphi = b\varphi$ if and only if $a(Ker\ \varphi) = b(Ker\ \varphi)$, where a,b are arbitrary elements of G.

8. Prove directly that any two cyclic groups of the same order are isomorphic.

9. Prove that any two dihedral groups of the same order are isomorphic.

10. Imitating Example 20.17(a), show that any dihedral group is isomorphic to a factor group of D_∞.

11. Show that a factor group of a dihedral group is either dihedral or cyclic.

12. Let $n \in \mathbb{N}$ and let X be a set with n elements. Prove that $S_X \cong S_n$.

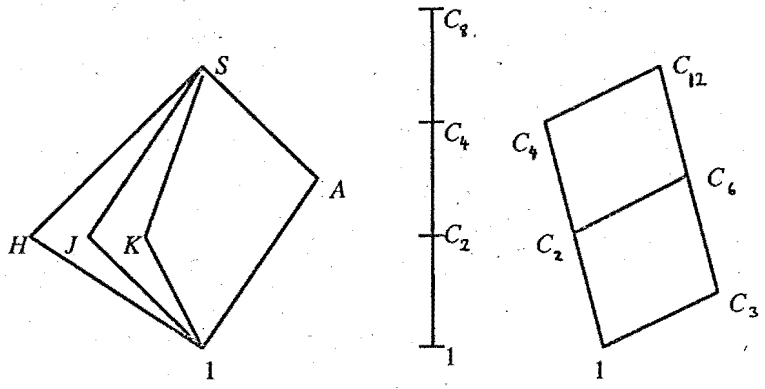
13. Prove that $(\mathbb{R}\backslash\{0\})/\mathbb{R}^+ \cong C_2$.

14. Let $\varphi: G \to G_1$ be a homomorphism, let K be a normal subgroup of G such that $K \leqslant Ker\,\varphi$, and let $\nu: G \to G/K$ be the associated natural homomorphism. Show that there is a homomorphism $\psi: G/K \to G_1$ such that $\nu\psi = \varphi$ and $Ker\,\psi = (Ker\,\varphi)/K$. What happens when we drop the condition $K \leqslant Ker\,\varphi$?

§21
Isomorphism Theorems

This paragraph is devoted to some very important theorems of group theory.

At this stage, it will be useful to introduce Hasse diagrams. A Hasse diagram is a convenient means to visualise inclusions holding between subgroups of a group. Subgroups are represented by points. Two points (subgroups) are joined by a line segment if and only if the lower subgroup is contained in the upper one. The line segments may be vertical or slate. The line segments may be thought of as factor groups when the lower subgroup is normal in the upper one. The Hasse diagrams of S_3, C_8 and C_{12} are depicted below.



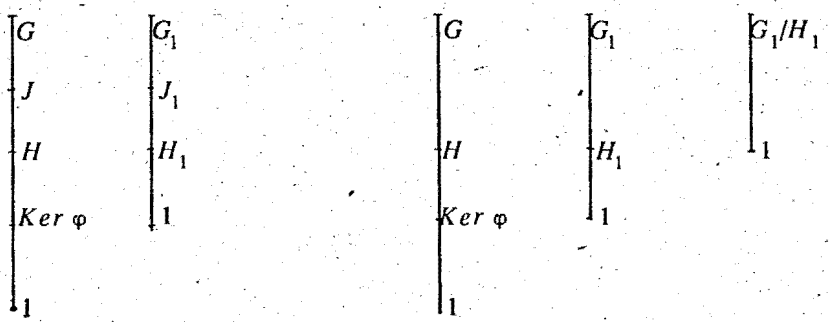
$H = \{\iota,(12)\}, J = \{\iota,(13)\}, K = \{\iota,(23)\}$.

21.1 Theorem: *Let $\varphi: G \to G_1$ be a group homomorphism from G onto G_1. Then there is a one-to-one correspondence between the set of all subgroups of G that contain $Ker\ \varphi$ and the set of all subgroups of G_1. This correspondence preserves inclusion. The normal subgroups of G that contain $Ker\ \varphi$ correspond to normal subgroups of G_1, and conversely. The factor groups by corresponding normal subgroups are isomorphic.*

In more detail and more precise language, the claim is the following.

(1) *For every* $H \leqslant G$ *with* $Ker\,\varphi \leqslant H$, *there is associated a unique subgroup of* G_1, *which will be denoted by* H_1.
(2) *If* $Ker\,\varphi \leqslant H \leqslant J \leqslant G$, *then* $H_1 \leqslant J_1$.
(3) *If* $Ker\,\varphi \leqslant H \leqslant G$, $Ker\,\varphi \leqslant J \leqslant G$ *and* $H_1 \leqslant J_1$, *then* $H \leqslant J$.
(4) *If* $Ker\,\varphi \leqslant H \leqslant G$, $Ker\,\varphi \leqslant J \leqslant G$ *and* $H_1 = J_1$, *then* $H = J$.
(5) *If* S *is any subgroup of* G_1, *then there is an* $H \leqslant G$ *such that* $Ker\,\varphi \leqslant H$ *and* $H_1 = S$.
(6) *For* $Ker\,\varphi \leqslant H \leqslant G$, *there holds* $H \trianglelefteq G$ *if and only if* $H_1 \trianglelefteq G_1$.
(7) *If* $Ker\,\varphi \leqslant H \trianglelefteq G$ *and* $H_1 \trianglelefteq G_1$, *then* $G/H \cong G_1/H_1$.

The situation is described in the accompanying diagrams.



Proof: (1) For each $H \leqslant G$ with $Ker\,\varphi \leqslant H$, we are to find a subgroup of G_1. How can we find it? Well, the subgroup we are looking for will be first of all a *subset* of G_1. How can we associate with H a subset of G_1? At our disposal, we have only one means of transportation from G to G_1, namely the mapping φ. The only thing we can do, then, is form the set of images of the elements of H under φ. Hence we put

$$H_1 := \{h\varphi \in G_1 : h \in H\}.$$

We now prove $H_1 \leqslant G_1$. We can do it by the subgroup criterion, but we prefer to use Theorem 20.6, which states that the image of a homomorphism is a subgroup of its range. We note that the restriction of φ to H is a homomorphism (Example 20.2(g)) and

$$H_1 = \{h\varphi \in G_1 : h \in H\} = \{h\varphi_{|H} \in G_1 : h \in H\} = Im\,\varphi_{|H}$$

by definition. Theorem 20.6 gives now $Im\ \varphi_H \leqslant G_1$, hence $H_1 \leqslant G_1$. The description $H_1 = Im\ \varphi_H$ will be useful.

(2) Suppose $Ker\ \varphi \leqslant H \leqslant J \leqslant G$. Under this assumption, we prove $H_1 \leqslant J_1$. This is easy: for any $h \in H$, we have $h \in J$, so $h\varphi \in Im\ \varphi_J = J_1$. Since $h\varphi \in J_1$ for all $h\varphi \in H_1$, we get $H_1 \leqslant J_1$.

(3) Suppose $Ker\ \varphi \leqslant H \leqslant G$, $Ker\ \varphi \leqslant J \leqslant G$ and $H_1 \leqslant J_1$. We want to prove $H \leqslant J$. Now $H_1 \leqslant J_1$ means $Im\ \varphi_H \leqslant Im\ \varphi_J$. Then, for every $h \in H$, we have $h\varphi \in Im\ \varphi_J$:

$$\text{for every } h \in H, \text{ there is a } j \in J \text{ such that } h\varphi = j\varphi.$$

We obtain, when h,j are as above

$$1 = h\varphi(j\varphi)^{-1} = h\varphi.j^{-1}\varphi = (hj^{-1})\varphi,$$
$$hj^{-1} \in Ker\ \varphi \leqslant J$$
$$h \in Jj = J.$$

Thus $h \in J$ for all $h \in H$. Therefore $H \leqslant J$.

(4) This is immediate from (3). If $H_1 = J_1$, we have $H_1 \leqslant J_1$ and $J_1 \leqslant H_1$, so $H \leqslant J$ and $J \leqslant H$ by (3), hence $H = J$. (This shows that the correspondence $H \to H_1$ is one-to-one.)

(5) For any $S \leqslant G_1$, we are to find an $H \leqslant G$ such that $Ker\ \varphi \leqslant H$ and $H_1 = S$. What can H be? As in part (1), there is only one thing we can do: take the preimages of the elements in S. Hence we put

$$H = \{a \in G: a\varphi \in S\}.$$

Thus $a \in H$ means $a\varphi \in S$. We show that $H \leqslant G$, that $Ker\ \varphi \leqslant H$ and that $H_1 = S$.

First $H \leqslant G$. From $1_G\varphi = 1_{G_1} \in S$ (Lemma 20.3(1)), we get $1 \in H$. So $H \neq \varnothing$. We apply the subgroup criterion.

(i) If $a,b \in H$, then $a\varphi, b\varphi \in S$, so $a\varphi b\varphi \in S$, so $(ab)\varphi \in S$, so $ab \in H$. Thus H is closed under multiplication.

(ii) If $a \in H$, then $a\varphi \in S$, then $(a\varphi)^{-1} \in S$, then $(a^{-1})\varphi \in S$, then $a^{-1} \in H$. Thus H is closed under the forming of inverses.

Thus H is a subgroup of G.

We prove next that H contains $Ker\ \varphi$. This is trivial. If $a \in Ker\ \varphi$, then $a\varphi = 1$, so $a\varphi \in S$, so $a \in H$. Hence $Ker\ \varphi \leqslant H$.

It remains to prove $H_1 = S$. We have

$$H_1 = Im\ \varphi_H = \{h\varphi \in G_1 : h \in H\} = \{h\varphi \in G_1 : h\varphi \in S\} = S$$

as claimed.

This completes the proof of (5). (Part (5) shows that the correspondence $H \to H_1$ is onto.)

(6) First we assume $H \trianglelefteq G$ and show that $H_1 \trianglelefteq G_1$. We are to show that $x^{-1}h_1x \in H_1$ for all $x \in G_1$ and for all $h_1 \in H_1$ (Lemma 18.2(1)). If $x \in G_1$ and $h_1 \in H_1$, then there are $a \in G$ with $a\varphi = x$ and $h \in H$ with $h\varphi = h_1$. This is so because φ is *onto* G_1 and H_1 is defined as $Im\ \varphi_H$. Then we are to show $(a\varphi)^{-1}(h\varphi)(a\varphi) \in H_1$. This is equivalent to $(a^{-1}ha)\varphi \in H_1$. Since $H \trianglelefteq G$, we know $a^{-1}ha \in H$, so $(a^{-1}ha)\varphi \in Im\ \varphi_H = H_1$. This proves $H_1 \trianglelefteq G_1$.

We assume now $H_1 \trianglelefteq G_1$ and prove $H \trianglelefteq G$. We can give an argument similar to the one above, but we prefer to use the fact that normal subgroups and kernels coincide. Our method will be used in the proof of part (7) as well.

The assumption is $H_1 \trianglelefteq G_1$. By Theorem 20.12, $H_1 = Ker\ \nu'$, where

$$\nu' : G_1 \to G_1/H_1$$

is the natural homomorphism. We get then the homomorphism

$$\varphi\nu' : G \to G_1/H_1 \qquad \text{(Theorem 20.4):}$$

$$G \xrightarrow{\varphi} G_1 \xrightarrow{\nu'} G_1/H_1. \tag{i}$$

We have

$$\begin{aligned} Ker\ \varphi\nu' &= \{a \in G : a(\varphi\nu') = H_1\} \\ &= \{a \in G : a\varphi \in Ker\ \nu'\} \\ &= \{a \in G : a\varphi \in H_1\} \end{aligned}$$

So $(Ker\ \varphi\nu')_1 = Im\ \varphi_{Ker\ \varphi\nu'} = \{a\varphi \in G_1 : a \in Ker\ \varphi\nu'\} = \{a\varphi \in G_1 : a\varphi \in H_1\} = H_1$ and we obtain

$$Ker\ \varphi\nu' = H \tag{ii}$$

by part (4). Theorem 20.6 gives $H \trianglelefteq G$, as was to be proved.

(7) We saw that any one of $H \trianglelefteq G$ and $H_1 \trianglelefteq G_1$ implies the other. Assume that one, and hence both of them are true. Then we have the homomorphism $\varphi\nu'$. From Theorem 20.16, we obtain

$$G/Ker\ \varphi\nu' \cong Im\ \varphi\nu' \qquad \text{(iii)}$$

We know $Ker\ \varphi\nu' = H$ by (ii). As for the image, since φ is onto G_1 by hypothesis and ν' is onto G_1/H_1 by Theorem 20.12, the composition $\varphi\nu'$ is onto by Theorem 3.11(1). Hence $Im\ \varphi\nu' = G_1/H_1$ and (iii) becomes

$$G/H \cong G_1/H_1.$$

The proof is complete. □

An important special case of Theorem 21.1 is the case of a natural homomorphism, recorded in the next theorem. It gives a complete description of the subgroups of a factor group. The last part of the theorem is known as the factor of a factor theorem.

21.2 Theorem: *Let $N \trianglelefteq G$. The subgroups of G/N are the factor groups S/N, where S runs through the subgroups of G satisfying $N \leqslant S$. More precisely, for each subgroup X of G/N, there is a unique subgroup S of G satisfying $N \leqslant S$ such that $X = G/N$. When X_1 and X_2 are subgroups of G/N, say $X_1 = S_1/N$ and $X_2 = S_2/N$, where $N \leqslant S_1 \leqslant G$ and $N \leqslant S_2 \leqslant G$, then $X_1 \leqslant X_2$ if and only if $S_1 \leqslant S_2$. Furthermore, $S/N \trianglelefteq G/N$ if and only if $S \trianglelefteq G$. In this case, there holds*

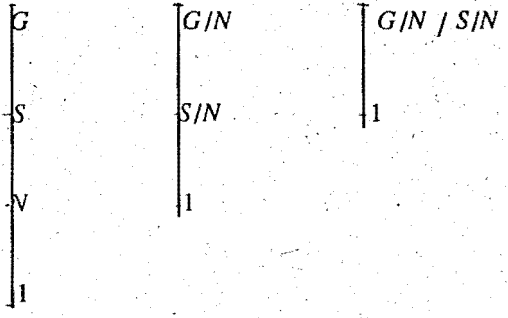
$$G/N \,/\, S/N \cong G/S.$$

Proof: Since $N \trianglelefteq G$, we can build the factor group G/N. The natural homomorphism $\nu: G \to G/N$ is onto by Theorem 20.12. We can therefore apply Theorem 21.1.

Theorem 21.1 states that any subgroup of G/N is of the form $Im\ \nu_S$ for some $S \leqslant G$ with $Ker\ \nu \leqslant S$ (here ν_S is the restriction of ν to S). Now

$$\begin{aligned} Im\ \nu_S &= \{s\nu \in G/N : s \in S\} \\ &= \{Ns \in G/N : s \in S\} = S/N \end{aligned}$$

and $Ker\ v = N$ by Theorem 20.12 (notice that S/N is meaningful, for $N \trianglelefteq G$ and $N \leq S$ imply $N \trianglelefteq S$; cf. Example 18.5(l)). Thus the subgroups of G/N are given by S/N, where $N \leq S \leq G$. By Theorem 21.1(2),(3),(4), $S_1/N \leq S_2/N$ if and only if $S_1 \leq S_2$ and $S_1/N \neq S_2/N$ whenever $S_1 \neq S_2$. Finally, $S_1/N \trianglelefteq G/N$ if and only if $S \trianglelefteq G$ by Theorem 21.1(6) and in this case $G/N \,/\, S/N \cong G/S$ by Theorem 21.1(7).This completes the proof. □



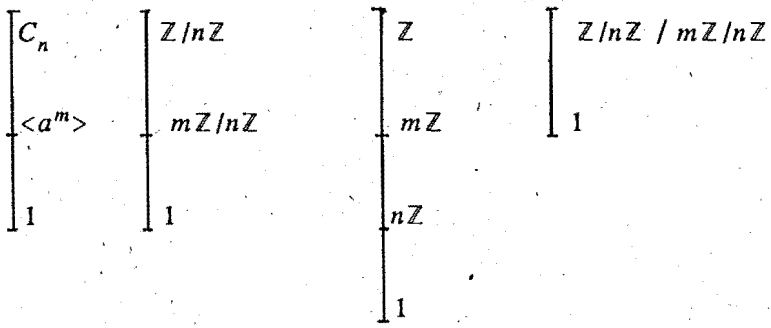
As an application of Theorem 21.2, we classify the factor groups of cyclic groups. We treat infinite and finite cyclic groups separately.

Any infinite cyclic group is isomorphic to $\mathbb{Z}$ under addition (Example 20.17(a)), so we need find the factor groups of $\mathbb{Z}$. As $\mathbb{Z}$ is abelian, any subgroup of $\mathbb{Z}$ is normal in $\mathbb{Z}$ (Example 18.5(b)) and we can build factor groups of $\mathbb{Z}$ by any subgroup of $\mathbb{Z}$. The subgroups of $\mathbb{Z}$ are 0 (see Example 18.5(a)) and $n\mathbb{Z}$, where $n \in \mathbb{N}$ (Theorem 11.8). For each $n \in \mathbb{N}$, the subgroup $n\mathbb{Z}$ is the unique subgroup of index n (Lemma 11.11). The factor group $\mathbb{Z}/0$ is isomorphic to $\mathbb{Z}$ (Example 20.10(c)). The factor groups $\mathbb{Z}/n\mathbb{Z}$ are known to be cyclic of order n (Example 20.17(a)). So all factor groups of $\mathbb{Z}$ are cyclic (cf. Lemma 18.9(3)). For each $m \in \mathbb{N} \cup \{\infty\}$, there is a unique factor group of order m of $\mathbb{Z}$, namely $\mathbb{Z}/m\mathbb{Z}$ if $m \in \mathbb{N}$ and $\mathbb{Z}/0 \cong \mathbb{Z}$ if $m = \infty$.
Now let $C_n = \langle a \rangle$ be a finite cyclic group of order $n \in \mathbb{N}$. As C_n is abelian, we can build factor groups of C_n by any subgroup of C_n. We know that $C_n \cong \mathbb{Z}/n\mathbb{Z}$ from Example 20.17(a). The subgroups of $\mathbb{Z}/n\mathbb{Z}$ are described in Theorem 21.2: any subgroup of $\mathbb{Z}/n\mathbb{Z}$ is of the form $M/n\mathbb{Z}$, where $n\mathbb{Z} \leq M \leq \mathbb{Z}$. Now $M \leq \mathbb{Z}$ means $M = m\mathbb{Z}$ for some $m \in \mathbb{N}$ or $M = 0$ (Theorem 11.8) and the condition $n\mathbb{Z} \leq M$ excludes $M = 0$. Hence $M = m\mathbb{Z}$ for some $m \in \mathbb{N}$, where furthermore $m|n$, because $n\mathbb{Z} \leq m\mathbb{Z}$. So the sub-

groups of $\mathbb{Z}/n\mathbb{Z}$ are given by $m\mathbb{Z}/n\mathbb{Z}$, where m runs through all positive divisors of n. For the factor group, we know

$$\mathbb{Z}/n\mathbb{Z}\,/\,m\mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z}$$

from Theorem 21.2. So all factor groups of $\mathbb{Z}/n\mathbb{Z}$ and of C_n are cyclic (cf. Lemma 18.9(3)). For each positive divisor m of n, there is a unique factor group of order m of $C_n = \langle a \rangle$, namely $\langle a \rangle / \langle a^m \rangle$, where $\langle a^m \rangle$ is the unique subgroup of order n/m of C_n (Lemma 11.10).



We end this paragraph with another important theorem of group theory.

21.3 Theorem: *Let $H \trianglelefteq G$ and $K \leqslant G$. Then $H \cap K \trianglelefteq K$ and*

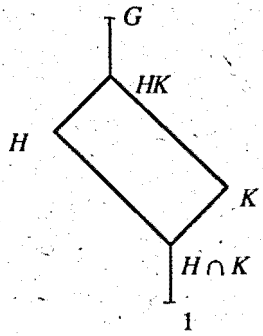
$$K/H \cap K \cong HK/H. \qquad (*)$$

Proof: Since $H \trianglelefteq G$, there is a group G/H and a homomorphism

$$\nu: G \to G/H$$

Let ν_K be the restriction of ν to K. This ν_K is a homomorphism (Example 20.2(g)). Hence $K/Ker\ \nu_K \cong Im\ \nu_K$ by Theorem 20.16. Here

$$\begin{aligned} Ker\ \nu_K &= \{k \in K: k\nu = 1\} \\ &= \{k \in K: k \in Ker\ \nu\} \\ &= K \cap Ker\ \nu \\ &= K \cap H. \end{aligned}$$



It remains to find $Im\ v_K$. We claim $Im\ v_K = HK/H$. First of all, $HK = KH$ is a subgroup of G because $H \trianglelefteq G$ (Lemma 19.4(2)), and $H \leqslant HK$, so $H \trianglelefteq HK$. So HK/H is meaningful. For any $k \in K$, we have $kv_K = Hk \in HK/H$, which shows that $Im\ v_K \subseteq HK/H$. Conversely, each element of HK/H is of the form Hhk, where $h \in H$, $k \in K$. But $Hhk = Hk = kv_K \in Im\ v_K$, so $HK/H \subseteq Im\ v_K$. Thus $Im\ v_K = HK/H$ and (*) yields

$$K/H \cap K \cong HK/H$$

as was to be proved. □

Exercises

1. Let $A \trianglelefteq C \leqslant G$ and $B \leqslant G$. Prove that $A \cap B \trianglelefteq C \cap B$ and $C \cap B / A \cap B \cong A(C \cap B)/A$.

2. Let $A \trianglelefteq C \leqslant G, B \leqslant G$ and let $\varphi: G \to H$ be a group homomorphism. Prove that $A\varphi \trianglelefteq C\varphi$. Choosing φ in particular to be the natural homomorphism $v: G \to G/B$, prove that $AB \trianglelefteq CB$.

§22
Direct Products

In this paragraph, we learn a method of constructing new groups from given ones. This method consists essentially in writing the groups one adjacent to the other.

22.1 Theorem: *Let H and K be groups. On the cartesian product $H \times K$, we define a binary operation by declaring*

$$(h,k)(h_1,k_1) = (hh_1,kk_1)$$

for all $(h,k),(h_1,k_1) \in H \times K$. With respect to this operation, $H \times K$ is a group.

Proof: Before beginning with the proof, it will not be amiss to formulate the theorem in a more precise way. Suppose $(H, \circ)$ and $(K, *)$ are groups. The claim is that $(H \times K, \Delta)$ is a group, where Δ is defined by

$$(h,k) \Delta (h_1,k_1) = (h \circ h_1, k * k_1)$$

for all $(h,k),(h_1,k_1) \in H \times K$.

The multiplication in $H \times K$ is carried out componentwise. Since H and K are groups themselves, it is natural to expect that $H \times K$ will be a group. We check the group axioms.

(i) For all $(h,k),(h_1,k_1) \in H \times K$, we have $h,h_1 \in H$, $k,k_1 \in K$, so $hh_1 \in H$ and $kk_1 \in K$ as H and K are closed under multiplication, and so $(hh_1,kk_1) \in H \times K$. So we have a binary operation on $H \times K$. In other words, $H \times K$ is closed under multiplication.

(ii) Associativity in $H \times K$ follows from associativity in H and K. For any $(h,k),(h_1,k_1),(h_2,k_2) \in H \times K$, we have

$$\begin{aligned}[(h,k)(h_1,k_1)](h_2,k_2) &= (hh_1,kk_1)(h_2,k_2) \\ &= ((hh_1)h_2,(kk_1)k_2) \\ &= (h(h_1h_2),k(k_1k_2))\end{aligned}$$

$$= (h,k)(h_1h_2,k_1k_2)$$
$$= (h,k)[(h_1,k_1)(h_2,k_2)]$$

and the operation on $H \times K$ is associative.

(iii) What can be the identity element of $H \times K$? The only reasonable guess would be $(1,1) = (1_H,1_K)$. We indeed have

$$(h,k)(1,1) = (h1,k1) = (h,k)$$

for all $(h,k) \in H \times K$. Thus $(1,1)$ is a right identity of $H \times K$.

(iv) What can be the inverse of $(h,k) \in H \times K$? Probably (h^{-1},k^{-1}). We indeed have

$$(h,k)(h^{-1},k^{-1}) = (hh^{-1},kk^{-1}) = (1,1)$$

for all $(h,k) \in H \times K$. So any $(h,k) \in H \times K$ has a right inverse in $H \times K$, namely $(h,k)^{-1} = (h^{-1},k^{-1})$.

Therefore, $H \times K$ is a group. □

22.2 Definition: Let H and K be groups. Then the group of Theorem 22.1 is called the *direct product of H and K*. It will be denoted by $H \times K$.

Thus the notation "$H \times K$" stands for the cartesian product of the sets H and K as well as the direct product of the groups H and K. This ambiguity will not lead to any confusion. The reader should be careful to distinguish between HK and $H \times K$. The former is defined only when H and K are subgroups of a common group G, whereas $H \times K$ is a meaningful group regardless of whether H and K are subgroups of a group. The elements of HK are elements of the group that contains H and K; the elements of $H \times K$ ore ordered pairs.

When the groups H and K are written additively, we write the group of Theorem 22.1 in the additive form, too. The operation is then given by

$$(h,k) + (h_1,k_1) = (h + h_1,k + k_1)$$

for all $(h,k),(h_1,k_1) \in H \times K$. The operation is called *addition* in this case, and the group is called the *direct sum of H and K*. We write the group as $H \oplus K$, to avoid confusion with $H + K$ (which is HK in additive notation, where H and K are subgroups of a group G).

22.3 Examples: (a) Consider $C_2 \times \mathbb{Q}^+$, where $C_2 = \{1,-1\} \subseteq \mathbb{R}$ and $\mathbb{Q}^+$ is the multiplicative group of the positive rational numbers. The elements of $C_2 \times \mathbb{Q}^+$ are ordered pairs $(\mp 1,q)$, where $q \in \mathbb{Q}^+$. Multiplication in $C_2 \times \mathbb{Q}^+$ is carried out according to the rule $(\varepsilon,q)(\varepsilon',q') = (\varepsilon\varepsilon',qq')$. We observe that the mapping

$$\begin{aligned} \varphi: \mathbb{Q}\backslash\{0\} &\to C_2 \times \mathbb{Q}^+ \\ q &\to (\operatorname{sgn} q, |q|) \end{aligned}$$

is a homomorphism, since $\begin{aligned}(qq')\varphi &= (\operatorname{sgn} qq', |qq'|) \\ &= (\operatorname{sgn} q.\operatorname{sgn} q', |q||q'|) \\ &= (\operatorname{sgn} q, |q|)(\operatorname{sgn} q', |q'|) \\ &= (q\varphi)(q'\varphi)\end{aligned}$

for all $q,q' \in \mathbb{Q}\backslash\{0\}$. Its kernel is

$$\begin{aligned} Ker\,\varphi = \{q \in \mathbb{Q}\backslash\{0\}: q\varphi = (1,1)\} &= \{q \in \mathbb{Q}\backslash\{0\}: \operatorname{sgn} q = 1, |q| = 1\} \\ &= \{q \in \mathbb{Q}\backslash\{0\}: q > 0, |q| = 1\} \\ &= \{1\}, \end{aligned}$$

which means that φ is one-to-one (Theorem 20.8). As any $(\varepsilon,q) \in C_2 \times \mathbb{Q}^+$ is the image of $\varepsilon|q| \in \mathbb{Q}\backslash\{0\}$, the homomorphism φ is onto. Hence φ is an isomorphism and

$$\mathbb{Q}\backslash\{0\} \cong C_2 \times \mathbb{Q}^+.$$

(b) Consider $\mathbb{R} \oplus \mathbb{R}$, where $\mathbb{R}$ is the additive group of real numbers. The elements of $\mathbb{R} \oplus \mathbb{R}$ are ordered pairs of real numbers. The operation on $\mathbb{R} \oplus \mathbb{R}$ is given by

$$(a,b) + (c,d) = (a + c, b + d)$$

for all $(a,b),(c,d) \in \mathbb{R} \oplus \mathbb{R}$. We leave it to the reader to prove that

$$\begin{aligned} \psi: \mathbb{C} &\to \mathbb{R} \oplus \mathbb{R} \\ a + bi &\to (a,b) \end{aligned}$$

is an isomorphism (where $\mathbb{C}$ is the group of complex numbers under addition). Hence

$$\mathbb{C} \cong \mathbb{R} \oplus \mathbb{R}.$$

22.4 Theorem: *Let H and K be groups and let $G := H \times K$ be the direct product of H and K. Then there are subgroups H_1 and K_1 of G such that*

$$\begin{aligned} H_1 &\cong H, & K_1 &\cong K \\ H_1 &\trianglelefteq G, & K_1 &\trianglelefteq G \\ H_1K_1 &= G, & H_1 \cap K_1 &= 1 \end{aligned}$$

Proof: We put $H_1 = \{(h,1) \in G: h \in H\}$ and $K_1 = \{(1,k) \in G: k \in K\}$. First we prove $H_1,K_1 \leqslant G$. Since

(i) $(h,1)(h',1) = (hh',1) \in H_1$ for all $(h,1),(h',1) \in H_1$ and

(ii) $(h,1)^{-1} = (h^{-1},1^{-1}) = (h^{-1},1) \in H_1$ for all $(h,1) \in H_1$,

H_1 is a subgroup of G. In the same way, $K_1 \leqslant G$.

H_1 and K_1 are in fact normal subgroups of G. To establish $K_1 \trianglelefteq G$, we show that $(h,k)^{-1}(1,k_0)(h,k) \in K_1$ for all $(h,k) \in G$, $(1,k_0) \in K_1$ (Lemma 18.2(1)). We indeed have

$$(h,k)^{-1}(1,k_0)(h,k) = (h^{-1},k^{-1})(1,k_0)(h,k) = (h^{-1}1h,k^{-1}k_0k) = (1,k^{-1}k_0k) \in K_1$$

as K is closed under multiplication. Hence $K_1 \trianglelefteq G$. One proves similarly $H_1 \trianglelefteq G$.

Next we show $H \cong H_1$ and $K \cong K_1$. The mapping $\mu_1: H \to H_1, h \to (h,1)$ is one-to-one (by the definition of equality of ordered pairs) and onto (by the definition of H_1), and is furthermore a homomorphism, since

$$(hh')\mu_1 = (hh',1) = (h,1)(h',1) = h\mu_1.h'\mu_1$$

for all $h,h' \in H$. Thus μ_1 is an isomorphism and $H \cong H_1$. An analogous argument shows that $\mu_2: K \to K_1, k \to (1,k)$ is an isomorphism, so $K \cong K_1$.

That $H_1K_1 = G$ follows immediately from the fact that any $(h,k) \in G$ can be written as $(h,1)(1,k)$ with $(h,1) \in H_1$, $(1,k) \in K_1$.

Finally, $H_1 \cap K_1 = 1$. Indeed, if $(h,k) \in H_1 \cap K_1$, then $h = 1$ as $(h,k) \in K_1$ and $k = 1$ as $(h,k) \in H_1$, thus $(h,k) = (1,1)$ and so $H_1 \cap K_1 \subseteq \{(1,1)\} = 1$, yielding $H_1 \cap K_1 = 1$.

This completes the proof. □

22.5 Theorem: *Let G be a group and let H,K be subgroups of G. The following statements are equivalent.*
(1) $H \trianglelefteq G, K \trianglelefteq G, G = HK$ *and* $H \cap K = 1$.
(2) *Every element of G can be expressed uniquely in the form hk, where $h \in H$ and $k \in K$; and every element of H commutes with every element of K.*

Proof: (1) $\Rightarrow$ (2) Suppose $H \trianglelefteq G, K \trianglelefteq G, G = HK$ and $H \cap K = 1$. Since $G = HK$, every element of G can be expressed as hk, with $h \in H, k \in K$. We must show that this representation is unique, i.e., when $hk = h'k'$ with $h,h' \in H$ and $k,k' \in K$, then necessarily $h = h'$ and $k = k'$. This follows from $H \cap K = 1$. Indeed, from $hk = h'k'$, we get $kk'^{-1} = h^{-1}h' \in H \cap K = 1$, so $kk'^{-1} = 1 = h^{-1}h'$, so $k = k'$ and $h = h'$.

It remains to prove that any element of H commutes with any element of K. Let $h \in H, k \in K$. We have to show $hk = kh$, or, equivalently, $h^{-1}k^{-1}hk = 1$. Now $h^{-1}k^{-1}h.k \in K$, since $h^{-1}k^{-1}h \in K$ (because $k^{-1} \in K$ and $K \trianglelefteq G$) and $h^{-1}.k^{-1}hk \in H$, since $k^{-1}hk \in H$ (because $h \in H$ and $H \trianglelefteq G$), so $h^{-1}k^{-1}hk \in H \cap K = 1$ and $h^{-1}k^{-1}hk = 1$, as claimed.

(2) $\Rightarrow$ (1) By hypothesis, every element of G can be written in the form hk, where $h \in H, k \in K$. So $G = HK$. We now prove $H \cap K = 1$. Let $a \in H \cap K$. If $a \neq 1$, then $1a = a1$ are two distinct representations of $a \in G$ with $1 \in H$, $a \in K$ and $a \in H$, $1 \in K$, contrary to the hypothesis that every element of G, in particular a, can be expressed *uniquely* in the form hk, with $h \in H$, $k \in K$. Thus $a = 1$. This proves $H \cap K = 1$.

In order to prove $H \trianglelefteq G$, we must show $g^{-1}hg \in H$ for all $h \in H, g \in G$ (Lemma 18.2(1)). Let $g \in G = HK$. Then $g = h'k'$ for some $h' \in H, k' \in K$. Thus

$$\begin{aligned} g^{-1}hg &= (h'k')^{-1}h(h'k') \\ &= k'^{-1}(h'^{-1}hh')k' \\ &= k'^{-1}.k'(h'^{-1}hh') \quad (h'^{-1}hh' \in H \text{ and } k' \in K \text{ commute}) \end{aligned}$$

$$= h'^{-1}hh' \in H$$

and therefore $H \trianglelefteq G$. The proof of $K \trianglelefteq G$ is similar and is left to the reader. □

22.6 Theorem: *Let G be a group and H,K be subgroups of G. Assume that $H \trianglelefteq G$, $K \trianglelefteq G$, $G = HK$ and $H \cap K = 1$. Then $G \cong H \times K$.*

Proof: We want to find an isomorphism $\varphi: H \times K \to G$. For each (h,k) in $H \times K$, this should give us an element $(h,k)\varphi$ of G. By hypothesis, $G = HK$. This suggests that $(h,k) \to hk$ might be an appropriate mapping from $H \times K$ into G. So we put $\varphi: H \times K \to G$. We show that φ is a homomorphism,

$$(h,k) \to hk$$

one-to-one and onto.

φ is a homomorphism if and only if

$$((h',k)(h,k'))\varphi = (h',k)\varphi.(h,k')\varphi \quad \text{for all } h,h' \in H,\ k,k' \in K,$$

that is, if and only if

$$h'hkk' = h'khk' \quad \text{for all } h,h' \in H,\ k,k' \in K,$$

which is equivalent to

$$hk = kh \quad \text{for all } h \in H,\ k \in K,$$

and this is true by Theorem 22.5. So φ is a homomorphism.

φ is one-to-one, for if $(h,k)\varphi = (h',k')\varphi$, then $hk = h'k'$; but every element in G can be expressed in the form hk with $h \in H$, $k \in K$ in a unique way by Theorem 22.5. So $h = h'$ and $k = k'$. Thus $(h,k) = (h',k')$. This proves that φ is one-to-one.

φ is onto because $HK = G$ by hypothesis.

Hence φ is an isomorphism and $H \times K \cong G$, and also $G \cong H \times K$. □

22.7 Theorem: (I) *A group G is isomorphic to the direct product of two subgroups H and K if and only if* (i) *every element of G can be expressed uniquely in the form hk, where $h \in H$ and $k \in K$ and* (ii) *every element of H commutes with every element of K.*

(2) *Let G be a group and $H,K \leqslant G$. If $G \cong H \times K$,* then $G/H \cong K$ and $G/K \cong H$.

Proof: (1) follows from Theorem 22.4, Theorem 22.5, Theorem 22.6. As for (2), we observe that $G \cong H \times K$ implies $H \trianglelefteq G, K \trianglelefteq G, G = HK, H \cap K = 1$, so that $G/H = HK/H \cong K/H \cap K = K/1 \cong K$ by Theorem 21.3. The proof of $G/K \cong H$ is similar. □

When the conditions of Theorem 22.7(1) are satisfied, G is said to be the internal direct product of H and K. The direct product of Definition 22.2 is called the external direct product of H and K. Theorem 22.5 and Theorem 22.6 state that the internal direct product of H and K is isomorphic to the external direct product $H \times K$. For this reason, we will not distinguish between external and internal direct products and refer to both of them simply as direct products.

As an illustration of Theorem 22.7(1), consider $\mathbb{Q}\backslash\{0\} \cong C_2 \times \mathbb{Q}^+$ (Example 22.3(a)). Theorem 22.7(1) asserts that every nonzero rational number can be written as $(\mp 1)q$, where $q \in \mathbb{Q}, q > 0$ in a unique way. This is of course well known to everybody.

Next we investigate the direct product of two finite cyclic groups of relatively prime orders. It will be sufficient to examine the direct sum of $\mathbb{Z}/m\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}$.

Let m and n be relatively prime natural numbers. For any integer a, we denote the residue class of a (mod m) by $\bar{a}$, and the residue class of a (mod n) by a^*. Hence $\bar{a} \in \mathbb{Z}_m$ and $a^* \in \mathbb{Z}_n$.

Consider the mapping $\varphi: \mathbb{Z} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$, $a \to (\bar{a}, a^*)$. It is easy to see that φ is a homomorphism:

$$(a+b)\varphi = (\overline{a+b}, (a+b)^*) = (\bar{a}+\bar{b}, a^*+b^*) = (\bar{a},a^*) + (\bar{b},b^*) = a\varphi + b\varphi$$

for all $a,b \in \mathbb{Z}$. So φ is a homomorphism and

$$\mathbb{Z}/Ker\,\varphi \cong Im\,\varphi$$

by Theorem 20.16. Now $a \in Ker\,\varphi$ if and only if $\bar{a} = \bar{0}$ and $a^* = 0^*$, that is, if and only if $m|a$ and $n|a$. Since m and n are relatively prime, the latter

condition is equivalent to $mn|a$. Hence $Ker\ \varphi = mn\mathbb{Z}$ and $\mathbb{Z}/mn\mathbb{Z} \cong Im\ \varphi$, where $Im\ \varphi$ is a subgroup of $\mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$. From

$$mn = |\mathbb{Z}/mn\mathbb{Z}| = |Im\ \varphi| \leqslant |\mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}| = |\mathbb{Z}/m\mathbb{Z}|\,|\mathbb{Z}/n\mathbb{Z}| = mn$$

we conclude $|Im\ \varphi| = mn$, hence $Im\ \varphi = \mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$. Therefore φ is onto and $\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$. Writing this multiplicatively, we get

22.8 Theorem: *If m and n are relatively prime natural numbers, then*

$$C_{mn} \cong C_m \times C_n. \qquad \square$$

We record an important result that we obtained as a bonus.

22.9 Theorem: *Let m and n be relatively prime natural numbers. Then the mapping*

$$\begin{array}{rcl} \varphi: \mathbb{Z} & \to & \mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z} \\ a & \to & (\bar{a}, a^*) \end{array}$$

is a group homomorphism onto $\mathbb{Z}/m\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$. $\square$

So far, we have examined the direct product of two groups. The construction extends immediately to n groups, where $n > 2$. We shall be content with enunciating the appropriate theorems. Their proofs consist in writing n-tuples in place of ordered pairs in the proofs above. The only novel point is extension of the previous condition $H \cap K = 1$. This is discussed in Theorem 22.12, whose proof we briefly sketch.

22.10 Theorem: *Let $H_1, H_2, \ldots, H_n$ be arbitrary groups. On the cartesian product $H_1 \times H_2 \times \ldots \times H_n$, we define a binary operation by declaring*

$$(h_1, h_2, \ldots, h_n)(h_1', h_2', \ldots, h_n') = (h_1h_1', h_2h_2', \ldots, h_nh_n')$$

for all $(h_1,h_2,\ldots,h_n),(h_1',h_2',\ldots,h_n') \in H_1 \times H_2 \times \ldots \times H_n$. *With respect to this operation,* $H_1 \times H_2 \times \ldots \times H_n$ *is a group.*

22.11 Definition: The group of Theorem 22.10 is called the *direct product of* $H_1,H_2,\ldots,H_n$ and is denoted by $H_1 \times H_2 \times \ldots \times H_n$. If the groups are written additively, we call the group of Theorem 22.10 the *direct sum of* $H_1,H_2,\ldots,H_n$ and denote is by $H_1 \oplus H_2 \oplus \ldots \oplus H_n$.

22.12 Theorem: *Let* $H_1,H_2,\ldots,H_n$ *be groups and* $G = H_1 \times H_2 \times \ldots \times H_n$. *Then there are subgroups* $G_1,G_2,\ldots,G_n$ *of* G *such that*

$$G_i \cong H_i \text{ and } G_i \trianglelefteq G \text{ for all } i = 1,2,\ldots,n,$$
$$G = G_1G_2\ldots G_n \text{ and } G_1G_2\ldots G_{j-1} \cap G_j = 1 \text{ for all } j = 2,\ldots,n.$$

Sketch of proof: Let G_i be the set $\{(1,\ldots,x,\ldots,1): x \in H_i\}$ of all n-tuples in G whose k-th components are equal to $1 \in H_k$ whenever $k \neq i$. It is easily verified that G_i is a subgroup of G, normal in G, isomorphic to H_i and that $G = G_1G_2\ldots G_n$. In fact, for all $j = 2,\ldots,n$,

$$G_1G_2\ldots G_{j-1} = \{(h_1,h_2,\ldots,h_{j-1},1,\ldots,1): h_1 \in H_1,h_2 \in H_2,\ldots,h_{j-1} \in H_{j-1}\}.$$

Finally, to prove $G_1G_2\ldots G_{j-1} \cap G_j = 1$ for all $j = 2,\ldots,n$, let $(u_1,u_2,\ldots,u_n) \in G_1G_2\ldots G_{j-1} \cap G_j$, where $j \in \{2,\ldots,n\}$. Here $u_k = 1$ for $k \neq j$, because $(u_1,u_2,\ldots,u_n) \in G_j$. Thus $(u_1,u_2,\ldots,u_n) = (1,\ldots,u_j,\ldots,1)$. But

$$(1,\ldots,u_j,\ldots,1) \in G_1G_2\ldots G_{j-1}$$
$$= \{(h_1,h_2,\ldots,h_{j-1},1,\ldots,1): h_1 \in H_1,h_2 \in H_2,\ldots,h_{j-1} \in H_{j-1}\},$$

hence $u_j = 1$ and

$(u_1,u_2,\ldots,u_n) = (1,\ldots,1,\ldots,1) = 1 \in G$. Thus $G_1G_2\ldots G_{j-1} \cap G_j = 1$. □

22.13 Theorem: *Let* G *be a group and let* $G_1,G_2,\ldots,G_n$ *be subgroups of* G. *The following statements are equivalent.*
(1) $G_i \trianglelefteq G$ *for all* $i = 1,2,\ldots,n$, $G = G_1G_2\ldots G_n$ *and* $G_1G_2\ldots G_{j-1} \cap G_j = 1$ *for all* $j = 2,\ldots,n$.

(2) *Every element of G can be expressed uniquely in the form $g_1g_2\ldots g_n$, where $g_1 \in G_1, g_2 \in G_2, \ldots, g_n \in G_n$; and every element of G_k commutes with every element of G_l $(k \neq l)$.*

22.14 Theorem: *Let G be a group and let $G_1,G_2,\ldots,G_n$ be subgroups of G. Assume that $G_i \trianglelefteq G$ for all $i = 1,2,\ldots,n$, $G = G_1G_2\ldots G_n$ and $G_1G_2\ldots G_{j-1} \cap G_j = 1$ for all $j = 2,\ldots,n$. Then $G \cong G_1 \times G_2 \times \ldots \times G_n$.*

If $n > 3$ and $G_1,G_2,\ldots,G_n$ are normal subgroups of a group G such that $G = G_1G_2\ldots G_n$ and $G_i \cap G_j = 1$ whenever $i \neq j$, then G is need not be isomorphic to the direct product of $G_1,G_2,\ldots,G_n$. By way of example, let $G = V_4$ and let $A = \{\iota,(12)(34)\}$, $B = \{\iota,(13)(24)\}$, $C = \{\iota,(14)(23)\}$. Then A,B,C are normal subgroups of G, and $G = ABC$, and $A \cap B = B \cap C = A \cap C = 1$. However, G is not isomorphic to $A \times B \times C$, because, for one thing, G has order 4, whereas $A \times B \times C$ has order 8. Thus the condition

$$G_1G_2\ldots G_{j-1} \cap G_j = 1 \text{ for all } j = 2,\ldots,n$$

cannot be relaxed to

$$G_i \cap G_j = 1 \text{ for all } i \neq j.$$

22.15 Theorem: *A group G is isomorphic to the direct product of n subgroups $G_1,G_2,\ldots,G_n$ if and only if* (i) *every element of G can be expressed uniquely in the form $g_1g_2\ldots g_n$, where $g_1 \in G_1, g_2 \in G_2, \ldots, g_n \in G_n$ and* (ii) *every element of G_k commutes with every element of G_l $(k \neq l)$.*

The last two elementary results will be needed in §28.

22.16 Lemma: *Let $G_1,G_2,\ldots,G_n,H_1,H_2,\ldots,H_n$ be groups and assume that $G_1 \cong H_1, G_2 \cong H_2, \ldots, G_n \cong H_n$. Then $G_1 \times G_2 \times \ldots \times G_n \cong H_1 \times H_2 \times \ldots \times H_n$.*

Proof: Let $\varphi_i: G_i \to H_i$ be an isomorphism $(i = 1,2,\ldots,n)$. The mapping

$$\begin{aligned} \varphi: G_1 \times G_2 \times \ldots \times G_n &\to H_1 \times H_2 \times \ldots \times H_n \\ (g_1,g_2,\ldots,g_n) &\to (g_1\varphi_1,g_2\varphi_2,\ldots,g_n\varphi_n) \end{aligned}$$

is a homomorphism, because $((g_1,g_2,\ldots,g_n)(g_1',g_2',\ldots,g_n'))\psi$

$$\begin{aligned}
&= (g_1g_1',g_2g_2',\ldots,g_ng_n')\psi \\
&= ((g_1g_1')\varphi_1,(g_2g_2')\varphi_2,\ldots,(g_ng_n')\varphi_n) \\
&= (g_1\varphi_1g_1'\varphi_1,g_2\varphi_2g_2'\varphi_2,\ldots,g_n\varphi_ng_n'\varphi_n) \\
&= (g_1\varphi_1,g_2\varphi_2,\ldots,g_n\varphi_n)(g_1'\varphi_1,g_2'\varphi_2,\ldots,g_n'\varphi_n) \\
&= (g_1,g_2,\ldots,g_n)\psi(g_1',g_2',\ldots,g_n')\psi
\end{aligned}$$

for all $(g_1,g_2,\ldots,g_n),(g_1',g_2',\ldots,g_n') \in G_1 \times G_2 \times \ldots \times G_n$. Since

$$\begin{aligned}
Ker\,\psi &= \{(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n : (g_1\varphi_1,g_2\varphi_2,\ldots,g_n\varphi_n) = (1,1,\ldots,1)\} \\
&= \{(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n : g_1\varphi_1 = 1, g_2\varphi_2 = 1, \ldots, g_n\varphi_n = 1\} \\
&= \{(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n : g_1 = 1, g_2 = 1, \ldots, g_n = 1\} \\
&= \{(1,1,\ldots,1)\} = 1,
\end{aligned}$$

ψ is one-to-one. Also, ψ is onto: given any $(h_1,h_2,\ldots,h_n) \in H_1 \times H_2 \times \ldots \times H_n$, there are $g_1 \in G_1, g_2 \in G_2, \ldots, g_n \in G_n$ with $g_1\varphi_1 = h_1, g_2\varphi_2 = h_2, \ldots, g_n\varphi_n = h_n$, thus $(h_1,h_2,\ldots,h_n)$ is the image, under ψ, of $(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n$.

So ψ is an isomorphism and $G_1 \times G_2 \times \ldots \times G_n \cong H_1 \times H_2 \times \ldots \times H_n$. □

22.17 Lemma: *Let $G_1,G_2,\ldots,G_n$ be groups and $H_1 \trianglelefteq G_1, H_2 \trianglelefteq G_2, \ldots, H_n \trianglelefteq G_n$. Then $H_1 \times H_2 \times \ldots \times H_n \trianglelefteq G_1 \times G_2 \times \ldots \times G_n$ and*

$$G_1 \times G_2 \times \ldots \times G_n \,/\, H_1 \times H_2 \times \ldots \times H_n \cong G_1/H_1 \times G_2/H_2 \times \ldots \times G_n/H_n.$$

Proof: The mapping $\varphi: G_1 \times G_2 \times \ldots \times G_n \longrightarrow G_1/H_1 \times G_2/H_2 \times \ldots \times G_n/H_n$

$$(g_1,g_2,\ldots,g_n) \longrightarrow (H_1g_1,H_2g_2,\ldots,H_ng_n)$$

is a homomorphism because $((g_1,g_2,\ldots,g_n)(g_1',g_2',\ldots,g_n'))\varphi$

$$\begin{aligned}
&= (g_1g_1',g_2g_2',\ldots,g_ng_n')\varphi \\
&= (H_1g_1g_1',H_2g_2g_2',\ldots,H_ng_ng_n') \\
&= (H_1g_1H_1g_1',H_2g_2H_2g_2',\ldots,H_ng_nH_ng_n') \\
&= (H_1g_1,H_2g_2,\ldots,H_ng_n)(H_1g_1',H_2g_2',\ldots,H_ng_n') \\
&= (g_1,g_2,\ldots,g_n)\varphi(g_1',g_2',\ldots,g_n')\varphi
\end{aligned}$$

for all $(g_1,g_2,\ldots,g_n),(g_1',g_2',\ldots,g_n') \in G_1 \times G_2 \times \ldots \times G_n$. Moreover, φ is onto: any $(H_1g_1,H_2g_2,\ldots,H_ng_n)$ in $G_1/H_1 \times G_2/H_2 \times \ldots \times G_n/H_n$ is the image, under φ, of $(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n$. Thus

$$Im\ \varphi = G_1/H_1 \times G_2/H_2 \times \ldots \times G_n/H_n.$$

To complete the proof, we need only show $Ker\ \varphi = H_1 \times H_2 \times \ldots \times H_n$ (Theorem 20.16). We indeed have

$$\begin{aligned} Ker\ \varphi &= \{(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n : (g_1,g_2,\ldots,g_n)\varphi = (H_1,H_2,\ldots,H_n)\} \\ &= \{(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n : H_1g_1 = H_1, H_2g_2 = H_2, \ldots, H_ng_n = H_n\} \\ &= \{(g_1,g_2,\ldots,g_n) \in G_1 \times G_2 \times \ldots \times G_n : g_1 \in H_1, g_2 \in H_2, \ldots, g_n \in H_n\} \\ &= H_1 \times H_2 \times \ldots \times H_n. \end{aligned}$$

□

Exercises

1. Prove that $V_4 \cong C_2 \times C_2$.

2. Show that C_{mn} is not isomorphic to $C_m \times C_n$ if $(m,n) \neq 1$.

3. Find three nonisomorphic abelian groups of order 8 and three nonisomorphic abelian groups of order 12.

4. Show that $G_1 \times G_2 \times \ldots \times G_n \cong G_{i_1} \times G_{i_2} \times \ldots \times G_{i_n}$ for any permutation $\begin{pmatrix} 1 & 2 & \ldots & n \\ i_1 & i_2 & \ldots & i_n \end{pmatrix}$ in S_n.

5. Prove that, if G is isomorphic to the direct product of its subgroups G_1, $G_2, \ldots, G_n$, then $G_1 \ldots G_{k-1}G_{k+1} \ldots G_n \cap G_k = 1$ for all $k = 1,2, \ldots, n$.

6. Let H,K be normal subgroups of G. Find a one-to-one homomorphism from $G/H \cap K$ into $G/H \times G/K$. Prove that $HK/H \cap K \cong H/H \cap K \times K/H \cap K$.

7. Let $\varphi_i : G_i \to H_i$ be group homomorphisms $(i = 1,2, \ldots, n)$. Define ψ by

$$\begin{aligned} \psi : G_1 \times G_2 \times \ldots \times G_n &\to H_1 \times H_2 \times \ldots \times H_n \\ (g_1,g_2,\ldots,g_n) &\to (g_1\varphi_1, g_2\varphi_2, \ldots, g_n\varphi_n) \end{aligned}$$

(ψ is sometimes denoted by $\varphi_1 \times \varphi_2 \times \ldots \times \varphi_n$). Show that ψ is a homomorphism and

$Ker\ \psi = Ker\ \varphi_1 \times Ker\ \varphi_2 \times \ldots \times Ker\ \varphi_n, Im\ \psi = Im\ \varphi_1 \times Im\ \varphi_2 \times \ldots \times Im\ \varphi_n$.

8. For any abelian group A, let $\hat{A}$ be the set of all homomorphisms from A into $\mathbb{C}\backslash\{0\}$. Prove that $\hat{A}$ is an abelian group under the multiplication

$$a(\varphi\psi) = a\varphi . a\psi \qquad \text{for all } a \in A, \varphi, \psi \in \hat{A}$$

and show that $\hat{A}_1 \times \hat{A}_2 \cong \widehat{A_1 \times A_2}$.

§23
Center and Automorphisms of Groups

We introduce an important subgroup of a group.

23.1 Definition: Let G be a group. We put

$$Z(G) = \{z \in G: zg = gz \text{ for all } g \in G\}$$

and call $Z(G)$ the *center of* G.

The center of G consists, therefore, of the elements of G that commute with every element of G. It is a subset of G. Since $1g = g1$ for all $g \in G$, the identity element belongs to $Z(G)$, so $Z(G) \neq \emptyset$. Obviously, $Z(G) = G$ if and only if G is abelian.

23.2 Theorem: *Let G be a group. Then $Z(G) \leqslant G$.*

Proof: We use our subgroup criterion (Lemma 9.2).

(i) Let $z_1, z_2 \in Z(G)$. We want to show $z_1z_2 \in Z(G)$. Thus we must show that $(z_1z_2)g = g(z_1z_2)$ for all $g \in G$. This follows easily from $z_1g = gz_1$, $z_2g = gz_2$ for all $g \in G$, which are true since $z_1, z_2 \in Z(G)$:

$$(z_1z_2)g = z_1(z_2g) = z_1(gz_2) = (z_1g)z_2 = (gz_1)z_2 = g(z_1z_2).$$

Hence $Z(G)$ is closed under multiplication.

(ii) Let $z \in Z(G)$. We want to show $z^{-1} \in Z(G)$. We know

$$g^{-1}z = zg^{-1} \text{ for any } g \in G;$$

so, taking inverses, we get

$$z^{-1}g = gz^{-1} \text{ for any } g \in G,$$

which means $z^{-1} \in Z(G)$. Hence $Z(G)$ is closed under the forming of inverses.

Thus $Z(G) \leqslant G$. □

As any two elements of $Z(G)$ commute, $Z(G)$ is an abelian subgroup of G. It is also a normal subgroup of G. We prove a slightly stronger result.

23.3 Theorem: *Let G be a group. If $H \leqslant Z(G)$, then $H \trianglelefteq G$.*

Proof: We are to show $g^{-1}hg \in H$ for all $g \in G, h \in H$. Now, if $g \in G, h \in H$ then $g^{-1}hg = g^{-1}(hg) = g^{-1}(gh) = (g^{-1}g)h = h \in H$, for $h \in Z(G)$ commutes with g. Thus $H \trianglelefteq G$. □

A subgroup of G which is contained in the center of G is called a *central subgroup of G*. With this terminology, Theorem 23.3 states that any central subgroup of G is normal in G. Central subgroups are abelian. Elements of $Z(G)$ are also called *central elements of G*.

23.4 Examples: (a) Let K be a field and let us put $G = GL(2,K)$ for brevity. We want to find $Z(G)$. Let $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$. Then $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Z(G)$ if and only if $\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x & y \\ z & u \end{pmatrix} = \begin{pmatrix} x & y \\ z & u \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ for all $\begin{pmatrix} x & y \\ z & u \end{pmatrix} \in G$. In particular,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

hence $a = a + c$, $a + b = b + d$ and $b = c$ $a = d$
$c = c$, $c + d = d$ $d = a$ $c = b$

for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in Z(G)$, so $\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$, where $a \neq 0$ since $det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq 0$.

Therefore $$Z(G) \subseteq \left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \in G : a \neq 0 \right\}$$

and conversely the set on the right hand side is contained in $Z(G)$, for

$$\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}\begin{pmatrix} x & y \\ z & u \end{pmatrix} = \begin{pmatrix} ax & ay \\ az & au \end{pmatrix} = \begin{pmatrix} xa & ya \\ za & ua \end{pmatrix} = \begin{pmatrix} x & y \\ z & u \end{pmatrix}\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \quad \text{for all } \begin{pmatrix} x & y \\ z & u \end{pmatrix} \in G.$$

Thus $Z(G) = \left\{ \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} \in G : a \neq 0 \right\}$. The elements of $Z(G)$ are called *scalar* matrices.

(b) Let $D_{4n} = \langle a,b : a^{2n} = 1, b^2 = 1, bab = a^{-1} \rangle$ be a dihedral group of order $4n > 4$. What is $Z(D_{4n})$? Well, let $x \in Z(D_{4n})$. Then $x = a^j$ or $x = a^j b$ for some $j \in \mathbb{Z}$, $0 \leqslant j \leqslant 2n - 1$. Since $xa = ax$ and $xb = bx$, we get

$a^j a = a a^j$	and	$a^j b = b a^j$	in case $x = a^j$,
$a^j b a = a a^j b$	and	$a^j b b = b a^j b$	in case $x = a^j b$.

These are equivalent to

(1)	$a^{j+1} = a^{j+1}$	and	$a^j b = a^{-j} b$	in case $x = a^j$,
(2)	$a^{j-1} b = a^{j+1} b$	and	$a^j = a^{-j}$	in case $x = a^j b$.

The equations in (1) are satisfied only when $n \mid 2j$, that is to say, only when $j = 0, n$, so only when $x = a^0, a^n$. The first equation in (2) is never satisfied, for $n > 1$ by hypothesis. Thus $Z(D_{4n}) \subseteq \{1, a^n\}$. The reader will easily show the reverse inclusion. Hence $Z(D_{4n}) = \{1, a^n\} = \langle a^n \rangle$.

(c) Let us find $Z(S_3)$. It is easy to see that ι and (12) are the only permutations in S_3 that commute with (12). Also, ι and (13) are the only permutations in S_3 that commute with (13). Hence ι is the only permutation in S_3 that commute with both (12) and (13). A fortiori, $Z(S_3) = 1$.

23.5 Lemma: *Let H be a central subgroup of G. If G/H is cyclic, then G is abelian.*

Proof: $H \trianglelefteq G$ by Theorem 23.3 and so G/H is meaningful. By hypothesis, $G/H = \langle Hg \rangle$ for some $g \in G$. Then, for any $x \in G$, there holds $Hx = (Hg)^m = Hg^m$ with a suitable $m \in \mathbb{Z}$. This means that any $x \in G$ can be written in the form hg^m, where $h \in H$, $m \in \mathbb{Z}$.

Let x, y be arbitrary elements of G. We write them as $x = hg^m$, $y = kg^n$, where $h, k \in H \leqslant Z(G)$ and $m, n \in \mathbb{Z}$. Then $xy = (hg^m)(kg^n) = h(g^m k)g^n = h(kg^m)g^n = (hk)(g^m g^n) = (hk)(g^{m+n}) = (hk)(g^{n+m}) = (kh)(g^n g^m) = k(hg^n)g^m = k(g^n h)g^m = (kg^n)(hg^m) = yx$ and G is commutative. □

The center of any group G is normal in G (Theorem 23.3) and is therefore the kernel of some homomorphism (Theorem 20.14). Now we construct a homomorphism whose kernel is $Z(G)$. We will need the concept of automorphisms.

23.6 Definition: Let G be a group. An isomorphism $\alpha: G \to G$ from G onto G itself is called an *automorphism of* G. The set of all automorphisms of G will be denoted by $Aut(G)$.

Since any isomorphism is one-to-one and onto, $\alpha \in Aut(G)$ implies $\alpha \in S_G$. Thus $Aut(G) \subseteq S_G$. The identity mapping ι_G on G is an isomorphism from G onto G, so $\iota_G \in Aut(G)$ and $Aut(G) \neq \emptyset$. We can form the composition $\alpha\beta$ of any $\alpha, \beta \in Aut(G)$. It turns out that $Aut(G)$ is a group.

23.7 Theorem: *Let G be a group. Then $Aut(G)$ is a group under the composition of mappings.*

Proof: We can check the group axioms, but there is a shorter way. We make use of $\emptyset \neq Aut(G) \subseteq S_G$. Now S_G is a group under the composition of mappings (Example 7.1(d)), so all we have to do is show that $Aut(G)$ is a subgroup of S_G.

(i) Let $\alpha, \beta \in Aut(G)$. Then $\alpha\beta$ is an isomorphism from G onto G by Lemma 20.11(1). Thus $\alpha\beta \in Aut(G)$ and $Aut(G)$ is closed under multiplication.

(ii) Let $\alpha \in Aut(G)$. Then α^{-1} is an isomorphism from G onto G by Lemma 20.11(2). Thus $\alpha^{-1} \in Aut(G)$ and $Aut(G)$ is closed under the forming of inverses.

By Lemma 9.2, $Aut(G) \leqslant S_G$. Thus $Aut(G)$ is a group. □

$Aut(G)$ is not a subgroup or a factor group of G, of course. The underlying set is neither a subset nor a set of cosets of a subgroup of G.

23.8 Example: Let G be a group. We fix an arbitrary element g of G. With each $x \in G$, we associate $g^{-1}xg$. This is a uniquely determined element of G, so we have a mapping $x \to g^{-1}xg$, which we denote by τ_g. So

$$\begin{aligned} \tau_g: G &\to G \\ x &\to g^{-1}xg \end{aligned}$$

We claim τ_g is a homomorphism. For all $x,y \in G$, we have

$$(xy)\tau_g = g^{-1}xyg = g^{-1}xg.g^{-1}yg = x\tau_g y\tau_g,$$

and τ_g is therefore a homomorphism.

We can build τ_g with any $g \in G$. Let us take the composition of two of them, τ_g and τ_h, say. For $g,h \in G$, we have

$$x(\tau_g\tau_h) = (x\tau_g)\tau_h = (g^{-1}xg)\tau_h = h^{-1}(g^{-1}xg)h = (h^{-1}g^{-1})x(gh) = (gh)^{-1}x(gh) = x\tau_{gh}$$

for all $x \in G$. Thus

$$\tau_g\tau_h = \tau_{gh} \qquad \text{for all } g,h \in G. \tag{1}$$

There holds $x\tau_1 = 1^{-1}x1 = x$ for all $x \in G$. Thus

$$\tau_1 = \iota. \tag{2}$$

For any $g \in G$, there holds $\tau_g\tau_{g^{-1}} = \tau_{gg^{-1}} = \tau_1 = \iota$ and $\tau_{g^{-1}}\tau_g = \tau_{g^{-1}g} = \tau_1 = \iota$ by (1) and (2). Thus τ_g is one-to-one and onto (Theorem 3.17(2)) and $\tau_{g^{-1}}$ is the inverse of τ_g:

$$(\tau_g)^{-1} = \tau_{g^{-1}} \tag{3}$$

So τ_g is an automorphism of G.

Such automorphisms deserve a name.

23.9 Definition: Let G be a group. An automorphism of G of the form τ_g, where $g \in G$, is called an *inner automorphism of* G. The set

$$\{\tau_g \in Aut(G): g \in G\}$$

of all inner automorphisms of G will be denoted by $Inn(G)$.

Inner automorphisms of a group form a group.

23.10 Theorem: *Let G be a group. Then $Inn(G) \leqslant Aut(G)$.*

Proof: $\iota = \tau_1 \in Inn(G)$ by (2), so $Inn(G) \neq \varnothing$. Now (i) the product of two inner automorphisms is an inner automorphism by (1); and (ii) the inverse of an inner automorphism is an inner automorphism by (3). So $Inn(G) \leqslant Aut(G)$. □

The relation (1) has a deep significance. It states that the mapping

$$\begin{aligned} \tau: G &\to Aut(G) \\ g &\to \tau_g \end{aligned}$$

is a homomorphism. Theorem 20.16 gives $G/Ker\ \tau \cong Im\ \tau$.

Here $Im\ \tau = \{\tau_g \in Aut(G): g \in G\} = Inn(G)$ by definition and

$$\begin{aligned} Ker\ \tau &= \{z \in G: \tau_z = \iota\} \\ &= \{z \in G: g\tau_z = g \text{ for all } g \in G\} \\ &= \{z \in G: z^{-1}gz = g \text{ for all } g \in G\} \\ &= \{z \in G: gz = zg \text{ for all } g \in G\} \\ &= Z(G). \end{aligned}$$

Thus $Z(G)$ is the kernel of $\tau: G \to Aut(G)$. We proved

23.11 Theorem: *Let G be a group. Then $G/Z(G) \cong Inn(G)$.* □

Next we prove that $Inn(G)$ is a normal subgroup of $Aut(G)$.

23.12 Lemma: *Let G be a group. Then $Inn(G) \trianglelefteq Aut(G)$.*

Proof: We know $Inn(G) \trianglelefteq Aut(G)$ from Theorem 23.10. We are to show $\sigma^{-1}\tau_g\sigma \in Inn(G)$ for any $\tau_g \in Inn(G)$, $\sigma \in Aut(G)$. For any $x \in G$, we have

$$\begin{aligned} x(\sigma^{-1}\tau_g\sigma) &= (x\sigma^{-1})(\tau_g\sigma) \\ &= ((x\sigma^{-1})\tau_g)\sigma \\ &= (g^{-1}(x\sigma^{-1})g)\sigma \\ &= (g^{-1}\sigma)((x\sigma^{-1})\sigma)(g\sigma) \\ &= (g\sigma)^{-1}x(g\sigma) \\ &= x\tau_{g\sigma}, \end{aligned}$$

thus $\sigma^{-1}\tau_g\sigma = \tau_{g\sigma}$ and $\sigma^{-1}\tau_g\sigma \in Inn(G)$. This proves $Inn(G) \trianglelefteq Aut(G)$. □

Let G be a group and let $H \trianglelefteq G$. According to Lemma 18.2(3), $H \trianglelefteq G$ if and only if $H\tau_g = H$ for all $\tau_g \in Inn(G)$. This suggests a way of strengthening the normality concept: Instead of requiring $H\sigma = H$ for all $\sigma \in Inn(G)$, we prescribe this to hold for all $\sigma \in Aut(G)$.

23.13 Definition: Let G be a group. A subgroup H of is said to be a *characteristic subgroup of* G or to be *characteristic in* G provided $H\sigma = H$ for all $\sigma \in Aut(G)$.

Here $H\sigma$ means the set $\{h\sigma : h \in H\} \subseteq G$ as usual. The equality $H\sigma = H$ is a set equality, of course. It does *not* mean that $h\sigma = h$ for all $h \in H$. It means that, $h\sigma \in H$ for any $h \in H$, and, for any $h \in H$, there is an $h' \in H$ such that $h'\sigma = h$. Cf. Example 18.5(b). As $Inn(G) \trianglelefteq Aut(G)$, any characteristic subgroup of G is normal in G, but the converse is not true in general.

Being characteristic is a transitive relation, a good property not shared by normality (Example 18.5(i)).

23.14 Lemma: *Let $K \leq H \leq G$. If K is characteristic in H and H is characteristic in G, then K is characteristic in G.*

Proof: We are to prove that $K\sigma = K$ for all $\sigma \in Aut(G)$. Let $\sigma \in Aut(G)$. We restrict σ to H. Then $\sigma_{|H}: H \to G$ is a one-to-one homomorphism onto $H\sigma$. Since H is characteristic in G,we have $H\sigma = H$ and $\sigma_{|H}$ is an automorphism of H. Then $K\sigma_{|H} = K$, because K is characteristic in H. Thus $K\sigma = K$ for all σ in $Aut(G)$ and K is a characteristic subgroup of G. □

Another useful result of this type is given in the next lemma.

23.15 Lemma: *Let $K \leqslant H \leqslant G$. If K is characteristic in H and H is normal in G, then K is normal in G.*

Proof: We are to prove that $K\tau_g = K$ for all $\tau_g \in Inn(G)$. Let $\tau_g \in Inn(G)$. We restrict τ_g to H. Then $\tau_{g|H}: H \to G$ is a one-to-one homomorphism onto $H\tau_g = g^{-1}Hg$. Since H is normal in G,we have $g^{-1}Hg = H$ and $\tau_{g|H}$ is an automorphism of H. Then $K\tau_{g|H} = K$, because K is characteristic in H. Thus $g^{-1}Kg = K\tau_{g|H} = K$ for all $g \in G$ and K is a normal subgroup of G. □

23.16 Theorem: *Let G be a group. Then $Z(G)$ is characteristic in G.*

Proof: We must show $Z(G)\sigma = Z(G)$ for all $\sigma \in Aut(G)$. If we can prove $Z(G)\sigma \subseteq Z(G)$ for all $\sigma \in Aut(G)$, then we will have $Z(G)\sigma^{-1} \subseteq Z(G)$, that is, $Z(G) \subseteq Z(G)\sigma$ for any $\sigma \in Aut(G)$ also (cf. the proof of (2) $\Longrightarrow$ (3) in Lemma 18.2). So we need only prove $Z(G)\sigma \subseteq Z(G)$. For any $z \in Z(G)$, we are to show that $(z\sigma)g = g(z\sigma)$ for all $g \in G$. As g runs through G, so does $g\sigma$, because σ is *onto* G. Thes we need only show $(z\sigma)(g\sigma) = (g\sigma)(z\sigma)$ for all $g \in G$. But this is obvious: $(z\sigma)(g\sigma) = (zg)\sigma = (gz)\sigma = (g\sigma)(z\sigma)$ since $z \in Z(G)$ and σ is a homomorphism. Consequently, $Z(G)$ is characteristic in G. □

We end this paragraph by finding the automorphism group of a finite cyclic group. In general, given a group G, it is quite difficult to find $Aut(G)$.

Let $C_n = \langle x: x^n = 1 \rangle$ be a cyclic group of order $n \in \mathbb{N}$. An automorphism of C_n is first of all a homomorphism of C_n. We claim that a homomorphism

from C_n into C_n is uniquely determined by its effect on the generator x. In other words, if α and β are homomorphisms from C_n into C_n and $x\alpha = x\beta$, then $\alpha = \beta$. To show this, we must prove $a\alpha = a\beta$ for all $a \in C_n$. But $a = x^m$ for some $m \in \mathbb{N}$, and $a\alpha = x^m\alpha = (x\alpha)^m = (x\beta)^m = x^m\beta = a\beta$. This proves the claim.

Let ψ be a homomorphism from C_n into C_n. Then $x\psi = x^m$ for some $m \in \mathbb{Z}$. Then $x^k\psi = (x\psi)^k = (x^m)^k = x^{mk} = (x^k)^m$ for any $k \in \mathbb{N}$. This shows $a\psi = a^m$ for any $a \in C_n$. Thus a homomorphism from C_n into C_n simply sends each element of C_n to its m-th power, m being a natural number depending only on the homomorphism. The homomorphism of taking m-th powers will be denoted by α_m. Hence

$$\begin{aligned} \alpha_m: \langle x\rangle &\to \langle x\rangle \\ a &\to a^m \end{aligned}$$

is a homomorphism from C_n into C_n, and any homomorphism from C_n into C_n is one of the α_m.

From the homomorphisms $\{\alpha_m: m \in \mathbb{Z}\}$, we want to select the automorphisms. These are the one-to-one α_m's onto C_n. Since C_n is a finite set, any one-to-one mapping from C_n into C_n is in fact onto C_n. So we need find only one-to-one α_m's. These and exactly these are the automorphisms of C_n.

Now α_m is one-to-one if and only if $Ker\ \alpha_m = 1$ (Theorem 20.8) and

$$\begin{aligned} Ker\ \alpha_m &= \{g \in C_n: g\alpha_m = 1\} \\ &= \{x^k: k \in \mathbb{Z} \text{ and } x^{km} = 1\} \\ &= \{x^k: k \in \mathbb{Z} \text{ and } n \mid km\} \\ &= \{x^k: k \in \mathbb{Z} \text{ and } n/(n,m) \mid km/(n,m)\} \\ &= \{x^k: k \in \mathbb{Z} \text{ and } n/(n,m) \mid k\} \\ &= \langle x^{n/(n,m)}\rangle, \end{aligned}$$

so $Ker\ \alpha_m = 1 = \langle x^n\rangle$ if and only if $(n,m) = 1$. Thus α_m is an automorphism of C_n if and only if $(n,m) = 1$.

Hence $$Aut(C_n) = \{\alpha_m: (n,m) = 1\}.$$

This description of $Aut(C_n)$ looks like an infinite set. $Aut(C_n)$ is finite of course. Therefore, there are repetitions among α_m. To see this more vividly, we remark that $\alpha_m = \alpha_k$ if and only if $m \equiv k \pmod{n}$. Indeed, α_m is equal to α_k if and only if $x\alpha_m = x\alpha_k$ by the claim above, thus if and

only if $x^m = x^k$, thus if and only if $x^{m-k} = 1$, thus if and only if $n \mid m - k$ by Lemma 11.6, thus if and only if $m \equiv k \pmod{n}$.

Hence, for any $\bar{m} \in \mathbb{Z}_n$, we may unambiguously write $\alpha_{\bar{m}}: C_n \to C_n$. With $a \to a^m$ this notation, we have

$$Aut(C_n) = \{\alpha_{\bar{m}}: \bar{m} \in \mathbb{Z}_n^\times\}$$

and $\bar{m} \neq \bar{k}$ implies $\alpha_{\bar{m}} \neq \alpha_{\bar{k}}$. In other words, the mapping

$$\alpha: \mathbb{Z}_n^\times \to Aut(C_n)$$

is one-to-one and onto. It is a homomorphism, because

$$x\alpha_{\overline{mk}} = x^{mk} = (x^m)^k = (x^m)\alpha_{\bar{k}} = (x\alpha_{\bar{m}})\alpha_{\bar{k}} = x(\alpha_{\bar{m}}\alpha_{\bar{k}})$$

and, by the claim at the beginning, $\alpha_{\overline{mk}} = \alpha_{\bar{m}}\alpha_{\bar{k}}$ for any $\bar{m}, \bar{k} \in \mathbb{Z}_n$. Hence α is an isomorphism and $\mathbb{Z}_n^\times \cong Aut(C_n)$. We proved

23.17 Theorem: *If G is a cyclic group of order $n \in \mathbb{N}$, then $Aut(G) \cong \mathbb{Z}_n^\times$.* □

Exercises

1. Let H,K be groups. Prove that $Z(H \times K) = Z(H) \times Z(K)$.

2. Let $K \trianglelefteq G$ and $|K| = 2$. Prove that K is a central subgroup of G.

3. Prove that $K \trianglelefteq G$ implies $Z(K) \trianglelefteq G$. Show by an example that $Z(K)$ is not necessarily characteristic in G.

4. Find groups K,G such that $K \trianglelefteq G$ and $Z(K) \not\trianglelefteq Z(G)$.

5. Let G be a group and $x,y \in G$. Prove that, if $xy \in Z(G)$, then $xy = yx$.

6. Find the centers of D_4, D_{2n} (n odd), $SL(2,\mathbb{C})$, $SL(2,\mathbb{Z})$.

7. Prove that $Z(S_n) = 1$ for $n \geq 3$ and $Z(A_n) = 1$ for $n \geq 4$.

8. Define a subgroup M of G by $M/Z(G) = Z(G/Z(G))$. Show that M is characteristic in G.

9. Let $\sigma \in Aut(G)$ and $H \leqslant G$. Prove that $H\sigma$ is a subgroup of G and is isomorphic to H.

10. Let $\emptyset \neq A \subseteq Aut(G)$ and $K \leqslant H \leqslant G$. Suppose that K is characteristic in H and $H\alpha = H$ for all $\alpha \in A$. Prove that $K\alpha = K$ for all $\alpha \in A$.

11. Show that, if $G \cong H$, then $Aut(G) \cong Aut(H)$.

12. Find all characteristic subgroups of D_8. Prove that $Inn(D_8) \neq 1$ and that $Aut(D_8) \cong D_8$.

13. Prove that $Aut(\mathbb{Z}) \cong C_2$, $Aut(V_4) \cong S_3$, $Aut(S_3) \cong S_3$, $Aut(Q_8) \cong S_4$ (see §17, Ex. 15).

14. Let H be a characteristic subgroup of G and put $N = \{\alpha \in Aut(G)\colon (x\alpha)x^{-1} \in H \text{ for all } x \in G\}$. Prove that $N \trianglelefteq Aut(G)$.

15. Find a one-to-one homomorphism from $Aut(H \times K)$ into $Aut(H) \times Aut(K)$.

16. Let $H \leqslant K \leqslant G$ and $\sigma \in Aut(G)$. Prove that $H\sigma \leqslant K\sigma$ and, if also $H \trianglelefteq K$, then $H\sigma \trianglelefteq K\sigma$.

§24
Generators and Commutators

We introduce an important subgroup which distingishes abelian factor groups from nonabelian ones. It is generated by the set of commutators. First we define 'generation'.

24.1 Definition: Let G be a group and let $X \subseteq G$. The intersection of all subgroups of G which contain X is called the *subgroup of G generated by X* and is denoted by $\langle X \rangle$.

Hence $\langle X \rangle = \bigcap_{X \subseteq H \leqslant G} H$. Here H runs through a nonempty set, since at least G is a subgroup of G that contains X. Note that $\langle \varnothing \rangle = 1$. When X is a finite set, for instance $X = \{x_1, x_2, \ldots, x_n\}$, we write $\langle x_1, x_2, \ldots, x_n \rangle$ rather than $\langle \{x_1, x_2, \ldots, x_n\} \rangle$. In particular, if $X = \{x\}$ consists of a single element, then $\langle x \rangle = \langle \{x\} \rangle$ is the cyclic group generated by x, as we introduced in Definition 11.1. Definitions 11.1 and 24.1 are consistent, as will be proved in Lemma 24.2, below. Our notation $\langle \rho, \sigma \rangle$ for dihedral groups is also consistent with Definition 24.1.

When $K \leqslant G$ and $X \subseteq K$, then $\langle X \rangle \subseteq K$ by definition. So $\langle X \rangle$ is the smallest subgroup of G containing X. In particular, if $H \leqslant G$, then $\langle H \rangle = H$.

The elements of $\langle X \rangle$ are described in the next lemma. See also Ex. 1 at the end of this paragraph.

24.2 Lemma: *Let X be a nonempty subset of a group G. Then*

$$\langle X \rangle = \{x_1^{m_1} x_2^{m_2} \ldots x_k^{m_k} \in G : k \in \mathbb{N}, x_i \in X \text{ and } m_i \in \mathbb{Z} \text{ for each } i = 1,2, \ldots, k\}.$$

Proof: Let Y be the set on the right hand side. We must show $Y \subseteq \langle X \rangle$ and $\langle X \rangle \subseteq Y$.

In order to prove $Y \subseteq \langle X \rangle$, we show that $Y \subseteq H$ for every $H \leqslant G$ such that $X \subseteq H$. This follows from the closure properties of subgroups. If $X \subseteq H$ and $H \leqslant G$, then, for any $x \in X$, there holds $x^n \in H$ for any $n \in \mathbb{N}$ since H is closed under multiplication, and also $x^n \in H$ for any $n \in \mathbb{Z}$ since H is closed under taking inverses and $x^0 = 1 \in H$. Hence, for any $k \in \mathbb{N}$, any $x_1, x_2, \ldots, x_k \in X$, any $m_1, m_2, \ldots, m_k \in \mathbb{Z}$, we have $x_1^{m_1}, x_2^{m_2}, \ldots x_k^{m_k} \in H$ and, from the closure of H under multiplication, we get $x_1^{m_1} x_2^{m_2} \ldots x_k^{m_k} \in H$. Thus $Y \subseteq H$ whenever $X \subseteq H \leqslant G$. This proves $Y \subseteq \langle X \rangle$.

Now we show $\langle X \rangle \subseteq Y$. By definition of Y, we have $X \subseteq Y$ (take $k = 1$ and $m_1 = 1$). So $\langle X \rangle \subseteq Y$ will be proved if we show that Y is a subgroup of G. But Y is closed under multiplication (because k runs through $\mathbb{N}$) and under the forming of inverses (because $-m_i \in \mathbb{Z}$ when $m_i \in \mathbb{Z}$). So $X \subseteq Y$ and $Y \leqslant G$, consequently $\langle X \rangle \subseteq Y$. □

24.3 Remark: $\langle X \rangle$ consists of all finite products of elements in X and the inverses of the elements in X. Notice that the set Y of Lemma 24.2 does not change if the elements of X are replaced by their inverses. Thus $\langle X \rangle = \langle Z \rangle$, where $Z = \{x^{-1} \in G : x \in X\}$.

24.4 Definition: Let G be a group. If $X \subseteq G$ and $\langle X \rangle = G$, then X is called *a set of generators of* G, and G is said to be *generated by* X. If G has a finite set of generators, G is said to be a *finitely generated* group.

24.5 Examples: (a) If $x \in G$, then $\langle x \rangle = \{x^n : n \in \mathbb{Z}\}$ by Lemma 24.2. So $\langle x \rangle$ is the cyclic group generated by x as in Definition 11.1.

(b) Any element of the dihedral group D_{2n} can be written in the form $\rho^m \sigma^j$, where $m, j \in \mathbb{Z}$. Hence $D_{2n} = \langle \rho, \sigma \rangle$. So the notation of §14 is consistent with Definition 24.1.

(c) Any permutation in S_n ($n \geqslant 2$) can be written as a product of transpositions (Theorem 16.2). Let T be the set of all transpositions in S_n. Then $S_n = \langle T \rangle$ by Lemma 24.2.

(d) $SL(2,\mathbb{Z})$ is generated by $\left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \right\}$. A proof of this is outlined in Ex. 9.

24.6 Lemma: *Let G be a group and let X be a nonempty subset of G. Suppose $x\sigma \in X$ for all $x \in X$ and for all $\sigma \in Aut(G)$ [respectively for all $\sigma \in Inn(G)$]. Then $\langle X \rangle$ is a characteristic [respectively normal] subgroup of G.*

Proof: Let $y \in \langle X \rangle$. Then $y = x_1^{m_1} x_2^{m_2} \ldots x_k^{m_k}$ for some suitable $k \in \mathbb{N}$, $x_1, x_2, \ldots, x_k \in X$, and $m_1, m_2, \ldots, m_k \in \mathbb{Z}$ (Lemma 24.2). Then, for any σ in $Aut(G)$ [respectively for any σ in $Inn(G)$],

$$y\sigma = (x_1^{m_1} x_2^{m_2} \ldots x_k^{m_k})\sigma = (x_1\sigma)^{m_1}(x_2\sigma)^{m_2}\ldots(x_k\sigma)^{m_k} \in X$$

by Lemma 24.2, for $x_1\sigma, x_2\sigma, \ldots, x_k\sigma \in X$ by hypothesis. Thus $\langle X \rangle\sigma \subseteq \langle X \rangle$ for any $\sigma \in Aut(G)$ [respectively for any $\sigma \in Inn(G)$]. But then we have $\langle X \rangle\sigma^{-1} \subseteq \langle X \rangle$ for any $\sigma \in Aut(G)$ [respectively for any $\sigma \in Inn(G)$], too. Then $\langle X \rangle = \langle X \rangle\sigma^{-1}\sigma \subseteq \langle X \rangle\sigma \subseteq \langle X \rangle$. Hence $\langle X \rangle\sigma = \langle X \rangle$ for all $\sigma \in Aut(G)$ [respectively for all $\sigma \in Inn(G)$] and $\langle X \rangle$ is a characteristic [respectively normal] subgroup of G. □

We are now in a position to introduce commutator subgroups.

24.7 Definition: Let G be a group and $x, y \in G$. Then

$$x^{-1}y^{-1}xy \in G$$

is called the *commutator of x and y* (in this order) and is denoted by $[x,y]$.

Some authors define $[x,y]$ to be $xyx^{-1}y^{-1}$. In this book, $[x,y]$ will always stand for $x^{-1}y^{-1}xy$. Clearly, $xy = yx[x,y]$ for any $x, y \in G$. In general, $xy \neq$

yx, and $[x,y]$ is that element z in G for which $xy = yx.z$, whence the name commutator.

24.8 Lemma: *Let G be a group and $x,y \in G$.*
(1) $[x,y]^{-1} = [y,x]$.
(2) $[x,y] = 1$ *if and only if x and y commute*: $xy = yx$.

Proof: (1) $[x,y]^{-1} = (x^{-1}y^{-1}xy)^{-1} = y^{-1}x^{-1}(y^{-1})^{-1}(x^{-1})^{-1} = y^{-1}x^{-1}yx = [y,x]$.

(2) $[x,y] = 1$ means $x^{-1}y^{-1}xy = 1$, and this means $xy = yx$. □

From Lemma 24.8(2), we understand that commutators measure, so to speak, how nonabelian a group is. When the set of commutators consists of 1 only, then the group is abelian. Rather sloppily, the more nonidentity commutators a group has, the more elements of G fail to commute with other elements of G, and the more nonabelian G is. This vague statement will acquire a precise meaning below (Lemma 24.12 and Theorem 24.14).

24.9 Definition: Let $H,K \leqslant G$. We define the *commutator subgroup corresponding to H and K* as

$$[H,K] = \langle [h,k] \in G: h \in H, k \in K \rangle.$$

We saw in Lemma 24.8(1) that the inverse of a commutator is a commutator. However, when H and K are subgroups of G, the inverse of a commutator of the form $[h,k]$, where $h \in H$, $k \in K$, need *not* be a commutator of the form $[h',k']$, with $h' \in H$, $k' \in K$. Also, the product of two commutators is not a commutator in general. The commutator subgroups are defined to be the subgroups *generated* by the set of appropriate commutators, *not* as the *set* of commutators.

24.10 Lemma: *Let $H,K \leqslant G$. Then $[H,K] = [K,H]$.*

Proof: We have $[H,K] = \langle [h,k] \in G : h \in H, k \in K \rangle$

$= \langle [h,k]^{-1} \in G : h \in H, k \in K \rangle$ (by Remark 24.3)

$= \langle [k,h] \in G : k \in K, h \in H \rangle$

$= [K,H]$. □

24.11 Lemma: *Let $H, K \leqslant G$. If H and K are characteristic [respectively normal] subgroups of G, then $[H,K]$ is characteristic [respectively normal] in G.*

Proof: We use Lemma 24.6, with $X = \{[h,k] : h \in H, k \in K\}$. It suffices to show that $x\sigma \in X$ for all $x \in X$ and for all $\sigma \in Aut(G)$ [respectively for all $\sigma \in Inn(G)$]. This follows from

$$x = [h,k] \text{ for some } h \in H, k \in K,$$

$$x\sigma = [h,k]\sigma = (h^{-1}k^{-1}hk)\sigma = (h\sigma)^{-1}(k\sigma)^{-1}(h\sigma)(k\sigma) = [h\sigma, k\sigma] \in X$$

as $h\sigma \in H$, $k\sigma \in K$ for any $\sigma \in Aut(G)$ [respectively for any $\sigma \in Inn(G)$] when H and K are characteristic [respectively normal] subgroups of G. □

24.12 Lemma: *Let $H \trianglelefteq G$, $K \trianglelefteq G$. Then $[H,K] \leqslant H \cap K$. In particular, if $H \cap K = 1$, then every element of H commutes with every element of K.*

Proof: It suffices to show that $[h,k] \in H \cap K$ for all $h \in H, k \in K$. For any $h \in H, k \in K$, we have indeed

$$[h,k] = h^{-1}.k^{-1}hk \in H \text{ since } H \trianglelefteq G$$

$$[h,k] = h^{-1}k^{-1}h.k \in K \text{ since } K \trianglelefteq G,$$

yielding $[h,k] \in H \cap K$.

If $H \cap K = 1$, then $[h,k] \in H \cap K = 1$ and $[h,k] = 1$, so $hk = kh$ for all $h \in H$, and $k \in K$. □

The preceding lemma supports our vague remark that commutators measure how nonabelian a group is. Suppose we treat, somehow, commutators like the identity. Then the group will be like an abelian group. The formal way of treating commutators like 1 is to define an equivalence relation on the group in such a way that all commutators will be

equivalent to 1. The most natural equivalence relation of this type is right congruence modulo the subgroup generated by all commutators (Definition 10.4). The equivalence classes are the right cosets of this subgroup, which is normal, form a factor group. We expect this factor group to be abelian. First we give a name to the subgroup.

24.13 Definition: Let G be a group. Then the subgroup

$$[G,G] = \langle [g,g']: g,g' \in G \rangle$$

generated by all commutators in G is called the *derived subgroup of* G, denoted by G'.

G is abelian if and only if $G' = 1$. Now G' is a characteristic subgroup of G (Lemma 24.11), hence we can build the factor group G/G'. We expect G/G' is abelian. In fact, much more is true.

24.14 Theorem: *Let $K \trianglelefteq G$. Then G/K is abelian if and only if $G' \leqslant K$.*

Proof: G/K is abelian $\Leftrightarrow$ $(xK)(yK) = (yK)(xK)$ for all $x,y \in G$

$\Leftrightarrow$ $xyK = yxK$ for all $x,y \in G$

$\Leftrightarrow$ $x^{-1}y^{-1}xyK = K$ for all $x,y \in G$

$\Leftrightarrow$ $x^{-1}y^{-1}xy \in K$ for all $x,y \in G$

$\Leftrightarrow$ $[x,y] \in K$ for all $x,y \in G$

$\Leftrightarrow$ $\langle [x,y]: x,y \in G \rangle \leqslant K$

$\Leftrightarrow$ $G' \leqslant K$. □

Exercises

1. Let G be a group and X a nonempty subset of G. Prove that $\langle X \rangle = \{x_1^{\varepsilon_1}x_2^{\varepsilon_2}\ldots x_k^{\varepsilon_k} \in G: k \in \mathbb{N}, x_i \in X \text{ and } \varepsilon_i = \mp 1 \text{ for all } i = 1,2,\ldots,k\}$.

2. Show that $S_n = \langle (12),(123\ldots n\text{-}1,n) \rangle$ when $n \geqslant 3$.

3. If $H \leqslant G$, $|G:H|$ is finite and G is finitely generated, show that H is also finitely generated.

4. If $H \trianglelefteq G$ and G is finitely generated, show that G/H is also finitely generated.

5. Show that every finitely generated subgroup of $\mathbb{Q}$ is cyclic.

6. Let $H_1 \leqslant H_2 \leqslant H_3 \leqslant \cdots$ be subgroups of G. Prove that $H := \bigcup_{i=1}^{\infty} H_i$ is a subgroup of G. Prove further that, if each H_i is a proper subgroup of G, then H is also a proper subgroup of G.

7. Let $\alpha: \mathbb{R} \to \mathbb{R}$ and $\beta: \mathbb{R} \to \mathbb{R}$ and put $G = \langle \alpha, \beta \rangle \leqslant S_{\mathbb{R}}$. Let $\alpha_n = \beta^n \alpha \beta^{-n}$ for
$u \to u+1$ $\quad u \to 2u$
$n \in \mathbb{N}$. Show that $\alpha_{n+1}^2 = \alpha_n$ for all $n \in \mathbb{N}$. Show that

$$\langle \alpha_1 \rangle < \langle \alpha_2 \rangle < \langle \alpha_3 \rangle < \cdots .$$

Prove that $\langle \alpha_n \rangle$ is a proper subgroup of $A := \bigcup_{i=1}^{\infty} \langle \alpha_i \rangle$ for all $n \in \mathbb{N}$. Using Ex. 6, conclude that A is not finitely generated. Thus a subgroup of a finitely generated group need not be finitely generated.

8. Let $M = \mathbb{Q} \setminus \{0,1\}$ and $\alpha: M \to M$ and $\beta: M \to M$. Prove that $\langle \alpha, \beta \rangle \leqslant S_M$
$x \to 1/x$ $\quad x \to 1/(1-x)$
and that $\langle \alpha, \beta \rangle$ is isomorphic to S_3.

9. Show that $SL(2,\mathbb{Z}) = \langle T,S \rangle$, where $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ by going through the following steps. Let $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z})$. If $c = 0$, then M is a power of T. Make induction: suppose a matrix in $SL(2,\mathbb{Z})$ belongs to $\langle T,S \rangle$ whenever its lower-left entry is positive and $< c$. If $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is a matrix whose lower-left entry is c, divide d by c, so that $d = qc + r$. Then $MT^{-q}S$ is in $\langle T,S \rangle$, and so is M. Thus $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \langle T,S \rangle$ whenever $c > 0$. If c is negative, $MS^2 \in \langle T,S \rangle$, and so $M \in \langle T,S \rangle$.

10. Let $H \leqslant G$. Prove that $[H,G] = 1$ if and only if $H \leqslant Z(G)$ and also that $[H,G] \leqslant H$ if and only if $H \trianglelefteq G$.

11. Show that, if $G' \leqslant N \leqslant G$, then N is a normal subgroup of G.

12. Let $K \trianglelefteq G$. Prove that $[xK,yK] = [x,y]K \in G/K$ for any $x,y \in G$. Then prove that $[HK/K,JK/K] = [H,J]K/K$ for all $H,J \leqslant G$.

13. Let $H_1,H_2 \leqslant H$ and $K_1,K_2 \leqslant K$. Show that $[H_1 \times K_1, H_2 \times K_2] = [H_1,H_2] \times [K_1,K_2]$ as subgroups of $H \times K$.

14. Show that $[xy,z] = y^{-1}[x,z]y.[y,z]$ and $[x,yz] = [x,z].z^{-1}[x,y]z$ for any elements x,y,z of a group G. Deduce that $[HJ,K] = [H,K][J,K]$ whenever H,J,K are normal subgroups of G.

15. For any elements x,y,z of a group G, show that

$$y^{-1}[[x,y^{-1}],z]y \,.\, z^{-1}[[y,z^{-1}],x]z \,.\, x^{-1}[[z,x^{-1}],y]x = 1.$$

16. Let H,K,L be subgroups of a group G and $N \trianglelefteq G$. If two of the subgroups $[[H,K],L]$, $[[K,L],H]$, $[[L,H],K]$ are contained in N, prove that the third is also contained in N.

17. Give an example of a group G and three subgroups H,K,L of G such that $[[H,K],L] \neq [H,[K,L]]$.

18. Prove: if $K \trianglelefteq G$, then $K' \trianglelefteq G$.

19. Find the derived subgroups of S_3, S_4, A_4, D_8, Q_8 (see §17, Ex. 15), $SL(2,\mathbb{Z}_3)$, $GL(2,\mathbb{Z}_3)$, S_n, A_n (for $n > 2$).

20. Let G be a group such that $G' \leqslant Z(G)$ and let a be a fixed element of G. Prove that the mapping $\varphi: G \to G$ is a homomorphism.

$$x \to [x,a]$$

§25
Group Actions

Many of the important groups we have examined so far are groups of functions. S_X is the group of one-to-one mappings on the set X; *Isom* E is the group of distance preserving functions on the Euclidean plane, *Aut*(G) is the group of multiplication preserving functions on a group G. You will see more examples later. In general, when X is a set with some structure on it (algebraic, geometric, analytic, topological or of some other type), the mappings on X that preserve this structure form a group. Up to now, we neglected the functional character of the elements of a group they might have. In this paragraph, we consider groups whose elements can be thought of as functions on a set X. This leads to the idea of group actions.

25.1 Definition: Let G be a group and let X be a nonempty set. We say that G *acts on* X provided, for all $x \in X$ and $g \in G$, there corresponds a uniquely determined element of X, denoted by xg, such that the following hold:

$$(xg_1)g_2 = x(g_1g_2) \quad \text{for all } x \in X, g_1,g_2 \in G,$$
$$x1 = x \quad \text{for all } x \in X.$$

More precisely, we say then that G acts on X on the right. We similarly define a left action of G on X by stipulating that $(g_1g_2)x = g_1(g_2x)$ and $1x = x$ for all $x \in X$, $g_1,g_2 \in G$, where gx is a uniquely determined element of X corresponding to the pair g,x.

25.2 Examples: (a) Let X be a nonempty set and $G = S_X$. Then G acts on X when we naturally interpret xg as the image of $x \in X$ under the mapping $g \in G$. The condition $(xg_1)g_2 = x(g_1g_2)$ is satisfied for all $x \in X$ and for all $g_1,g_2 \in G$, for it is nothing else than the definition of composition of mappings. The condition $x1 = x$ holds, too, since it is the definition of the identity mapping $1 \in G$ on X. More generally, if $G \leqslant S_X$, then G acts on X.

(b) Let $X = \mathbb{R} \times \mathbb{R}$ and let $G = GL(2,\mathbb{R})$. Then G acts on X if we put

$$(x,y)\begin{pmatrix} a & b \\ c & d \end{pmatrix} = (xa + yc,\ xb + yd).$$

We have indeed $\left((x,y)\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)\begin{pmatrix} e & f \\ g & h \end{pmatrix} = (xa + yc,\ xb + yd)\begin{pmatrix} e & f \\ g & h \end{pmatrix}$

$$= ((xa + yc)e + (xb + yd)g,\ (xa + yc)f + (xb + yd)h)$$
$$= (xae + yce + xbg + ydg,\ xaf + ycf + xbh + ydh)$$

and $$(x,y)\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix}\right) = (x,y)\begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

$$= (x(ae + bg) + y(ce + dg),\ x(af + bh) + y(cf + dh))$$
$$= (xae + xbg + yce + ydg,\ xaf + xbh + ycf + ydh)$$

and so $\left((x,y)\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right)\begin{pmatrix} e & f \\ g & h \end{pmatrix} = (x,y)\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} e & f \\ g & h \end{pmatrix}\right)$ for all $(x,y) \in X$ and $\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \begin{pmatrix} e & f \\ g & h \end{pmatrix} \in G$.

One proves analogously that G acts on $Y = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} : x,y \in \mathbb{R} \right\}$ on the left when we put $\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax+by \\ cx+dy \end{pmatrix}$ for all $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G$, $\begin{pmatrix} x \\ y \end{pmatrix} \in Y$.

Clearly, the field $\mathbb{R}$ can be replaced by any field in this example.

(c) Let $X = \mathbb{Z} \times \mathbb{Z} \times \mathbb{Z}$ and $G = SL(2,\mathbb{Z}) = \left\{ \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in Mat_2(\mathbb{Z}) : \alpha\delta - \beta\gamma = 1 \right\}$.

Then G acts on X when we define $(a,b,c)\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ to be

$$(a\alpha^2 + b\alpha\gamma + c\gamma^2,\ 2a\alpha\beta + b(\alpha\delta + \beta\gamma) + 2c\gamma\delta,\ a\beta^2 + b\beta\delta + c\delta^2).$$

The verification is left to the reader.

(d) Suppose G acts on X on the left and we denote the element of X corresponding to the pair g,x $(g \in G, x \in X)$ by $g*x$. Then G acts on X on the right when we put $xg := g^{-1}*x$, because

$$(xg_1)g_2 = (g_1^{-1}*x)g_2 = g_2^{-1}*(g_1^{-1}*x) = g_2^{-1}g_1^{-1}*x = (g_1g_2)^{-1}*x = x(g_1g_2)$$

and

$$x1 = 1^{-1}*x = 1*x = x$$

for all $x \in X$, $g_1, g_2 \in G$. We could not write $xg := g*x$, for then we would get $(xg_1)g_2 = x(g_2g_1)$ instead of $(xg_1)g_2 = x(g_1g_2)$. However, if G is commutative, G acts on X on the right when we put $xg := g*x$.

(e) Let F be a nonempty subset of the Euclidean plane E. Then $Sym\ F$ acts on F, because $f\sigma \in F$ for all $f \in F$, $\sigma \in Sym\ F$ and

$$f(\sigma_1\sigma_2) = (f\sigma_1)\sigma_2$$
$$f\iota = f$$

for all $f \in F$, $\sigma_1, \sigma_2 \in Sym\ F$.

(f) Let G be a group. Then $Aut(G)$ acts on G, because $g(\alpha_1\alpha_2) = (g\alpha_1)\alpha_2$ and $g\iota = g$ for all $g \in G$ and for all $\alpha_1, \alpha_2 \in Aut(G)$.

(g) Let $\mathfrak{F}$ be the set of all nonempty subsets of the Euclidean plane E. Then S_E acts on $\mathfrak{F}$ since $(F\alpha)\beta = F(\alpha\beta)$ and $F\iota = F$ for all $F \in \mathfrak{F}$ and $\alpha, \beta \in S_E$ (Lemma 14.1).

(h) Assume that a group G acts on a set X. Then any subgroup of G also acts on X.

In the next two theorems, we shall show that any group action on a set X is essentially a homomorphism into S_X.

25.3 Theorem: *Let G act on X. For each $g \in G$, consider $x \to xg$ as a function and put $\rho_g : X \to X$. Then $\rho_g \in S_X$ and the mapping*

$$x \to xg$$

$$\rho: G \to S_X$$
$$g \to \rho_g$$

is a homomorphism (called the *permutation representation of G* corresponding to the action).

Proof: Let $g \in G$. Since G acts on X, to each $x \in X$, there corresponds a uniquely determined element xg of X. Hence $\rho_g : x \to xg$ is indeed a function from X into X.

For any $x \in X$, $g_1, g_2 \in G$, we have

$$x\rho_{g_1g_2} = x(g_1g_2) = (xg_1)g_2 = (xg_1)\rho_{g_2} = (x\rho_{g_1})\rho_{g_2} = x(\rho_{g_1}\rho_{g_2}),$$

so

$$\rho_{g_1g_2} = \rho_{g_1}\rho_{g_2}. \tag{1}$$

Furthermore, $x\rho_1 = x1 = x$ for all $x \in X$, hence

$$\rho_1 = \iota_X \in S_X. \tag{2}$$

From (1) and (2), we obtain

$$\rho_g\rho_{g^{-1}} = \rho_{gg^{-1}} = \rho_1 = \iota_X = S_X; \quad \rho_{g^{-1}}\rho_g = \rho_{g^{-1}g} = \rho_1 = \iota_X = S_X$$

and thus ρ_g is one-to-one and onto (Theorem 3.17(2)). So $\rho_g \in S_X$ for all g in G.

So we have a mapping $\rho: G \to S_X$ and it is a homomorphism by (1). □
$g \to \rho_g$

25.4 Theorem: *Let X be a nonempty set and let $\sigma: G \to S_X$ be a group homomorphism. Then G acts on X when we put*

$$xg = x(g\sigma)$$

for all $x \in X$, $g_1, g_2 \in G$. Furthermore, the permutation representation of G corresponding to this action is σ.

Proof: The proof consists in observing that σ is a homomorphism. We have

$$(xg_1)g_2 = (xg_1)(g_2\sigma) = (x(g_1\sigma))(g_2\sigma) = x((g_1\sigma)(g_2\sigma)) = x((g_1g_2)\sigma) = x(g_1g_2)$$

and

$$x1 = x(1\sigma) = x$$

for all $x \in X$, $g_1, g_2 \in G$. Here we use the fact that $1\sigma \in S_X$ is the identity element of the group S_X (Lemma 20.3(a)), which is the identity mapping on X. Thus setting $xg = x(g\sigma)$ does define a group action.

Let us find the permutation representation of G corresponding to this action. This is $\rho: G \to S_X$, where ρ_g is the mapping $x \to xg$ on G. Since

$$g \to \rho_g$$

$$x\rho_g = xg = x(g\sigma)$$

for all $x \in X$, $g \in G$, we have $\rho_g = g\sigma$ for all $g \in G$. Hence $\rho = \sigma$ by the definition of equality of mappings. □

We now show that group actions define an equivalence relation on the underlying set X. The number of elements in an equivalence class can be expressed in group theoretical terms. This gives some arithmetical information about groups.

25.5 Lemma: *Let G act on X. For any $x,y \in X$, we put $x \sim y$ if and only if there is an element $g \in G$ such that $xg = y$. Then $\sim$ is an equivalence relation on X.*

Proof: (cf. Lemma 15.7.) (i) Since $1 \in G$ and $x1 = x$ for all $x \in X$, we have $x \sim x$ for all $x \in X$. Thus $\sim$ is reflexive.

(ii) If $x,y \in X$ and $x \sim y$, then there is a $g \in G$ such that $xg = y$, so $yg^{-1} = (xg)g^{-1} = x(gg^{-1}) = x1 = x$. From $g^{-1} \in G$ and $yg^{-1} = x$, we conclude $y \sim x$. Thus $\sim$ is symmetric.

(iii) Suppose $x,y,z \in X$ and $x \sim y$, $y \sim z$. Then there are $g,h \in G$ such that $xg = y$ and $yh = z$. Then $x(gh) = (xg)h = yh = z$. From $gh \in G$ and $x(gh) = z$, we conclude $x \sim z$. Thus $\sim$ is transitive.

So $\sim$ is an equivalence relation on X. □

25.6 Definition: Let G act on X. The equivalence classes of the equivalence relation in Lemma 25.5 are called *orbits*. The equivalence class $\{xg \in X: g \in G\}$ of $x \in X$ is called the *orbit of x*.

25.7 Lemma: *Let G act on X. For $x \in X$, we write*

$$Stab_G(x) = \{g \in G : xg = x\}.$$

Then $Stab_G(x)$ is a subgroup of G (called the *stabilizer of x in G*).

Proof: The proof is a routine application of our subgroup criterion.

(i) Let $g,h \in Stab_G(x)$. Then $xg = x$ and $xh = x$. So $x(gh) = (xg)h = xh = x$, so $gh \in Stab_G(x)$. Hence $Stab_G(x)$ is closed under multiplication.

(ii) Let $g^{-1} \in Stab_G(x)$. Then $xg = x$. So $xg^{-1} = (xg)g^{-1} = x(gg^{-1}) = x1 = x$, so $g^{-1} \in Stab_G(x)$. Hence $Stab_G(x)$ is closed under the forming of inverses.

Thus $Stab_G(x) \leqslant G$. □

Stabilizers of elements in the same orbit are closely related.

25.8 Lemma: *Let G act on X. Let $x \in X$ and $g \in G$. Then*

$$Stab_G(xg) = g^{-1}Stab_G(x)g.$$

Proof: As
$$\begin{aligned} h \in Stab_G(xg) &\Leftrightarrow (xg)h = xg \\ &\Leftrightarrow x(gh) = xg \\ &\Leftrightarrow (x(gh))g^{-1} = x \\ &\Leftrightarrow x(ghg^{-1}) = x \\ &\Leftrightarrow ghg^{-1} \in Stab_G(x) \\ &\Leftrightarrow h \in g^{-1}Stab_G(x)g, \end{aligned}$$

$Stab_G(xg) = g^{-1}Stab_G(x)g$. □

The kernel of the permutation representation can be expressed in terms of the stabilizers.

25.9 Lemma: *Assume G acts on X and let $\rho: G \to S_X$ be the permutation representation. Then* $\operatorname{Ker} \rho = \bigcap_{x \in X} Stab_G(x)$.

Proof: For $g \in G$, we have $\rho: g \to \rho_g \in S_X$, where $\rho_g: x \to xg$. Hence

$$\begin{aligned} Ker\ \rho &= \{g \in G: \rho_g = 1 \in S_X\} \\ &= \{g \in G: x\rho_g = x \text{ for all } x \in X\} \\ &= \{g \in G: xg = x \text{ for all } x \in X\} \\ &= \bigcap_{x \in X} \{g \in G: xg = x\} \\ &= \bigcap_{x \in X} Stab_G(x). \end{aligned}$$

□

The following elementary counting principle has many applications.

25.10 Lemma: *Let G act on X. For any $x \in X$, we have*

$$|orbit\ of\ x| = |G:Stab_G(x)|.$$

Proof: The orbit of x is the set $\{xg \in X: g \in G\}$. The index $|G:Stab_G(x)|$ is the number of right cosets of $Stab_G(x)$ in G, more precisely, the cardinal number of $\mathcal{R} = \{Stab_G(x)g: g \in G\}$. We must find a one-to-one correspondence between the orbit $\{xg \in X: g \in G\}$ of x and the set $\mathcal{R} = \{Stab_G(x)g: g \in G\}$ of the right cosets of $Stab_G(x)$ in G. The description of these sets leads us to consider the mapping

$$\begin{aligned} \alpha: \text{orbit of } x &\to \mathcal{R}, \\ xg &\to Sg \end{aligned}$$

where we put $S = Stab_G(x)$ for brevity. Let us see if α is one-to-one and onto.

Before that, however, we must check that α is well defined, for one and the same element in the orbit of x can have representations xg, xh with $g \neq h$. We must prove that $xg = xh$ implies $Sg = Sh$. If $xg = xh$, then $x(gh^{-1}) = (xg)h^{-1} = (xh)h^{-1} = x(hh^{-1}) = x1 = x$, so $gh^{-1} \in S$ and therefore $Sg = Sh$ by Lemma 10.2(5). Thus α is well defined.

That α is one-to-one follows by reversing the argument above. If $(xg)\alpha = (xh)\alpha$, then $Sg = Sh$, then $gh^{-1} \in S$, then $x(gh^{-1}) = x$, then $(x(gh^{-1}))h = xh$, so $xg = xh$. Therefore α is one-to-one.

α is certainly onto, since any $Sg \in \mathcal{R}$ is the image of xg in the orbit of x.

Thus α is a one-to-one mapping from the orbit of x onto $\mathcal{R}$. This gives

$$|orbit\ of\ x| = |G:Stab_G(x)|. \qquad \square$$

25.11 Definition: Let G act on X. We say G *acts transitively on* X or the action of G on X is said to be a *transitive* action if, for any $x,y \in X$, there is a $g \in G$ such that $xg = y$. If G does not act transitively on X, then G is said to act *intransitively* on X.

Thus G acts transitively on X if and only if there is one and only one orbit. The whole set X is the single orbit of the action.

25.12 Examples: (a) A group G acts on itself by right multiplication: to the pair $x,g \in G$, there corresponds the product $xg \in G$. The conditions $(xg_1)g_2 = x(g_1g_2)$ and $x1 = x$ (for all $x,g_1,g_2 \in G$) are immediate from the associativity of multiplication and from the definition of the identity element. This action is transitive, because, given any $x,y \in G$, there is an element g in G, namely $g = x^{-1}y$, such that $xg = y$. Hence, for any $x \in G$, we have $|G| = |\text{orbit of } x| = |G:Stab_G(x)|$, thus $Stab_G(x) = 1$, as can be seen also from $Stab_G(x) = \{g \in G: xg = x\} = \{g \in G: g = 1\} = \{1\} = 1$. This action is called the *regular action of* G *on* G. The kernel of the permutation representation $\rho: G \to S_X$ is $Ker\ \rho = \bigcap_{x \in X} Stab_G(x) = 1$ by Lemma 25.9. Thus ρ is one-to-one and Theorem 20.16 gives $G \cong G/1 = G/Ker\ \rho \cong Im\ \rho \leqslant S_G$.

(b) The preceding example can be generalized. Let $H \leqslant G$ and let $\mathcal{R} = \{Ha: a \in G\}$ be the set of all right cosets of H in G. Then G acts on $\mathcal{R}$ by right multiplication, where, to the pair Ha, g, there corresponds the coset Hag, because

$$((Ha)g_1)g_2 = (Hag_1)g_2 = H((ag_1)g_2) = H(a(g_1g_2)) = (Ha)(g_1g_2)$$

and

$$(Ha)1 = Ha1 = Ha$$

for all $Ha \in \mathcal{R}, g_1,g_2 \in G$.

This action is transitive, because, given any $Ha, Hb \in \mathcal{R}$, there is an element g in G, namely $g = a^{-1}b$, such that $(Ha)g = Hb$.

We have $$Stab_G(H) = \{g \in G: Hg = H\} = \{g \in G: g \in H\} = H$$
and $$Stab_G(Ha) = a^{-1}Stab_G(H)a = a^{-1}Ha$$
by Lemma 25.8.

The kernel of the permutation representation $\rho: G \to S_{\mathcal{R}}$ is, by Lemma 25.9,

$$Ker\ \rho = \bigcap_{Ha \in \mathcal{R}} Stab_G(Ha) = \bigcap_{a \in G} Stab_G(Ha) = \bigcap_{a \in G} a^{-1}Ha.$$

The intersection $\bigcap_{a \in G} a^{-1}Ha$ is called the *core of H in G*, and is designated by H_G. Theorem 20.16 gives now $G/H_G = G/Ker\ \rho \cong Im\ \rho \leqslant S_{\mathcal{R}}$.

25.13 Theorem : *Let G be a group.*
(1) (Cayley's theorem) *G is isomorphic to a subgroup of S_G.*
(2) *Let $H \leqslant G$ be of index $|G:H| = n$. Then G/H_G is isomorphic to a subgroup of S_n.*

Proof: (1) This follows from Example 25.12(a).

(2) From Example 25.12(b), it follows that G/H_G is isomorphic to a subgroup of $S_{\mathcal{R}}$, where $\mathcal{R}$ is a set with n elements. Let $\mu: \mathcal{R} \to \{1,2, \ldots, n\}$ be a one-to-one mapping from $\mathcal{R}$ onto $\{1,2, \ldots, n\}$. Then, for each $f \in S_{\mathcal{R}}$, the mapping $\mu^{-1}f\mu$ is a one-to-one mapping from $\{1,2, \ldots, n\}$ onto $\{1,2, \ldots, n\}$, so $\mu^{-1}f\mu \in S_n$. Now the function

$$\begin{aligned} M: S_{\mathcal{R}} &\to S_n, \\ f &\to \mu^{-1}f\mu \end{aligned}$$

is easily verified to be a homomorphism: $fgM = \mu^{-1}fg\mu = \mu^{-1}f\mu\mu^{-1}g\mu = fMgM$ for all $f, g \in S_{\mathcal{R}}$; and M is one-to-one and onto, because the mapping

$$\begin{aligned} N: S_n &\to S_{\mathcal{R}} \\ \sigma &\to \mu\sigma\mu^{-1} \end{aligned}$$

is such that MN = identity mapping on $S_{\mathcal{R}}$ and NM = identity mapping on S_n (Theorem 3.17(2)). Hence M is an isomorphism and $S_{\mathcal{R}} \cong S_n$. Together with $G/H_G \cong S_{\mathcal{R}}$, this gives $G/H_G \cong S_n$. □

25.14 Example: Another important group action is *conjugation*. For any $x,g \in G$, we call $g^{-1}xg$ the *conjugate of x by g*. In order to avoid any confusion with right multiplication, we shall write x^g for $g^{-1}xg$. This notation is standard. Since

$$(x^{g_1})^{g_2} = (g_1^{-1}xg_1)^{g_2} = g_2^{-1}(g_1^{-1}xg_1)g_2 = g_2^{-1}g_1^{-1}xg_1g_2 = (g_1g_2)^{-1}x(g_1g_2) = x^{(g_1g_2)}$$

and

$$x^1 = 1^{-1}x1 = x$$

for all $x,g_1,g_2 \in G$, conjugation is indeed an action of G on G.

The orbit $\{x^g: g \in G\} = \{g^{-1}xg: g \in G\}$ of $x \in G$ is called the *conjugacy class* of x. We have

$$Stab_G(x) = \{g \in G: x^g = x\} = \{g \in G: g^{-1}xg = x\} = \{g \in G: xg = gx\};$$

so $Stab_G(x)$ consists of the all those elements in G which commute with x. It is called the *centralizer of x in G* in this case and is denoted by $C_G(x)$.

The permutation representation is $\tau: G \to S_G$, where $\tau_g: G \to G$. Hence τ_g is

$$g \to \tau_g \qquad x \to x^g$$

the inner automorphism of G induced by g. We get

$$Ker\ \tau = \bigcap_{x \in G} C_G(x) = \bigcap_{x \in G} \{g \in G: xg = gx\} = \{g \in G: xg = gx \text{ for all } x \in G\} = Z(G)$$

as we know also from the proof of Theorem 23.10. In this case, Lemma 25.10 assumes the following form.

25.15 Lemma: *Let G be a group and $x \in G$. Then*

$$|\text{conjugacy class of } x| = |G:C_G(x)|. \qquad \square$$

25.16 Lemma (Class equation): *Let G be a finite group. Assume G has k distinct conjugacy classes and let $x_1,x_2,\ldots,x_k$ be representatives of these classes. Then*

$$|G| = \sum_{i=1}^{k} |G:C_G(x_i)|.$$

Proof: Conjugacy is an equivalence relation on G and gives rise to a partition of G (Theorem 2.5):

$$G = \bigcup_{i=1}^{k} \text{conjugacy class of } x_i,$$

the union being disjoint. Counting the number of elements on both sides, and using Lemma 25.15, we obtain

$$|G| = \sum_{i=1}^{k} |\text{conjugacy class of } x_i| = \sum_{i=1}^{k} |G:C_G(x_i)|. \qquad \square$$

We give an important application of the class equation.

25.17 Theorem: *Let G be a group of order p^n, where p is prime and n is a natural number. Then $Z(G) \neq 1$.*

Proof: Let k be the number of conjugacy classes in G, and let $x_1, x_2, \ldots, x_k$ be representatives of these classes. Then, in the class equation

$$|G| = \sum_{i=1}^{k} |G:C_G(x_i)|,$$

each summand on the right hand side is a divisor of p^n by Lagrange's theorem. So $|G:C_G(x_i)| = p^{m_i}$ with suitable nonnegative integers m_i (for each $i = 1,2,\ldots,k$). Thus the class equation is

$$p^n = p^{m_1} + p^{m_2} + \cdots + p^{m_k}.$$

Here $p^{m_i} = 1$ if and only if $|G:C_G(x_i)| = 1$, so if and only if $C_G(x_i) = G$, and so if and only if $x_i \in Z(G)$. Thus exactly $|Z(G)|$ summands on the right hand side are equal to 1, and the class equation gives

$$p^n = |Z(G)| + (\text{a sum of powers of } p \text{ greater than } p^0 = 1)$$

(The second term is absent in case $|G| = |Z(G)|$; in this case $Z(G) = G \neq 1$). The last equation tells us that $|Z(G)|$ is divisible by p, so $|Z(G)| \neq 1$, hence $Z(G) \neq 1$. $\square$

25.18 Lemma: *Let p be a prime number. If G is a group of order p^2, then G is abelian.*

Proof: We must show $Z(G) = G$, or, equivalently, $|Z(G)| = p^2$. We know $|Z(G)| = 1$ or p or p^2 by Lagrange's theorem, and $|Z(G)| \neq 1$ by Theorem 25.17. We suppose, by way of contradiction, that $|Z(G)| = p$. Since $Z(G) \trianglelefteq G$ (Theorem 23.3), we can build the factor group $G/Z(G)$, which has order $p^2/p = p$ and which is therefore cyclic by Theorem 11.13. Then G must be abelian by Lemma 23.5, and $|Z(G)| = p^2$, contrary to the assumption $|Z(G)| = p$. Thus $|Z(G)| = p$ is impossible and there remains only the possibility $|Z(G)| = p^2$. Hence G is abelian. □

We wish to present the basic idea in the proof of Theorem 25.17 in its purest form. We need a definition.

25.19 Definition: Let G act on X. If $x \in X$, $g \in G$ and $xg = x$, we say that g *fixes* x. The set

$$\{x \in X: xg = x \text{ for all } g \in G\} = \{x \in X: Stab_G(x) = G\}$$

of all elements in X which are fixed by each element of G is called the *fixed point subset of* X and denoted by $Fix_X(G)$.

Thus $Fix_X(G)$ consists of all those elements in X which form an orbit with only one element in it. When we count the number of elements in X as the sum of the number of elements in each orbit, each element in $Fix_X(G)$ contributes 1 to this sum. Notice that, under the action of a group G on itself by conjugation, $Fix_G(G)$ is nothing else than $Z(G)$.

25.20 Lemma: *Let G act on X. If G has order p^n, where p is a prime number and $n \in \mathbb{N}$, and X is a finite set, then*

$$|X| \equiv |Fix_X(G)| \pmod{p}.$$

Proof: We consider the equivalence relation $\sim$ of Lemma 25.5 on X. Under this equivalence relation, X is partitioned into finitely many disjoint orbits, say

$$X = \bigcup_{i=1}^{k} \text{orbit of } x_i.$$

Counting the number of elements on both sides, we get

$$|X| = \sum_{i=1}^{k} |\text{orbit of } x_i|.$$

Hence, by Lemma 25.10, $|X| = \sum_{i=1}^{k} |G:Stab_G(x_i)|$.

Now each of the indices $|G:Stab_G(x_i)|$ is a divisor of $|G| = p^n$, hence is equal to some power p^{m_i} of p with a nonnegative integer m_i. Here $p^{m_i} = p^0 = 1$ if and only if $G = Stab_G(x_i)$, that is to say, if and only if $x_i \in Fix_X(G)$. Thus there are exactly $|Fix_X(G)|$ summands equal to 1, and the sum above becomes

$$|X| = (\underbrace{1 + 1 + \cdots + 1}_{|Fix_X(G)| \text{ times}}) + (\text{sum of } p^{m_i} \text{ with } m_i > 0)$$

(the second term is missing in case there is no p^{m_i} with $m_i > 0$). So

$$|X| = |Fix_X(G)| + (\text{a number divisible by } p)$$

and therefore $|X| \equiv |Fix_X(G)| \pmod{p}$, as was to be proved. □

We end this paragraph with a generalization of conjugation.

25.21 Example: Let G be a group and let $\mathfrak{S}$ be the set of all nonempty subsets of G. For any $U \in \mathfrak{S}$ and $g \in G$, we put

$$U^g = \{u^g \in G: u \in U\} = \{g^{-1}ug \in G: u \in U\} = g^{-1}Ug.$$

U^g consists therefore of conjugates by g of the elements of U and is called the *conjugate of U by g*. With this definition, G acts on $\mathfrak{S}$, because

$$(U^{g_1})^{g_2} = \{u^{g_1} \in G: u \in U\}^{g_2} = \{(u^{g_1})^{g_2} \in G: u \in U\} = \{u^{(g_1g_2)} \in G: u \in U\} = U^{(g_1g_2)}$$

and

$$U^1 = 1^{-1}U1 = U$$

for all $U \in \mathfrak{S}, g_1, g_2 \in G$.

The orbit $\{U^g: g \in G\} = \{g^{-1}Ug: g \in G\}$ of $U \in \mathfrak{X}$ is called the *conjugacy class* of U. We have

$$Stab_G(U) = \{g \in G: U^g = U\} = \{g \in G: g^{-1}Ug = U\} = \{g \in G: Ug = gU\};$$

so $Stab_G(U)$ consists of the all those elements in G which fix U as a set. It is called the *normalizer of U in G* in this case and is denoted by $N_G(U)$. The set

$$\{g \in G: u^g = u \text{ for all } u \in U\}$$
$$= \{g \in G: g^{-1}ug = u \text{ for all } u \in U\} = \{g \in G: ug = gu \text{ for all } u \in U\}$$

of all those elements in G which fix each element of U under conjugation, or, what is the same, which commute with every element of U, is called the *centralizer of U in G* and is denoted by $C_G(U)$. So $C_G(U)$ is the intersection of the centralizers of the elements of U:

$$C_G(U) = \bigcap_{u \in U} C_G(u).$$

In particular, $C_G(U)$ is a subgroup of G. We have $C_G(U) \leqslant N_G(U) \leqslant G$.

The orbit of $U \in \mathfrak{X}$ is

$$\{U^g \in \mathfrak{X}: g \in G\} = \{g^{-1}Ug \in \mathfrak{X}: g \in G\}$$

and is called the *conjugacy class of U in G*. We have

$$|\text{conjugacy class of } U| = |G:N_G(U)|$$

by Lemma 25.10.

In general, U neither contains nor is contained in $C_G(U)$ or $N_G(U)$. However, if U happens to be a subgroup of G, we have $g^{-1}ug \in U$ for all u,g in U, so $U^g = \{u^g \in G: u \in U\} = \{g^{-1}ug \in G: u \in U\} \subseteq U$ and, for any g in U, we get $U \subseteq (U^{g^{-1}})^g \subseteq U^g \subseteq U$, thus $U^g = U$ and $U \leqslant N_G(U)$.

We collect the last two remarks in a theorem.

25.22 Theorem: *Let G be a group and let H be a subgroup of G. Then*

$$H \leqslant N_G(H) \leqslant G \text{ and } |\text{conjugacy class of } H| = |G:N_G(H)|. \quad \square$$

Exercises

1. Prove that $SL(2,\mathbb{Z})$ acts on $X := \left\{ \begin{pmatrix} a & b/2 \\ b/2 & c \end{pmatrix} : a,b,c \in \mathbb{Z} \right\}$ when we associate the matrix $g^t x g$ with the pair $(x,g) \in X \times SL(2,\mathbb{Z})$.

2. Let G act on X and let H be a subgroup of G. Show that

$$Stab_H(x) = Stab_G(x) \cap H$$

for any $x \in X$.

3. Let G act on X and H act on Y. Prove that the direct product $G \times H$ acts on the cartesian product $X \times Y$.

4. Give an examples of groups G and subsets U of G such that $U \not\subseteq N_G(U)$, $N_G(U) \not\subseteq U$, $U \not\subseteq C_G(U)$, $C_G(U) \not\subseteq U$.

5. Prove that $C_G(U) \trianglelefteq N_G(U)$ for any nonempty subset U of a group G.

6. Let $H \leqslant G$. Show that $N_G(H)$ acts on H by conjugation. Considering the permutation representation of this action, prove that $N_G(H)/C_G(H)$ is isomorphic to a subgroup of $Aut(H)$.

7. Assume G acts on X, and let K be the kernel of the permutation representation of this action. Suppose $H \trianglelefteq G$ and $H \leqslant K$. Show that G/H acts on X when we put $x(Hg) = xg$ for all $x \in X$, $Hg \in G/H$. What is the kernel of the permutation representation?

§26
Sylow's Theorem

Let G be a finite group. Lagrange's theorem asserts that, if G has a subgroup of order k, then k is a divisor of $|G|$. The converse of Lagrange's theorem, like the converses of many theorems, is wrong. If k is a divisor of $|G|$, then G need not have a subgroup of order k. For instance, A_4 has order 12, 12 is divisible by 6, yet A_4 has no subgroups of order 6.

The converse of Lagrange's theorem becomes true if we impose the additional condition that k be a prime power such that k and $|G|/k$ are relatively prime. In other words, if $|G| = p^a m$, where p is a prime number and $p \nmid m$, then G does have a subgroup H of order p^a. Then any conjugate H^g of H, too, is a subgroup of order p^a and the question arises as to whether G has subgroups of order p^a other than the conjugates of H. The answer turns out to be negative. The conjugates of H are the only subgroups of order p^a.

This theorem was proved by the Norwegian mathematician L. Sylow in 1872. It is a very important tool in the theory of finite groups. We present here a very elegant proof due to H. Wielandt (1959).

26.1 Theorem (Sylow's Theorem): *Let G be a finite group of order $|G| = p^a m$, where p is a prime number and $p \nmid m$ (that is, let p^a be the highest power of p dividing $|G|$). Then the following assertions hold.*

(1) *G has a subgroup H of order p^a.*
(2) *If J is any subgroup of G whose order $|J|$ is a power of p, then there is an $x \in G$ such that $J \leqslant H^x$.*
(3) *If n_p denotes the number of subgroups of order p^a, then $n_p | m$ and $n_p \equiv 1 \pmod{p}$.*

Some remarks will now be in order. If $p^a || G|$ and $p^{a+1} \nmid |G|$, then a subgroup of G of order p^a is called a *Sylow p-subgroup of G*. Part (1) of Sy-

low's theorem states that every finite group has a Sylow p-subgroup, for all prime numbers p.

If H is a Sylow p-subgroup of G, so is H^g for any $g \in G$. Part (2) of Sylow's theorem states that any subgroup of p-power-order of G is a subgroup of a suitable conjugate of H. In particular, any Sylow p-subgroup of G is contained in a suitable H^x for some $x \in G$, and, since the orders of that Sylow p-subgroup and of H^x coincide, that Sylow p-subgroup must be H^x itself. So any Sylow p-subgroup of G is a conjugate of H.

If a Sylow p-subgroup H of G is normal in G, then all conjugates of H are equal to H, hence H is the unique Sylow p-subgroup of G. Then, for any automorphism α of G, $H\alpha$ is a subgroup of order p^a, and therefore is equal to H. So H is in fact a characteristic subgroup of G in this case.

Part (3) of Sylow's theorem gives us arithmetical information about the possible number Sylow p-subgroups. Two applications of this is given in Lemma 26.5 and in Lemma 26.6.

Proof of Sylow's theorem: The basic idea of the proof is as follows. If there is a Sylow p-subgroup H of G, then H is first of all a *subset* of G having exactly p^a elements and is furthermore such that $Hh = H$ for all $h \in H$. So $H = \{h \in G: Uh = U\}$ for some subset U of G with $|U| = p^a$. In order to find a subgroup of order p^a, so we look at the sets $\{h \in G: Uh = U\}$, for each $U \subseteq G$ with $|U| = p^a$. Such sets are the stabilizers of U's under the group action described below. A juidicious choice of U will produce a subgroup of order p^a.

Step 1. Let $\mathcal{Y} = \{U \subseteq G: |U| = p^a\}$. Then the number $|\mathcal{Y}|$ of elements of $\mathcal{Y}$ (= subsets of G in $\mathcal{Y}$) is not divisible by p:

There are clearly $\binom{p^a m}{p^a}$ subsets of G in $\mathcal{Y}$. We are to prove $p \nmid \binom{p^a m}{p^a}$.

We have $\binom{p^a m}{p^a} = \frac{(p^a m)!}{p^a!(p^a m - p^a)!} = \frac{p^a m}{p^a}\frac{p^a m - 1}{p^a - 1}\frac{p^a m - 2}{p^a - 2} \cdots \frac{p^a m - (p^a - 1)}{1}$. Now consider each one of the factors $\frac{p^a m - s}{p^a - s} (s = 1,2, \ldots, p^a - 1)$. We write $s = p^b t$,

with $t \in \mathbb{Z}$ and $p \nmid t$, and observe that neither the numerator nor the denominator of these numbers

$$\frac{p^a m - s}{p^a - s} = \frac{p^a m - p^b t}{p^a - p^b t} = \frac{p^{a-b} m - t}{p^{a-b} - t}$$

contain p after cancellations are made. Hence their product $\binom{p^a m}{p^a}$ is not divisible by p.

As an example, note that all 3's are cancelled in

$$\binom{18}{9} = \binom{3^2 2}{3^2} = \frac{18}{9}\frac{17}{8}\frac{16}{7}\frac{15}{6}\frac{14}{5}\frac{13}{4}\frac{12}{3}\frac{11}{2}\frac{10}{1} = \frac{2}{1}\frac{17}{8}\frac{16}{7}\frac{5}{2}\frac{14}{5}\frac{13}{4}\frac{4}{1}\frac{11}{2}\frac{10}{1}.$$

Step 2. G acts on $\mathcal{Y}$ when we put $Ug = \{ug: u \in U\}$ for $U \in \mathcal{Y}$, $g \in G$:

The mapping $\begin{matrix} U \to Ug \\ u \to ug \end{matrix}$ is one-to-one (Lemma 8.1(2)) and onto (by definition of Ug). Hence $|Ug| = |U| = p^a$ and Ug is an element of $\mathcal{Y}$. Now $(Ug_1)g_2 = U(g_1 g_2)$ for all $U \in \mathcal{Y}$, $g_1, g_2 \in G$ by Lemma 19.2 and also $U1 = \{u1: u \in U\} = \{u: u \in U\} = U$ for all $U \in \mathcal{Y}$. Thus G acts on $\mathcal{Y}$.

Step 3. There is an orbit of $\mathcal{Y}$ under the action of Step 2 such that the number of elements (of $\mathcal{Y}$; equivalently, the number of subsets of G) in it is not divisible by p:

The orbit of any $U \in \mathcal{Y}$ is $\{Ug \in \mathcal{Y}: g \in G\}$. Now $\mathcal{Y}$ is partitioned into disjoint orbits. If $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_k$ are the orbits, then

$$\mathcal{Y} = \mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_k.$$

Counting the number of elements and keeping in mind that the orbits are pairwise disjoint, we get

$$|\mathcal{Y}| = |\mathcal{Y}_1| + |\mathcal{Y}_2| + \cdots + |\mathcal{Y}_k|.$$

If $|\mathcal{Y}_1|, |\mathcal{Y}_2|, \ldots, |\mathcal{Y}_k|$ were all divisible by p, their sum $|\mathcal{Y}|$ would be divisible by p, too, contrary to Step 1. Thus at least one of the numbers $|\mathcal{Y}_1|, |\mathcal{Y}_2|, \ldots, |\mathcal{Y}_k|$ is not divisible by p, as contended.

Let $U_0 \in \mathcal{Y}$ be such that the number of elements (of $\mathcal{Y}$) in its orbit is not divisible by p. This is the juidicious choice we have alluded to. We put $H = Stab_G(U_0)$.

Step 4. $H \leqslant G$ and $|H| = p^a$:

H is a subgroup of G by Lemma 25.7. As to the second assertion, first we note that the orbit of U_0 is equal to $|G:H|$ (Lemma 25.10) and, by the choice of U_0, this index $|G:H|$ is not divisible by p. So $p \nmid |G|/|H|$, so $p \nmid p^a m/|H|$. Writing $|H| = p^b n$, where $n \in \mathbb{N}$, $p \nmid n$ and, by Lagrange's theorem, $b \leqslant a$ and $n|m$, we get $p \nmid p^{a-b} m/n$. This is possible only in case $p^{a-b} = p^0$. Hence $a = b$ and $|H| = p^a n \geqslant p^a$. On the other hand, if $U_0 = \{u_1, u_2, \ldots, u_{p^a}\}$, then, for any $h \in H = Stab_G(U_0)$, we have

$$\begin{aligned} u_1 h \in U_0 h = U_0 &= \{u_1, u_2, \ldots, u_{p^a}\} \\ h \in \{u_1^{-1}u_1, u_1^{-1}u_2, \ldots, u_1^{-1}u_{p^a}\} \\ H \subseteq \{u_1^{-1}u_1, u_1^{-1}u_2, \ldots, u_1^{-1}u_{p^a}\} \\ |H| \leqslant p^a. \end{aligned}$$

From $|H| \geqslant p^a$ and $|H| \leqslant p^a$, we get $|H| = p^a$.

By Step 4, H is a Sylow p-subgroup of G. This completes the proof of part (1). We proceed to the proof of part (2). Let $J \leqslant G$ be such that $|J| = p^b$, where $b \geqslant 0$.

Step 5. There is an $x \in G$ such that $J \leqslant H^x$:

Let $\mathcal{R} = \{Ha : a \in G\}$ be the set of all right cosets of H in G. Then G acts on $\mathcal{R}$ by right multiplication (Example 25.12(b)) and its subgroup J also acts on $\mathcal{R}$. Since the order of J is a power of p, we can apply Lemma 25.20 and conclude

$$|\mathcal{R}| \equiv |Fix_{\mathcal{R}}(J)| \pmod{p},$$

hence

$$|Fix_{\mathcal{R}}(J)| \equiv |\mathcal{R}| = |G:H| = m \not\equiv 0 \pmod{p}$$

$$|Fix_{\mathcal{R}}(J)| \neq 0$$

$$Fix_{\mathcal{R}}(J) \neq \emptyset.$$

So there is a right coset Hx in $Fix_{\mathcal{R}}(J)$. Thus $Stab_J(Hx) = J$. But $Stab_J(Hx) = J \cap Stab_G(Hx) = J \cap H^x$ by Example 25.12(b). So we obtain $J \cap H^x = J$, which means $J \leqslant H^x$.

This completes the proof of part (2). In view of the remarks preceding the proof, all Sylow p-subgroups of G are conjugate; and a normal Sylow p-subgroup of a finite group is the unique Sylow p-subgroup of that group.

Let $N := N_G(H) = \{g \in G: H^g = H\}$ be the normalizer of H in G. Then $H \trianglelefteq N$, $N \leqslant G$ and, since $p \nmid |N:H|$, H is a Sylow p-subgroup of N. Thus H is the unique Sylow p-subgroup of N.

We now prove part (3). Let n_p be the number of Sylow p-subgroups of G.

Step 6. $n_p = |G:N|$:

Let $\mathfrak{S} = \{H^x \leqslant G: x \in G\}$. Then $\mathfrak{S}$ is the set of all Sylow p-subgroups of G. We want to evaluate $n_p = |\mathfrak{S}|$. Here G acts on $\mathfrak{S}$ by conjugation, because $(H^x)^g = H^{xg} \in \mathfrak{S}$; $(H^{g_1})^{g_2} = H^{(g_1 g_2)}$; and $H^1 = 1^{-1}H1 = H$ for all $H \in \mathfrak{S}, g_1, g_2 \in G$. Lemma 25.10 gives now

$$|\text{orbit of } H| = |G:Stab_G(H)|.$$

But the orbit of $H = \{H^x \leqslant G: x \in G\} = \mathfrak{S}$ and $Stab_G(H) = N_G(H) = N$. Thus

$$n_p = |\mathfrak{S}| = |G:N|,$$

as was to be proved.

Step 7. $n_p \mid m$ and $n_p \equiv 1 \pmod{p}$:

Of course $n_p = |G:N|$ divides $|G:N||N:H| = |G:H| = m$.

Now we want to prove $n_p \equiv 1 \pmod{p}$. This will be done by applying Lemma 25.20. In order to apply Lemma 25.20, we need the action of a group of p-power order on a finite set. Our group of p-power order will be H, as this is the only group of p-power order available to us. H acts on $\mathcal{R}_1 = \{Na: a \in G\}$ the set of all right cosets of N in G by right multiplication (Example 25.12(b), Example 25.2(h)). Lemma 25.20 yields

$$|\mathcal{R}_1| \equiv |Fix_{\mathcal{R}_1}(H)| \pmod{p}.$$

Since $n_p = |G:N| = |\mathcal{R}_1|$, the claim will be established when we show that $|Fix_{\mathcal{R}_1}(H)| = 1$.

From the equivalences

$$\begin{aligned}
Na \in Fix_{\mathcal{R}_1}(H) &\Leftrightarrow Stab_H(Na) = H \\
&\Leftrightarrow (Na)h = Na \text{ for all } h \in H \\
&\Leftrightarrow Naha^{-1} = N \text{ for all } h \in H \\
&\Leftrightarrow aha^{-1} \in N \text{ for all } h \in H \\
&\Leftrightarrow h^{a^{-1}} \in N \text{ for all } h \in H \\
&\Leftrightarrow H^{a^{-1}} \subseteq N \\
&\Leftrightarrow H^{a^{-1}} \text{ is a Sylow } p\text{-subgroup of } N \\
&\Leftrightarrow H^{a^{-1}} \text{ the unique Sylow } p\text{-subgroup } H \text{ of } N \\
&\Leftrightarrow H^{a^{-1}} = H \\
&\Leftrightarrow a^{-1} \in N_G(H) = N \\
&\Leftrightarrow a \in N \\
&\Leftrightarrow Na = N,
\end{aligned}$$

it follows that $Fix_{\mathcal{R}_1}(H) = \{N\}$. Thus $|Fix_{\mathcal{R}_1}(H)| = 1$ and $n_p \equiv 1 \pmod{p}$.

This completes the proof. □

26.2 Definition: Let p be a prime number. A finite group G is called a *finite p-group* if $|G| = p^a$ for some integer $a \geqslant 0$.

26.3 Theorem: *Let G be a finite p-group, with $|G| = p^a > 1$.*
(1) G has a normal subgroup of order p.
(2) There are normal subgroups H_i of G such that $|H_i| = p^i$ $(i = 0,1,2, \ldots, a)$ and $1 = H_0 \leqslant H_1 \leqslant H_2 \leqslant \cdots \leqslant H_{a-1} \leqslant H_a = G.$

Proof: (1) From Theorem 25.17, we know $Z(G) \neq 1$. Let $z \in Z(G)$ with $z \neq 1$, and let $o(z) = p^k$ $(1 \leqslant k \leqslant a)$. Then $o(z^{p^{k-1}}) = p$. Thus $\langle z^{p^{k-1}} \rangle$ is a subgroup of order p and is normal in G (Theorem 23.3).

(2) We make induction on a. If $a = 1$, then $|G| = p$ and G has normal subgroups H_0 and H_1, namely $H_0 = 1$ and $H_1 = G$, with $|H_0| = 1$ and $|H_1| = p$ such that $H_0 \leqslant H_1$.

Assume now that $a \geqslant 2$ and that the claim is true for any finite p-group of order p^{a-1}. By part (1), there is $H_1 \trianglelefteq G$ with $|H_1| = p$. We consider the factor group G/H_1, which has order $|G/H_1| = |G|/|H_1| = p^a/p = p^{a-1}$. By induction, there are normal subgroups, say H_{i+1}/H_1, of G/H_1 with $|H_{i+1}/H_1| = p^i$ $(i = 0,1, \ldots,a-1)$ and

$$1 = H_1/H_1 \leqslant H_2/H_1 \leqslant H_3/H_1 \leqslant \cdots \leqslant H_{a-1}/H_1 \leqslant H_a/H_1 = G/H_1.$$

By Theorem 21.2, each $H_i \trianglelefteq G$ $(i = 1,2, \ldots,a)$ and

$$H_1 \leqslant H_2 \leqslant \cdots \leqslant H_{a-1} \leqslant H_a = G.$$

Here $|H_{i+1}| = |H_{i+1}/H_1||H_1| = p^i p = p^{i+1}$ for $i = 0,1, \ldots,a-1$. Thus, when we put $H_0 = 1$, the claim is proved for finite p-groups of order p^a. □

26.4. Theorem: *Let G be a finite group and let p be a prime number. Suppose $p^b \| |G|$, where $b \geqslant 0$. Then G has a subgroup of order p^b.*

Proof: Let us write $|G| = p^a m$, with $m \in \mathbb{N}$ and $p \nmid m$. Then G has a Sylow p-subgroup H of order p^a, and, by Theorem 26.3(2), H has a subgroup J of order p^b. Hence J is a subgroup of G with $|J| = p^b$. □

Theorem 26.4 generalizes Sylow's theorem (1) to the case where p^b is any prime power divisor of $|G|$ (not necessarily the highest power of p dividing $|G|$). Part (2) of Sylow's theorem does not generalize: two subgroups J_1 and J_2, of the same order p^b, are not necessarily conjugate in G, or even isomorphic. Part (3) of Sylow's theorem, however, is true in the more general case: if $p^b \| |G|$, then the number of subgroups of order p^b in G is congruent to 1 modulo p.

We close this paragraph with two applications of Sylow's theorem.

26.5 Lemma: *Let p and q be distinct prime numbers and let G be a group of order pq. Then either a Sylow p-subgroup or a Sylow q-subgroup of G is normal in G. In fact, if $p > q$, then a Sylow p-subgroup of G is normal in G.*

Proof: Suppose $p > q$ and let n_p be the number of Sylow p-subgroups of G. Then n_p divides $|G|/p = q$, so $n_p = 1$ or q, and $n_p \equiv 1 \pmod{p}$. So $n_p = q$ implies $p \mid q-1$, which is not compatible with $p > q$. Thus $n_p = q$ is impossible and $n_p = 1$. Then there is a unique Sylow p-subgroup of G, and it is normal in G. □

26.6 Lemma: *Let p and q be distinct prime numbers and let G be a group of order p^2q. Then either a Sylow p-subgroup or a Sylow q-subgroup of G is normal in G.*

Proof: Let n_p, n_q be the number of Sylow p and Sylow q-subgroups of G, respectively. The claim is that either $n_p = 1$ or $n_q = 1$. Suppose, by way of contradiction, that $n_p > 1$ and $n_q > 1$.

Since n_p divides $|G|/p^2 = q$, and since q is prime, we have $n_p = q$. From $q = n_p \equiv 1 \pmod{p}$, we get $p \leq q - 1$, so $q > p$. Besides, n_q divides $|G|/q = p^2$, so $n_q = p$ or p^2. Here $n_q = p$ is impossible, because $n_q \equiv 1 \pmod{q}$ and $q > p$. Thus $n_q = p^2$.

Let $Q_1, Q_2, \ldots, Q_{p^2}$ be the Sylow q-subgroups of G. An element of order q is a nonidentity element in one of these subgroups, and any two distinct of them have a trivial intersection: $Q_i \cap Q_j = 1$. Hence

$$\{g \in G: o(g) = q\} = (Q_1 \setminus \{1\}) \cup (Q_2 \setminus \{1\}) \cup \ldots \cup (Q_{p^2} \setminus \{1\}),$$

where the union is taken over pairwise disjoint sets. Counting the number of elements on the right hand side, we see that there are exactly $p^2(q-1)$ elements of order q in G. So there are exactly $|G| - p^2(q-1) = p^2$ elements in

$$G \setminus \{g \in G: o(g) = q\} = \{g \in G: o(g) \neq q\}.$$

Let P be a Sylow p-subgroup of G. Then $P \subseteq \{g \in G: o(g) \neq q\}$, and, since both of these sets have p^2 elements, we have $P = \{g \in G: o(g) \neq q\}$.

Therefore $\{g \in G: o(g) \neq q\}$ is the unique Sylow p-subgroup of G and n_p is equal to 1, a contradiction. So G has either a normal Sylow p-subgroup or a normal Sylow q-subgroup. □

Exercises

1. Find Sylow 2- and Sylow 3-subgroups of S_4, A_4, $SL(2,\mathbb{Z}_3)$, $GL(2,\mathbb{Z}_3)$.

2. Find a Sylow p-subgroup of D_{2n} $(n \in \mathbb{N})$.

3. Let G be a finite group with exactly one Sylow p-subgroup. Prove that every subgroup and every factor group of G, too, has exactly one Sylow p-subgroup.

4. Let G be a finite group and $K \trianglelefteq G$. If P is a Sylow p-subgroup of G, show that $P \cap K$ is a Sylow p-subgroup of K and PK/K is a Sylow p-subgroup of G/K.

5. Let G be a finite group and $H \leqslant G$. Show that, if P is a Sylow p-subgroup of G, then $P \cap H$ is not necessarily a Sylow p-subgroup of H.

6. Let G be a finite group, $H \leqslant G$ and let P_1 be a Sylow p-subgroup of H. Show that there is a Sylow p-subgroup P of G such that $P \cap H = P_1$.

7. Let $P \leqslant K \trianglelefteq G$, where G is a finite group and P is a Sylow p-subgroup of K. Show that $G = N_G(P)K$.

8. Let G be a finite group and $H, J \leqslant G$. Suppose J is a finite p-group and $|H| \not\equiv 1 \pmod{p}$, where p is a prime number. Prove that $H \cap C_G(J) \neq 1$.

9. Let G be a finite p-group. Show that, if $1 \neq H \trianglelefteq G$, then $H \cap Z(G) \neq 1$.

10. Let G be a finite p-group and $H < G$. Prove that $H < N_G(H)$.

11. Let p,q,r be distinct prime numbers and let G be a group of order pqr. Show that G has a nontrivial proper normal subgroup.

12. Let G be a finite p-group, with $|G| = p^a > 1$, and let $K \leqslant G$. Prove that there are subgroups H_i of G such that

(i) $|H_i| = p^i$ for all $i = 0,1,2, \ldots,a$,

(ii) $K = H_j$ for some $i = 0,1,2, \ldots,a$,

(iii) $1 = H_0 \lhd H_1 \lhd H_2 \lhd \cdots \lhd H_{a-1} \lhd H_a = G$.

§27
Series

In this paragraph, we study series of groups. The celebrated Jordan-Hölder theorem is proved and the class of solvable groups is introduced.

27.1 Definition: A nontrivial group G is called a *simple* group if G has no nontrivial proper normal subgroup.

Thus a group G is simple if and only if $G \neq 1$ and 1 and G are the only normal subgroups of G. This resembles the definition of prime numbers. Just as prime numbers are the building blocks of integers, simple groups are the building blocks of certain groups, as will be seen below. Moreover, the fundamental theorem of arithmetic has a counterpart, namely the Jordan-Hölder theorem. This theorem states that, for a class of groups G satisfying certain conditions that will be specified later, the building blocks of G are uniquely determined. However, this analogy should not be pushed too far. For one thing, the building blocks may be combined in various ways to produce different groups. Stated otherwise, different groups may have the same building blocks. In fact, the problem of determining a group from its building blocks, known as the extension problem, still awaits its solution.

It is an easy matter to find all *abelian* simple groups. Any subgroup of an abelian group is normal in that group, so an abelian group is simple if and only if it has no subgroups except 1 and itself. If G is an abelian simple group, then $G \neq 1$ by definition, and so there is an $x \in G, x \neq 1$. Then $\langle x \rangle$ is a nontrivial subgroup of G and, since G is simple, $\langle x \rangle$ has to be G. Thus $G = \langle x \rangle$ is cyclic. Now an infinite cyclic group has subgroups of every index (Lemma 11.11) and cannot be simple. Therefore G is a finite cyclic group, say $|G| = n > 1$. Then, for every positive divisor m of n, the group G has a subgroup of order m (Lemma 11.10). But the order of any subgroup of G is either 1 or n. Hence 1 and n are the only positive divisors of n and n is prime. Thus an abelian simple group is a cyclic group

of prime order. Conversely, a cyclic group of prime order has no non-trivial proper subgroup by Lagrange's theorem, and is therefore an abelian simple group. We proved the following theorem.

27.2 Theorem: *An abelian group G is simple if and only if G is cyclic of prime order.* □

We prove next that the alternating groups A_n, where $n \geqslant 5$, are simple. We need a lemma. Let us recall that a 3-cycle is a permutation of the form (abc), $a \neq b \neq c \neq a$.

27.3 Lemma: *If $n \geqslant 3$, then A_n is generated by the set of all 3-cycles in A_n.*

Proof: We must prove that every element of A_n can be written as a product of 3-cycles (Lemma 24.2). Every element of A_n can be written as a product of an even number of transpositions and, taking the transpositions in pairs, we see that every element of A_n can be written as a product of permutations of the form $(ab)(cd)$, where $a \neq b$ and $c \neq d$. Hence it suffices to prove that every permutations of the form $(ab)(cd)$ can be written as a product of 3-cycles.

There are three cases to consider, in which two or one or none of c,d is in the set $\{a,b\}$. In the first case, $\{c,d\} = \{a,b\}$, hence $(cd) = (ab)$ and $(ab)(cd) = (ab)(ab) = \iota = (abe)(abe)(abe)$ is a product of 3-cycles, where e is distinct from a and b (here we use the assumption $n \geqslant 3$). In the second case, we may assume $c = a$ without loss of generality. Then $(ab)(cd) = (ab)(ad) = (abd)$ is a product of one 3-cycle. In the third case, a,b,c,d are all distinct and $(ab)(cd) = (abc)(adc)$ is a product of two 3-cycles. The proof is complete. □

27.4 Theorem: *If $n \geqslant 5$, then A_n is simple.*

Proof: Let $1 \leqslant N \lhd A_n$. We will prove $N = 1$.

First we prove that there can be no 3-cycle in N. Assume, by way of contradiction, that there is a 3-cycle (abc) in N. Let $(a'b'c')$ be any 3-cycle and choose two distinct numbers e,f from $\{1,2,\ldots,n\}\backslash\{a',b',c'\}$. This is possible because $n \geqslant 5$. Let σ be a permutation in S_n such that $a\sigma = a'$, $b\sigma = b'$ and $c\sigma = c'$ and put $\pi = \sigma(ef)$. Then $a\pi = a'$, $b\pi = b'$ and $c\pi = c'$ as well and $\sigma^{-1}(abc)\sigma = (a'b'c') = \pi^{-1}(abc)\pi$. Since the signs of σ and $\pi = \sigma(ef)$ are different, either σ or π is in A_n. Then, since $(abc) \in N$ and $N \lhd A_n$, either $\sigma^{-1}(abc)\sigma$ or $\pi^{-1}(abc)\pi$ is in N. So $(a'b'c') \in N$ and N contains *all* 3-cycles. From Lemma 27.3, we conclude $A_n \leqslant N$, contrary to $N \lhd A_n$. Therefore there can be no 3-cycle in N.

Secondly, there can be no permutation in N involving a cycle of length greater than or equal to 4 when written out as a product of disjoint cycles. Indeed, if $\sigma = (abcd\ldots)\pi \in N$, where $(abcd\ldots)$ and π are disjoint permutations, then

$$\begin{aligned}\sigma^{-1}.(abc)^{-1}\sigma(abc) &= \pi^{-1}(\ldots dcba)(cba)(abcd\ldots)\pi(abc)\\ &= \pi^{-1}\pi(\ldots dcba)(cba)(abcd\ldots)(abc)\\ &= (\ldots dcba)(cba)(abcd\ldots)(abc)\\ &= (abd)\end{aligned}$$

would be in N, contrary to what we proved above. So the disjoint cycles of a nonidentity permutation in N have lengths (1, in which case we do not write them, or) 2 or 3.

Thirdly, in the disjoint cycle decomposition of any nonidentity element of N, there can be no 3-cycle. To prove this, we first note that, if there were only *one* 3-cycle in the disjoint cycle decomposition of a nonidentity element of N, so that its disjoint cycle decomposition is a 3-cycle times a product of transpositions, then the square of that element would be a 3-cycle in N, which is impossible. Thus, if there is a σ in N whose disjoint cycle decomposition involves a 3-cycle at all, then there are at least *two* 3-cycles in the disjoint cycle decomposition of σ. Then we have $\sigma = (abc)(def)\pi$, say, where $(abc),(def),\pi$ are disjoint permutations and

$$\begin{aligned}\sigma.(dec)^{-1}\sigma(dec) &= (abc)(def)\pi(ced)(abc)(def)\pi(dec)\\ &= (abc)(def)(ced)(abc)(def)(dec)\pi^2\\ &= (adcbf)\pi^2\end{aligned}$$

is in N, which is impossible, since there is a cycle of length 5 in its disjoint cycle decomposition ($(adcbf)$ and π^2 are disjoint permutations). Hence, in the disjoint cycle decomposition of any nonidentity element of N, there is no cycle of length 3. Combining this with what we proved

above, we conclude that any nonidentity element in N must be a product of (an even number of) disjoint transpositions.

Fourthly, a product of $2k$ disjoint transpositions cannot belong to N if k is greater than or equal to 2, for if $\sigma = (ab)(cd)(ef)(gh)\pi$ belonged to N, where $\pi = \iota$ or π is a product of disjoint transpositions and disjoint from $(ab)(cd)(ef)(gh)$, then $\sigma.(de)^{-1}(bc)^{-1}\sigma(bc)(de)$

$$= (ab)(cd)(ef)(gh)\pi(ed)(cb)(ab)(cd)(ef)(gh)\pi(bc)(de)$$
$$= (aed)(bcf)\pi^2$$

would also belong to N. This possibility was excluded above.

Since we assume $N \neq 1$, there is a $\sigma \in N$, $\sigma \neq \iota$. Here σ is necessarily a product of two disjoint transpositions, say $\sigma = (ab)(cd)$. We choose a number e from $\{1,2, \ldots, n\}\backslash\{a,b,c,d\}$. Then the 3-cycle $\sigma.(aeb)^{-1}\sigma(aeb) = (ab)(cd)(bea)(ab)(cd)(aeb) = (abe)$ belongs to N as well, the final contradiction. This shows that the assumption $N \neq 1$ is untenable. Thus $N = 1$ and A_n is simple. □

27.5 **Definition:** Let G be a nontrivial group. A proper normal subgroup M of G is said to be a *maximal normal subgroup* of G if there is no subgroup L of G such that $M < L \lhd G$. Equivalently, M is a maximal normal subgroup of G if $M \lhd G$, and $M \leqslant K \lhd G$ implies that either $M = K$ or $K = G$.

27.6 **Lemma:** *Let $M \lhd G$. Then G/M is a simple group if and only if M is a maximal normal subgroup of G.*

Proof: Since $M \lhd G$, we have $G/M \neq 1$. If G/M is not simple, there is a normal subgroup of G/M, say N/M, which is distinct from M/M and G/M, so $M/M < N/M \lhd G/M$. By Theorem 21.2, $M < N \lhd G$ and M is not a maximal normal subgroup of G. Conversely, if M is not a maximal normal subgroup of G, there is an N such that $M < N \lhd G$ and, by Theorem 21.2, $M/M < N/M \lhd G/M$. So G/M has a nontrivial proper normal subgroup N/M and G/M is not simple. □

27.7 Definitions: Let $H \leqslant G$. A finite sequence of subgroups of G, including H and G, is called a *series from H to G*, or a *series between H and G*, if each group in the sequence is a normal subgroup of the next one. Thus a series from H to G can be written

$$H = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{n-1} \trianglelefteq H_n = G. \tag{1}$$

The subgroups $H_0, H_1, \ldots, H_{n-1}, H_n$ are called the *terms* of the series (1). The factor groups $H_1/H_0, H_2/H_1, \ldots, H_n/H_{n-1}$ are called the *factors* of the series (1). A series from 1 to G will be called shortly a *series of G*.

If each term $H_0, H_1, \ldots, H_{n-1}, H_n$ of the series (1) happens to be normal (characteristic) in G, the series (1) will be called a *normal (characteristic) series*.

There may be repetitions in (1). If, however, $H_{i-1} \triangleleft H_i$ for each $i = 1,2,\ldots,n$, the series (1) will be called a *proper series*.

A series

$$H = J_0 \trianglelefteq J_1 \trianglelefteq \cdots \trianglelefteq J_{m-1} \trianglelefteq J_m = G \tag{2}$$

from H to G is said to be a *refinement of* (1) if every term of (1) is also a term of (2). Thus a refinement of (1) is obtained from (1) by inserting additional groups between some consecutive terms of (1). These additional terms need not be distinct from the terms of (1). For example, $A \trianglelefteq B \trianglelefteq B \trianglelefteq C$ is a refinement of $A \trianglelefteq B \trianglelefteq C$. If (2) is a refinement of (1) and if there is at least one term in (2) which is not a term of (1), then (2) is called a *proper refinement of* (1).

27.8 Definition: Let G be a group. A series of G is called a *composition series of G* if it is a proper series of G and has no proper refinement. A factor of a composition series of G is called a *composition factor of G*.

27.9 Lemma: *A series*

$$1 = G_0 \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{n-1} \trianglelefteq G_n = G$$

of a group G is a composition series of G if and only if all factors G_i/G_{i-1} $(i = 1,2,\ldots,n)$ are simple.

Proof: Suppose first that the given series is a composition series of G. By definition, it is a proper series. So $G_{i-1} \lhd G_i$ and all factors G_i/G_{i-1} are distinct from the trivial group ($i = 1,2, \ldots,n$). If one of the factors, say G_j/G_{j-1}, were not simple, G_j/G_{j-1} would have a nontrivial proper normal subgroup, which may be written as H/G_{j-1}, where $G_{j-1} < H < G_j$ by Theorem 21.2. Hence $G_{j-1} \lhd H \lhd G_j$ (in fact $G_{j-1} \lhd G_j$) and the given series has a proper refinement which is obtained by inserting H between G_{j-1} and G_j, contrary to our hypothesis that the given series is a composition series. Hence G_i/G_{i-1} are all simple ($i = 1,2, \ldots,n$).

Conversely, let us assume that all factors G_i/G_{i-1} are simple ($i = 1,2, \ldots,n$). Then G_i/G_{i-1} is not trivial and so $G_{i-1} \lhd G_i$ for all $i = 1,2, \ldots,n$. Thus the given series is proper. If it were not a composition series, it would have a proper refinement. To fix the ideas, let us assume that such a refinement has a term H between G_j and G_{j-1}, so that $G_{j-1} \lhd H \lhd G_j$. By Theorem 21.2, H/G_{j-1} would be a nontrivial proper normal subgroup of G/G_{j-1}, contrary to the hypothesis that all factors, including G/G_{j-1}, are simple. Hence the given series is a composition series. □

27.10 Examples: (a) $1 \lhd S_3$ is a series of S_3 and $1 \lhd A_3 \lhd S_3$ is a refinement thereof. The latter is a composition series of S_3, because the factors $A_3/1 \cong C_3$ and $S_3/A_3 \cong C_2$ are simple (Theorem 27.2, Lemma 27.9). It is easily seen that $1 \lhd A_3 \lhd S_3$ is the unique composition series of S_3 (cf. §15, Ex.10).

(b) $1 \lhd V_4 \lhd A_4 \lhd S_4$ is a normal series of S_4 (it is a chief series of S_4; see Ex. 4). It is not a composition series of S_4, for it can be refined by inserting one of the subgroups $U_1 = \{\iota,(12)(34)\}$, $U_2 = \{\iota,(13)(24)\}$, and $U_3 = \{\iota,(14)(23)\}$ between 1 and $V_4 = \{\iota,(12)(34),(13)(24),(14)(23)\}$. Each one of the three series $1 \lhd U_i \lhd V_4 \lhd A_4 \lhd S_4$ is a composition series of S_4 ($i = 1,2,3$). The reader will easily verify that these are the only composition series of S_4.

(c) We want to find all composition series of S_n for $n \geq 5$. For this purpose, we determine all normal subgroups of S_n.

Let $n \geq 5$ and $1 < N \trianglelefteq S_n$. Then $N \cap A_n \trianglelefteq A_n$ by Theorem 21.3 and, since A_n is simple (Theorem 27.4), either $N \cap A_n = A_n$ or $N \cap A_n = 1$.

In case $N \cap A_n = A_n$, we have $A_n \leq N \leq S_n$. Thus $|N:A_n|$ divides $|S_n:A_n| = 2$ and $|N:A_n| = 1$ or $|N:A_n| = 2$. Hence $N = A_n$ or $N = S_n$.

In case $N \cap A_n = 1$, we have $N \not\leq A_n$ (because $1 < N$), so $A_n < A_n N \leq S_n$, so $A_n N = S_n$ and $|N| = |N:1| = |N:N \cap A_n| = |A_n N:A_n| = |S_n:A_n| = 2$. Thus $N = \{\iota,\sigma\}$ for some $\sigma \in S_n \backslash A_n$. Since $N \trianglelefteq S_n$, we obtain

$$\{\iota,\sigma\} = N = N^\tau = \{\iota,\sigma\}^\tau = \{\iota^\tau,\sigma^\tau\} = \{\iota,\sigma^\tau\} \text{ for all } \tau \in S_n,$$

hence

$$\sigma^\tau = \sigma \qquad \text{for all } \tau \in S_n. \tag{*}$$

From $o(\sigma) = |\langle\sigma\rangle| = |N| = 2$, we see that the disjoint cycle decomposition of σ involves transpositions only (Theorem 15.17), say

$$\sigma = (a_1 b_1)(a_2 b_2)\ldots(a_m b_m)$$

for some odd number m. If $m \geq 3$, then $\pi := (a_3 b_3)\ldots(a_m b_m)$ is disjoint from $(a_1 b_1)(a_2 b_2)$ and

$$\sigma^{(a_1 a_2)} = (a_1 b_1)^{(a_1 a_2)}(a_2 b_2)^{(a_1 a_2)}\pi^{(a_1 a_2)} = (a_2 b_1)(a_1 b_2)\pi \neq (a_1 b_1)(a_2 b_2)\pi = \sigma$$

contrary to (*). Hence $m = 1$ and $\sigma = (a_1 b_1)$. Let $c \in \{1,2,\ldots,n\}\backslash\{a_1,b_1\}$. Now

$$\sigma^{(a_1 c)} = (c a_1)(a_1 b_1)(a_1 c) = (b_1 c) \neq (a_1 b_1) = \sigma,$$

again contradicting (*). Hence there is no nontrivial normal subgroup N of S_n such that $N \cap A_n = 1$.

Consequently, $1, A_n, S_n$ are the only normal subgroups of S_n when $n \geq 5$. Thus if $1 = G_0 \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{k-1} \trianglelefteq G_k = S_n$ is a composition series of S_n, here G_{k-1} has to be A_n and G_{k-2} has to be 1. Therefore the series must be $1 \triangleleft A_n \triangleleft S_n$, which is indeed a composition series of S_n, for $A_n/1 \cong A_n$ and $S_n/A_n \cong C_2$ are simple groups (Lemma 27.9).

Thus $1 \triangleleft A_n \triangleleft S_n$ is the unique composition series of S_n when $n \geq 5$.

(d) Not every group has a composition series. For example, $\mathbb{Z}$ has no composition series. Indeed, any series of $\mathbb{Z}$ is of the form

$$0 \triangleleft m_1\mathbb{Z} \trianglelefteq m_2\mathbb{Z} \trianglelefteq \ldots \trianglelefteq m_n\mathbb{Z} \trianglelefteq \mathbb{Z}, \tag{3}$$

where $m_2 | m_1$, $m_3 | m_2$, ..., $m_n | m_{n-1}$. If m_0 is a multiple of m_1 and $m_0 \neq m_1$, then

$$0 \trianglelefteq m_0\mathbb{Z} \triangleleft m_1\mathbb{Z} \trianglelefteq m_2\mathbb{Z} \trianglelefteq \ldots \trianglelefteq m_n\mathbb{Z} \trianglelefteq \mathbb{Z}$$

is a proper refinement of (3). Thus any series of $\mathbb{Z}$ has a proper refinement. Consequently, no series of $\mathbb{Z}$ can be a composition series of $\mathbb{Z}$.

(e) Let $\langle a \rangle$ be a cyclic group of order 12. Then
$1 \lhd \langle a^6 \rangle \lhd \langle a^2 \rangle \lhd \langle a \rangle$; $1 \lhd \langle a^6 \rangle \lhd \langle a^3 \rangle \lhd \langle a \rangle$; $1 \lhd \langle a^4 \rangle \lhd \langle a^2 \rangle \lhd \langle a \rangle$
are the composition series of $\langle a \rangle$. The composition factors are isomorphic to C_2, C_3, C_2; C_2, C_2, C_3; C_3, C_2, C_2.
Thus, aside from order, the composition factors arising from different composition series are isomorphic groups.

27.11 Definition: Let G be a group. Two series

$$1 = G_0 \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{n-1} \trianglelefteq G_n = G$$
$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{m-1} \trianglelefteq H_m = G$$

of G are said to be *equivalent* if $n = m$ and if the factors G_i/G_{i-1} are, in some order, isomorphic to the factors H_j/H_{j-1} $(i,j = 1,2,\ldots,n)$.

Here it is not stipulated that $G_i/G_{i-1} \cong H_i/H_{i-1}$ for all $i = 1,2,\ldots,n$. The condition in Definition 27.11 is that $G_i/G_{i-1} \cong H_{i\sigma}/H_{i\sigma-1}$ for some $\sigma \in S_n$. Clearly, Definition 27.11 introduces an equivalence relation on the set of all series of G. The three series in Example 27.10(e) are equivalent. We will prove that any two composition series of a group are equivalent, provided G does have a composition series (Jordan-Hölder theorem). In fact, a much stronger theorem is true (see Schreier's theorem below). We need some elementary results.

27.12 Lemma (Dedekind's modular law): *Let G be a group and let A,B,C be subgroups of G such that $A \leqslant C$. Then*

$$A(B \cap C) = AB \cap C.$$

($A(B \cap C)$ and $AB \cap C$ are not necessarily subgroups of G.)

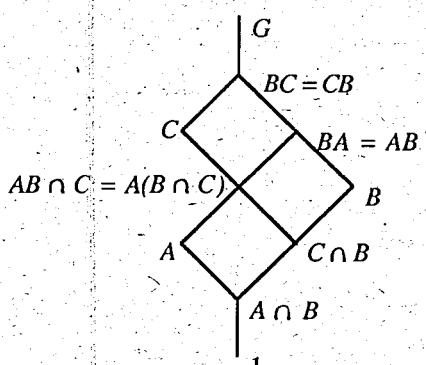
Proof: Let $x \in A(B \cap C)$. Then $x = ab$ for some $a \in A$, $b \in B \cap C$. Thus $x = ab \in AB$ and $x = ab \in AC = C$, so $x \in AB \cap C$. This gives $A(B \cap C) \subseteq AB \cap C$. To show the reverse inclusion, let $c \in AB \cap C$. Then $c = a_1 b_1$ for some a_1

in A and b_1 in B. From $b_1 = a_1^{-1}c \in AC = C$, we conclude $b_1 \in B \cap C$, hence $c = a_1 b_1 \in A(B \cap C)$. This gives $AB \cap C \subseteq A(B \cap C)$. So $A(B \cap C) = AB \cap C$. □

27.13 Lemma: *Let $A \trianglelefteq C \leqslant G$ and $B \leqslant G$. Then*

$$A \cap B \trianglelefteq C \cap B \quad \text{and} \quad C \cap B \,/\, A \cap B \cong A(B \cap C)/A.$$

Proof: If G is a group, $H \trianglelefteq G$ and $K \leqslant G$, then $H \cap K \trianglelefteq K$ and $K \,/\, H \cap K$ is isomorphic to HK/H by Theorem 21.3. Using this theorem with G,H,K replaced by $C,A,C \cap B$, respectively, we obtain $A \cap (C \cap B) \trianglelefteq C \cap B$ and $C \cap B \,/\, A \cap (C \cap B) \cong A(C \cap B)/A$. Since $A \cap (C \cap B) = A \cap B$ and $A(C \cap B) = A(B \cap C)$, the claim follows. □



27.14 Lemma: *Let $A \trianglelefteq C \leqslant G$ and $B \trianglelefteq G$. Then*

$$BA \trianglelefteq BC \quad \text{and} \quad BC/BA \cong C/A(B \cap C).$$

Proof: Since $B \trianglelefteq G$, we know from Lemma 19.4 that $AB = BA \leqslant G$ and that $CB = BC \leqslant G$. Thus $BA \leqslant BC$. We prove next that BA is normal in BC. We observe

$$B \leqslant BA \leqslant N_G(BA)$$

hence $$(BA)^b = BA \qquad \text{for all } b \in B.$$

Then, for any $b \in B, c \in C$, we obtain

$$(BA)^{bc} = [(BA)^b]^c = (BA)^c = B^c A^c = BA^c = BA$$

since $B \trianglelefteq G$ and $A \trianglelefteq C$. Thus $(BA)^x = BA$ for all $x \in BC$ and $BA \trianglelefteq BC$.

Using Theorem 21.3 with BC, BA, C in place of G,H,K, respectively, we get

$$AB \cap C \trianglelefteq C \text{ and } C/AB \cap C \cong C(AB)/AB.$$

Since $AB \cap C = A(B \cap C)$ and $C(AB) = (CA)B = CB = BC$, this isomorphism means $C/A(B \cap C) \cong BC/BA$. □

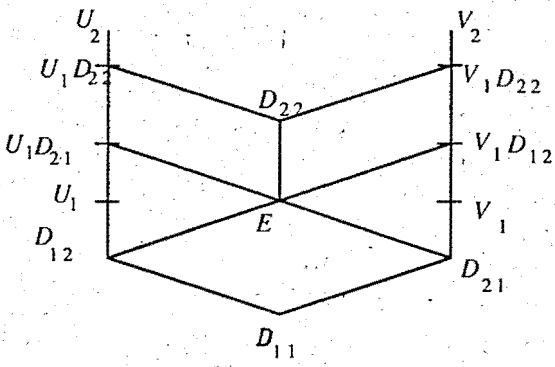
27.15 Lemma (Zassenhaus' lemma): *Let G be a group,*

$$U_1 \trianglelefteq U_2 \leqslant G \text{ and } V_1 \trianglelefteq V_2 \leqslant G.$$

Then $U_1(U_2 \cap V_1) \trianglelefteq U_1(U_2 \cap V_2)$, $V_1(U_1 \cap V_2) \trianglelefteq V_1(U_2 \cap V_2)$ *and*

$$U_1(U_2 \cap V_2)/U_1(U_2 \cap V_1) \cong V_1(U_2 \cap V_2)/V_1(U_1 \cap V_2).$$

Proof: We put $D_{ij} := U_i \cap V_j$ $(i,j = 1,2)$. Since $U_1 \trianglelefteq U_2$, we have $U_1 \cap V_2 \trianglelefteq U_2 \cap V_2$ by Lemma 27.13, so $D_{12} \trianglelefteq D_{22}$. Similarly, $V_1 \trianglelefteq V_2$ and Lemma 27.13 gives $U_2 \cap V_1 \trianglelefteq U_2 \cap V_2$, so $D_{21} \trianglelefteq D_{22}$. Now $D_{12} \trianglelefteq D_{22}$ and $D_{21} \trianglelefteq D_{22}$ and writing $E = D_{12}D_{21} = D_{21}D_{12}$ for brevity, we get $E \trianglelefteq D_{22}$ by Lemma 19.4(3).



Since $E \trianglelefteq D_{22} \leqslant U_2$ and $U_1 \trianglelefteq U_2$, Lemma 27.14 gives

$$U_1E \trianglelefteq U_1D_{22} \text{ and } U_1D_{22}/U_1E \cong D_{22}/E(U_1 \cap D_{22}). \tag{4}$$

Here $U_1E = U_1(D_{12}D_{21}) = (U_1D_{12})D_{21} = U_1D_{21}$ and $E(U_1 \cap D_{22}) = E$ (because $U_1 \cap D_{22} = D_{12} \subseteq E$), so (4) becomes

$$U_1D_{21} \trianglelefteq U_1D_{22} \text{ and } U_1D_{22}/U_1D_{21} \cong D_{22}/E. \tag{5}$$

Repeating the same argument with U's replaced by V's, we get

$$V_1D_{12} \trianglelefteq V_1D_{22} \text{ and } V_1D_{22}/V_1D_{12} \cong D_{22}/E. \tag{6}$$

The claim follows from (5) and (6). □

27.16 Theorem (Schreier): *Any two series of a group have equivalent refinements. More precisely, if*

$$1 = G_0 \trianglelefteq G_1 \trianglelefteq \cdots \trianglelefteq G_{n-1} \trianglelefteq G_n = G \qquad (g)$$
$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{m-1} \trianglelefteq H_m = G \qquad (h)$$

are series of G, then there are series (g') and (h') of G such that (g') is a refinement of (g), (h') is a refinement of (h), and (g') and (h') are equivalent.

Proof: We will try to build a series between each G_{i-1} and G_i $(i = 1,2,\ldots,n)$ by so modifying the terms of (h) that the modified series begin from G_{i-1} and terminate at G_i. There are two natural ways of doing this. Either we multiply each term of (h) by G_{i-1} (the resulting series will thus begin from G_{i-1}) and intersect the products with G_i (the modified series will thus terminate at G_i); or we intersect each term of (h) by G_i (the resulting series will thus terminate at G_i) and multiply the intersections by G_{i-1} (the modified series will thus begin from G_{i-1}). By Dedekind's modular law (Lemma 27.12), these two series between G_{i-1} and G_i are identical.

$$\begin{array}{lll} H_m & G_{i-1}H_m & G_{i-1}H_m \cap G_i = G_i \\ H_{m-1} & G_{i-1}H_{m-1} & G_{i-1}H_{m-1} \cap G_i \\ \vdots & \vdots & \vdots \\ H_1 & G_{i-1}H_1 & G_{i-1}H_1 \cap G_i \\ H_0 & G_{i-1}H_0 & G_{i-1}H_0 \cap G_i = G_{i-1} \end{array}$$

$$\begin{array}{lll} H_m & H_m \cap G_i & G_{i-1}(H_m \cap G_i) = G_i \\ H_{m-1} & H_{m-1} \cap G_i & G_{i-1}(H_{m-1} \cap G_i) \\ \vdots & \vdots & \vdots \\ H_1 & H_1 \cap G_i & G_{i-1}(H_1 \cap G_i) \\ H_0 & H_0 \cap G_i & G_{i-1}(H_0 \cap G_i) = G_{i-1} \end{array}$$

We put $G_{ij} = G_{i-1}H_j \cap G_i = G_{i-1}(H_j \cap G_i)$ $(i = 1,2, \ldots,n; j = 1,2, \ldots,m)$.

Similarly, we put $H_{ij} = H_{j-1}G_i \cap H_j = H_{j-1}(G_i \cap H_j)$ $(i = 1,2, \ldots,n; j = 1,2, \ldots,m)$.

Here $G_{i-1} \trianglelefteq G_i$, hence $G_{i-1}(H_j \cap G_i) \leqslant G_i$ by Lemma 19.4(2). Thus G_{ij} is a subgroup of G_i. In the same way, H_{ij} is a subgroup of H_i. So

$$G_{i-1} = G_{i0} \leqslant G_{i1} \leqslant G_{i2} \leqslant \cdots \leqslant G_{i,m-1} \leqslant G_{im} = G_i \qquad (g_i)$$

and
$$H_{j-1} = H_{0j} \leqslant H_{1j} \leqslant H_{2j} \leqslant \cdots \leqslant H_{n-1,j} \leqslant H_{nj} = H_j. \qquad (h_j)$$

Using Zassenhaus' lemma (Lemma 27.15) with

$$U_1 = G_{i-1},\ U_2 = G_i,\ V_1 = H_{j-1},\ V_2 = H_j,$$

we obtain, for each $i = 1,2, \ldots,n$, $j = 1,2, \ldots,m$:

$$G_{i-1}(G_i \cap H_{j-1}) \trianglelefteq G_{i-1}(G_i \cap H_j),\ H_{j-1}(G_{i-1} \cap H_j) \trianglelefteq H_{j-1}(G_i \cap H_j)$$

and
$$G_{i-1}(G_i \cap H_j) / G_{i-1}(G_i \cap H_{j-1}) \cong H_{j-1}(G_i \cap H_j) / H_{j-1}(G_{i-1} \cap H_j).$$

Thus $G_{i,j-1} \trianglelefteq G_{ij}$, $H_{i-1,j} \trianglelefteq H_{ij}$ and $G_{ij}/G_{i,j-1} \cong H_{ij}/H_{i-1,j}$.

Therefore (g_i) is a series between G_{i-1} and G_i, and (h_j) is a series between H_{j-1} and H_j. Writing the terms of $(g_1),(g_2),(g_3), \ldots,(g_n)$ consecutively, we obtain a series (g') of G with nm factors; and writing the terms of $(h_1),(h_2),(h_3), \ldots,(h_m)$ consecutively, we obtain a series (h') of H with mn factors. Here (g') is a refinement of (g) and (h') is a refinement of (h). Finally, in view of the isomorphisms $G_{ij}/G_{i,j-1} \cong H_{ij}/H_{i-1,j}$, the series (g') and (h') are equivalent. □

27.17 Theorem: *Let G be a group and assume that G has a composition series.*
(1) *Every proper series of G has a refinement which is a composition series.*
(2) (Jordan-Hölder Theorem) *Any two composition series of G are equivalent.*

Proof: Let

$$1 = G_0 \lhd G_1 \lhd \cdots \lhd G_{n-1} \lhd G_n = G \qquad (g)$$

be a proper series of G and let

$$1 = H_0 \lhd H_1 \lhd \cdots \lhd H_{m-1} \lhd H_m = G \qquad (h)$$

be a composition series of G. By Schreier's theorem (Theorem 27.16), there are equivelent series (g') and (h') of G such that (g') is a refinement of (g) and (h') is a refinement of (h). From (g') and (h'), we delete repeated factors and thereby obtain two equivalent proper series, say (g'') and (h''), respectively. Here (g'') is a refinement of (g) and (h'') is a refinement of (h), because both (g) and (h) are proper series.

(1) (h'') is a proper series and is a refinement of (h). But (h) has no proper refinement, because (h) is a composition series. Hence (h'') is identical with (h). Thus (g) has a refinement (g'') which is equivalent to the composition series $(h'') = (h)$. Then the factors of (g''), being isomorphic to the composition factors in (h), are all simple groups and (g'') itself is a composition series by Lemma 27.9. Therefore any proper series (g) of G has a refinement (g'') which is a composition series.

(2) Assume now (g) is also a composition series of G. By the same argument as above, (g'') must be identical with (g). Then $(g) = (g'')$ and $(h'') = (h)$ are equivalent. Thus any two composition series of G are equivalent. □

We now discuss the class of solvable groups.

27.18 Definition: A series

$$H = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{m-1} \trianglelefteq H_m = G$$

from H to G is said to be an *abelian* series if all the factors

$$H_1/H_0, H_2/H_1, \ldots, H_m/H_{m-1}$$

are abelian groups.

27.19 Definition: A group G is called a *solvable* (or *soluble*) group if G has an abelian series (from 1 to G).

Clearly, any abelian group A is solvable: $1 \trianglelefteq A$ is an abelian series of A. S_3 is an example of a nonabelian solvable group. Not every group is solvable. For example nonabelian simple groups are certainly not solvable. In particular, A_n is not solvable for $n \geqslant 5$.

27.20 Lemma: *If G is a solvable group, then all subgroups and factor groups of G are solvable.*

Proof: Being solvable, G has an abelian series

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{m-1} \trianglelefteq H_m = G.$$

Let K be an arbitrary subgroup of G. Then, by Lemma 27.13,

$$1 = H_0 \cap K \trianglelefteq H_1 \cap K \trianglelefteq \cdots \trianglelefteq H_{m-1} \cap K \trianglelefteq H_m \cap K = G \cap K = K \qquad (6)$$

is a series of K and $H_i \cap K / H_{i-1} \cap K \cong H_{i-1}(K \cap H_i)/H_i \leqslant H_i/H_{i-1}$ for all $i = 1,2,\ldots,m$. Since H_i/H_{i-1} are abelian, $H_i \cap K / H_{i-1} \cap K$ are also abelian and (6) is an abelian series of K. Hence K is solvable.

Now let N be an arbitrary normal subgroup of G. By Lemma 27.14 and Theorem 21.2,

$$N/N = H_0N/N \trianglelefteq H_1N/N \trianglelefteq \cdots \trianglelefteq H_{m-1}N/N \trianglelefteq H_mN/N = G/N \qquad (7)$$

is a series of G/N, and, for all $i = 1,2,\ldots,m$, $H_iN/N / H_{i-1}N/N \cong H_iN/H_{i-1}N \cong H_i/H_{i-1}(N \cap H_i) \cong H_i/H_{i-1} / H_{i-1}(N \cap H_i)/H_{i-1}$ is a factor group of the

abelian group H_i/H_{i-1} and therefore $H_iN/N \;/\; H_{i-1}N/N$ is abelian (Lemma 18.9(2)). So (7) is an abelian series of G/N and G/N is solvable. □

27.21 Lemma: *Let $N \trianglelefteq G$. If N and G/N are both solvable, then G is solvable.*

Proof: By hypothesis, there are an abelian series

$$1 = N_0 \trianglelefteq N_1 \trianglelefteq \cdots \trianglelefteq N_{m-1} \trianglelefteq N_m = N$$

of N and an abelian series

$$N/N = H_0/N \trianglelefteq H_1/N \trianglelefteq \cdots \trianglelefteq H_{k-1}/N \trianglelefteq H_k/N = G/N$$

of G/N. By Theorem 21.2,

$$1 = N_0 \trianglelefteq N_1 \trianglelefteq \cdots \trianglelefteq N_{m-1} \trianglelefteq N_m = N = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{k-1} \trianglelefteq H_k = G$$

is a series of G. Since N_j/N_{j-1} is abelian for $j = 1,2,\ldots,m$ and $H_i/H_{i-1} \cong H_i/N \;/\; H_{i-1}/N$ is abelian for $i = 1,2,\ldots,k$, this is an abelian series of G. Thus G is solvable. □

27.22 Lemma: *Let H and K be normal solvable subgroups of a group G. Then HK is a normal solvable subgroup of G.*

Proof: HK is a normal subgroup of G by Lemma 19.4(3). Also, since K is solvable, $K/H \cap K$ is solvable by Lemma 27.20, so HK/H is solvable by Theorem 21.3. So H and HK/H are solvable and consequently HK is solvable by Lemma 27.21. □

27.23 Theorem: *If G is a finite p-group, then G is solvable.*

Proof: If G is a finite p-group of order $|G| = p^a$, then there is a series

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{a-1} \trianglelefteq H_a = G$$

of G whose factors H_i/H_{i-1} $(i = 1,2,\ldots,a)$ are cyclic of order p (Theorem 26.3(2)). Thus G has an abelian series and G is solvable. □

The series in Theorem 27.23 is a composition series of G. We now want to prove more generally that a finite group G is solvable if and only if every composition factor of G is cyclic of prime order. A finite group does have a composition series, of course.

27.24 Lemma: *A solvable group G is simple if and only if G is cyclic of prime order.*

Proof: Let G be a simple solvable group. Then $G \neq 1$ and G has an abelian series

$$1 = H_0 \trianglelefteq H_1 \trianglelefteq \cdots \trianglelefteq H_{m-1} \trianglelefteq H_m = G.$$

After deleting repetitions, we may assume that this is a proper series. Then $H_{m-1} \lhd G$ and G/H_{m-1} is a nontrivial abelian group. Thus $G' \leqslant H_{m-1}$ (Theorem 24.14) and G' is a proper normal subgroup of G. Since G is simple, $G' = 1$ and G is abelian. Thus G is cyclic of prime order by Theorem 27.2. Conversely, a cyclic group of prime order is simple; and abelian, hence solvable. □

27.25 Theorem: *Let G be a finite group. G is solvable if and only if every composition factor of G has prime order.*

Proof: If G is solvable, then any composition factor of G is solvable by Lemma 27.20, simple by Lemma 27.9, and so has prime order by Lemma 27.24. Conversely, if every composition factor of G has prime order, then a composition series of G is an abelian series of G and therefore G is solvable. □

The following result will play a crucial role in proving that a polynomial equation of degree greater than four cannot be solved by radicals.

27.26 Theorem: *If $n \geqslant 5$, then S_n is not solvable.*

Proof: Otherwise the subgroup A_n of S_n would be solvable (Lemma 27.20), whereas A_n, being a nonabelian simple group (Lemma 27.3),

cannot have an abelian series. The conclusion follows also from Theorem 27.25, since A_n is a composition factor of S_n ($1 \lhd A_n \lhd S_n$ is the unique composition series of S_n by Example 27.10(c)). □

Exercises

1. Let $\{G_i: i \in \mathbb{N}\}$ be a collection of simple groups such that $G_i \leqslant G_{i+1}$ for all $i \in \mathbb{N}$ and let $G = \bigcup_{i=1}^{\infty} G_i$. Prove that G is a simple group.

2. Let $S_{(\mathbb{N})} = \{\sigma \in S_{\mathbb{N}}: k\sigma \neq k \text{ for at most finitely many } k \in \mathbb{N}\}$ and, for each $n \in \mathbb{N}$, let $S(n) = \{\sigma \in S_{\mathbb{N}}: k\sigma \neq k \text{ for all } k \geqslant n+1\}$. Show that $S_n \cong S(n) \leqslant S_{(\mathbb{N})} \leqslant S_{\mathbb{N}}$. Let $A(n)$ denote the image of A_n under the isomorphism $S_n \cong S(n)$ for $n \geqslant 2$ and show that $A := \bigcup_{i=5}^{\infty} A(n)$ is a simple group. (A is called the *infinite alternating group of degree* $|\mathbb{N}|$.)

3. Let $M \lhd G$ and $|G:M|$ be prime. Prove that M is a maximal normal subgroup of G.

4. A normal series of a group G is called a *chief series of* G if it is a proper series and if it has no proper refinement which is a normal series of G. A factor of a chief series of G is called a *chief factor of* G.

Let G be a nontrivial group. A nontrivial normal subgroup M of G is called a *minimal normal subgroup of* G if there is no $L \trianglelefteq G$ such that $1 < L < M$.

Prove the following statements.

(a) H/K is a chief factor of G if and only if H/K is a minimal subgroup of G/K.

(b) If M is a minimal normal subgroup of G, then M has no characteristic subgroup except 1 and M.

(c) If G has a composition series, then G has a chief series.

(d) $1 \lhd V_4 \lhd A_4 \lhd S_4$ is the unique chief series of S_4.

5. Suppose G has a finite abelian group having no characteristic subgroups except 1 and G. Show that there is a prime number p such that $g^p = 1$ for all $g \in G$.

6. Prove that an abelian group has a composition series if and only if it is finite.

7. Find an infinite abelian subgroup of the infinite alternating group (see Ex. 2). Conclude that a subgroup of a group with a composition series does not necessarily have a composition series.

8. Let $H \trianglelefteq G$. Prove that, if G has a composition series, so does G/H.

9. Let $H \trianglelefteq G$. Prove that, if H and G/H have composition series, so does G.

10. Repeat the proof of Schreier's theorem for the two series $1 \triangleleft C_{18} \triangleleft C_{36}$ and $1 \triangleleft C_4 \triangleleft C_{12} \triangleleft C_{36}$ of the cyclic group C_{36}.

11. Prove that, if H and K are solvable, so is $H \times K$.

12. Prove that, if $H,K \trianglelefteq G$ and G/H, G/K are solvable, so is $G/H \cap K$.

13. For each $n \in \mathbb{N}$, we define a subgroup $G^{(n)}$ of G recursively by $G^{(n+1)} = (G^{(n)})' = [G^{(n)},G^{(n)}]$. The series

$$G \geqslant G^{(1)} \geqslant G^{(2)} \geqslant \cdots$$

is called the *derived series of* G. Show that each $G^{(n)}$ is characteristic in G and that, if

$$G_r \trianglelefteq G_{r-1} \trianglelefteq G_{r-2} \trianglelefteq \ldots \trianglelefteq G_1 \trianglelefteq G_0 = G$$

is an abelian series between G_r and G, then $G^{(n)} \leqslant G_n$ for each $n = 1,2, \ldots,r$. Prove that G is solvable if and only if $G^{(r)} = 1$ for some $r \in \mathbb{N}$.

14. Prove that a solvable group has a composition series if and only if it is finite (cf. Ex. 6).

§28
Finitely Generated Abelian Groups

In this last paragraph of Chapter 2, we determine the structure of finitely generated abelian groups. A complete classification of such groups is given. Complete classification theorems are very rare in mathematics and, in general, they require sophisticated machinery. However, the main theorems in this paragraph are proved by quite elementary methods, chiefly by induction! This is due to the fact that commutativity is a very strong condition.

This paragraph is not needed in the sequel.

28.1 Lemma: *Let G be an abelian group. We write*

$$T(G) := \{g \in G: o(g) \text{ is finite}\}.$$

(1) *$T(G)$ is a subgroup of G* (called the *torsion subgroup of G*).
(2) *In $G/T(G)$, every nonidentity element is of infinite order.*

Proof: (1) Since $o(1) = 1 \in \mathbb{N}$, $1 \in T(G)$ and $T(G) \neq \emptyset$. Suppose now a,b are in $T(G)$, say $o(a) = n$, $o(b) = m$ $(n,m \in \mathbb{N})$. Then $(ab)^{nm} = a^{nm}b^{nm} = 1.1 = 1$, so $o(ab) \leqslant nm$, thus $ab \in T(G)$; and $o(a^{-1}) = n \in \mathbb{N}$, thus $a^{-1} \in T(G)$. By the subgroup criterion, $T(G) \leqslant G$.

(2) Since G is abelian, we can build the factor group $G/T(G)$. If $T(G)x$ in $G/T(G)$ has finite order, say $n \in \mathbb{N}$, then $(T(G)x)^n = T(G)$, so $T(G)x^n = T(G)$, so $x^n \in T(G)$, so $o(x^n)$ is finite. Let $o(x^n) = m \in \mathbb{N}$. Then $x^{nm} = (x^n)^m = 1$, so $o(x) \leqslant nm$. Thus $o(x)$ is finite and $x \in T(G)$. It follows that $T(G)x = T(G)$ is the identity element of $G/T(G)$. Hence every nonidentity element of $G/T(G)$ has infinite order. □

28.2 Definition: A group G is called a *torsion group* if every element of G has finite order. A group is said to be *without torsion*, or *torsion-free* if every nonidentity element of G has infinite order.

Thus 1 is the only group which is both a torsion group and torsion-free.

Every finite group is a torsion group, but there are also infinite torsion groups, for example $\mathbb{Q}/\mathbb{Z}$.

In view of Lemma 28.1, we are led to investigate two classes of abelian groups: torsion abelian groups and torsion-free abelian groups. When this is done, we will know the structure of $T(G)$ and $G/T(G)$, where G is an abelian group. We must then investigate how $T(G)$ and $G/T(G)$ are combined to build G.

We cannot expect to carry out this ambitious program without imposing additional conditions on G. We will assume that G is finitely generated (Definition 24.4). Under this assumption, $T(G)$ turns out to be a finite group (Theorem 28.15). The study of finite abelian groups reduces to the study of finite abelian p-groups, p being a prime number, whose structures are described in Theorem 28.10. After that, we turn our attention to torsion-free abelian groups (Theorem 28.13). The next step in our program is to put the pieces $T(G)$ and $G/T(G)$ together in the appropriate way to form G. The appropriate way proves to be the simplest way: G is isomorphic to the direct product of $T(G)$ and $G/T(G)$. The structure of G will be completely determined by a set of integers.

28.3 Definition: Let G be an abelian group and let $S = \{g_1, g_2, \ldots, g_r\}$ be a finite, nonempty subset of G. If, for any integers $a_1, a_2, \ldots, a_r$, the relation

$$g_1^{a_1} g_2^{a_2} \cdots g_r^{a_r} = 1$$

implies that $g_1^{a_1} = g_2^{a_2} = \cdots = g_r^{a_r} = 1$, then S is said to be *independent*. If S is independent and generates G, and if $1 \notin S$, then S is called a *basis of* G.

In the following lemma, we will prove, among other things, that $S = \{g_1, g_2, \ldots, g_r\}$ is a basis of G if and only if G is the direct product of the cyclic groups $\langle g_1 \rangle, \langle g_2 \rangle, \ldots, \langle g_r \rangle$. Lemma 28.4(2) is of especial importance: it states that a finitely generated abelian torsion group is in fact a finite group.

28.4 Lemma: *Let G be an abelian group and $g_1,g_2,\ldots,g_r$ be finitely many elements of G, not necessarily distinct ($r \geqslant 1$). Let $B \leqslant G$.*

(1) $\langle g_1,g_2,\ldots,g_r\rangle = \langle g_1\rangle\langle g_2\rangle\ldots\langle g_r\rangle$.

(2) *If each g_i has finite order, then* $|\langle g_1,g_2,\ldots,g_r\rangle| \leqslant o(g_1)o(g_2)\ldots o(g_r)$.

(3) *If* $G = \langle g_1,g_2,\ldots,g_r\rangle$ *and* $\varphi: G \to A$ *is a homomorphism onto A, then* $A = \langle g_1\varphi,g_2\varphi,\ldots,g_r\varphi\rangle$.

(4) *If* $G = \langle g_1,g_2,\ldots,g_r\rangle$, *then* $G/B = \langle Bg_1,Bg_2,\ldots,Bg_r\rangle$.

(5) *If* $G/B = \langle Bg_1,Bg_2,\ldots,Bg_r\rangle$, *then* $G = B\langle g_1,g_2,\ldots,g_r\rangle$. *If, in addition,* $b_1,\ldots,b_s \in B$ *and* $B = \langle b_1,\ldots,b_s\rangle$, *then* $G = \langle b_1,\ldots,b_s,g_1,g_2,\ldots,g_r\rangle$.

(6) *If* $B = \langle g_1\rangle$ *and* $G/B = \langle Bg_2,\ldots,Bg_r\rangle$, *then* $G = \langle g_1,g_2,\ldots,g_r\rangle$.

(7) $\{g_1,g_2,\ldots,g_r\}$ *is an independent subset of G and* $G = \langle g_1,g_2,\ldots,g_r\rangle$ *if and only if* $G = \langle g_1\rangle\times\langle g_2\rangle\times\ldots\times\langle g_r\rangle$. *In particular, in case $g_1,g_2,\ldots,g_r$ are all distinct from 1, the subset $\{g_1,g_2,\ldots,g_r\}$ is a basis of G if and only if* $G = \langle g_1\rangle\times\langle g_2\rangle\times\ldots\times\langle g_r\rangle$.

Proof: (1) Certainly $\{g_1,g_2,\ldots,g_r\} \subseteq \langle g_1\rangle\langle g_2\rangle\ldots\langle g_r\rangle \leqslant G$ by repeated use of Lemma 19.4(3), and so $\langle g_1,g_2,\ldots,g_r\rangle \leqslant \langle g_1\rangle\langle g_2\rangle\ldots\langle g_r\rangle$ by the definition of $\langle g_1,g_2,\ldots,g_r\rangle$. Also, any element of $\langle g_1\rangle\langle g_2\rangle\ldots\langle g_r\rangle$, necessarily of the form $g_1^{m_1}g_2^{m_2}\ldots g_r^{m_r}$ with suitable integers $m_1,m_2,\ldots,m_r$, is in $\langle g_1,g_2,\ldots,g_r\rangle$ by Lemma 24.2 and so $\langle g_1\rangle\langle g_2\rangle\ldots\langle g_r\rangle \leqslant \langle g_1,g_2,\ldots,g_r\rangle$. Hence $\langle g_1,g_2,\ldots,g_r\rangle = \langle g_1\rangle\times\langle g_2\rangle\ldots\langle g_r\rangle$.

(2) Suppose $o(g_i) = k_i \in \mathbb{N}$ for each $i = 1,2,\ldots,r$. If $g \in \langle g_1,g_2,\ldots,g_r\rangle$, then, by part (1), $g = g_1^{m_1}g_2^{m_2}\ldots g_r^{m_r}$ with suitable integers m_i. Dividing m_i by k_i, we may write $m_i = k_iq_i + t_i$, where $q_i,t_i \in \mathbb{Z}$ and $0 \leqslant t_i < k_i$. Then $g_i^{m_i} = (g_i^{k_i})^{q_i}g_i^{t_i} = g_i^{t_i}$ and $g = g_1^{t_1}g_2^{t_2}\ldots g_r^{t_r}$. Thus

$$\langle g_1,g_2,\ldots,g_r\rangle \subseteq \{g_1^{t_1}g_2^{t_2}\ldots g_r^{t_r} : 0 \leqslant t_i < k_i \text{ for all } i = 1,2,\ldots,r\}$$

and

$$|\langle g_1,g_2,\ldots,g_r\rangle| \leqslant k_1k_2\ldots k_r.$$

(3) If $a \in A$, then $a = g\varphi$ for some $g \in G$ since φ is onto and $g = g_1^{m_1}g_2^{m_2}\ldots g_r^{m_r}$ with suitable integers m_i since $G = \langle g_1,g_2,\ldots,g_r\rangle$. Thus

$$a = g\varphi = (g_1^{m_1}g_2^{m_2}\ldots g_r^{m_r})\varphi = (g_1\varphi)^{m_1}(g_2\varphi)^{m_2}\ldots(g_r\varphi)^{m_r} \in \langle g_1\varphi,g_2\varphi,\ldots,g_r\varphi\rangle$$

and $A \subseteq \langle g_1\varphi,g_2\varphi,\ldots,g_r\varphi\rangle$.

(4) This follows from part (3) when we take A to be G/B and φ to be the natural homomorphism $\nu: G \to G/B$.

(5) Suppose $G/B = \langle Bg_1, Bg_2, \ldots, Bg_r \rangle$. Let $g \in G$. Then $Bg \in G/B$ and, by part (1) with G/B in place of G and Bg_i in place of g_i, we have

$$Bg = (Bg_1)^{m_1}(Bg_2)^{m_2}\ldots(Bg_r)^{m_r} = Bg_1^{m_1}g_2^{m_2}\ldots g_r^{m_r} \text{ for some integers } m_i.$$

Hence $g = bg_1^{m_1}g_2^{m_2}\ldots g_r^{m_r}$ for some $b \in B$ and $g \in B\langle g_1, g_2, \ldots, g_r \rangle$. So $G = B\langle g_1, g_2, \ldots, g_r \rangle$. If, in addition, $B = \langle b_1, \ldots, b_s \rangle$, then

$$\begin{aligned} G &= \langle b_1, \ldots, b_s \rangle \langle g_1, g_2, \ldots, g_r \rangle = \langle b_1 \rangle \ldots \langle b_s \rangle \langle g_1 \rangle \langle g_2 \rangle \ldots \langle g_r \rangle \\ &= \langle b_1, \ldots, b_s, g_1, g_2, \ldots, g_r \rangle. \end{aligned}$$

(6) This follows from part (5) with a slight change in notation.

(7) Since G is abelian, $G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle$ if and only if every element of G can be expressed in the form $u_1u_2\ldots u_r$, where $u_i \in \langle g_i \rangle$, in a unique manner (Theorem 22.15).

Every element of G has at least one such representation if and only if $G = \langle g_1 \rangle \langle g_2 \rangle \ldots \langle g_r \rangle$, that is, if and only if $G = \langle g_1, g_2, \ldots, g_r \rangle$.

We want to show that every element of G has at most one such representation if and only if $\{g_1, g_2, \ldots, g_r\}$ is independent. Equivalently, we will prove that there is an element in G with two different representations if and only if $\{g_1, g_2, \ldots, g_r\}$ is not independent. Indeed, there is an element in G with two different representations if and only if $g_1^{m_1}g_2^{m_2}\ldots g_r^{m_r} = g_1^{n_1}g_2^{n_2}\ldots g_r^{n_r}$ for some integers such that $g_i^{m_i} \neq g_i^{n_i}$ for at least one $i \in \{1, 2, \ldots, r\}$. The latter condition holds if and only if

$$g_1^{m_1-n_1}g_2^{m_2-n_2}\ldots g_r^{m_r-n_r} = 1,$$

where not all of $g_1^{m_1-n_1}, g_2^{m_2-n_2}, \ldots, g_r^{m_r-n_r}$ are equal to 1, that is, if and only if $\{g_1, g_2, \ldots, g_r\}$ is not independent. □

28.5 Lemma: *Let G be a group and $g_1, g_2, \ldots, g_r$ elements of G. Let $B = \langle g_1 \rangle$ and suppose $o(g_i) = o(Bg_i)$ for $i = 2, \ldots, r$.*

(1) *If $\{Bg_2, \ldots, Bg_r\}$ is an independent subset of G/B, then $\{g_1, g_2, \ldots, g_r\}$ is an independent subset of G.*

(2) *Assume $g_1, g_2, \ldots, g_r$ are all distinct from 1. If $G/B = \langle Bg_2 \rangle \times \ldots \times \langle Bg_r \rangle$, then $G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle$.*

Proof: (1) If $m_1, m_2, \ldots, m_r$ are integers such that

$$g_1^{m_1} g_2^{m_2} \ldots g_r^{m_r} = 1, \qquad (*)$$

then $B = Bg_1^{m_1} g_2^{m_2} \ldots g_r^{m_r} = (Bg_1)^{m_1} (Bg_2)^{m_2} \ldots (Bg_r)^{m_r} = (Bg_2)^{m_2} \ldots (Bg_r)^{m_r}$, so $(Bg_2)^{m_2} = \cdots = (Bg_r)^{m_r} = B$ since $\{Bg_2, \ldots, Bg_r\}$ is independent. Thus $o(g_i) = o(Bg_i)$ divides m_i in case $o(g_i)$ is finite and $m_i = 0$ in case $o(g_i) = o(Bg_i)$ is infinite ($i = 2, \ldots, r$). In both cases $g_i^{m_i} = 1$ ($i = 2, \ldots, r$), and, because of $(*)$, $g_1^{m_1} = 1$ as well. Hence $\{g_1, g_2, \ldots, g_r\}$ is independent.

(2) If $G/B = \langle Bg_2 \rangle \times \ldots \times \langle Bg_r \rangle$, then $G/B = \langle Bg_2, \ldots, Bg_r \rangle$ and

$\{Bg_2, \ldots, Bg_r\}$ is independent (Lemma 28.4(7)),

$G = \langle g_1, g_2, \ldots, g_r \rangle$ (Lemma 28.4(6)),

$\{g_1, g_2, \ldots, g_r\}$ is independent (Lemma 28.5(1)),

$G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle$ (Lemma 28.4(7)). □

We now examine the structure of finite abelian groups. A finite abelian group is a direct product of its Sylow p-subgroups. This follows immediately if the existence of Sylow p-subgroups is granted. In order to keep this paragraph independent of §26, we give another proof, from which the existence of Sylow p-subgroups (of finite abelian groups) follows as a bonus. We need a lemma.

28.6 Lemma: *Let A be a finite abelian group and let q be a prime number. If q divides $|A|$, then A has an element of order q.*

Proof: Let $|A| = n$ and let $a_1, a_2, \ldots, a_n$ be the n elements of A. We write $m_i = o(a_i)$ for $i = 1, 2, \ldots, n$. We list all products

$$a_1^{k_1} a_2^{k_2} \ldots a_n^{k_n}$$

where each k_i runs through $0, 1, \ldots, m_i - 1$. Our list has thus $m_1 m_2 \ldots m_n$ entries. Every element of A appears in our list. Two entries $a_1^{k_1} a_2^{k_2} \ldots a_n^{k_n}$ and $a_1^{s_1} a_2^{s_2} \ldots a_n^{s_n}$ are equal if and only if the entry $a_1^{r_1} a_2^{r_2} \ldots a_n^{r_n}$, where r_i is such that $0 \leq r_i \leq m_i - 1$ and $k_i - s_i \equiv r_i \pmod{m_i}$, is equal to the

identity element of A. Thus any element of A appears in our list as many times as 1 does, say t times. The number of entries is therefore $m_1 m_2 \ldots m_n = nt$. Since q divides n, we see $q \mid m_1 m_2 \ldots m_n$ and q divides one of the numbers $m_1, m_2, \ldots, m_n$ (Lemma 5.16), say $q \mid m_1$. Let us put $m_1 = qh$, $h \in \mathbb{N}$. By Lemma 11.9(2), a_1^h has order

$$o(a_1^h) = o(a_1)/(o(a_1),h) = m_1/(m_1,h) = qh/(qh,h) = qh/h = q. \qquad \square$$

28.7 Theorem: *Let G be a finite abelian group and let $|G| = p_1^{a_1} p_2^{a_2} \ldots p_s^{a_s}$ be the canonical decomposition of $|G|$ into prime numbers ($a_i > 0$).*

(1) *For $n \in \mathbb{N}$, we put $G[n] := \{g \in G : g^n = 1\}$. Then $G[n] \leqslant G$ for any $n \in \mathbb{N}$.*
(2) *Let $G_i = G[p_i^{a_i}]$ for $i = 1,2,\ldots,s$. Then $G = G_1 \times G_2 \times \ldots \times G_s$.*
(3) *$|G_i| = p_i^{a_i}$ (and G_i is called a Sylow p_i-subgroup of G).*
(4) *Let H be an abelian group with $|H| = |G|$ and $H_i = H[p_i^{a_i}]$ $(i = 1,2,\ldots,s)$. Then $G \cong H$ if and only if $G_i \cong H_i$ for all $i = 1,2,\ldots,s$.*

Proof: (1) Let $n \in \mathbb{N}$. From $1^n = 1$, we get $1 \in G[n]$, so $G[n] \neq \emptyset$. We use our subgroup criterion.

(i) If $x,y \in G[n]$, then $x^n = 1 = y^n$ and $(xy)^n = x^n y^n = 1.1 = 1$ and so $xy \in G[n]$.

(ii) If $x \in G[n]$ then $x^n = 1$ and $(x^{-1})^n = (x^n)^{-1} = 1^{-1} = 1$ and so $x^{-1} \in G[n]$.

Thus $G[n] \leqslant G$.

(2) We must show that $G = G_1 G_2 \ldots G_s$ and $G_1 \ldots G_{j-1} \cap G_j = 1$ for all $j = 2,\ldots,s$ (Theorem 22.12). We put $|G|/p_i^{a_i} = m_i$ $(i = 1,2,\ldots,s)$. Here the integers $m_1, m_2, \ldots, m_s$ are relatively prime and there are integers $u_1, u_2, \ldots, u_s$ such that $u_1 m_1 + u_2 m_2 + \cdots + u_s m_s = 1$.

We show $G = G_1 G_2 \ldots G_s$. If $g \in G$, then $g = g^{u_1 m_1} g^{u_2 m_2} \ldots g^{u_s m_s}$, with $g^{u_i m_i} \in G_i$ since $(g^{u_i m_i})^{p_i^{a_i}} = g^{u_i |G|} = 1$ $(i = 1,2,\ldots,s)$. Thus $G \subseteq G_1 G_2 \ldots G_s$ and $G = G_1 G_2 \ldots G_s$. Secondly, let $j \in \{2,\ldots,s\}$ and $g \in G_1 \ldots G_{j-1} \cap G_j$. Then $g = g_1 \ldots g_{j-1}$, where $g_1^{p_1^{a_1}} = \cdots = g_{j-1}^{p_{j-1}^{a_{j-1}}} = 1$, therefore $g^{p_1^{a_1} \ldots p_{j-1}^{a_{j-1}}} = 1$ and $o(g) \mid p_1^{a_1} \ldots p_{j-1}^{a_{j-1}}$. On the other hand, $g \in G_j$, so $g^{p_j^{\beta_j}} = 1$ and $o(g) \mid p_j^{a_j}$. Thus $o(g) = 1$ and $g = 1$. Thus $G_1 \ldots G_{j-1} \cap G_j \subseteq 1$ and $G_1 \ldots G_{j-1} \cap G_j = 1$. This proves $G = G_1 \times G_2 \times \ldots \times G_s$.

(3) By the very definition of $G_i = G[p_i^{a_i}]$, the order of any element in G_i is a divisor of $p_i^{a_i}$. Then, by Lemma 28.6, $|G_i|$ is not divisible by any prime number q distinct from p_i. Thus $|G_i| = p_i^{b_i}$ for some b_i, $0 \leqslant b_i \leqslant a_i$. From $p_1^{b_1}p_2^{b_2}\ldots p_s^{b_s} = |G_1|\,|G_2|\ldots|G_s| = |G_1 \times G_2 \times \ldots \times G_s| = p_1^{a_1}p_2^{a_2}\ldots p_s^{a_s}$, we get $p_i^{b_i} = |G_i| = p_i^{a_i}$ for all $i = 1,2,\ldots,s$.

(4) Let $\varphi: G \to H$ be an isomorphism. For any $g \in G_i$, we have $g^{p_i^{a_i}} = 1$, so $(g\varphi)^{p_i^{a_i}} = (g^{p_i^{a_i}})\varphi = 1\varphi = 1$. Thus $g\varphi \in H_i$ and $G_i\varphi \leqslant H_i$. Also, if $h \in H_i$, then $h = g\varphi$ for some $g \in G$ and $(g^{p_i^{a_i}})\varphi = (g\varphi)^{p_i^{a_i}} = h^{p_i^{a_i}} = 1$. Thus $g^{p_i^{a_i}} \in Ker\,\varphi = 1$, so $g^{p_i^{a_i}} = 1$, so $g \in G_i$ and $h = g\varphi \in G_i\varphi$. Hence $H_i \leqslant G_i\varphi$. We obtain $G_i\varphi = H_i$. Consequently, $\varphi_{G_i}: G_i \to H_i$ is an isomorphism and $G_i \cong H_i$ for all $i = 1,2,\ldots,s$.

Conversely, assume $|G| = |H|$ and $G_i \cong H_i$ for all $i = 1,2,\ldots,s$. From part (2), we get $G = G_1 \times G_2 \times \ldots \times G_s$ and $H = H_1 \times H_2 \times \ldots \times H_s$ and Lemma 22.16 gives $G \cong H$. □

According to Theorem 28.7, the structure of a finite abelian group is completely determined by the structure of its Sylow subgroups. Consequently, we focus our attention on finite abelian p-groups. After two prepatory lemmas, the structure of finite abelian p-groups will be described in Theorem 28.10.

28.8 Lemma: *Let G be an abelian group and $g_1, g_2, \ldots, g_r$ elements of G. Let $n \in \mathbb{N}$. We write $G^n = \{g^n : g \in G\}$.*

(1) $G^n \leqslant G$.
(2) *If $G = \langle g_1, g_2, \ldots, g_r\rangle$, then $G^n = \langle g_1^n, g_2^n, \ldots, g_r^n\rangle$.*
(3) *If $G = \langle g_1\rangle \times \langle g_2\rangle \times \ldots \times \langle g_r\rangle$, then $G^n = \langle g_1^n\rangle \times \langle g_2^n\rangle \times \ldots \times \langle g_r^n\rangle$ and $G/G^n \cong \langle g_1\rangle/\langle g_1^n\rangle \times \langle g_2\rangle/\langle g_2^n\rangle \times \ldots \times \langle g_r\rangle/\langle g_r^n\rangle$.*
(4) *Let H be an abelian group. If $G \cong H$, then $G^n \cong H^n$ and $G/G^n \cong H/H^n$.*

Proof: (1) and (2) Since $(ab)^n = a^n b^n$ for all $a,b \in G$, the mapping

$$\psi: G \to G^n$$
$$a \to a^n$$

is a homomorphism onto G^n. So $G^n = Im\ \psi \leqslant G$ by Theorem 20.6. Also, if $G = \langle g_1, g_2, \ldots, g_r \rangle$, then $G^n = \langle g_1\psi, g_2\psi, \ldots, g_r\psi \rangle = \langle g_1^n, g_2^n, \ldots, g_r^n \rangle$ by Lemma 28.4(3).

(3) If $G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle$, then $G = \langle g_1, g_2, \ldots, g_r \rangle$ and $\{g_1, g_2, \ldots, g_r\}$ is independent (Lemma 28.4(7)). Then $G^n = \langle g_1^n, g_2^n, \ldots, g_r^n \rangle$ by part (2). Moreover, $\{g_1^n, g_2^n, \ldots, g_r^n\}$ is independent, for if $m_1, m_2, \ldots, m_r$ are integers and $(g_1^n)^{m_1}(g_2^n)^{m_2}\ldots(g_r^n)^{m_r} = 1$, then $g_1^{nm_1}g_2^{nm_2}\ldots g_r^{nm_r} = 1$, so $(g_i^n)^{m_i} = g_i^{nm_i} = 1$ because $\{g_1, g_2, \ldots, g_r\}$ is independent. From Lemma 28.4(7), we obtain that $G^n = \langle g_1^n \rangle \times \langle g_2^n \rangle \times \ldots \times \langle g_r^n \rangle$. The second assertion follows from Lemma 22.17.

(4) Assume $\varphi: G \to H$ is an isomorphism. For any $g \in G$, $g^n\varphi = (g\varphi)^n \in H^n$, and therefore $G^n\varphi \leqslant H^n$. Also, if $h_1 \in H^n$, then $h_1 = h^n$ for some $h \in H$ and $h = g\varphi$ for some $g \in G$, so $h_1 = h^n = (g\varphi)^n = g^n\varphi \in G^n\varphi$ and thus $H^n \leqslant G^n\varphi$. Hence $H^n = G^n\varphi$ and $\varphi_{G^n}: G^n \to H^n$ is an isomorphism: $G^n \cong H^n$. By Theorem 21.1(7), we have also $G/G^n \cong G\varphi/G^n\varphi = H/H^n$. □

28.9 Lemma: *Let p be a prime number and G a finite abelian p-group. Let $g_1 \in G$ be such that $o(g_1) \geqslant o(a)$ for all $a \in G$ and put $B = \langle g_1 \rangle$. If $Bx \in G/B$ and $o(Bx) = p^m$, then $Bx = Bg$ for some $g \in G$ satisfying $o(g) = p^m$.*

Proof: Let $o(g_1) = p^s$, $o(Bx) = p^m$ and $o(x) = p^u$. Since $(Bx)^{p^u} = Bx^{p^u} = B1 = B$, we have $p^m | p^u$ by Lemma 11.6. Also, $Bx^{p^m} = (Bx)^{p^m} = B$, thus $x^{p^m} \in B = \langle g_1 \rangle$ and $x^{p^m} = g_1^n$ for some $n \in \mathbb{Z}$ with $1 \leqslant n \leqslant p^s$. We write $n = p^k t$, where k and t are integers, $k \geqslant 0$ and $(p,t) = 1$. Then $p^k \leqslant p^k t = n \leqslant p^s$ and, by Lemma 11.9,

$$\begin{aligned} p^{u-m} &= p^u/p^m = p^u/(p^u,p^m) = o(x)/(o(x),p^m) = o(x^{p^m}) \\ &= o(g_1^n) = o(g_1^{p^k t}) = o(g_1)/(o(g_1),p^k t) = p^s/(p^s,tp^k) = p^s/p^k = p^{s-k} \end{aligned}$$

So $p^{s+m-k} = p^u = o(x) \leqslant o(g_1) = p^s$ by hypothesis and $m \leqslant k$.

We put $z = g_1^{tp^{k-m}}$ and $g = z^{-1}x$. Then $z \in \langle g_1 \rangle = B$ and $Bg = Bx$ (Lemma 10.2(5)). From

$$x^{p^m} = g_1^n = g_1^{tp^k} = (g_1^{tp^{k-m}})^{p^m} = z^{p^m},$$

$$g^{p^m} = (z^{-1}x)^{p^m} = (z^{p^m})^{-1}x^{p^m} = 1,$$

we obtain $o(g)|p^m$. Also $p^m = o(Bx) = o(Bg) \leqslant o(g)$. Thus $o(g) = p^m$. This completes the proof.

We can now describe finite abelian groups.

28.10 Theorem: (1) *Let p be a prime number and let G be a nontrivial finite abelian p-group. Then G has a basis, that is, there are elements $g_1, g_2, \ldots, g_r$ in $G \setminus 1$ such that*

$$G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ \ldots \times \langle g_r \rangle.$$

(2) *The number of elements in a basis of G, as well as the orders of the elements in a basis of G, are uniquely determined by G. More precisely, let $\{g_1, g_2, \ldots, g_r\}$ and $\{h_1, h_2, \ldots, h_s\}$ be bases of G, let $o(g_i) = p^{m_i}$ $(i = 1,2,\ldots,r)$ and $o(h_j) = p^{n_j}$ $(j = 1,2,\ldots,s)$, and suppose the notation is so chosen that $m_1 \geqslant m_2 \geqslant \ldots \geqslant m_r > 0$ and $n_1 \geqslant n_2 \geqslant \ldots \geqslant n_s > 0$. Then $r = s$ and the r-tuple $(p^{m_1}, p^{m_2}, \ldots, p^{m_r})$ is equal to the s-tuple $(p^{n_1}, p^{n_2}, \ldots, p^{n_s})$. The r-tuple $(p^{m_1}, p^{m_2}, \ldots, p^{m_r})$ is called the* type of G.

(3) *Let H be a nontrivial finite abelian p-group. Then $G \cong H$ if and only if G and H have the same type.*

Proof: (1) We make induction on u, where $|G| = p^u$. If $u = 1$, then $|G| = p$, so G is cyclic (Theorem 11.13) and the claim is true. Assume now G is a finite abelian p-group, $|G| \geqslant p^2$ and assume that, whenever G_1 is a finite abelian p-group with $1 < |G_1| < |G|$, then G_1 is a direct product of certain nontrivial cyclic subgroups.

We choose an element g_1 of G such that $o(g_1) \geqslant o(a)$ for all $a \in G$ and put $\langle g_1 \rangle = B$. Since $G \neq 1$, we have $B \neq 1$. If $G = B = \langle g_1 \rangle$, the claim is established, so we suppose $B < G$. Then G/B is a finite abelian p-group with $1 < |G/B| < |G|$. By induction, there are elements $Bx_2, \ldots, Bx_r$ of G/B, distinct from $B1$, such that

$$G/B = \langle Bx_2 \rangle \times \ldots \times \langle Bx_r \rangle.$$

Let us put $o(Bx_i) = p^{m_i}$ for $i = 2, \ldots, r$. Using Lemma 28.9, we find $g_i \in G$ such that $Bx_i = Bg_i$ and $o(g_i) = p^{m_i}$ $(i = 2, \ldots, r)$. Let us write $o(g_1) = p^{m_1}$. Then $G/B = \langle Bg_2 \rangle \times \ldots \times \langle Bg_r \rangle$ and, by Lemma 28.5(2),

$$G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ \ldots \times \langle g_r \rangle,$$

where $g_2, \ldots, g_r$ are distinct from 1 since $Bg_2, \ldots, Bg_r$ are distinct from B and g_1 is distinct from 1 since $o(g_1) \geq o(a)$ for all $a \in G$ and $G \neq 1$. This completes the proof of part (1).

(2) and (3) For convenience, a t-tuple $(p^{a_1}, p^{a_2}, \ldots, p^{a_t})$ will be called a *type* of a nontrivial finite abelian p-group if $a_1 \geq a_2 \geq \ldots \geq a_s > 0$ and if A has a basis $\{f_1, f_2, \ldots, f_t\}$ with $o(f_k) = p^{a_k}$ $(k = 1,2, \ldots, t)$. We cannot say *the* type of A, for part (2) is not proved yet. The claim in part (2) is that all types of a nontrivial finite abelian p-group (arising from different bases) are equal.

Let G and H be nontrivial finite abelian p-groups, let $(p^{m_1}, p^{m_2}, \ldots, p^{m_r})$ be a type of G, arising from a basis $\{g_1, g_2, \ldots, g_r\}$ of G and let $(p^{n_1}, p^{n_2}, \ldots, p^{n_s})$ be a type of H, arising from a basis $\{h_1, h_2, \ldots, h_s\}$ of H.

If $r = s$ and $(p^{m_1}, p^{m_2}, \ldots, p^{m_r}) = (p^{n_1}, p^{n_2}, \ldots, p^{n_s})$, then $\langle g_i \rangle \cong C_{p^{m_i}} \cong \langle h_i \rangle$ for $i = 1,2, \ldots, r$ and $G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle \cong \langle h_1 \rangle \times \langle h_2 \rangle \times \ldots \times \langle h_r \rangle = H$ (Lemma 22.16). This proves the "if" part of (3).

Now the "only if" part of (3), which includes (2) as a particular case (when $G = H$): we will prove that $G \cong H$ implies $r = s$ and $(p^{m_1}, p^{m_2}, \ldots, p^{m_r}) = (p^{n_1}, p^{n_2}, \ldots, p^{n_s})$.

Suppose $G \cong H$. We make induction on u, where $|G| = p^u$. If $u = 1$, then $|G| = p = |H|$, so G and H are both cyclic, hence $G = \langle g_1 \rangle$ and $H = \langle h_1 \rangle$. Thus $r = 1 = s$ and $p^{m_1} = o(g_1) = p = o(h_1) = p^{n_1}$. The claim is therefore established when $u = 1$. Now suppose $|G| \geq p^2$ and suppose inductively that, if G_1 and H_1 are isomorphic finite abelian p-groups with $1 < |G_1| < |G|$, and if $(p^{a_1}, p^{a_2}, \ldots, p^{a_{r'}})$ is a type of G_1 and $(p^{b_1}, p^{b_2}, \ldots, p^{b_{s'}})$ is a type of H_1, then $r' = s'$ and $(p^{a_1}, p^{a_2}, \ldots, p^{a_{r'}}) = (p^{b_1}, p^{b_2}, \ldots, p^{b_{s'}})$. We distinguish two cases: the case when $G^p = 1$ and the case $G^p \neq 1$.

In case $G^p = 1$, we have $g^p = 1$ for all $g \in G$, in particular $p^{m_i} = o(g_i) = p$ for all $i = 1,2, \ldots, r$. Also $H^p = 1$ (Lemma 28.8(4)) and $p^{n_j} = o(h_j) = p$ for all $j = 1,2, \ldots, s$. Hence $p^r = |\langle g_1 \rangle| |\langle g_2 \rangle| \ldots |\langle g_r \rangle| = |G| = |H| = |\langle h_1 \rangle| |\langle h_2 \rangle| \ldots |\langle h_r \rangle| = p^s$, so $r = s$ and $(p^{m_1}, p^{m_2}, \ldots, p^{m_r}) = (p, p, \ldots, p) = (p^{n_1}, p^{n_2}, \ldots, p^{n_s})$, as claimed.

Suppose now $G^p \neq 1$. Then $H^p \neq 1$. Thus there are elements in G and H of order $\geq p$, so $p^{m_i} \geq p$ and $p^{n_j} \geq p$. Assume k is the greatest index in

$\{1,2,\ldots,r\}$ with $p^{m_k} > p$, so that (when $k < r$) $p^{m_{k+1}} = \cdots = p^{m_r} = p$. Let the index $l \in \{1,2,\ldots,s\}$ have a similar meaning for the group H. Then

$$(p^{m_1},p^{m_2},\ldots,p^{m_r}) = (p^{m_1},\ldots,p^{m_k},\underbrace{p,\ldots,p}_{r-k \text{ times}}) \qquad (\dagger)$$

$$(p^{n_1},p^{n_2},\ldots,p^{n_r}) = (p^{n_1},\ldots,p^{n_l},\underbrace{p,\ldots,p}_{s-l \text{ times}}), \qquad (\dagger\dagger)$$

it being understood that the entries p should be deleted when $k = r$ or $s = l$. By Lemma 28.8(3),

$$\begin{aligned} G^p &= \langle g_1^p\rangle \times \langle g_2^p\rangle \times \cdots \times \langle g_r^p\rangle \\ &= \langle g_1^p\rangle \times \cdots \times \langle g_k^p\rangle \times \underbrace{\langle 1\rangle \times \cdots \times \langle 1\rangle}_{r-k \text{ times}} \\ &= \langle g_1^p\rangle \times \cdots \times \langle g_k^p\rangle \end{aligned}$$

with $o(g_i^p) = p^{m_i-1} > 1$ for $i = 1,\ldots,k$. Hence $\{g_1^p,\ldots,g_k^p\}$ is a basis and $(p^{m_1-1},\ldots,p^{m_k-1})$ is a type of G^p. In the same way, $(p^{n_1-1},\ldots,p^{n_l-1})$ is a type of H^p. Here G^p is an abelian p-group with $1 < |G^p| = p^{(m_1-1)+\cdots+(m_k-1)} < p^{m_1+m_2+\cdots+m_s} = |G|$. Since $G^p \cong H^p$ by Lemma 28.8(4), our inductive hypothesis gives

$$k = l \quad \text{and} \quad (p^{m_1-1},\ldots,p^{m_k-1}) = (p^{n_1-1},\ldots,p^{n_l-1}).$$

Then $p^{m_i} = p^{n_i}$ for $i = 1,\ldots,k$. From

$$p^{m_1+\cdots+m_k}p^{r-k} = |G| = |H| = p^{n_1+\cdots+n_l}p^{s-l} = p^{m_1+\cdots+m_k}p^{s-l}$$

we get $r - k = s - l = s - k$. Thus $r = s$ and a glance at $(\dagger),(\dagger\dagger)$ shows

$(p^{m_1},p^{m_2},\ldots,p^{m_r}) = (p^{n_1},p^{n_2},\ldots,p^{n_r})$. This completes the proof. □

28.11 Examples: (a) We find all abelian groups of order p^5, where p is a prime number. An abelian group A of order p^5 is determined by its type $(p^{m_1},\ldots,p^{m_r})$, where of course $p^{m_1+\cdots+m_r} = |A| = p^5$. Since $m_i > 0$ and $m_1 + \cdots + m_r = 5$, the only possible types are

$$(p^5),\ (p^4,p),\ (p^3,p^2),\ (p^3,p,p),\ (p^2,p^2,p),\ (p^2,p,p,p),\ (p,p,p,p,p)$$

and any abelian group of order p^5 is isomorphic to one of

$$C_{p^5},\quad C_{p^4}\times C_p,\quad C_{p^3}\times C_{p^2},\quad C_{p^3}\times C_p\times C_p,\quad C_{p^2}\times C_{p^2}\times C_p,$$
$$C_{p^2}\times C_p\times C_p\times C_p,\qquad C_p\times C_p\times C_p\times C_p\times C_p.$$

In particular, there are exactly seven nonisomorphic abelian groups of order p^5.

(b) The number of nonisomorphic abelian groups of order p^n (p prime) can be found by the same argument. This number is clearly the number of ways of writing n as a sum of positive integers $m_1, \ldots, m_r$. If $n \in \mathbb{N}$, an equation of the form $n = m_1 + \cdots + m_r$, where $m_1, m_2, \ldots, m_r$ are natural numbers and $m_1 \geqslant m_2 \geqslant \ldots \geqslant m_r > 0$, is called a *partition of* n. Thus the number of nonisomorphic abelian groups of order p^n is the number of partitions of n. Notice that this number depends only on n, not on p.

The partitions of 6 are

6, 5+1, 4+2, 4+1+1, 3+3, 3+2+1, 2+2+2, 2+2+1+1, 2+1+1+1+1, 1+1+1+1+1+1

and an abelian group of order p^6 is isomorphic to one of

C_{p^6}, $C_{p^5} \times C_p$, $C_{p^4} \times C_{p^2}$, $C_{p^4} \times C_p \times C_p$, $C_{p^3} \times C_{p^3}$, $C_{p^3} \times C_{p^2} \times C_p$,
$C_{p^2} \times C_{p^2} \times C_{p^2}$, $C_{p^2} \times C_{p^2} \times C_p \times C_p$, $C_{p^2} \times C_p \times C_p \times C_p \times C_p$,
$C_p \times C_p \times C_p \times C_p \times C_p \times C_p$.

(c) Let us find all abelian groups of order $324\,000 = 2^5 3^4 5^3$ (to within isomorphism). An abelian group A of this order is the direct product $A_2 \times A_3 \times A_5$, where A_p denotes the Sylow p-subgroup of A ($p = 2,3,5$). Here A_2 has order 2^5 and is isomorphic to one of the seven groups of type

$$(2^5),\ (2^4,2),\ (2^3,2^2),\ (2^3,2,2),\ (2^2,2^2,2),\ (2^2,2,2,2),\ (2,2,2,2,2).$$

Likewise there are five possibilities for A_3:

$$(3^4),\ (3^3,3),\ (3^2,3^2),\ (3^2,3,3),\ (3,3,3,3)$$

and three possibilities for A_5:

$$(5^3),\ (5^2,5),\ (5,5,5).$$

The 7·3·5 various direct products $A_2 \times A_3 \times A_5$ gives us a complete list of nonisomorphic abelian groups of order 324 000.

Now that we obtained a complete classification of finite abelian groups, we turn our attention to torsion-free ones.

28.12 Lemma: *Let G be an abelian group, B a subgroup of G and assume that G/B is a direct product of k infinite cyclic groups ($k \geqslant 1$), say*

$$G/B = \langle By_1 \rangle \times \langle By_2 \rangle \times \ldots \times \langle By_k \rangle$$

$(y_1, y_2, \ldots, y_k \in G)$. *Then* $\langle y_1\rangle, \langle y_2\rangle, \ldots, \langle y_k\rangle$ *are infinite cyclic groups and*

$$G = B \times \langle y_1\rangle \times \langle y_2\rangle \times \ldots \times \langle y_k\rangle.$$

Proof: Let $Y := \langle y_1, y_2, \ldots, y_k\rangle \leqslant G$. Then $G/B = \langle By_1, By_2, \ldots, By_k\rangle$ and, from Lemma 28.4(5), we obtain $G = BY$. We will show that $G = B \times Y$ and $Y = \langle y_1\rangle \times \langle y_2\rangle \times \ldots \times \langle y_k\rangle$.

To establish $G = B \times Y$, we need only prove $B \cap Y = 1$. Let $g \in B \cap Y$. Then $g = y_1^{a_1} y_2^{a_2} \ldots y_k^{a_k}$ for some integers $a_1, a_2, \ldots, a_k$ (Lemma 28.4(1)) and $B = Bg = (By_1)^{a_1}(By_2)^{a_2} \ldots (By_k)^{a_k}$. Since $\{By_1, By_2, \ldots, By_k\}$ is an independent subset of G/B (Lemma 28.4(7)), we get $(By_1)^{a_1} = (By_2)^{a_2} = \cdots = (By_k)^{a_k} = B$. But $o(By_1) = o(By_2) = \cdots = o(By_k) = \infty$ by hypothesis, so $a_1 = a_2 = \cdots = a_k = 0$ and thus $g = y_1^0 y_2^0 \ldots y_k^0 = 1$. This proves $B \cap Y = 1$. Hence $G = B \times Y$.

We now prove $Y = \langle y_1\rangle \times \langle y_2\rangle \times \ldots \times \langle y_k\rangle$. In view of Lemma 28.4(7), we must only show that $\{y_1, y_2, \ldots, y_k\}$ is an independent subset of Y. Suppose $m_1, m_2, \ldots, m_k$ are integers with

$$y_1^{m_1} y_2^{m_2} \ldots y_k^{m_k} = 1.$$

Then $$(By_1)^{m_1}(By_2)^{m_2} \ldots (By_k)^{m_k} = B,$$

so $$m_1 = m_2 = \cdots = m_k = 0$$

and $$y_1^{m_1} = y_2^{m_2} = \cdots = y_k^{m_k} = 1.$$

Hence $\{y_1, y_2, \ldots, y_k\}$ is independent and $Y = \langle y_1\rangle \times \langle y_2\rangle \times \ldots \times \langle y_k\rangle$.

Finally, since By_i has infinite order, we see that y_i has also infinite order and $\langle y_i\rangle$ is an infinite cyclic group $(i = 1, 2, \ldots, k)$. □

28.13 Theorem: *Let G be a finitely generated nontrivial torsion-free abelian group.*

(1) *G has a basis, that is, there are elements $g_1, g_2, \ldots, g_r$ in $G\setminus 1$ such that*

$$G = \langle g_1\rangle \times \langle g_2\rangle \times \ldots \times \langle g_r\rangle.$$

(2) *The number of elements in a basis of G is uniquely determined by G. More precisely, if $\{g_1, g_2, \ldots, g_r\}$ and $\{h_1, h_2, \ldots, h_s\}$ are bases of G, then $r = s$. The number of elements in a basis of G is called the* rank of G.

(3) *Let H be a finitely generated nontrivial torsion-free abelian group. Then $G \cong H$ if and only if G and H have the same rank.*

Proof: (1) Let G be a nontrivial torsion-free abelian group and assume that $G = \langle u_1, u_2, \ldots, u_n \rangle$. We prove the claim by induction on n. If $n = 1$, then $G = \langle u_1 \rangle$ is a nontrivial cyclic group and the claim is true (with $r = 1$, $g_1 = u_1$).

Suppose now $n \geqslant 2$ and suppose inductively: if G_1 is a nontrivial torsion-free abelian group generated by a set of m elements, where $m \leqslant n - 1$, then G_1 is a direct product of a finitely many cyclic subgroups of G.

If $u_1 = 1$, then $G = \langle u_1, u_2, \ldots, u_n \rangle = \langle u_2, \ldots, u_n \rangle$ is generated by a set of $n - 1$ elements and, by induction, G has a basis. Let us assume therefore $u_1 \neq 1$. Then $o(u_1) = \infty$. We put $B/\langle u_1 \rangle := T(G/\langle u_1 \rangle)$.

For any $b \in B$, the element $\langle u_1 \rangle b$ of $B/\langle u_1 \rangle$ has finite order, thus there is a natural number n with $b^n \in \langle u_1 \rangle$. Consequently, for any $b \in B$, there is an $n \in \mathbb{N}$ and $m \in \mathbb{Z}$ such that $b^n = u_1^m$.

We define a mapping $\varphi: B \to \mathbb{Q}$ by declaring $b\varphi = m/n$ for any $b \in B$, where $n \in \mathbb{N}$, $m \in \mathbb{Z}$ are such that $b^n = u_1^m$. This mapping is well defined, for if $n' \in \mathbb{N}$ and $m' \in \mathbb{Z}$ are also such that $b^{n'} = u_1^{m'}$, then $u_1^{m'n - n'm} = (u_1^{m'})^n [(u_1^m)^{n'}]^{-1} = b^{n'n}(b^{nn'})^{-1} = 1$, so $m'n - n'm = 0$ (because $o(u_1) = \infty$) and $m/n = m'/n'$.

φ is in fact a homomorphism. To see this, let $b, c \in B$ and $b\varphi = m/n$, $c\varphi = m'/n'$ (where $n, n' \in \mathbb{N}$, $m, m' \in \mathbb{Z}$). Then $b^n = u_1^m$ and $c^{n'} = u_1^{m'}$, so

$$(bc)^{nn'} = (b^n)^{n'}(c^{n'})^n = u_1^{mn'} u_1^{m'n} = u_1^{mn' + m'n}$$

and $$(bc)\varphi = (mn' + m'n)/nn' = m/n + m'/n' = b\varphi + c\varphi.$$

Thus φ is a homomorphism.

Since $Ker\, \varphi = \{b \in B: b\varphi = 0/1\} = \{b \in B: b^1 = u_1^0\} = 1$, the homomorphism φ is one-to-one and $\varphi: B \to Im\, \varphi$ is an isomorphism: $B \cong Im\, \varphi$.

Claim: if B is finitely generated, then B is cyclic. To prove this, assume $B = \langle b_1, b_2, \ldots, b_t \rangle$ and let $b_i\varphi = m_i/n_i$ $(i = 1,2, \ldots, t)$. Using Lemma 28.4(3), we see that $Im\, \varphi = \langle b_1\varphi, b_2\varphi, \ldots, b_t\varphi \rangle = \langle m_1/n_1, m_2/n_2, \ldots, m_t/n_t \rangle$ is a subgroup of the additive cyclic group $\langle 1/n_1 n_2 \ldots n_t \rangle$. Hence $Im\, \varphi$ is cyclic and B is cyclic.

If $B = G$, then B is finitely generated by hypothesis, so $B = G$ is cyclic and (1) is proved. We assume therefore $B \neq G$. Then

$$G/B = \langle Bu_1, Bu_2, \ldots, Bu_n \rangle = \langle Bu_2, \ldots, Bu_n \rangle$$

(see Lemma 28.4(4)) is a nontrivial abelian group, generated by $n-1$ elements. Moreover, $G/B = G/\langle u_1 \rangle \,/\, B/\langle u_1 \rangle = G/\langle u_1 \rangle \,/\, T(G/\langle u_1 \rangle)$ is torsion-free by Lemma 28.1(2). So, by induction,

$$G/B = \langle Bg_2 \rangle \times \ldots \times \langle Bg_r \rangle$$

with suitable $g_i \in G$, where Bg_i is distinct from B $(i = 2, \ldots, r)$. Thus $o(Bg_i) = \infty$, and this forces $o(g_i) = \infty$ $(i = 2, \ldots, r)$. Lemma 28.12 yields

$$G = B \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle.$$

We put $\langle g_2, \ldots, g_r \rangle = A$. Then $G = B \times A$ and $B \cong G/A$ is finitely generated by Theorem 22.7(2), Lemma 28.4(4). Hence, by the claim above, B is cyclic, say $B = \langle g_1 \rangle$. Since $1 \neq \langle u_1 \rangle \leqslant \langle g_1 \rangle$, we have $o(g_1) \neq 1$, so $o(g_1) = \infty$ and

$$G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle.$$

This completes the proof of (1).

(2) and (3) For convenience, a natural number r will be called a *rank* of a finitely generated nontrivial torsion-free abelian group A if A has a basis of r elements. We cannot say *the* rank of A, for part (2) is not proved yet. The claim in part (2) is that all ranks of a finitely generated nontrivial torsion-free abelian group (arising from different bases) are equal.

Let G and H be finitely generated nontrivial torsion-free abelian groups, let r be a rank of G and s be a rank of H, say $G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle$ and $H = \langle h_1 \rangle \times \langle h_2 \rangle \times \ldots \times \langle h_s \rangle$.

If $r = s$, then $\langle g_i \rangle \cong \mathbb{Z} \cong \langle h_i \rangle$ for $i = 1, 2, \ldots, r$ and

$$G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle \cong \langle h_1 \rangle \times \langle h_2 \rangle \times \ldots \times \langle h_r \rangle \cong H$$

by Lemma 22.16. This proves the "if" part of (3).

Now the "only if" part of (3), which includes (2) as a particular case (when $G = H$): we will prove that $G \cong H$ implies $r = s$. This is easy. Now

$$G/G^2 \cong \langle g_1 \rangle/\langle g_1^2 \rangle \times \langle g_2 \rangle/\langle g_2^2 \rangle \times \ldots \times \langle g_r \rangle/\langle g_r^2 \rangle \cong C_2 \times C_2 \times \ldots \times C_2$$

is a finite group of order 2^r by Lemma 28.8(3). Also

$$H/H^2 \cong \langle h_1\rangle/\langle h_1^2\rangle \times \langle h_2\rangle/\langle h_2^2\rangle \times \ldots \times \langle h_s\rangle/\langle h_s^2\rangle \cong C_2 \times C_2 \times \ldots \times C_2$$

is a finite group of order 2^s. If $G \cong H$, then $G/G^2 \cong H/H^2$ (Lemma 28.8(4)), so $2^r = |G/G^2| = |H/H^2| = 2^s$. Hence $r = s$. □

28.14 Remark: Theorem 28.13 states essentially that a direct sum $\mathbb{Z}^r := \mathbb{Z} \oplus \mathbb{Z} \oplus \ldots \oplus \mathbb{Z}$ of r copies of $\mathbb{Z}$ cannot be isomorphic to a direct sum $\mathbb{Z}^s := \mathbb{Z} \oplus \mathbb{Z} \oplus \ldots \oplus \mathbb{Z}$ of s copies of $\mathbb{Z}$ unless $r = s$. This is not obvious: there are many one-to-one mappings from $\mathbb{Z}^r$ onto $\mathbb{Z}^s$, and there is no a priori reason why one of these mappings should not be an isomorphism. The proof of $\mathbb{Z}^r \cong \mathbb{Z}^s \implies r = s$ does not and cannot consist in cancelling one $\mathbb{Z}$ at a time from both sides of $\mathbb{Z}^r \cong \mathbb{Z}^s$. In general, it does *not* follow from $A \times B \cong A \times C$ that $B \cong C$. As a matter of fact, there are abelian groups G such that $G \cong G \times G \times G$ but $G \ncong G \times G$!

28.15 Theorem: *Let G be a finitely generated abelian group. Then $T(G)$ is a finite group and there is a subgroup I of G such that $G = T(G) \times I$.*

Proof: $G/T(G)$ is a finitely generated abelian group (Lemma 28.4(4)), and is torsion-free (Lemma 28.1(2)). Thus either $G/T(G) \cong 1$; or $G/T(G) \cong \langle T(G)g_1\rangle \times \langle T(G)g_2\rangle \times \ldots \times \langle T(G)g_r\rangle$ with suitable $g_1, g_2, \ldots, g_r \in G$ (Theorem 28.13(1)) and therefore $G = T(G) \times \langle g_1\rangle \times \langle g_2\rangle \times \ldots \times \langle g_r\rangle$ (Lemma 28.12). Putting $I = 1$ in the first case and $I = \langle g_1\rangle \times \langle g_2\rangle \times \ldots \times \langle g_r\rangle$ in the second case, we obtain $G = T(G) \times I$.

Then $T(G) \cong G/I$ by Theorem 22.7(2). Since G is finitely generated, so is G/I (Lemma 28.4(4)) and $T(G)$ is also finitely generated. From Lemma 28.4(2), it follows that $T(G)$ is a finite group. □

The subgroup I in Theorem 28.15 is not uniquely determined by G. However, its rank $r(I)$, which is the rank of $G/T(G)$ is completely determined by G when $G/T(G) \ncong 1$. Let us define the *rank of the trivial group* 1 to be 0 and let us call $\varnothing$ a *basis of* 1. Then the rank of any finitely generated torsion-free abelian group is the number of elements in a

basis of that group, and $r(I)$ is completely determined by G, also in case $G/T(G) \cong 1$.

As $G = T(G) \times I$, the finitely generated abelian group G is determined uniquely to within isomorphism by $T(G)$ and I. Now I is determined uniquely to within isomorphism by the integer $r(I)$ (Theorem 28.13.(3) and the definition $r(1) = 0$); and $T(G)$, being a finite abelian group (Theorem 28.15), is determined uniquely to within isomorphism by its Sylow subgroups (Theorem 28.7(4)). Let s be the number of distinct prime divisors of $|T(G)|$ (so $s = 0$ when $T(G) \cong 1$). Each one of the s Sylow subgroups (corresponding to the s distinct prime divisors) is determined uniquely to within isomorphism by its type (Theorem 28.10(3)). Thus the finitely generated abelian group G gives rise to the following system of nonnegative integers.

(i) A nonnegative integer r, namely the rank of $G/T(G)$. Here $r = 0$ means that G is a finite group. If $r > 0$, then $T(G) \times I$, where I is a direct product of r cyclic groups of infinite order. The subgroup I is not, but its isomorphism type *is* uniquely determined by G.

(ii) A nonnegative integer s, namely the number of distinct prime divisors of $|T(G)|$. Here $s = 0$ means that $T(G) \cong 1$ and G is a torsion-free group.

(iii) In case $s > 0$, a system $p_1, p_2, \ldots, p_s$ of prime numbers, namely the distinct prime divisors of $|T(G)|$; and for each $i = 1,2, \ldots, s$, a positive integer t_i and t_i positive integers $m_{i1}, m_{i2}, \ldots, m_{it_i}$, so that $(p_i^{m_{i1}}, p_i^{m_{i2}}, \ldots, p_i^{m_{it_i}})$ is the type of the Sylow p_i-subgroup of $T(G)$.

With this information, G is a direct product of $r + t_1 + t_2 + \cdots + t_s$ cyclic subgroups. r of them are infinite cyclic; and (in case $s > 0$) t_i of them have orders equal to a prime number p_i, more specifically, t_i of them have orders $p_i^{m_{i1}}, p_i^{m_{i2}}, \ldots, p_i^{m_{it_i}}$. Furthermore, two finitely generated abelian groups are isomorphic if and only if they give rise to the same system of integers.

Exercises

1. Let G be an abelian group and $H \leqslant G$. Prove that
   (a) $T(H) = T(G) \cap H$,
   (b) $T(G)/T(H) \cong HT(G)/H \leqslant T(G/H)$

and that $HT(G)/H$ need not be equal to $T(G/H)$.

2. Let G be an abelian group. Show that

(a) if G is finite, then $G/G^n \cong G[n]$ for all $n \in \mathbb{N}$;

(b) if G is infinite, then $G/G^n \cong G[n]$ need not hold for any $n \in \mathbb{N}\setminus\{1\}$.

3. Let G be a finite abelian group. The *exponent of* G is defined to be the largest number in $\{o(a): a \in G\}$, i.e., the largest possible order of the elements in G. Show that

(a) the exponent of G divides $|G|$;

(b) for any $g \in G$, $o(g)$ divides the exponent of G;

(c) the exponent of G is the least common multiple of the order of the elements in G;

(d) G is cyclic if and only if the exponent of G is $|G|$.

4. Let G be a finite abelian group and $H \leqslant G$. Let $K \leqslant G$ such that $H \cap K = 1$ and $H \cap L \neq 1$ for any $L \leqslant G$ satisfying $K < L$. Let $g \in G$.

(a) Assume $g^p \in K$ for some prime number p. Prove that, if $g \notin K$, then there are $h \in H$, $k \in K$ and an integer r relatively prime to p such that $h = kg^r$. Conclude that $g \in HK$.

(b) Prove that $G = H \times K$ if and only if, for any prime number p and elements $g \in G$, $h \in H$, $k \in K$ such that $g^p = hk$, there is an element $h' \in H$ satisfying $h = (h')^p$.

5. Let G be a finite abelian group of exponent e and let $g \in G$ be of order e, so that $o(g) = e$. Put $H = \langle g \rangle$. Show that $G = H \times K$ for some $K \leqslant G$. (Hint: Use Ex. 4. Consider the cases $p|e$ and $p\nmid e$ separately.

6. Let G be a nontrivial finite abelian group. Using Ex. 5, prove by induction on $|G|$ that there are non-trivial elements $g_1, g_2, \ldots, g_r$ in G such that $G = \langle g_1 \rangle \times \langle g_2 \rangle \times \ldots \times \langle g_r \rangle$ and (in case $r > 1$) $o(g_i)$ divides $o(g_{i+1})$ for $i = 1, 2, \ldots, r-1$.

7. Keep the notation of Ex. 6. Prove that the integers $o(g_1), o(g_2), \ldots, o(g_r)$ determine the types of the Sylow p-subgroups of G uniquely, and conversely the types of the Sylow p-subgroups of G completely determine the integers $o(g_1), o(g_2), \ldots, o(g_r)$. (The integers $o(g_1), o(g_2), \ldots, o(g_r)$ are called the *invariant factors of* G. Two finite abelian groups are thus isomorphic if and only if they have the same invariant factors.)

8. Find the invariant factors of the finite abelian groups $C_6 \times C_9$, $C_6 \times C_8 \times C_{15} \times C_{30}$, $C_4 \times C_6 \times C_{15} \times C_{20}$.

CHAPTER 3

Rings

§29
Basic Definitions

In the preceding chapter, we have examined groups. Groups are sets with one binary operation on them. In this chapter, we want to study sets with two binary operations defined on them. The most fundamental algebraic structure with two binary operations is called a ring.

29.1 Definition: Let R be a nonempty set and let $+$ and $\cdot$ be two binary operations defined on R. The ordered triple $(R,+,\cdot)$ is called a *ring* if the following conditions (ring axioms) are satisfied.

(i) For all $a,b \in R$, $a + b \in R$.

(ii) For all $a,b,c \in R$, $(a + b) + c = a + (b + c)$.

(iii) There is an element in R, denoted by 0, such that

$$a + 0 = a \text{ for all } a \in R.$$

(iv) For each $a \in R$, there is an element in R, denoted by $-a$, such that

$$a + (-a) = 0.$$

(v) For all $a,b \in R$, $a + b = b + a$.

(1) For all $a,b \in R$, $a \cdot b \in R$.

(2) For all $a,b,c \in R$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

(D) For all $a,b,c \in R$, there hold

$$a\cdot(b+c) = a\cdot b + a\cdot c \quad \text{and} \quad (b+c)\cdot a = b\cdot a + c\cdot a.$$

The conditions (i) and (1) assert that two binary operations + and · are defined on R. We shall refer to + as *addition* and to · as *multiplication*. Further, we shall call the element $a + b$ the *sum of a and b*, and the element ab the *product of a and b*. The conditions (i)-(v) say that R forms a group with respect to addition. The identity element 0 of this group will be called the *zero element*, or simply the *zero of R*. So 0 is an element of the set R and not neccessarily the number zero. The inverse element $-a$ of $a \in R$ is called the *opposite of a*.

The condition (2) states that the multiplication on R is associative. The condition (D) relates the two binary operations + and ·. It is called the distributivity of multiplication over addition. Here it should be noted that $a\cdot b + a\cdot c$ stands for $(a\cdot b) + (a\cdot c)$ and similarly $b\cdot a + c\cdot a$ for $(b\cdot a) + (c\cdot a)$. Notice that there are two equations in (D), and we must check both of them when we want to show that a given ordered triple $(R,+,\cdot)$ is a ring. In general, neither of them implies the other, and it is not enough to check one of them. There are ordered triples $(R,+,\cdot)$ for which all the conditions above are satisfied, except for one of the equations in (D), and they fail to be a ring just for that reason.

For ease of notation, we shall frequently denote multiplication by juxtaposition and thus write ab in place of $a\cdot b$. Also, we shall write $a - b$ for $a + (-b)$. Since multiplication in a ring is associative, the products of elements in a ring are independent of the mode of inserting parentheses and the usual exponentiation rules are valid (see §8). We shall use the results of §8 without explicit mention.

29.2 Examples: (a) Let $(R,+)$ be any commutative group, whose identity element we shall denote as 0. We define a multiplication on R by declaring

$$ab = 0 \quad \text{for all } a,b \in R.$$

It is easily seen that $(R,+,\cdot)$ is a ring.

(b) A more interesting ring is $(\mathbb{Z},+,\cdot)$, where $+$ and $\cdot$ are the usual addition and multiplication of integers.

(c) Let $2\mathbb{Z}$ denote the set of even integers. Then $(2\mathbb{Z},+,\cdot)$, where $+$ and $\cdot$ are the usual addition and multiplication of integers, is a ring. In the same way, if $n \in \mathbb{N}$ and $n\mathbb{Z}$ is the set of integers divisibile by n, then $(n\mathbb{Z},+,\cdot)$ is a ring.

(d) $(\mathbb{Q},+,\cdot)$, $(\mathbb{R},+,\cdot)$, $(\mathbb{C},+,\cdot)$, $(\mathbb{Z}_n,+,\cdot)$ are rings under the usual addition and multiplication.

(e) Let $R := \{a/b \in \mathbb{Q}: (a,b) = 1 \text{ and } 5 \nmid b\}$. With respect to the usual addition and multiplication of rational numbers, $(R,+,\cdot)$ is a ring.

(f) Let $S := \{a/b \in \mathbb{Q}: (a,b) = 1 \text{ and } 6 \nmid b\}$. With respect to the usual addition and multiplication of rational numbers, $(S,+,\cdot)$ is a not ring. The very first property (i) is not satisfied. For example

$$\frac{1}{2} \in S,\ \frac{1}{3} \in S,\ \text{but } \frac{1}{2}+\frac{1}{3}=\frac{5}{6} \notin S.$$

(g) Let p be a prime number and put $T = \{a/b \in \mathbb{Q}: (a,b) = 1 \text{ and } p \nmid b\}$. With respect to the usual addition and multiplication of rational numbers, $(T,+,\cdot)$ is a ring.

(h) Let R be a ring. A *matrix over* R is an array $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ of four elements a,b,c,d of R, arranged in two rows and two columns and enclosed within parentheses. The set of all matrices over R will be denoted by $Mat_2(R)$. If $A,B \in Mat_2(R)$, we say A is equal to B provided the corresponding entries in A and B are equal and write $A = B$ in this case. This is clearly an equivalence relation on $Mat_2(R)$.

Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, $B = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \in Mat_2(R)$. The sum $A + B$ of A and B is defined to be the matrix $\begin{pmatrix} a+e & b+f \\ c+g & d+h \end{pmatrix}$ and the product AB of A and B is defined to be the matrix $\begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$.

The proof of Theorem 17.4 remains valid and shows that $Mat_2(R)$ is a commutative group under addition. The proof of Theorem 17.6(1),(2),(4) is also valid and establishes the ring axioms (1),(2),(D). So $(Mat_2(R),+,\cdot)$ is a ring.

(i) Let K be the set of all real-valued functions defined on the closed interval [0,1]. We define operations $+$ and $\cdot$ on K by

$$(f+g)(x) = f(x) + g(x), \quad (f\cdot g)(x) = f(x)g(x) \quad \text{for all } x \in [0,1]$$

$(f,g \in K)$. So $f+g$ is that function that maps any $x \in [0,1]$ to the sum of the values $f(x)$ and $g(x)$ of the functions f and g at x; and $f\cdot g$ is that function that maps any $x \in [0,1]$ to the product of the values $f(x)$ and $g(x)$. In "$f+g$", the sign "+" stands for the binary operation + we just defined, and in "$f(x)+g(x)$", the sign "+" stands for the usual addition of real numbers. It is easily verified that $(K,+,\cdot)$ is a ring. The sum $f+g$ and the product $f\cdot g$ are said to be defined *pointwise*. The operations $+$ and $\cdot$ are called *pointwise addition* and *pointwise multiplication*.

(j) Let S be any set and let $(R,+,\cdot)$ be any ring. Let L denote the set of all functions from S into R. For $f, g \in L$, we put

$$(f+g)(s) = f(s) + g(s), \quad (f\cdot g)(s) = f(s)g(s) \quad \text{for all } s \in S.$$

On the right, we have the sum (product) of elements $f(s),g(s)$ in R, on the left, we have the operations on L. The operations $+$ and $\cdot$ on L are called *pointwise addition* and *pointwise multiplication*. With these operations, $(L,+,\cdot)$ is a ring.

Let us find the zero elements of the rings in Example 29.2. This is the identity element of the commutative group R in Example 29.2(a); the number zero in the Examples 29.2(b),(c),(d),(e),(f),(g) except in the case $\mathbb{Z}_n$ of Example 29.2(d), where the zero element is the residue class $\overline{0} \in \mathbb{Z}_n$ of $0 \in \mathbb{Z}$; the so-called zero matrix $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in Mat_2(R)$, where 0 is the zero element of R in Example 29.2(h). In the ring of Example 29.2(i), the zero element is the function $\zeta\colon [0,1] \to \mathbb{R}$ for which $\zeta(x) = 0$ for all $x \in [0,1]$; and in the ring of Example 29.2(j), the zero element is the function $u\colon S \to R$ for which $u(s) = 0$ for all $s \in S$.

We make a convention. As in the case of groups, if $(R,+,\cdot)$ is a ring, and if it is clear from what the binary operations + and $\cdot$ are, we shall call the *set* R a ring. Hence we shall speak of the ring $\mathbb{Z}$ instead of using the more correct but more cumbersome expression "the ring $(\mathbb{Z},+,\cdot)$", etc.

The addition in a ring has all the desirable properties one could wish for: it is associative, there is an identity element, all elements possess inverses, and it is also commutative. As for multiplication, only one of these properties, namely the associativity, is assumed to be satisfied. It may happen, of course, that multiplication in a ring has some of these properties. Then we make the following definitions.

29.3 Definition: A ring R is called a *commutative* ring if $ab = ba$ for all $a,b \in R$.

29.4 Definition: A ring R is called a *ring with identity* if there is an element e in R such that $ae = ea = a$ for all $a \in R$.

Thus a ring is a commutative ring if the multiplication on it is commutative. This is a natural definition: since addition is commutative in any ring, commutativity can refer only to multiplication. $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $2\mathbb{Z}$ and $\mathbb{Z}_n$ are examples of commutative rings. $Mat_2(\mathbb{Z})$ is not a commutative ring because, for instance,

$$\begin{pmatrix}1&0\\1&1\end{pmatrix}\cdot\begin{pmatrix}0&0\\1&1\end{pmatrix} = \begin{pmatrix}0&0\\1&1\end{pmatrix} \neq \begin{pmatrix}0&0\\2&1\end{pmatrix} = \begin{pmatrix}0&0\\1&1\end{pmatrix}\cdot\begin{pmatrix}1&0\\1&1\end{pmatrix}.$$

Likewise, a ring with identity is a ring with a multiplicative identity. The additive identity exists in any ring anyway. Notice that e in Definition 29.4 must be both a right identity and a left identity. Since multiplication in a ring is not necessarily commutative, we cannot conclude, say, from

$$ae = a \quad \text{for all } a \in R$$

that the other condition

$$ea = a \quad \text{for all } a \in R$$

also holds. In the case of groups, we proved that a right identity is also a left identity, but in the proof we made use of the existence of inverse elements. We cannot use the same argument in the case of rings, for we

do not know anything about the existence of inverse elements. They may or may not exist for all $a \in R$. It is possible that a ring R has an element f such that

$$af = a \quad \text{for all } a \in R$$

but

$$fb \neq b \quad \text{for some } b \in R.$$

In short, R may have a multiplicative right identity which is not a left identity. If each right identity in a ring fails to be a left identity, then the ring is not a ring with identity.

These remarks make sense only for noncommutative rings. Of course, in a commutative ring, any right (left) identity is also a left (right) identity.

A ring may be commutative without having an identity: $2\mathbb{Z}$ is an example. A ring may have an identity without being commutative: $Mat_2(\mathbb{Z})$ is an example. An identity of this ring is the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. More generally, if R is a ring with an identity e, then $Mat_2(R)$ is a ring with an identity $\begin{pmatrix} e & 0 \\ 0 & e \end{pmatrix}$. The proof of Theorem 17.6(3) works here without change.

29.5 Lemma: *Let R be a ring with identity. Then its multiplicative identity is unique (i.e., there is one and only one element e such that $ea = ae = a$ for all $a \in R$).*

Proof: If e and f are identity elements of R, then $e = ef$ since f is a right identity and $ef = f$ since e is a left identity, so $e = ef = f$. □

In view of this lemma, we can speak of *the* identity. We shall follow the convention of writing 1 for the multiplicative identity of a ring with identity. 1 is therefore an element of the ring under study, and not necessarily the number one. For instance, in the ring $Mat_2(\mathbb{Z})$, the element 1 is the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, the identity matrix. The ring K of Example 29.2(i) is a ring with identity, and one checks easily that 1 here is the function $h: [0,1] \to \mathbb{R}$ such that $h(x) = 1$ (real number one) for all $x \in [0,1]$.

What about the existence of multiplicative inverses? Of course the ring must be a ring with identity if we are to speak about multiplicative

inverses. We will see presently that the additive identity 0 of a ring cannot have a multiplicative inverse unless the ring is idiosyncratic.

29.6 Lemma: *Let R be a ring and 0 its zero element.*
(1) $a0 = 0$ *for all* $a \in R$.
(2) $0a = 0$ *for all* $a \in R$.
(3) $a(-b) = -(ab)$ *for all* $a,b \in R$.
(4) $(-a)b = -(ab)$ *for all* $a,b \in R$.
(5) $(-a)b = a(-b)$ *for all* $a,b \in R$.
(6) $(-a)(-b) = ab$ *for all* $a,b \in R$.

Proof: (1) Since 0 is the additive identity of R, we have $0 + 0 = 0$. Thus

$$a(0 + 0) = a0 \quad \text{for all } a \in R,$$
$$a0 + a0 = a0 \quad \text{for all } a \in R.$$

By Lemma 7.3(1), $a0$ must be the identity of the group $(R, +)$. Thus $a0 = 0$.

(2) This is proved by the same agument, using $0a + 0a = (0 + 0)a = 0a$.

(3) For any $a,b \in R$, we have

$$0 = a0 = a(b + (-b)) = ab + a(-b).$$

So $a(-b)$ is the additive inverse of ab. The additive inverse of ab is $-(ab)$ by definition. Hence $a(-b) = -(ab)$.

(4) For any $a,b \in R$, we have

$$0 = 0b = (a + (-a))b = ab + (-a)b.$$

So $(-a)b$ is the additive inverse of ab. The additive inverse of ab is $-(ab)$. Hence $(-a)b = -(ab)$.

(5) This follows from (3) and (4).

(6) This follows from (5) on writing $-b$ for b and observing $-(-b) = b$. □

29.7 Lemma: *Let R be a ring with identity 1. If the zero element 0 of R has an inverse (i.e., if there is an element $t \in R$ such that $0t = t0 = 1$), then R has only one element.*

Proof: If $r \in R$, then $r = r1 = r(0t) = (r0)t = 0t = 0$, so $R \subseteq \{0\}$, so $R = \{0\}$. □

The set $\{0\}$ can be made into a ring if we define $+$ and $\cdot$ in the only possible way: $0 + 0 = 0$ and $0 \cdot 0 = 0$. This is a commutative ring with identity, the multiplicative identity being the additive identity 0. This ring is called the *null ring*.

29.8 Lemma: *Let R be a ring with identity 1. If R is not the null ring, then $1 \neq 0$.*

Proof: If R is not the null ring, then there is an $r \in R$, $r \neq 0$. Then the assumption $1 = 0$ leads to the contradiction $r = r1 = r0 = 0$. So $1 \neq 0$. □

Lemma 29.7 states that 0 in a ring cannot possess a multiplicative inverse unless the ring is the null ring. We now want to show that divisors of 0 cannot possess a multiplicative inverses, either.

29.9 Definition: Let R be a ring. If $a \neq 0$, $b \neq 0$ are elements of R such that $ab = 0$, then a is called a *left zero divisor* and b is called a *right zero divisor*.

It may very well happen that $a \neq 0$, $b \neq 0$, but $ab = 0$ in a ring. For example, in the ring $Mat_2(\mathbb{Q})$ of matrices over $\mathbb{Q}$,

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \neq 0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \neq 0, \text{ but } \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = 0.$$

As a second example, consider the ring K of real-valued functions on $[0,1]$ with respect to pointwise addition and multiplication (Example 29.2(i)). The zero element in this ring is the function ζ, where $\zeta(x) = 0 \in \mathbb{R}$ for all $x \in [0,1]$. The functions a and b, where

$$a(x) = \begin{cases} 0 \text{ if } 0 \leq x \leq 1/2 \\ 1 \text{ if } 1/2 < x \leq 1 \end{cases}, \quad b(x) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1/2 \\ 0 \text{ if } 1/2 < x \leq 1 \end{cases}$$

are thus distinct from ζ, but their pointwise product is ζ, as $a(x)b(x) = 0$ for all $x \in [0,1]$.

In a commutative ring, there is no distinction between right and left zero divisors. But in a non commutative ring, an element $a \neq 0$ may be a right zero divisor without being a left zero divisor, and vice versa.

29.10 Lemma: *Let R be a ring with identity. If a is a left zero divisor, then a does not have a multiplicative left inverse. If a is a right zero divisor, then a does not have a multiplicative right inverse.*

Proof: Let 1 be the identity of R. If a is left zero divisor, then $a \neq 0$ and there is a $b \neq 0$ in R such that $ab = 0$. Now if a had a left inverse x, so that $xa = 1$, we would obtain $b = 1b = (xa)b = x(ab) = x0 = 0$, a contradiction. So a has no left inverse. The second statement is proved analogously. □

We know that the zero element in a ring distinct from the null ring cannot have an inverse and we understand from Lemma 29.10 that being a zero divisor is the very opposite of having an inverse. So if we want a ring to have the property that every nonzero element in it has a multiplicative inverse, the ring has to be free from zero divisors.

29.11 Definition: A commutative ring with identity, which is distinct from the null ring, and which has no zero divisors, is called an *integral domain*.

29.12 Definition: A ring with identity, which is distinct from the null ring, and in which every nonzero element has a right inverse, is called a *division ring*.

An integral domain is therefore a ring in which we may expect that nonzero elements have inverses, but nothing is said about the actual existence of inverses. The necessary condition that zero divisors be absent

is satisfied in an integral domain, plus commutativity. Whether the nonzero elements do in fact have inverses is not relevant in the definition of integral domains.

In a division ring, every nonzero element does have a right inverse; more precisely, a right inverse. But this means that the nonzero elements in a division ring form a group under multiplication. We know that, in any group, right inverses are also left inverses and that they are unique (Lemma 7.3). Hence, in a division ring, every nonzero element has a left inverse as well, and the right and left inverse of an arbitrary element coincide. This will be called *the* inverse of that element.

$\mathbb{Z}$ is an integral domain. In fact, $\mathbb{Z}$ is the prototype of all integral domains. $2\mathbb{Z}$ is not an integral domain, because $2\mathbb{Z}$ is not a ring with identity, although $2\mathbb{Z}$ is commutative and has no zero divisors. An example of division rings is given in Ex. 9.

A ring which is both an integral domain and a division ring deserves a name.

29.13 Definition: A commutative ring with identity, which is distinct from the null ring, and in which every nonzero element has a multiplicative inverse, is called a *field*.

Thus a field is a commutative division ring. Also, a field is an integral domain in which every nonzero element does have an inverse. A field is a ring in which the nonzero elements form a commutative group under multiplication.

$\mathbb{Z}$ is not a field, since $2 \in \mathbb{Z}$, for instance, does not have an inverse in $\mathbb{Z}$ (there is no $z \in \mathbb{Z}$ such that $2z = 1$). Thus $\mathbb{Z}$ is an integral domain which is not a field. The rings $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Z}_p$ (where p is a prime number) are example of fields, so Definition 17.1 is consistent with Definition 29.13. There are fields with finitely many elements as well as with infinitely many elements.

29.14 Definition: Let R be a ring with identity. An element $a \in R$ of R is sait to be a *unit of* R if a has both a right inverse and a left inverse in R. The set of all units in R will be denoted by $R^\times$.

For example, the units of $\mathbb{Z}$ are 1 and -1, so $\mathbb{Z}^\times = \{1,-1\}$. The units in $\mathbb{Z}_n$ are the residue classes $\bar{a}$ for which there is a $\bar{b} \in \mathbb{Z}_n$ such that $\bar{a}\,\bar{b} = \bar{1}$, and this holds if and only if $(a,n) = 1$. Hence $\mathbb{Z}_n^\times = \{\bar{a} \in \mathbb{Z}_n : (a,n) = 1\}$, as in §11. We know that $\mathbb{Z}^\times = \{1,-1\}$ and $\mathbb{Z}_n^\times$ are groups under multiplication (Theorem 12.4). This are a special cases of the following theorem.

29.15 Theorem: *Let R be a ring with identity. Then $R^\times$ is a group under multiplication.*

Proof: We denote the identity of R by 1. Since $1 \cdot 1 = 1$, we have $1 \in R^\times$ and so $R^\times \neq \emptyset$. We now show that any unit of R has a unique right inverse, which is also the unique left inverse of that unit. Let $a \in R^\times$, let x be *any* right inverse of a and let y be *any* left inverse of a. Then $ax = 1 = ya$ and

$$y = y1 = y(ax) = (ya)x = 1x = x.$$

Thus any right inverse of a is equal to y. Hence there is only one right inverse of a, namely x. Then any left inverse of a is also equal to x. Hence there is a unique left inverse of a, namely the unique right inverse x of a.

We check the group axioms.

(i) If $a,b \in R^\times$, then there are uniquely determined elements x, z in R with $ax = 1 = xa$ and $bz = 1 = zb$. From

$$(ab)(zx) = a(bz)x = a1x = ax = 1, \quad (zx)(ab) = z(xa)b = z1b = zb = 1,$$

we see that zx is both a right inverse and a left inverse of ab. Hence $ab \in R^\times$ and $R^\times$ is closed under multiplication.

(ii) The multiplication on $R^\times$ is associative since R is a ring.

(iii) Since $a1 = a = 1a$ for all $a \in R$, and since $1 \in R^\times$, we see that 1 is the identity element of $R^\times$.

(iv) If $a \in R^\times$, then there is an $x \in R$ with $ax = 1 = xa$. This x is in fact an element of $R^\times$: it follows from $ax = 1 = xa$ that a is a left and right inverse of x, so $x \in R^\times$. So any $a \in R^\times$ has an inverse in $R^\times$.

Thus $R^\times$ is a group under multiplication. □

The reader will check easily that, if R is a ring with identity, distinct from the null ring, then R is a division ring if and only if $R^\times = R\backslash\{0\}$. Likewise, if K is a commutative ring with identity, distinct from the null ring, then K is a field if and only if $K^\times = K\backslash\{0\}$.

From now on we will write $\mathbb{Q}^\times$, $\mathbb{R}^\times$, $\mathbb{C}^\times$ for the multiplicative groups $\mathbb{Q}\backslash\{0\}$, $\mathbb{R}\backslash\{0\}$, $\mathbb{C}\backslash\{0\}$ of nonzero rational, real, complex numbers, respectively.

We conclude this paragraph with the binomial theorem.

29.16 Theorem (Binomial Theorem) : *Let R be a ring and $a,b \in R$. If $ab = ba$, then* $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$.

Proof: First, we remark that $a^n b^0$ and $a^0 b^n$ are to be interpreted as a^n and b^n respectively, even if R has no identity. As usual, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ and $0! = 1$. We use the formula $\binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}$ for $1 \leqslant k \leqslant n-1$.

We make induction on n. The formula $(a+b)^1 = \binom{1}{0} a^1 b^0 + \binom{1}{1} a^0 b^1$ is clearly true. We suppose that the formula is proved when the exponent of $a+b$ is n. Then

$$(a+b)^{n+1} = (a+b)(a+b)^n = (a+b) \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$$

$$= \sum_{k=0}^{n} \binom{n}{k} a^{n+1-k} b^k + \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^{k+1}$$

$$= \binom{n}{0} a^{n+1}b^0 + \sum_{k=1}^{n} \binom{n}{k} a^{n+1-k}b^k + \sum_{k=0}^{n-1} \binom{n}{k} a^{n-k}b^{k+1} + \binom{n}{n} a^0 b^{n+1}$$

$$= \binom{n+1}{0} a^{n+1}b^0 + \sum_{k=1}^{n} \binom{n}{k} a^{n+1-k}b^k + \sum_{k=1}^{n} \binom{n}{k-1} a^{n-(k-1)}b^k + \binom{n+1}{n+1} a^0 b^{n+1}$$

$$= \binom{n+1}{0} a^{n+1}b^0 + \sum_{k=1}^{n} \left(\binom{n}{k} + \binom{n}{k-1} \right) a^{n+1-k}b^k + \binom{n+1}{n+1} a^0 b^{n+1}$$

$$= \binom{n+1}{0} a^{n+1}b^0 + \sum_{k=1}^{n} \binom{n+1}{k} a^{n+1-k}b^k + \binom{n+1}{n+1} a^0 b^{n+1}$$

$$= \sum_{k=0}^{n+1} \binom{n+1}{k} a^{n+1-k}b^k$$

and the formula is true when the exponent of $a + b$ is $n + 1$. This completes the proof. □

Exercises

1. Let $X = \{a + b\sqrt{2} : a,b \in \mathbb{Z}\}$ and $Y = \{a + b\sqrt[3]{2} : a,b \in \mathbb{Z}\}$. Determine whether X and Y are rings under the usual addition and multiplication of real numbers.

2. Let $(R,+,\cdot)$ be a ring. On the group $(R,+)$, we define an operation $\circ$ by declaring $a \circ b = ba$ for all $a,b \in R$. Show that $(R,+,\circ)$ is a ring (called the *opposite* ring of $(R,+,\cdot)$).

3. On the group $\mathbb{Z} \oplus \mathbb{Z}$, we define a multiplication by

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

for all $(a,b), (c,d) \in \mathbb{Z} \oplus \mathbb{Z}$. Does $\mathbb{Z} \oplus \mathbb{Z}$ become a ring with this multiplication?

4. Show that the set $A = \{a/b \in \mathbb{Q}: (a,b) = 1, n \nmid b\}$ is not a ring (under the usual addition and multiplication of rational numbers) if n is a composite number.

5. Prove that $\mathbb{Z}_n$ has zero divisors if n is composite, and that $\mathbb{Z}_n$ is a field if n is prime.

6. On the group $R = \mathbb{Z} \oplus \mathbb{Z}$, we define a multiplication by

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

for all $(a,b), (c,d) \in \mathbb{Z} \oplus \mathbb{Z}$. Prove that, with this multiplication, R becomes a ring. Show that (1,0) is a left identity in R, but not a right identity; and that (1,0) is a right zero divisor, but not a left zero divisor. Is R a ring with identity?

7. Let R be a ring without identity, and let $S = R \oplus \mathbb{Z}$. On the commutative group S, we define a multiplication by

$$(r,a)\cdot(r',b) = (rr' + ar' + br, ab)$$

for all $(r,a), (r',b) \in S$. Prove that S is a ring with identity.

8. On the group $R = \mathbb{Z}_n \oplus \mathbb{Z}_n$, we define a multiplication by

$$(\bar{a},\bar{b})\cdot(\bar{c},\bar{d}) = (\overline{ac - bd}, \overline{ad + bc})$$

for all $(\bar{a},\bar{b}),(\bar{c},\bar{d}) \in R$. Show that R is a commutative ring with identity. Prove that R is a field when $n = 3,7,11$ and that R is not an integral domain if $n = 5,13,17$.

9. Let $H = \left\{ \begin{pmatrix} a & -b \\ b & a \end{pmatrix}: a,b \in \mathbb{C} \right\} \subseteq Mat_2(\mathbb{C})$. Prove that, under the usual matrix addition and multiplication, H is a division ring (cf. §17, Ex. 14).

10. Let $R_1, R_2, \ldots, R_n$ be rings. Prove that the group $R_1 \oplus R_2 \oplus \ldots \oplus R_n$ becomes a ring if multiplication is defined by

$$(r_1, r_2, \ldots, r_n)(s_1, s_2, \ldots, s_n) = (r_1 s_1, r_2 s_2, \ldots, r_n s_n)$$

for all $(r_1, r_2, \ldots, r_n), (s_1, s_2, \ldots, s_n) \in R_1 \oplus R_2 \oplus \ldots \oplus R_n$. Moreover, prove that $R_1 \oplus R_2 \oplus \ldots \oplus R_n$ is a commutative ring if and only if each R_k is; and that $R_1 \oplus R_2 \oplus \ldots \oplus R_n$ is a ring with identity if and only if each R_k is. The ring $R_1 \oplus R_2 \oplus \ldots \oplus R_n$ is called the *direct sum of* $R_1, R_2, \ldots, R_n$.

§30
Subrings, Ideals and Homomorphisms

As in the case of groups, we give a name to subsets of a ring which are themselves rings.

30.1 Definition: Let R be a ring. A nonempty subset S of R is called a *subring* of R if S itself is a ring with respect to the operations on R.

Thus a nonempty subset S of a ring R is a subring of R if and only if S satisfies all the ring axioms in Definition 29.1. As in the case of groups, we can dispense with some of them.

Let $(R,+,\cdot)$ be a ring and $\varnothing \neq S \subseteq R$. If S is a subring of R, then $(S,+)$ is a commutative group, thus $(S,+)$ is a subgroup of $(R,+)$; and $(S,+)$ is a subgroup of $(R,+)$ if and only if

(i) $a + b \in S$ for all $a,b \in S$,

(ii) $-a \in S$ for all $a \in S$,

as we know from Lemma 9.2. Let us now consider multiplication. If $(S,+,\cdot)$ is to be a ring, the the restriction of the operation $\cdot$ to S must be a binary operation on S; and this holds if and only if

(1) $a \cdot b \in S$ for all $a \in S$.

So, if a nonempty subset S of a ring R is a subring of R, then (i),(ii),(1) hold. Conversely, if S is a nonempty subset of a ring R and (i),(ii),(1) hold, then $(S,+)$ is a subgroup of $(R,+)$, so $(S,+)$ is a a commutative group, and $\cdot$ is a binary operation on S, and the associativity of multiplication and the disributivity of multiplication over addition holds in S since they hold in fact in R. Thus $(S,+,\cdot)$ is a subring of $(R,+,\cdot)$. We proved the following lemma.

30.2 Lemma (Subring criterion): *Let $(R,+,\cdot)$ be a ring and let S be a nonempty subset of R. Then $(S,+,\cdot)$ is a subring of R if and only if*

(i) $a + b \in S$ *for all* $a,b \in S$,

(ii) $-a \in S$ *for all* $a \in S$,
(iii) $a \cdot b \in S$ *for all* $a,b \in S$. □

30.2′ Examples: (a) $\{0\}$ and R are subrings of any ring R.

(b) If R is a ring and S_i is an arbitrary collection of subrings of R, then it follows immediately from Lemma 30.2 that $\bigcap_{i \in I} S_i$ is a subring of R.

(c) If R is a ring and X is a subset of R, the intersection of all subrings of R that contain X is a subring of R by Example 30.2′(b). It is called the *subring generated by X.*

Some properties of multiplication are inherited by subrings.

30.3 Lemma: (1) A subring of a commutative ring is a commutative ring.
(2) A subring of a noncommutative ring can be commutative.
(3) A subring of a ring with identity can be a ring without identity.
(4) A subring of a ring without identity can be a ring with identity.
(5) A subring of a ring without zero divisors is a ring without zero divisors.
(6) A subring of a ring with zero divisors can be a ring without zero divisors.
(7) A subring of a division ring is not necessarily a division ring.
(8) A subring of a field is not necessarily a field.
(9) A subring, distinct from $\{0\}$, of an integral domain is an integral domain if and only if it contains the identity.

Proof: Let R be a ring and S a subring of R.

(1) If R is commutative, then $ab = ba$ for all $a,b \in R$ and, a fortiori, $ab = ba$ for all $a,b \in S$. Hence S is commutative.

(2) Assume R is not commutative. Then there are $a,b \in R$ with $ab \neq ba$. The point is that all such pairs a,b may be outside S, and that $st = ts$ may

hold for all $s,t \in S$. For example, $R = Mat_2(\mathbb{Q})$ is not commutative, but $S = \left\{ \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} : a,b \in \mathbb{Q} \right\}$ is a subring of R and S is commutative.

(3) The point is that the identity 1 of R need not belong to S. For example, $\mathbb{Z}$ is a ring with identity, $2\mathbb{Z}$ is a subring of $\mathbb{Z}$ and $2\mathbb{Z}$ has no identity.

(4) The point is that there may be an e in S such that $es = se = s$ for all s in S, but $er = re = r$ need not be true for all $r \in R$, i.e., $er_0 \neq r_0$ or $r_0 e \neq r_0$ for a particular r_0 in R. As an example, consider $R = \mathbb{Z} \times \mathbb{Z}$, on which addition and multiplication are defined by declaring

$$(a,b) + (c,d) = (a + c, b + d)$$
$$(a,b)(c,d) = (ac, ad)$$

for all $(a,b) \in R$ and which is easily verified to be a ring with respect to these operations. If $(a,b) \in R$ is a left identity element of R so that $(a,b)(x,y) = (x,y)$ for all $(x,y) \in R$, then $(ax,ay) = (x,y)$ for all $(x,y) \in R$, thus $a = 1$. But $(1,b)$ is not a right identity element of R, because $(x,y)(1,b) = (x,xb) \neq (x,y)$ for any $(x,y) \in R$ with $y \neq xb$. Thus R is a ring without an identity. However, $S = \{(a,0): a \in \mathbb{Z}\}$ is a subring of R with an identity $(1,0) \in S$, as $(1,0)(a,0) = (a,0) = (a,0)(1,0)$ for any $(a,0) \in S$.

(5) If R has no zero divisors, then

$$\text{for all } a,b \in R, \qquad a \neq 0 \neq b \implies ab \neq 0.$$

But this holds for all $a,b \in S$, too. Hence S has no zero divisors.

(6) If R has no zero divisors, it may happen that all zero divisors fall outside S, and in this case S has no zero divisors. For instance, The ring $R = Mat_2(\mathbb{Q})$ has zero divisors, but its subset $S = \left\{ \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix} : a \in \mathbb{Q} \right\}$ is a subring of R with no zero divisors. For if $s,t \in S$ and $st = 0$, then $s = \begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$ and $t = \begin{pmatrix} b & 0 \\ 0 & 0 \end{pmatrix}$ for some $a,b \in \mathbb{Q}$, and $st = 0$ means $\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} b & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, which is possible only if $a = 0$ or $b = 0$ (in $\mathbb{Q}$), that is, only if $s = 0$ or $t = 0$ (in S).

(7) and (8) Consider the division ring $\mathbb{Q}$, which is a field as well. Its subring $\mathbb{Z}$ is neither a division ring nor a field.

(9) A subring $S \neq \{0\}$ of an integral domain R is commutative by (1) and has no zero divisors by (5). Hence S is an integral domain if and only if S has an identity. We claim S has an identity if and only if the identity of R belongs to S. Indeed, if S contains the identity element 1_R of R, then of course 1_R is an identity element of S. Conversely, if S has an identity element e, then $ee = e = 1_R e$, so $ee - 1_R e$, so $(e - 1_R)e = 0$ and, since $e \neq 0$ (for $S \neq \{0\}$ by assumption) and R has no zero divisors, $e - 1_R = 0$ and hence e must be equal to 1_R. □

The claim in the proof of Lemma 30.3(9) is not self-evident. If R is a ring with identity and S is a subring of R, then it is possible that S is a ring with identity and the identity of S is *distinct* from the identity of R. Can you give some examples?

Just as in the case of groups, we want to define factor rings by subrings. We take our factor group construction as a model. For a group G and a subgroup H of G, the factor group G/H is the set of all right cosets of H in G, on which the multiplication is defined by the rule $Ha \cdot Hb = Hab$. In order that this multiplication be well defined, it is necessary and sufficient that H be normal in G (Theorem 18.4).

Now let R be a ring and S a subring of R. Then R is an abelian group with respect to addition and S is a subgroup of R. Using our results in group theory, we build the factor group R/S. This is possible because S is a normal subgroup of R (any subgroup of an abelian group is normal in that group). The elements of R/S are the (right or left) cosets $r + S$, where r ranges over R. Of course we must write the cosets as $r + S$ or as $S + r$, not as rS or as Sr, for the group R is an additive group. We now wish to define a multiplication on R/S and make R/S into a ring.

The most natural way to define a multiplication on R/S is to put

$$(r + S)(u + S) = ru + S \qquad \text{for all } r,u \in R.$$

Let us see if this multiplication is well defined. Once we show that this multiplication is well defined, it is routine to prove that R/S becomes a ring with this multiplication. This multiplication is well defined if and only if the implication

$$r_1 + S = r_2 + S,\ t_1 + S = t_2 + S \implies r_1 t_1 + S = r_2 t_2 + S. \quad (\text{for all } r_1, r_2, t_1, t_2 \in R)$$

holds, and it holds if and only if

$r_1 = r_2 + s_1,\ t_1 = t_2 + s_2,\ s_1,s_2 \in S \implies r_1t_1 - r_2t_2 \in S$ (for all $r_1,r_2,t_1,t_2 \in R$), i.e., if and only if

$$s_1,s_2 \in S \implies (r_2 + s_1)(t_2 + s_2) - r_2t_2 \in S \quad \text{(for all } r_2,t_2 \in R),$$

i.e., if and only if

$$s_1,s_2 \in S \implies r_2s_2 + s_1t_2 + s_1s_2 \in S \quad \text{(for all } r_2,t_2 \in R),$$

that is, since $s_1s_2 \in S$ when $s_1,s_2 \in S$, if and only if

$$s_1,s_2 \in S \implies rs_2 + s_1t \in S \qquad \text{(for all } r,t \in R) \qquad (*)$$

is true. We dropped the subscripts of r_2 and t_2.

Assume $(*)$ holds. Then, choosing $t = 0$, we see $rs_2 \in S$ whenever $r \in R$, $s_2 \in S$; and choosing $r = 0$, we see $s_1t \in S$ whenever $s_1 \in S$, $t \in R$. Conversely, if $rs_2 \in S$ and $s_1t \in S$ whenever $r \in R$, $s_2 \in S$ and $s_1 \in S$, $t \in R$, then $rs_2 + s_1t \in S$ for all $r,t \in R$, $s_2,s_1 \in S$, since S is a subgroup of R with respect to addition. Thus $(*)$ is equivalent to, and the multiplication on R/S is well defined if and only if:

for all $s \in S$, $r \in R$, there hold $rs \in S$ and $sr \in S$. $(**)$

Subrings with this property have a name.

30.4 Definition: A nonempty subset S of a ring R is called an *ideal of* R if the following two conditions are satisfied.

(i) S is a subgroup of R under addition.

(ii) For all $s \in S$, $r \in R$, we have $rs \in S$ and $sr \in S$.

According to this definition, an ideal of a ring R is a subring of R, since it is closed under multiplication by (ii). The condition (ii) tells more than simply that the product of an element in S by an element in S is in S. It tells that the product of any element in R by any element in S, as well as the product of any element in S by an element in R, are both in S. Thus S "swallows" or "absorbs" products by elements in R.

The condition (ii) consists of two subconditions: $rs \in S$ and $sr \in S$. In a commutative ring, these subconditions are identical. But when R is not commutative, neither of them implies the other in general, and one of them is not enough to make S an ideal: both of them ought to hold.

Definition 30.4 and the discussion preceding it give us the following theorem (cf. Theorem 18.4).

30.5 **Theorem:** *Let R be a ring and S a subgroup of R under addition. The multiplication on the set R/S of right (and left) cosets of S, given by*

$$(r + S)(u + S) = ru + S \quad \text{for all } r,u \in R$$

is well defined if and only if S is an ideal of R. □

After giving some examples of ideals, we will prove that the multiplication on R/S makes R/S into a ring.

30.6 **Examples:** (a) In any ring R, the set $\{0\}$ is an ideal (Lemma 29.6(1) and (2)). The set R itself is also an ideal of R since R is closed under multiplication.

(b) In the ring $\mathbb{Z}$ of integers, $2\mathbb{Z}$ is a subring and in fact an ideal of $\mathbb{Z}$, since the product of an even integer by an arbitrary integer is always an even integer. In the same way, the set $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$ ($n \in \mathbb{N}$).

(c) Let K be the ring of real-valued functions on [0,1] (Example 29.2(i)). Its subset $\{f \in K: f(1/2) = 0\}$ is an ideal of K. Similarly, when Y is a subset of [0,1], the subset $\{f \in K: f(y) = 0 \text{ for all } y \in Y\}$ is an ideal of K.

(d) Let $T = \{a/b \in \mathbb{Q}: (a,b) = 1, p \nmid b\}$ be the ring in Example 29.2(g). Then its subsets

$$A = \{a/b \in \mathbb{Q}: (a,b) = 1, p \nmid b, p \mid a\} \text{ and } \{a/b \in \mathbb{Q}: (a,b) = 1, p \nmid b, p^2 \mid a\}$$

are ideals of T.

(e) $\mathbb{Z}$ is not an ideal of $\mathbb{Q}$, since for example, $1 \in \mathbb{Z}$, $\frac{1}{2} \in \mathbb{Q}$, but $\frac{1}{2} \cdot 1 \notin \mathbb{Z}$.

(f) Consider the subset $S = \left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix}: a,b \in \mathbb{Q} \right\}$ of $Mat_2(\mathbb{Q})$. Then S is a subring of $Mat_2(\mathbb{Q})$. Also, one sees easily that $rs \in S$ for all $r \in Mat_2(\mathbb{Q})$, $s \in S$. Nevertheless, S is not an ideal of $Mat_2(\mathbb{Q})$, since it is not true that

$sr \in S$ for all $r \in Mat_2(\mathbb{Q})$, $s \in S$: for example $\left(\begin{smallmatrix}1&0\\1&0\end{smallmatrix}\right) \in S$, $\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right) \in Mat_2(\mathbb{Q})$, but $\left(\begin{smallmatrix}1&0\\1&0\end{smallmatrix}\right)\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right) = \left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right) \notin S$.

Now let $S_1 = \left\{ \left(\begin{smallmatrix}c&0\\0&0\end{smallmatrix}\right) : c \in \mathbb{Q} \right\}$ and $S_2 = \left\{ \left(\begin{smallmatrix}a&0\\0&b\end{smallmatrix}\right) : a,b \in \mathbb{Q} \right\}$. It is easy to see that S_1 and S_2 are subrings of $Mat_2(\mathbb{Q})$ and of course $S_1 \subseteq S_2$. Here S_1 is an ideal of S_2, because $\left(\begin{smallmatrix}a&0\\0&b\end{smallmatrix}\right)\left(\begin{smallmatrix}c&0\\0&0\end{smallmatrix}\right) = \left(\begin{smallmatrix}c&0\\0&0\end{smallmatrix}\right)\left(\begin{smallmatrix}a&0\\0&b\end{smallmatrix}\right) = \left(\begin{smallmatrix}ac&0\\0&0\end{smallmatrix}\right) \in S_1$ for any $\left(\begin{smallmatrix}a&0\\0&b\end{smallmatrix}\right) \in S_2$ and $\left(\begin{smallmatrix}c&0\\0&0\end{smallmatrix}\right) \in S_1$. On the other hand, S_1 is not an ideal of $Mat_2(\mathbb{Q})$, because, for example, $\left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right) \in S_1$, $\left(\begin{smallmatrix}1&1\\0&0\end{smallmatrix}\right) \in Mat_2(\mathbb{Q})$ and yet $\left(\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right)\left(\begin{smallmatrix}1&1\\0&0\end{smallmatrix}\right) = \left(\begin{smallmatrix}1&1\\0&0\end{smallmatrix}\right) \notin S_1$. Thus S_1 is an ideal of S_2 but not an ideal of $Mat_2(\mathbb{Q})$. This shows that "idealness" is not an intrinsic property of a subring. A subring is not merely an ideal, but an ideal of a ring that has to be clearly specified. Compare this with Example 18.5(i).

(g) Intersection of ideals in a ring is an ideal. More precisely, if R is a ring and S_i are ideals of R ($i \in I$), then $S := \bigcap_{i \in I} S_i$ is an ideal of R: we know that S is an additive subgroup of R (Example 9.4(f)) and whenever $r \in R$, $s \in S$, we have $s \in S_i$ for all $i \in I$, hence $rs \in S_i$ and $sr \in S_i$ for all $i \in I$, hence $rs \in S$ and $sr \in S$, and S is therefore an ideal of R.

(h) Let R be a ring and X a subset of R. There are ideals of R which contain X, for example R itself. The intersection of all ideals that contain X is an ideal of R by Example 30.6(g). This ideal is called the *ideal generated by X*. Compare this with Definition 24.1. When X consists of a single element only, say when $X = \{a\}$, the ideal generated by X is said to be a *principal ideal*, more exactly the *principal ideal generated by a*. It is easy to verify that the principal ideal generated by a is

$$\{za + ua + at + \sum_{i=1}^{n} r_i a s_i : z \in \mathbb{Z}, u,t,r_i,s_i \in R, n \in \mathbb{N}\}$$

(cf. Lemma 24.2). If R has an identity, this ideal can be written more simply as

$$\{\sum_{i=1}^{n} r_i a s_i : r_i, s_i \in R, n \in \mathbb{N}\}.$$

If R is commutative, the principal ideal generated by a is

$$\{za + ra : z \in \mathbb{Z}, r \in R\}.$$

If R is a commutative ring with identity, in particular, if R is an integral domain,

$$\{ra : r \in R\} = \{ar : r \in R\}$$

is the principal ideal generated by a. This is usually written as Ra, or as aR, or as (a).

30.7 Theorem: *Let R be a ring and A an ideal of R. On the set R/A of right cosets of A in R, we define two operations $+$ and $\cdot$ by*

$$(r + A) + (s + A) = (r + s) + A, \quad (r + A)\cdot(s + A) = rs + A$$

for all $r,s \in R$. With respect to these operations, R/A is a ring.

Proof: The addition on R/A is well defined since A is a normal additive subgroup of R and the multiplication on R/A is well defined since A is an ideal of R (Theorem 30.5).

R/A is a commutative group under addition (Theorem 18.7, Lemma 18.9(2)). We must now check the associativity of multiplication and the distributivity laws.

For all $r + A, s + A, t + A \in R/A$, we have

$$\begin{aligned}[(r + A)\cdot(s + A)]\cdot(t + A) &= (rs + A)\cdot(t + A)\\ &= (rs)t + A\\ &= r(st) + A\\ &= (r + A)\cdot(st + A)\\ &= (r + A)\cdot[(s + A)\cdot(t + A)],\end{aligned}$$

so multiplication is associative; and we also have

$$\begin{aligned}(r + A)\cdot[(s + A) + (t + A)] &= (r + A)\cdot[(s + t) + A)]\\ &= r(s + t) + A\\ &= (rs + rt) + A\\ &= (rs + A) + (rt + A)\\ &= (r + A)\cdot(s + A) + (r + A)\cdot(t + A)\end{aligned}$$

and

$$\begin{aligned}[(s + A) + (t + A)]\cdot(r + A) &= [(s + t) + A)]\cdot(r + A) \\ &= (s + t)r + A \\ &= (sr + tr) + A \\ &= (sr + A) + (tr + A), \\ &= (s + A)\cdot(r + A) + (t + A)\cdot(r + A).\end{aligned}$$

Hence R/A is a ring. □

30.8 Definition: Let A be an ideal of a ring R. The ring R/A of Theorem 30.7 is called the *factor ring of R with respect to A*, or the *factor ring R by A*, or the *factor ring R mod(ulo) A*. Other names for R/A are: "quotient ring", "difference ring", "residue class ring".

30.9 Examples: (a) In the ring $\mathbb{Z}$ of integers, the multiples $n\mathbb{Z}$ of an integer n form an ideal, the principal ideal generated by n (Example 30.6(b) and (h)). The factor ring $\mathbb{Z}/n\mathbb{Z}$ is exactly the ring $\mathbb{Z}_n$ of integers mod n.

(b) Let T and A be as in Example 30.6(d). Then A is an ideal of T and we can build the factor ring T/A. This factor ring has precisely p elements. What are they?

(c) Let R be a ring and A an ideal of R. If R is commutative, so is R/A, for then $(r + A)\cdot(s + A) = rs + A = sr + A = (s + A)\cdot(r + A)$ for all $(r + A),(s + A)$ in R/A; and if R is a ring with identity, so is R/A, for if 1 is an identity of R, then $1 + A \in R/A$ is an identity of R/A, because

$$(r + A)\cdot(1 + A) = r1 + A = r + A = r1 + A = (1 + A)\cdot(r + A)$$

for all $r + A \in R/A$.

Ideals are the subrings with respect to which we can build factor rings, just as normal subgroups are the subgroups with respect to which we can build factor groups. We know that normal subgroups are exactly the kernels of homomorphisms. We now show that ideals, too, are the kernels of homomorphisms.

30.10 Definition: Let R and R_1 be rings and let $\varphi: R \to R_1$ be a mapping from R into R_1. If

$$(a+b)\varphi = a\varphi + b\varphi \qquad \text{and} \qquad (ab)\varphi = a\varphi \cdot b\varphi$$

for all $a,b \in R$, then φ is called a *(ring) homomorphism.*

The operations on the left hand sides are the operations on R, and those on the right hand side are the operations on R_1. If the operations on R_1 were denoted by $\oplus$ and $\otimes$, the equations would read $(a+b)\varphi = a\varphi \oplus b\varphi$ and $(ab)\varphi = a\varphi \otimes b\varphi$.

If $\varphi: R \to R_1$ is a ring homomorphism and S is a subring of R, then the restriction φ_S of φ to S is also a ring homomorphism.

A ring homomorphism is a homomorphism of additive groups which preserves products as well. This remark enables us to use the properties of group homomorphisms whenever we investigate ring homomorphisms.

30.11 Lemma: *Let $\varphi: R \to R_1$ be a ring homomorphism.*

(1) $0\varphi = 0$.

(2) $(-a)\varphi = -(a\varphi)$ *for all* $a \in R$.

(3) $(a_1 + a_2 + \cdots + a_n)\varphi = a_1\varphi + a_2\varphi + \cdots + a_n\varphi$ *for all* $a_1, a_2, \ldots, a_n \in R$, $n \in \mathbb{N}$, $n \geqslant 2$. *(In particular,* $(na)\varphi = n(a\varphi)$ *for all* $a \in R$*).*

(4) $(a_1 a_2 \ldots a_n)\varphi = a_1\varphi a_2\varphi \ldots a_n\varphi$ *for all* $a_1, a_2, \ldots, a_n \in R$, $n \in \mathbb{N}$, $n \geqslant 2$. *(In particular,* $(a^n)\varphi = (a\varphi)^n$ *for all* $a \in R$*).*

Proof: (1),(2),(3) follow immediately from Lemma 20.3, since φ is a group homomorphism. (4) is proved by the same argument as in the proof of Lemma 20.3(3). □

We now establish the ring theoretical analogues of theorems about group homomorphisms.

30.12 Theorem: *Let* $\varphi: R \to R_1$ *and* $\psi: R_1 \to R_2$ *be a ring homomorphisms. Then the composition mapping*

$$\varphi\psi: R \to R_2$$

is a ring homomorphism from R *into* R_2.

Proof: We regard φ and ψ as group homomorphisms. We know from Theorem 20.4 that $\varphi\psi$ is an additive group homomorphism. It remains to show that $\varphi\psi$ preserves multiplication. Since

$$\begin{aligned}(rs)\varphi\psi &= ((rs)\varphi)\psi \\ &= (r\varphi \cdot s\varphi)\psi \\ &= (r\varphi)\psi \cdot (s\varphi)\psi \\ &= r(\varphi\psi) \cdot s(\varphi\psi)\end{aligned}$$

for all $r,s \in R$, $\varphi\psi$ does preserve multiplication and hence $\varphi\psi$ is a ring homomorphism. □

Since any ring homomorphism $\varphi: R \to R_1$ is a group homomorphism, we can talk about the image and kernel of φ. Of course

$$Im\,\varphi = \{r\varphi \in R_1 : r \in R\} \subseteq R_1 \text{ and } Ker\,\varphi = \{r \in R : r\varphi = 0\} \subseteq R.$$

30.13 Theorem: *Let* $\varphi: R \to R_1$ *be a ring homomorphism. Then* $Im\,\varphi$ *is a subring of* R_1 *and* $Ker\,\varphi$ *is an ideal of* R (cf. Theorem 20.6).

Proof: $Im\,\varphi$ is a subgroup of R_1 by Theorem 20.6. We must show that $Im\,\varphi$ is closed under multiplication (Lemma 30.2). Let $x,y \in Im\,\varphi$. Then $x = r\varphi$, $y = s\varphi$ for some $r,s \in R$. Then $xy = r\varphi \cdot s\varphi = (rs)\varphi$ is the image, under φ, of an element of R, namely of $rs \in R$. So $xy \in Im\,\varphi$ and $Im\,\varphi$ is closed under multiplication. This proves that $Im\,\varphi$ is a subring of R_1.

$Ker\,\varphi$ is a subgroup of R by Theorem 20.6. We must only show that $Ker\,\varphi$ has the "absorbing" property (Definition 30.4). For any $r \in R$ and $a \in Ker\,\varphi$, we have $a\varphi = 0$ and so

$$(ra)\varphi = r\varphi \cdot a\varphi = r\varphi \cdot 0 = 0 \text{ and } (ar)\varphi = a\varphi \cdot r\varphi = 0 \cdot r\varphi = 0$$

by Lemma 29.6(1),(2). Thus $ra \in Ker\,\varphi$ and $ar \in Ker\,\varphi$. Therefore $Ker\,\varphi$ is an ideal of R. □

We prove conversely that every ideal is the kernel of some homomorphism.

30.14 Theorem: *Let R be a ring and let A be an ideal of R. Then*

$$\nu: R \to R/A$$
$$r \to r + A$$

is a ring homomorphism from R onto R/A and $Ker\ \nu = A$ (ν is called the natural or canonical homomorphism).

Proof: The natural mapping $\nu: R \to R/A$ is a group homomorphism from R onto R/A and $Ker\ \nu = A$ by Theorem 20.12. So we need only show that ν is a ring homomorphism, i.e., that ν preserves multiplication. This follows from the very definition of multiplication in R/A: we have

$$(rs)\nu = rs + A = (r + A)(s + A) = r\nu \cdot s\nu$$

for all $r,s \in R$. So ν is a ring homomorphism. □

30.15 Definition: A ring homomorphism $\varphi: R \to R_1$ is called a (*ring*) *isomorphism* if it is one-to-one and onto. In this case, we say R is *isomorphic* to R_1 and write $R \cong R_1$. If R is not isomorphic to R_1, we put $R \ncong R_1$.

So a ring isomorphism is a group isomorphism that preserves multiplication. We use the same sign "$\cong$" for isomorphic rings as for isomorphic groups. This should not lead to any confusion. When confusion is likely, we state explicitly whether we mean ring isomorphism or group isomorphism.

30.16 Lemma: *Let $\varphi: R \to R_1$ and $\psi: R_1 \to R_2$ be ring isomorphisms.*
(1) *$\varphi\psi: R \to R_2$ is a ring isomorphism.*
(2) *$\varphi^{-1}: R_1 \to R$ is a ring isomorphism.*

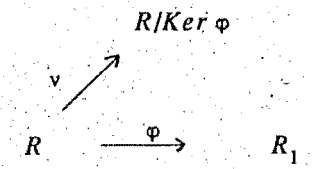
Proof: (1) We know that $\varphi\psi$ is a group isomorphism (Lemma 20.11(1)) and a ring homomorphism (Theorem 30. 12), so $\varphi\psi$ is a ring isomorphism. This proves (1).

(2) We know that φ^{-1} is a group isomorphism (Lemma 20.11(2)). We must also show that φ^{-1} preserves products. For any $x,y \in R_1$, we must show $(xy)\varphi^{-1} = x\varphi^{-1}.y\varphi^{-1}$. Since φ is onto, there are $a,b \in R$ such that $a\varphi = x$ and $b\varphi = y$. Now a and b are unique with this property, for φ is one-to-one, and $a = x\varphi^{-1}$, $b = y\varphi^{-1}$. This is the definition of the inverse mapping. Since φ is a homomorphism, we have

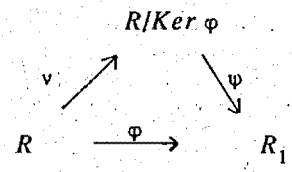
$$\begin{aligned} (ab)\varphi &= a\varphi.b\varphi \\ (\,a\;b\,)\,\varphi &= xy \\ ab &= (xy)\varphi^{-1} \\ x\varphi^{-1}.y\varphi^{-1} &= (xy)\varphi^{-1} \end{aligned}$$

for all $x,y \in R_1$. So $\varphi^{-1}: R_1 \to R$ is a ring homomorphism and consequently φ^{-1} is a ring isomorphism. □

30.17 Theorem (Fundamental theorem on homomorphisms): *Let $\varphi: R \to R_1$ be a ring homomorphism and let $\nu: R \to R/Ker\,\varphi$ be the natural homomorphism.*



(a)



(b)

Then there is a one-to-one ring homomorphism $\psi: R/Ker\,\varphi \to R_1$ such that $\nu\psi = \varphi$.

Proof: From Theorem 20.15 and its proof, we know that the mapping

$$\begin{aligned} \psi: R/Ker\,\varphi &\to R_1 \\ r + Ker\,\varphi &\to r\varphi \end{aligned}$$

is a well defined one-to-one group homomorphism such that $\nu\psi = \varphi$. It only remains to check that ψ preserves multiplication. For all $r,s \in R$, we have

$$\begin{aligned} ((r + Ker\,\varphi)\cdot(s + Ker\,\varphi))\psi &= (rs + Ker\,\varphi)\psi \\ &= (rs)\varphi \\ &= r\varphi \cdot s\varphi \end{aligned}$$

$$= (r + Ker\,\varphi)\psi \cdot (s + Ker\,\varphi)\psi,$$

so ψ preserves products and ψ is a ring homomorphism. □

30.18 Theorem: *Let* $\varphi: R \to R_1$ *be a ring homomorphism. Then*

$$R/Ker\,\varphi \cong Im\,\varphi \qquad (ring\ isomorphism).$$

Proof: The mapping $\psi: R/Ker\,\varphi \to R_1$ is a one-to-one ring homomorphism

$$r + Ker\,\varphi \to r\varphi$$

(Theorem 30.17) and $Im\,\psi = Im\,\varphi$ (see the proof of Theorem 20.16). Thus ψ is a one-to-one ring homomorphism onto $Im\,\varphi$ and therefore

$$R/Ker\,\varphi \cong Im\,\varphi. \qquad \square$$

30.19 Theorem: *Let* $\varphi: R \to R_1$ *be a ring homomorphism from* R *onto* R_1.

(1) *Each subring* S *of* R *with* $Ker\,\varphi \subseteq S$, *is mapped to a subring of* R_1, *which will be denoted by* $S_{\varphi,S}$.

(2) *If* S *and* T *are subrings of* R *and* $Ker\,\varphi \subseteq S \subseteq T$, *then* $S_1 \subseteq T_1$.

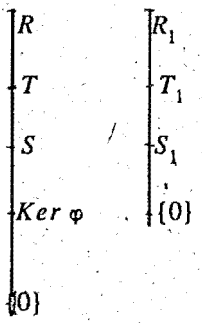
(3) *If* S *and* T *are subrings of* R *containing* $Ker\,\varphi$ *and if* $S_1 \subseteq T_1$, *then* $S \subseteq T$.

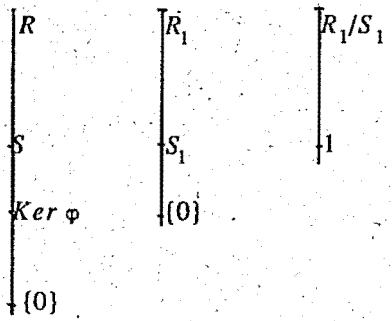
(4) *If* S *and* T *are subrings of* R *containing* $Ker\,\varphi$ *and if* $S_1 = T_1$, *then* $S = T$.

(5) *For any subring* U *of* R_1, *there is a subring* S *of* R *such that* $Ker\,\varphi \subseteq S$ *and* $S_1 = U$.

(6) *Let* S *be a subring of* R *containing* $Ker\,\varphi$. *Then* S *is an ideal of* R *if and only if* S_1 *is an ideal of* R_1.

(7) *If* S *is an ideal of* R *containing* $Ker\,\varphi$, *then* $R/S \cong R_1/S_1$.





Proof: (1) As in Theorem 21.1, we put $S_1 = Im\ \varphi_S$. By Theorem 30.13, S_1 is a subring of R_1. (The restriction of a ring homomorphism to a subring is also a ring homomorphism.)

(2),(3),(4) We regard φ merely as a group homomorphism and apply Theorem 21.1(2),(3),(4).

(5) Let U be a subring of R_1. Consider U as an additive subgroup of R_1. From Theorem 21.1(5), we know that there is a subgroup S of R, namely

$$S = \{r \in R: r\varphi \in U\}$$

with $Ker\ \varphi \subseteq S$ and $S_1 = U$. It remains to show that S is a subring of R. We need only check that S is closed under multiplication, and this is easy: if $r,s \in S$, then $r\varphi, s\varphi \in U$, then $r\varphi \cdot s\varphi \in U$, then $(rs)\varphi \in U$, then $rs \in S$ and S is multiplicatively closed.

(6) Let S be a subring of R, with $Ker\ \varphi \subseteq S$. First we assume that S is an ideal of R and prove that S_1 is an ideal of R_1. We must show that $r_1 s_1 \in S_1$ and $s_1 r_1 \in S_1$ for all $r_1 \in R_1$, $s_1 \in S_1$. Well, if $r_1 \in R_1$, $s_1 \in S = Im\ \varphi_S$, then there are $r \in R$ with $r\varphi = r_1$ and $s \in S$ with $s\varphi = s_1$, and so

$$r_1 s_1 = r\varphi \cdot s\varphi = (rs)\varphi \in Im\ \varphi_S = S_1 \text{ since } rs \in S \text{ as } S \text{ is an ideal of } R,$$

$$s_1 r_1 = s\varphi \cdot r\varphi = (sr)\varphi \in Im\ \varphi_S = S_1 \text{ since } sr \in S \text{ as } S \text{ is an ideal of } R.$$

This proves that S_1 is an ideal of R_1 if S is an ideal of R.

Next we suppose S_1 is an ideal of R_1. By Theorem 30.14, $S_1 = Ker\ \nu'$, where $\nu': R_1 \to R_1/S_1$ is the natural homomorphism. Then $\varphi\nu': R \to R_1/S_1$ is a ring homomorphism (Theorem 30.12) with

$$Ker\ \varphi\nu' = S, \qquad (*)$$

as follows from (ii) on page 225. By Theorem 30.13, S is an ideal of R.

(7) Assume that S is an ideal of R and S_1 is an ideal of R_1. From the ring homomorphism $\varphi\nu': R \to R_1/S_1$, we get

$$R/Ker\ \varphi\nu' \cong Im\ \varphi\nu' \qquad \text{(ring isomorphism)}$$

by Theorem 30.18. Here $Ker\ \varphi\nu' = S$ by (*) and $Im\ \varphi\nu' = R_1/S_1$, for φ and ν' are both onto. Thus

$$R/S \cong R_1/S_1 \qquad \text{(ring homomorphism).} \qquad \square$$

30.20 Theorem: *Let A be an ideal of R. The subrings of R/A are given by S/A, where S runs through the subrings of R containing A. In other words, for each subring U of R/A, there is a unique subring S of R such that $A \subseteq U$ and $U = S/A$. When U_1 and U_2 are subrings of R/A, say with $U_1 = S_1/A$ and $U_2 = S_2/A$, where S_1, S_2 are subrings of R containing A, then $U_1 \subseteq U_2$ if and only if $S_1 \subseteq S_2$. Furthermore, S/A is an ideal of R/A if and only if S is an ideal of R. In this case*

$$R/A \,/\, S/A \cong R/S \qquad \textit{(ring isomorphism)}.$$

Proof: The natural homomorphism $v: R \to R/A$ is onto by Theorem 30.14. Now we may apply Theorem 30.19, which states that any subring of R/A is of the form $S_1 = Im\ v_S = \{sv \in R/A: s \in S\} = \{s + A \in R/A: s \in S\} = S/A$ for some subring S of R containing $Ker\ v = A$ (notice that S/A is meaningful, for A is an ideal of S when $A \subseteq S$ and S is a subring of R). We know that $U_1 = Im\ v_{S_1} \subseteq Im\ v_{S_2} = U_2$ if and only if $S_1 \subseteq S_2$ (Theorem 30.19 (2),(3)). Finally, $S/A = Im\ v_S$ is an ideal of R/A if and only if S is an ideal of R, in which case $R/A \,/\, S/A \cong R/S$ (Theorem 30.19(6),(7)). □

30.21 Theorem: *Let R be a ring, S a subring of R and A an ideal of R.*
(1) *$S + A$ is a subring of R* (here $S + A$ denotes $\{s + a \in R: s \in S, a \in A\} \subseteq R$ in accordance with Definition 19.1).
(2) *A is an ideal of $S + A$, and $S \cap A$ is an ideal of S.*
(3) $S + A \,/\, A \cong S \,/\, S \cap A$ (ring isomorphism).

Proof: (1) $S + A$ is an additive subgroup of R (Lemma 19.4), and it is also closed under multiplication since

$$(s + a)(s' + a') = ss' + sa' + as' + aa' \in S + A$$

for all $s,s' \in S$, $a,a' \in A$, because then $ss' \in S$; and $sa', as', aa' \in A$, consequently $sa' + as' + aa' \in A$. So $S + A$ is a subring of R.

(2) A is an ideal of R and a subset of $S + A$, so, a fortiori, A is an ideal of $S + A$. Also, $S \cap A$ is a subgroup of S and, for all $a \in S \cap A$, $s \in S$,

$$sa \in S \text{ and } sa \in A, \text{ so } sa \in S \cap A,$$
$$as \in S \text{ and } as \in A, \text{ so } as \in S \cap A$$

since S is closed under multiplication and A is an ideal of R. This shows that $S \cap A$ is an ideal of S.

(3) We have a ring homomorphism $v_S: S \to R/A$; the restriction of the natural homomorphism $v: R \to R/A$. Hence $S/Ker\ v_S \cong Im\ v_S$. From the proof of Theorem 21.3, we know $Ker\ v_S = S \cap A$ and $Im\ v_S = S + A\ /A$. So

$$S + A\ /A \ \cong \ S\ /\ S \cap A$$

as contended. □

Exercises

1. Let R be a ring. The *center of* R is defined to be the set

$$Z(R) = \{z \in R: za = az \text{ for all } a \in R\}.$$

Is $Z(R)$ a subring or an ideal of R?

2. Given a ring R, find $Z(Mat_2(R))$.

3. Prove that, if D is a division ring, then $Z(D)$ is a field.

4. Let R be a ring and $b \in R$. Is the *centralizer*

$$C_R(b) := \{r \in R: rb = br\}$$

of b a subring of R?

5. Let R be a ring with identity. Prove or disprove that $Z(R^\times) = (Z(R))^\times$.

6. Show that, if K is a field, then $\{0\}$ and K are the only ideals of K.

7. Let D be a division ring. Find all ideals of $Mat_2(D)$.

8. Let R be a ring and let $End(R)$ be the set of all ring homomorphisms from R into R. For any $\varphi, \psi \in End(R)$, we define $\varphi + \psi: R \to R$ by

$$r(\varphi + \psi) = r\varphi + r\psi.$$

Show that $\varphi + \psi \in End(R)$ and that $(End(R), +, \circ)$ is a ring ($\circ$ is the composition of functions).

9. Let R be a ring and A an ideal of R. Prove that

$$\{r \in R: rx \in A \text{ for all } x \in R\}$$

is an ideal of R.

10. Let $(R,+,\cdot)$ be a ring. If A,B are nonempty subsets of R, we define AB to be the nonempty subset

$$\{a_1b_1 + a_2b_2 + \cdots + a_nb_n \in R: n \in \mathbb{N}, a_i \in A, b_i \in B\}$$

of R. A subgroup A of $(R,+)$ is called a *right* (resp. *left*) *ideal of* R provided $ar \in A$ (resp. $ra \in A$) for all $a \in A, r \in R$. Prove that, if A,B,C are arbitrary right (resp. left) ideals of R, then

(a) $A + B, AB, A \cap B$ are right (resp. left) ideals of R,

(b) $(AB)C = A(BC)$,

(c) $A(B + C) = AB + AC$ and $(B + C)A = BA + CA$.

11. Let R be a ring. An ideal P of R is said to be *prime* if $P \neq R$ and if, for any two ideals A,B of R, the implication

$$AB \subseteq P \implies A \subseteq P \text{ or } B \subseteq P$$

is valid (see Ex. 10). Prove the following statements.

(a) Let P be an ideal of R and $P \neq R$. If, for any $a,b \in R$,

$$ab \in P \implies a \in P \text{ or } b \in P$$

then P is a prime ideal of R.

(b) Let R be commutative. If P is a prime ideal of R, then

$$ab \in P \implies a \in P \text{ or } b \in P$$

for any $a,b \in R$.

(c) $\{0\}$ is a prime ideal of any integral domain.

(d) Let R be a commutative ring with identity and P an ideal of R. Then P is a prime ideal of R if and only if R/P is an integral domain.

12. Let R be a ring. An ideal (resp. right ideal, resp. left ideal) M of R is said to be *maximal ideal* (resp. *right ideal*, resp. *left ideal*)*of* R if $M \neq R$ and if there is no ideal (resp. right ideal, resp. left ideal) N of R such that $M \subset N \subset R$. Prove the following statements.

(a) If R is a commutative ring with identity, then every maximal ideal of R is prime.

(b) If R is a ring with identity, distinct from the null ring, and if M is an ideal of R such that R/M is a division ring, then M is maximal.

(c) If R is a ring with identity and M a maximal ideal of R, then R/M is a field.

(d) Find a noncommutative ring R with identity and a maximal ideal M of R such that R/M is not a division ring.

13. An element a in a ring R is said to be nilpotent if $a^n = 0$ for some $n \in \mathbb{N}$. Prove that, if a,b are nilpotent elements in a ring, and if $ab = ba$, then $a + b$ is also nilpotent.

14. Let R be a commutative ring. Show that the set N of all nilpotent elements in R is an ideal of R and that the factor ring R/N has no nilpotent elements other than 0.

15. Find rings R,S with identities 1_R, 1_S respectively and a ring homomorphism $\varphi: R \to S$ such that $(1_R)\varphi \neq 1_S$.

16. If R,S are rings with identities $1_R,1_S$ respectively, and if $\varphi: R \to S$ is a ring homomorphism onto S, prove that $(1_R)\varphi = 1_S$.

17. The notation being as in §29, Ex. 7, prove that the mapping $r \to (r,0)$ is a one-to-one ring homomorphism from R into S.

§31
Field of Fractions of an Integral Domain

Let D be an integral domain, i.e., a commutative ring with identity which has zo zero divisors, distinct from the null ring. Let a,b,c be elements of D such that

$$a \neq 0, \qquad ab = ac. \tag{i}$$

If a has a multiplicative inverse a^{-1} in D, we could multiply both sises of this equation by a^{-1} and obtain

$$b = c. \tag{ii}$$

But we do not know whether a has an inverse in D and we cannot argue in this way. Nevertheless, it is true that (i) implies (ii) in an integral domain: from (i), we get

$$ab - ac = 0$$
$$a(b - c) = 0,$$

and, since $a \neq 0$ and D has no zero divisors,

$$b - c = 0$$
$$b = c.$$

Hence the cancellation law holds in an integral domain D just as if the nonzero elements in D had inverses in D, i.e., as if $D\backslash\{0\}$ were a group under multiplication.

It is the objective of this paragraph to show that any integral domain is in fact a subring of a ring F such that $F\backslash\{0\}$ is a commutative multiplicative group, i.e., a subring of a *field* F. We can then say that the nonzero elements in D do have inverses, perhaps not in D, but certainly in F.

First we show that finite integral domains are always fields.

31.1 Theorem: *If an integral domain has finitely many elements, then it is a field.*

Proof: (cf. Lemma 9.3) Let D be an integral domain with finitely many elements. We are to show that every nonzero element of D has a multiplicative inverse in D.

Let $a \in D$, $a \neq 0$. Since $|D|$ is finite, the elements

$$a, a^2, a^3, \ldots, a^n, \ldots$$

of D cannot be all distinct. So there are natural numbers $k, l \in \mathbb{N}$ such that $a^k = a^l$, with $k < l$, say. We obtain then

$$\begin{aligned} a^k - a^l &= 0 \\ a^k - a^k a^{l-k} &= 0 \\ a^k(1 - a^{l-k}) &= 0, \end{aligned}$$

where 1 is the identity of D. Since D has no zero divisors and $a \neq 0$, we conclude $a^k = a \cdot a \cdot \ldots \cdot a \neq 0$, which yields

$$\begin{aligned} 1 - a^{l-k} &= 0 \\ a^{l-k} &= 1, \\ a \cdot a^{l-k-1} &= 1. \end{aligned}$$

Thus $a^{l-k-1} \in D$ is an inverse of a. So D is a field. □

Starting from an integral domain D, we now construct, without any hypothesis on $|D|$, a field F which contains D as a subring. This construction is an immediate generalization of the construction of $\mathbb{Q}$ from $\mathbb{Z}$, whose basic moments we recollect: every rational number is a fraction $\frac{a}{b}$ of integers a, b, with $b \neq 0$; different fractions can represent the same rational number, in fact $\frac{a}{b} = \frac{c}{d}$ if and only if $ad = bc$ (where $a, b, c, d \in \mathbb{Z}$ and $b \neq 0 \neq c$); the addition of two rational numbers $\frac{a}{b}, \frac{c}{d}$ is carried out by writing them with a common denominator and adding the numerators $\left(\frac{a}{b} + \frac{c}{d} = \frac{ad}{bd} + \frac{bc}{bd} = \frac{ad + bc}{bd}\right)$; the multiplication is carried out by multiplying the numerators and denominators separately $\left(\frac{a}{b}\frac{c}{d} = \frac{ac}{bd}\right)$; an integer a is considered to be equal to the rational number $\frac{a}{1}$.

All these carry over to the more general case of an arbitrary integral domain D in place of $\mathbb{Z}$, and give rise to a field F which is related to D in the same way as $\mathbb{Q}$ is related to $\mathbb{Z}$. The elements of F will be like "fractions" of elements of D. We introduce them in the next two lemmas.

31.2 Lemma: *Let D be an integral domain and put*

$$S := \{(a,b): a,b \in D, b \neq 0\} = D \times (D\backslash\{0\}).$$

We define a relation $\sim$ on S by declaring

$$(a,b) \sim (c,d) \text{ if and only if } ad = bc$$

for all $(a,b), (c,d) \in S$. Then $\sim$ is an equivalence relation on S.

Proof: (i) For all $(a,b) \in S$, we have $(a,b) \sim (a,b)$ since $ab = ba$. So $\sim$ is reflexive.

(ii) If $(a,b), (c,d) \in S$ and $(a,b) \sim (c,d)$, then

$$\begin{aligned} ad &= bc \\ da &= cb \\ cb &= da \\ (c,d) &\sim (a,b) \end{aligned}$$

and $\sim$ is symmetric.

(iii) If $(a,b), (c,d), (e,f) \in S$ and $(a,b) \sim (c,d), (c,d) \sim (e,f)$, then

$$\begin{aligned} ad = bc &\text{ and } cf = de \\ adf = bcf &\text{ and } bcf = bde \\ daf &= dbe \\ d(af - be) &= 0. \end{aligned}$$

From $(c,d) \in S$, we know $d \neq 0$, and, since D has no zero divisors, we obtain $af - be = 0$. Thus $af = be$, so $(a,b) \sim (e,f)$ and $\sim$ is transitive.

So $\sim$ is an equivalence relation on S. □

31.3 Lemma: *Let D be an integral domain, $S = D \times (D\backslash\{0\})$, and let $\sim$ be the equivalence relation of Lemma 31.2. For $(a,b) \in S$, we designate the equivalence class of (a,b) by $[a:b]$. Thus $[a:b] = \{(c,d) \in S: (c,d) \sim (a,b)\}$.*

Let $F = \{[a:b]: (a,b) \in S\}$ be the set of all equivalence classes of the elements in S. For all $[a:b], [c:d] \in F$, we put

$$\begin{aligned} [a:b] + [c:d] &= [ad + bc : bd], \\ [a:b] \cdot [c:d] &= [ac : bd]. \end{aligned}$$

Then $+$ and $\cdot$ are well defined operations on F.

Proof: First we remark that, if $(a,b), (c,d) \in S$, then $b,d \neq 0$, and so $bd \neq 0$ since D has no zero divisors. Thus $(ad + bc,bd), (ac,bd) \in S$ and therefore $[ad + bc:bd], [ac:bd] \in F$.

We must show that + and · are well defined operations on F. This means we must show that the implication

$$[a:b] = [x:y],\ [c:d] = [z:u] \implies [ad + bc:bd] = [xu + yz:yu],\ [ac:bd] = [xz:yu]$$

is valid. This implication is equivalent to

$$(a,b) \sim (x,y),\ (c,d) \sim (z,u) \implies (ad + bc,bd) \sim (xu + yz,yu),\ (ac,bd) \sim (xz,yu)$$

which, in turn, is equivalent to

$$ay = bx,\ cu = dz \implies (ad + bc)yu = bd(xu + yz),\ ac{\cdot}yu = bd{\cdot}xz,$$

where $b,d,y,u \neq 0$. But certainly, when $ay = bx$, $cu = dz$, we have

$(ad + bc)yu = adyu + bcyu = ay{\cdot}du + by{\cdot}cu = bx{\cdot}du + by{\cdot}cu = bx{\cdot}du + by{\cdot}dz$
$= bd{\cdot}xu + bd{\cdot}yz = bd(xu + yz)$ and $ac{\cdot}yu = ay{\cdot}cu = bx{\cdot}dz = bd{\cdot}xz$. □

31.4 Theorem: *With the notation of Lemma* 31.3, $(F,+,\cdot)$ *is a field.*

Proof: (i) According to Lemma 31.3, + is a binary operation on F.

(ii) + is associative since for any $[a:b], [c:d], [e:f] \in F$, we have

$$\begin{aligned}([a:b] + [c:d]) + [e:f] &= [ad + bc:bd] + [e:f] \\ &= [(ad + bc)f + (bd)e : (bd)f] \\ &= [a(df) + b(cf + de): b(df)] \\ &= [a:b] + [cf + de:df] \\ &= [a:b] + ([c:d] + [e:f]).\end{aligned}$$

(iii) $[0:1]$ is a right additive identity since $[a:b] + [0:1] = [a1 + b0: b1] = [a:b]$ for any $[a:b] \in F$. (Notice that $[0:1] = [0:d]$ for all $d \in D, d \neq 0$.)

(iv) Any $[a:b] \in F$ has a right additive inverse: $[-a:b]$ is the opposite of $[a:b]$, for $[a:b] + [-a:b] = [ab + b(-a): b^2] = [0:b^2] = [0:1]$.

(v) + is commutative since for any $[a:b], [c:d] \in F$, we have

$$[a:b] + [c:d] = [ad + bc:bd] = [cb + da:db] = [c:d] + [a:b].$$

We proved that $(F,+)$ is a commutative group. We now check the remaining ring axioms.

(1) According to Lemma 31.3, $\cdot$ is a binary operation on F.

(2) $\cdot$ is associative since for any $[a:b]$, $[c:d]$, $[e:f] \in F$, we have

$$\begin{aligned}([a:b]\cdot[c:d])\cdot[e:f] &= [ac:bd]\cdot[e:f] \\ &= [(ac)e:(bd)f] \\ &= [a(ce):b(df)] \\ &= [a:b]\cdot[ce:df] \\ &= [a:b]\cdot([c:d]\cdot[e:f]).\end{aligned}$$

(D) For all $[a:b]$, $[c:d]$, $[e:f] \in F$, we have

$$\begin{aligned}[a:b]\cdot([c:d] + [e:f]) &= [a:b]\cdot[cf + de:df] \\ &= [a(cf + de):b(df)] \\ &= [acf + ade:bdf] \\ &= [bacf + bade:bbdf] \qquad \text{(why?)} \\ &= [ac\cdot bf + bd\cdot ae:bd\cdot bf] \\ &= [ac:bd] + [ae:bf] \\ &= [a:b]\cdot[c:d] + [a:b]\cdot[e:f]\end{aligned}$$

and one of the distributivity laws hold in F. We must prove the other distributivity law. We can give an argument similar to the above, but we show presently that $\cdot$ is commutative, and this will give the other distributivity law as a bonus.

We have *not* yet proved that $(F,+,\cdot)$ is a ring.

(3) $\cdot$ is commutative since for any $[a:b]$, $[c:d] \in F$, we have

$$[a:b]\cdot[c:d] = [ac:bd] = [ca:db] = [c:d]\cdot[a:b].$$

As we have already remarked above, this yields the distributivity law we have not checked:

$$\begin{aligned}([c:d] + [e:f])\cdot[a:b] &= [a:b]\cdot([c:d] + [e:f]) \\ &= [a:b]\cdot[c:d] + [a:b]\cdot[e:f] \\ &= [c:d]\cdot[a:b] + [e:f]\cdot[a:b]\end{aligned}$$

for all $[a:b]$, $[c:d]$, $[e:f] \in F$.

We now proved that $(F,+,\cdot)$ is a commutative ring.

(4) $[1:1]$ is the multiplicative identity because

$$[a:b]\cdot[1:1] = [a1:b1] = [a:b]$$

for all $[a:b] \in F$. Since multiplication is commutative, there holds also $[1:1]\cdot[a:b] = [a:b]$ for any $[a:b] \in F$. (Notice that $[1:1] = [d:d]$ for all $d \in D$ with $d \neq 0$.)

Thus $(F,+,\cdot)$ is a commutative ring with identity. It remains to show that every nonzero element in F has a multiplicative inverse in F.

(5) For all $[a:b] \in F\backslash\{0\}$, we show that $[b:a]$ is a multiplicative inverse of $[a:b]$. First of all, since $[a:b] \neq [0:1]$ in F,

$$\begin{aligned} (a,b) \text{ is not equivalent to } (0,1) &\text{ in } S \\ a1 &\neq b0 \\ a &\neq 0 \\ (b,a) &\in S \end{aligned}$$

and $[b:a]$ is an element of F. Secondly, $[a:b]\cdot[b:a] = [ab:ba] = [ab:ab] = [1:1] =$ multiplicative identity of F. Thus $[b:a]$ is a multiplicative inverse of $[a:b] \neq [0:1]$ in F.

This proves that $(F,+,\cdot)$ is a field. □

31.5 Theorem: *Let D be an integral domain and let $(F,+,\cdot)$ be the field of Theorem* 31.4. *Then D is isomorphic to a subring of F.*

Proof: Let $\varphi: D \to F$, $a \to [a:1]$. We demonstrate that φ is a one-to-one ring homomorphism. For all $a,b \in D$,

$$a\varphi + b\varphi = [a:1] + [b:1] = [a1 + 1b:1\cdot 1] = [a + b:1] = (a + b)\varphi$$
$$a\varphi\cdot b\varphi = [a:1]\cdot[b:1] = [ab:1\cdot 1] = [ab:1] = (ab)\varphi,$$

thus φ is a homomorphism. Also

$$\begin{aligned} Ker\,\varphi &= \{a \in D: a\varphi = \text{zero element of } F\} \\ &= \{a \in D: [a:1] = [0:1]\} \\ &= \{a \in D: (a,1) \sim (0,1)\} \\ &= \{a \in D: a\cdot 1 = 1\cdot 0\} \\ &= \{a \in D: a = 0\} \\ &= \{0\} \end{aligned}$$

and hence φ is one-to-one. By Theorem 30.18, D is isomorphic to the subring $Im\,\varphi = D\varphi = \{[a:1] \in F: a \in D\}$ of F. □

31.6 Definition: Let D be an integral domain. Then the field of Theorem 31.4 is called the *field of fractions* or the *field of quotients of* D.

From now on, we shall write $\frac{a}{b}$ for $[a:b]$. The elements of F will be called *fractions* (of elements from D). Furthermore, we identify the integral domain D with its image $D\varphi$ under the mapping in Theorem 31.5. Thus we write a instead of $\frac{a}{1}$ and regard D as a subring of F. Then the inverse of $b \in D \subseteq F$ is $\frac{1}{b} \in F$ and that of $\frac{a}{b}$ is $\frac{b}{a}$ (here $a,b \in D$, $a,b \neq 0$). With these notations, calculations are carried out in the usual way.

Starting from an integral domain D, we constructed the field F of fractions of D. Now this field F is also an integral domain, too, and we may repeat our construction and obtain the field of fractions of F, say F_1. However, nothing is gained by this repetition, for F_1 is not essentially distinct from F.

31.7 Theorem: *Let K be a field and let K_1 be the field of fractions of K. Then K_1 is isomorphic to K.*

Proof: From Theorem 31.6, we know that $\varphi: K \to K_1$ is an isomorphism

$$a \to \frac{a}{1}$$

from K onto $K\varphi = Im\ \varphi$. We will show that $Im\ \varphi = K_1$.

Any element of K_1 can be written as $\frac{a}{b}$, where $a,b \in K$ and $b \neq 0$. Since K is a field and $b \neq 0$, there is an inverse $b^{-1} \in K$ of b in K. Thus $ab^{-1} \in K$ and

$$\frac{a}{b} = [a:b] = [ab^{-1}:1] = \frac{ab^{-1}}{1} = (ab^{-1})\varphi \in Im\ \varphi.$$

This proves $K_1 \subseteq Im\ \varphi$, so $Im\ \varphi = K_1$ and K_1 is isomorphic to K. □

Next we show that the field of fractions of an integral domain is the smallest field containing that integral domain.

31.8 Theorem: *Let D be an integral domain and F the field of fractions of D. If K is any field that contains D, then K contains a subring isomorphic to F.*

Proof: We construct an isomorphism From F onto a subring of K. The elements of F are fractions $\frac{a}{b}$, where $a,b \in D$ and $b \neq 0$. Regarded as an element of K, b has an inverse b^{-1} in K, so $ab^{-1} \in K$. Let $\psi: F \to K$.

$$\frac{a}{b} \to ab^{-1}$$

ψ is a well defined mapping, for if $\frac{a}{b} = \frac{c}{d}$ $(a,b,c,d \in D, b \neq 0 \neq d)$, then $ad = bc$, so $ad \cdot b^{-1}d^{-1} = bc \cdot b^{-1}d^{-1}$, so $ab^{-1} = cd^{-1}$, so $(\frac{a}{b})\psi = (\frac{c}{d})\psi$. Now

$$(\frac{a}{b} + \frac{c}{d})\psi = (\frac{ad+bc}{bd})\psi = (ad + bc)(bd)^{-1}$$

$$= (ad + bc)d^{-1}b^{-1} = ab^{-1} + cd^{-1} = (\frac{a}{b})\psi + (\frac{c}{d})\psi$$

and $(\frac{a}{b} \cdot \frac{c}{d})\psi = (\frac{ac}{bd})\psi = (ac)(bd)^{-1} = ac \cdot d^{-1}b^{-1} = ab^{-1} \cdot cd^{-1} = (\frac{a}{b})\psi \cdot (\frac{c}{d})\psi$

for any two fractions $\frac{a}{b}, \frac{c}{d}$ in F and ψ is a ring homomorphism. Here ψ is one-to-one because $Ker\ \psi = \{0\}$, for $\frac{a}{b} \in Ker\ \psi$ implies $(\frac{a}{b})\psi = 0$, so $ab^{-1} = 0$, so $a = abb^{-1} = 0b^{-1} = 0$, so $\frac{a}{b} = \frac{0}{b} = \frac{0}{1} = 0$. Hence F is isomorphic to the subring $Im\ \psi$ of K (Theorem 30.18). □

Exercises

1. Let $D_1 = \{a + bi \in \mathbb{C}: a,b \in \mathbb{Z}\}$, $D_2 = \{a + 2bi \in \mathbb{C}: a,b \in \mathbb{Z}\}$ and

$E = \{a + b\sqrt[3]{2} + c\sqrt[3]{4} \in \mathbb{R} : a,b,c \in \mathbb{Z}\}$. Show that D_1, D_2, E are integral domains and describe, as simply as you can, the elements in the field of fractions of these integral domains.

2. Let R be a commutative ring and let M be a nonempty multiplicatively closed subset of R. For ordered pairs in $R \times M$, we put

$(r,m) \sim (r',m')$ if and only if there is an $n \in M$ such that $n(rm' - r'm)$

Show that $\sim$ is an equivalence relation on $R \times M$. Denote the equivalence class of (r,m) by $\frac{r}{m}$ and define, on the set $M^{-1}R$ of all equivalence classes, addition and multiplication by

$$\frac{r}{m} + \frac{r'}{m'} = \frac{rm' + mr'}{mm'} \quad \text{and} \quad \frac{r}{m} \cdot \frac{r'}{m'} = \frac{rr'}{mm'}$$

Prove that $M^{-1}R$ is a commutative ring with identity under these operations.

3. Keep the notation of Ex. 2. Prove that, if A is an ideal of R, then $M^{-1}A = \{\frac{a}{m} \in M^{-1}R : a \in A, m \in M\}$ is an ideal of $M^{-1}R$. If A,B are ideals of R, then $M^{-1}(A + B) = M^{-1}A + M^{-1}B$ and $M^{-1}(A \cap B) = M^{-1}A \cap M^{-1}B$. Does every ideal of $M^{-1}R$ have the form $M^{-1}A$ for some ideal A of R?

4. Let R be a commutative ring with identity and let P be a prime ideal of M (see §30, Ex. 11). Show that $M := R \backslash P$ is a multiplicatively closed subset of R. Prove that $M^{-1}R$, in the notation of Ex. 2, has a unique maximal ideal (see §30, Ex. 12).

5. Discuss the rings in Example 29.2(e),(g) under the light of Ex. 3 and 4.

§32
Divisibility Theory in Integral Domains

As we have already mentioned, the ring $\mathbb{Z}$ of integers is the prototype of integral domains. There is a divisibility relation on $\mathbb{Z}^*$: an integer b is said to be divisible by a nonzero integer a when there is an integer c such that $ac = b$. Integers with no nontrivial divisors are called prime, and every nonzero integer that is not a unit can be written as a product of prime numbers in a unique way.

We want to investigate whether there are similar results in other integral domains. More generally, one can ask whether there are similar results in an arbitrary ring. However, in an arbitrary ring, one has to distinguish between left divisors and right divisors: if a,b,c are elements of a ring and $ab = c$, then a is called a left divisor, and b is called a right divisor of c. Here a may be a left divisor of c without being a right divisor of c, and vice versa. Furthermore, the existence of zero divisors in a ring complicates the theory. For these reasons, in this introductory book, we confine ourselves to integral domains.

32.1 Definition: Let D be an integral domain and let $\alpha,\beta \in D$. If $\alpha \neq 0$ and if there is a $\gamma \in D$ such that $\alpha\gamma = \beta$, then α is called a *divisor* or a *factor of* β and β is said to be *divisible by* α. We also say α *divides* β.

We write $\alpha|\beta$ when α divides β, and $\alpha\nmid\beta$ when $\alpha \neq 0$ and α does not divide β.

32.2 Lemma: *Let D be an integral domain and let $\alpha,\beta,\gamma,\mu,\nu,\mu_1,\mu_2,\ldots,\mu_s$, $\beta_1,\beta_2,\ldots,\beta_s$ be elements of D.*

(1) *If $\alpha|\beta$, then $\alpha|-\beta$, $-\alpha|-\beta$, $-\alpha|\beta$.*

(2) *If $\alpha|\beta$ and $\beta|\gamma$, then $\alpha|\gamma$.*

(3) *If $\alpha|\beta$ and $\gamma \neq 0$, then $\alpha\gamma|\beta\gamma$.*

---

* More precisely, on $\mathbb{Z}\setminus\{0\}$.

(4) *If* $\alpha\gamma|\beta\gamma$, *then* $\alpha|\beta$.
(5) *If* $\alpha|\beta$ *and* $\alpha|\gamma$, *then* $\alpha|\beta+\gamma$.
(6) *If* $\alpha|\beta$ *and* $\alpha|\gamma$, *then* $\alpha|\beta-\gamma$.
(7) *If* $\alpha|\beta$*and* $\alpha|\gamma$, *then* $\alpha|\mu\beta+\nu\gamma$.
(8) *If* $\alpha|\beta_1, \alpha|\beta_2, \ldots, \alpha|\beta_s$, *then* $\alpha|\mu_1\beta_1+\mu_2\beta_2+\cdots+\mu_s\beta_s$.
(9) *If* $\alpha \neq 0$, *then* $\alpha|0$.
(10) $1|\alpha$ *and* $-1|\alpha$.

Proof: The claims are proved exactly as in the proof of Lemma 5.2. □

We know that the units 1, −1 of $\mathbb{Z}$ divide every integer. This is true in any arbitrary integral domain (see Definition 29.14).

32.3 Lemma: *Let D be an integral domain and $\varepsilon \in D$. Then ε is a unit of D (i.e., $\varepsilon \in D^\times$) if and only if $\varepsilon|\alpha$ for all $\alpha \in D$.*

Proof: If ε is a unit, then $\varepsilon\beta = 1$ for some $\beta \in D$; in particular, since D is an integral domain, $\varepsilon\beta = 1 \neq 0$ and $\varepsilon \neq 0$. For any $\alpha \in D$, we have $\varepsilon(\beta\alpha) = (\varepsilon\beta)\alpha = 1\alpha = \alpha$, with $\beta\alpha \in D$. Thus $\varepsilon|\alpha$ for any $\alpha \in D$. Conversely, if $\varepsilon|\alpha$ for all α in D, then $\varepsilon|1$, so $\varepsilon\beta = 1$ for some $\beta \in D$. Thus ε has an inverse in D and ε is a unit of D. □

32.4 Definition: Let D be an integral domain and $\alpha, \beta \in D$. Then α is said to be *associate to* β if there is a unit $\varepsilon \in D^\times$ such that $\alpha = \beta\varepsilon$. In this case, we write $\alpha \approx \beta$.

32.5 Lemma: *Let D be an integral domain. Then $\approx$ is an equivalence relation on D.*

Proof: (i) For any $\alpha \in D$, we have $\alpha = \alpha 1$ and 1 is a unit. Hence $\alpha \approx \alpha$ and $\approx$ is reflexive. (ii) If $\alpha, \beta \in D$ and $\alpha \approx \beta$, then $\alpha = \beta\varepsilon$ for some $\varepsilon \in D^\times$, then $\beta = \alpha\varepsilon^{-1}$ with $\varepsilon^{-1} \in D^\times$ (for $D^\times$ is a group by Theorem 29.15) and $\beta \approx \alpha$, so $\approx$ is symmetric. (iii) If $\alpha, \beta, \gamma \in D$ and $\alpha \approx \beta$, $\beta \approx \gamma$, then $\alpha = \beta\varepsilon$, $\beta = \gamma\varepsilon'$, where $\varepsilon, \varepsilon'$

are units in D. So $\alpha = \gamma\varepsilon'\varepsilon$, with $\varepsilon'\varepsilon \in D^\times$ (Theorem 29.15), thus $\alpha \approx \gamma$ and $\approx$ is transitive. □

Since $\approx$ is a symmetic relation, it is legitimate to say that α and β are associate when α is associate to β. The alert reader will have noticed that the group $D^\times$ acts on the set D in the sense of Definition 25.1, and the orbit of any $\alpha \in D$ consists of the associates of α (that is, elements of D which are associate to α). Lemma 32.5 is thus merely a special case of Lemma 25.5.

For any $\alpha \in D$, the units and associates of α are divisors of α. A divisor of α, which is neither a unit nor an associate of α, is called a proper divisor of α. An element need not have proper divisors; for instance, a unit has no proper divisors.

The relation $\alpha|\beta$ holds if and only if the relation $\alpha_1|\beta_1$ holds for any associate α_1 of α and for any associate β_1 of β. In other words, as far as divisibility is concerned, associate elements play the same role.

32.6 Examples: (a) The theory of divisibility in $\mathbb{Z}$ was discussed in §5. The units in $\mathbb{Z}$ are 1 and -1, and the associates of $a \in \mathbb{Z}$ are a and $-a$. The terminology in this paragraph is consistent with that of §5.

(b) Let D be a field. Then $\alpha|\beta$ for any $\alpha,\beta \in D, \alpha \neq 0$ since $\alpha \neq 0$ implies that there is an inverse α^{-1} of α in D and $\alpha(\alpha^{-1}\beta) = \beta$ with $\alpha^{-1}\beta \in D$. In particular, $\alpha|1$ for any $\alpha \in D, \alpha \neq 0$. Hence any nonzero element in D is a unit and any two nonzero elements are associate. The divisibility theory is not very interesting in a field.

(c) Let $R = \{a/b \in \mathbb{Q}: (a,b) = 1, 5 \nmid b\}$ be the ring of Example 29.2(e). It is easily seen that R is an integral domain. Let us find the units of R. The multiplicative inverse of $a/b \in R \subseteq \mathbb{Q}$ $((a,b) = 1)$ is $b/a \in \mathbb{Q}$, and $b/a \in R$ if and only if $5 \nmid a$. Thus $R^\times = \{a/b \in R: (a,b) = 1, 5 \nmid b, 5 \nmid a\}$. The associates of $a/b \in R$ are the numbers x/y with $(x,y) = 1$, where a and x are exactly divisible by the same power of of 5.

(d) We put $\mathbb{Z}[i] := \{a + bi \in \mathbb{C} : a,b \in \mathbb{Z}\}$. One easily checks that $\mathbb{Z}[i]$ is a subring of $\mathbb{C}$ and that $\mathbb{Z}[i]$ is an integral domain. The elements of $\mathbb{Z}[i]$ are called *gaussian integers* (after C. F. Gauss (1777-1855) who introduced them in his investigations about the so-called biquadratic reciprocity law).

Since $\mathbb{Z}[i]$ is a subring of $\mathbb{C}$, each element $\alpha = a + bi$ in $\mathbb{Z}[i]$ has a conjugate and a norm. The conjugate of $\alpha = a + bi \in \mathbb{Z}[i]$ is defined to be $\bar{\alpha} = a - bi$ in $\mathbb{Z}[i]$ $(a,b \in \mathbb{Z})$. Notice that $\overline{\alpha\beta} = \bar{\alpha}\,\bar{\beta}$ for any $\alpha,\beta \in \mathbb{Z}[i]$. The norm $N(\alpha)$ of $a + bi \in \mathbb{Z}[i]$ is defined by $N(\alpha) = \alpha\bar{\alpha}$; hence $N(a + bi) = a^2 + b^2$ $(a,b \in \mathbb{Z})$. Thus $N(\alpha)$ is a nonnegative integer for any $\alpha \in \mathbb{Z}[i]$, and equals 0 if and only if $\alpha = 0 + 0i = 0$. Moreover, $N(\beta\gamma) = \beta\gamma \cdot \overline{\beta\gamma} = \beta\gamma \cdot \bar{\beta}\,\bar{\gamma} = \beta\bar{\beta}\,\gamma\bar{\gamma} = N(\beta)N(\gamma)$ for any $\beta,\gamma \in \mathbb{Z}[i]$.

Using this, it is easy to determine the units in $\mathbb{Z}[i]$. We claim $\varepsilon \in \mathbb{Z}[i]$ is a unit in $\mathbb{Z}[i]$ if and only if $N(\varepsilon) = 1$. Indeed, if ε is a unit in $\mathbb{Z}[i]$, then $\varepsilon\varepsilon^{-1} = 1$, then $N(\varepsilon)N(\varepsilon^{-1}) = 1$, where $N(\varepsilon),N(\varepsilon^{-1})$ are positive integers. This forces $N(\varepsilon) = 1$, as claimed. Conversely, if $N(\varepsilon) = 1$, then $\varepsilon\bar{\varepsilon} = 1$, where $\bar{\varepsilon} \in \mathbb{Z}[i]$, and this yields $\varepsilon | 1$, which means ε is a unit.

Thus $\varepsilon \in a + bi$ is a unit if and only if $N(\varepsilon) = a^2 + b^2 = 1$ (here $a,b \in \mathbb{Z}$) and $a^2 + b^2 = 1$ if and only if $a^2 = 1$, $b^2 = 0$ or $a^2 = 0$, $b^2 = 1$. Therefore ε is a unit if and only if $\varepsilon = 1,-1,i,-i$, so that $\mathbb{Z}[i]^\times = \{1,-1,i,-i\}$. The associates of α in $\mathbb{Z}[i]$ are the numbers $\alpha, -\alpha, \alpha i, -\alpha i$.

(e) We put $\omega = \dfrac{-1 + \sqrt{3}i}{2} \in \mathbb{C}$. Thus $\omega = \cos\dfrac{2\pi}{3} + i\sin\dfrac{2\pi}{3}$. By de Moivre's theorem, $\omega^2 = \cos 2\dfrac{2\pi}{3} + i\sin 2\dfrac{2\pi}{3} = e^{\frac{4\pi}{3}} = \dfrac{-1 - \sqrt{3}i}{2} = \bar{\omega}$ and

$\omega^3 = \cos 3\dfrac{2\pi}{3} + i\sin 3\dfrac{2\pi}{3} = 1$. So $\omega^3 - 1 = 0$, so $(\omega - 1)(\omega^2 + \omega + 1) = 0$. Since $\omega - 1 \neq 0$, we conclude $\omega^2 + \omega + 1 = 0$, which can also be verified directly. From $\omega^3 = 1$, we obtain $\omega^4 = \omega$, whence $(\omega^2)^2 + \omega^2 + 1 = 0$.

We put $\mathbb{Z}[\omega] := \{a + b\omega \in \mathbb{C} : a,b \in \mathbb{Z}\}$. One easily checks that $\mathbb{Z}[\omega]$ is a subring of $\mathbb{C}$ and that $\mathbb{Z}[\omega]$ is an integral domain. The closure of $\mathbb{Z}[\omega]$ under multiplication follows from $\omega^2 = -1 - \omega$:

$$\begin{aligned}(a + b\omega)(c + d\omega) &= ac + ad\omega + bc\omega + bd\omega^2 \\ &= ac + ad\omega + bc\omega + bd(-1 - \omega) \\ &= (ac - bd) + (ad + bc - bd)\omega\end{aligned}$$

$\in \mathbb{Z}[\omega]$

for all $a + b\omega, c + d\omega \in \mathbb{Z}[\omega]$. The ring $\mathbb{Z}[\omega]$ was introduced independently by C. G. J. Jacobi (1804-1851) and by G. Eisenstein (1823-1852) in their investigations about the so-called cubic reciprocity law.

Repeating the proof for $\mathbb{Z}[i]$, we see that $a + b\omega \in \mathbb{Z}[\omega]$ is a unit in $\mathbb{Z}[\omega]$ if and only if $N(a + b\omega) = 1$. This is equivalent to $a^2 - ab + b^2 = 1$, so equivalent to $4a^2 - 4ab + 4b^2 = 4$, so to $(2a - b)^2 + 3b^2 = 4$. The last equation holds if and only if $2a - b = \mp 2$, $b = 0$ or $2a - b = \mp 1$, $b = \mp 1$ $(a,b \in \mathbb{Z})$. In this way, we get $a + b\omega = \mp 1, \mp\omega, \mp(-1 - \omega) = \mp\omega^2$. The units in $\mathbb{Z}[\omega]$ are $\mp 1, \mp\omega, \mp(-1 - \omega) = \mp\omega^2$; the associates of $\alpha \in \mathbb{Z}[\omega]$ are the numbers $\mp\alpha, \mp\omega\alpha$, $\mp(-1 - \omega)\alpha = \mp\omega^2\alpha$.

(f) We put $\mathbb{Z}[\sqrt{5}i] = \{a + b\sqrt{5}i \in \mathbb{C} : a,b \in \mathbb{Z}\}$. Again, it is easily verified that $\mathbb{Z}[\sqrt{5}i]$ is an integral domain, and $\alpha \in \mathbb{Z}[\sqrt{5}i]$ is a unit if and only if $N(\alpha) = 1$. Now $N(a + b\sqrt{5}i) = a^2 + 5b^2 = 1$ if and only if $a = \mp 1$, $b = 0$ (here $a,b \in \mathbb{Z}$). Thus ∓ 1 are the only units in $\mathbb{Z}[\sqrt{5}i]$ and the associates of a number $\alpha \in \mathbb{Z}[\sqrt{5}i]$ are the numbers $\mp\alpha$.

So far, the divisibility theory in an arbitrary integral domain has been completely analogous to the theory in $\mathbb{Z}$, which culminates in the fundamental theorem of arithmetic asserting that every integer, not a zero or a unit, can be written as a product of prime numbers in a unique way. We proceed to investigate if a similar theorem is true in an arbitrary integral domain. First we introduce the counterparts of prime numbers.

32.7 Definition: Let D be an integral domain and $\alpha \in D$. Then α is said to be *irreducible* if α is neither zero nor a unit, and if, in any factorization $\alpha = \beta\gamma$ of α, where $\beta,\gamma \in D$, either β or γ is a unit in D. When α is neither zero nor a unit, and when α is not irreducible in D, α is said to be *reducible*.

An irreducible element in D is therefore one which has no proper divisors. Clearly, when α and β are associates, α is irreducible if and only if β is irreducible. One might expect that such elements be called prime

rather than irreducible, but the term "prime" is reserved for another property (Definition 32.20).

We now ask if every nonzero, nonunit element in an integral domain D can be expressed as a product of finitely many irreducible elements (cf. Theorem 5.13) Let us try to argue as in Theorem 5.13. Given $\alpha \in D$ ($\alpha \neq 0$, not a unit), α is either irreducible or not. In the former case, α is a product of *one* irreducible element. In the latter case, $\alpha = \alpha_1\beta$ for some suitable proper divisors of α. Here α_1 is either irreducible or not. In the former case, α_1 is an irreducible divisor of α. In the latter case, $\alpha_1 = \alpha_2\beta$ for some suitable proper divisors of α_1. Here α_2 is either irreducible or not. Repeating this procedure, we get a sequence

$$\alpha = \alpha_0, \alpha_1, \alpha_2, \ldots \qquad (s)$$

of elements in D, where α_{i+1} is a proper divisor of α_i ($i = 0,1,2, \ldots$).

When the sequence (s) stops after a finite number of steps, we obtain an irreducible divisor of α. However, we do not know that the sequence (s) ever terminates. In the case of $\mathbb{Z}$, the *absolute values* of α_i, which are nonnegative integers, get smaller and smaller and, since there are finitely many nonnegative integers less than $|\alpha|$, the sequence (s) does come to an end. But this argument cannot be extended to the general case, for there is no absolute value concept. Let us suppose, however, that there is associated a nonnegative integer $d(\alpha_i)$ to each α_i in such a way that $d(\alpha_{i+1}) < d(\alpha_i)$. If this is possible, we can conclude that sequence (s) does terminate.

For example, when D is one of $\mathbb{Z}[i]$, $\mathbb{Z}[\omega]$, $\mathbb{Z}[\sqrt{5}i]$, we may consider the norm $N(\alpha_i)$ of α_i. The norm $N(\alpha_i)$ is a nonnegative integer, and also $N(\alpha_{i+1}) < N(\alpha_i)$ whenever α_{i+1} is a proper divisor of α_i. In fact, with the norm function, there is a division algorithm in $\mathbb{Z}[i]$ and in $\mathbb{Z}[\omega]$.

32.8 Theorem: *Let α, β be elements of $\mathbb{Z}[i]$ (resp. of $\mathbb{Z}[\omega]$), with $\beta \neq 0$. Then there are two elements κ and ρ in $\mathbb{Z}[i]$ (resp. in $\mathbb{Z}[\omega]$) such that*

$$\alpha = \kappa\beta + \rho \quad \text{and} \quad N(\rho) < N(\beta).$$

Proof: (Cf. Theorem 5.3; note that κ and ρ are not claimed to be unique.) The elements α, β of $\mathbb{Z}[i]$ (resp. of $\mathbb{Z}[\omega]$) are complex numbers, and $\beta \neq 0$. Thus $\alpha/\beta \in \mathbb{C}$. Let us write

$$\frac{\alpha}{\beta} = x + yi \quad \text{(resp. } \frac{\alpha}{\beta} = x + y\omega\text{)}$$

with $x,y \in \mathbb{R}$. We want $\frac{\alpha}{\beta}$ to be "approximately equal" to κ, with an "error" $\frac{\rho}{\beta}$ so small that $N(\frac{\rho}{\beta}) < 1$. So we approximate $x + yi$ (resp. $x + y\omega$) by an element κ in $\mathbb{Z}[i]$ (resp. in $\mathbb{Z}[\omega]$) as closely as we can. To this end, we choose integers $a,b \in \mathbb{Z}$ such that

$$|x - a| \leqslant \frac{1}{2} \qquad |y - b| \leqslant \frac{1}{2}$$

and put $\kappa = a + bi$ (resp. $\kappa = a + b\omega$). This is possible since the distance between x and the integer closest to x is less than or equal to $\frac{1}{2}$. When x is half an odd integer, there are two choices for a, and therefore there can be no hope for uniqueness. In this case, we have in fact $|x - a| = \frac{1}{2}$ and the approximation above is the best possible one. The same remarks apply to y and b.

We now put $\rho = \alpha - \kappa\beta$. Then $\alpha = \kappa\beta + \rho$. It remains to show that $N(\frac{\rho}{\beta}) < 1$. We have indeed

$$N(\frac{\rho}{\beta}) = N(\frac{\alpha - \kappa\beta}{\beta}) = N(\frac{\alpha}{\beta} - \kappa) = \begin{cases} N((x + yi) - (a + bi)) \\ N((x + y\omega) - (a + b\omega)) \end{cases}$$

$$= \begin{cases} N((x - a) + (y - b)i) \\ N((x - a) + (y - b)\omega) \end{cases}$$

$$= \begin{cases} (x - a)^2 + (y - b)^2 \\ (x - a)^2 - (x - a)(y - b) + (y - b)^2 \end{cases}$$

$$= \begin{cases} |x - a|^2 + |y - b|^2 \\ |x - a|^2 + |x - a||y - b| + |y - b|^2 \end{cases}$$

$$\leqslant \begin{cases} (1/2)^2 + (1/2)^2 \\ (1/2)^2 + (1/2)(1/2) + (1/2)^2 \end{cases}$$

$$= \begin{cases} 2/4 \\ 3/4 \end{cases} < 1.$$

This completes the proof. □

What happens in $\mathbb{Z}[\sqrt{5}i]$? For $a,b \in \mathbb{Z}[\sqrt{5}i]$, $\beta \neq 0$, we write $\frac{\alpha}{\beta} = x + y\sqrt{5}i$, with $x,y \in \mathbb{R}$. The best approximation to $\frac{\alpha}{\beta}$ is $\kappa = a + b\sqrt{5}i$, where a,b

are integers such that $|x - a| \leqslant \frac{1}{2}$, $|y - b| \leqslant \frac{1}{2}$. Putting $\rho = \alpha - \kappa\beta$, we can conclude only

$$N(\tfrac{\rho}{\beta}) = N(\tfrac{\alpha-\kappa\beta}{\beta}) = N(\tfrac{\alpha}{\beta}-\kappa) = N((x - a) + (y - b)\sqrt{5}i)$$
$$= (x - a)^2 + 5(y - b)^2 \leqslant (1/2)^2 + 5(1/2)^2 = 3/2 \qquad (\dagger)$$

instead of $N(\frac{\rho}{\beta}) < 1$ as in $\mathbb{Z}[i]$, $\mathbb{Z}[\omega]$.

32.9 Theorem: *In* $\mathbb{Z}[\sqrt{5}i]$, *there are elements* α_0, β_0 *with* $\beta_0 \neq 0$ *such that*

$$N(\alpha_0 - \kappa\beta_0) > N(\beta_0) \quad \textit{for all } \kappa \in \mathbb{Z}[\sqrt{5}i].$$

Proof: We choose α_0, β_0 in such a way that equality holds in (†) above. This will be the case when x and y are half odd integers. So we set $\alpha_0 = 1 + \sqrt{5}i$, $\beta_0 = 2$. Then, for any $\kappa = a + b\sqrt{5}i \in \mathbb{Z}[\sqrt{5}i]$ (with $a,b \in \mathbb{Z}$), we have

$$N(\alpha_0 - \kappa\beta_0) = N(\beta_0)N(\tfrac{\alpha_0-\kappa\beta_0}{\beta_0}) = N(\beta_0)N(\tfrac{\alpha_0}{\beta_0} - \kappa) = N(\beta_0)N\big(\tfrac{1+\sqrt{5}i}{2} - (a + b\sqrt{5}i)\big)$$
$$= N(\beta_0)\Big[(\tfrac{1}{2} - a)^2 + 5(\tfrac{1}{2} - b)^2\Big] \geqslant N(\beta_0)[(1/2)^2 + 5(1/2)^2] = N(\beta_0)\tfrac{3}{2} > N(\beta_0),$$

as claimed. □

The integral domains on which there is a division algorithm are called Euclidean domains. The formal definition is as follows.

32.10 Definition: Let D be an integral domain. D is called a *Euclidean domain* if there is a function $d: D\backslash\{0\} \to \mathbb{N} \cup \{0\} \subseteq \mathbb{Z}$ such that

(i) $d(\alpha) \leqslant d(\alpha\beta)$ for all $\alpha,\beta \in D\backslash\{0\}$,

(ii) for any $\alpha,\beta \in D\backslash\{0\}$, there are $\kappa,\rho \in D$ satisfying

$$\alpha = \kappa\beta + \rho \quad \text{and} \quad \rho = 0 \text{ or } d(\rho) < d(\beta).$$

The first condition (i) assures that the d-value of a divisor of $\gamma \in D\backslash\{0\}$ is less than or equal to the d-value of γ. It follows that $d(\rho) = d(\rho')$ whenever ρ and ρ' are associate. Using (ii) repeatedly, the analog of the Euclidean algorithm is seen to be valid, and the last nonzero remainder

is a greatest common divisor. It will be a good exercise for the reader to prove this result.

$\mathbb{Z}$ is a Euclidean domain, with the absolute value function working as the function d of Definition 32.10. This follows from Theorem 5.3, with b replaced by $|b|$. Also, $\mathbb{Z}[i]$ and $\mathbb{Z}[\omega]$ are Euclidean domains, with the norm function working as the function d of Definition 32.10, as Theorem 32.8 shows. On the other hand, we do not yet know whether $\mathbb{Z}[\sqrt{5}i]$ is a Euclidean domain. It does *not* follow from Theorem 32.9 that $\mathbb{Z}[\sqrt{5}i]$ is not Euclidean. From Theorem 32.9, it follows only that either $\mathbb{Z}[\sqrt{5}i]$ is not Euclidean, or $\mathbb{Z}[\sqrt{5}i]$ is a Euclidean domain with a function d that is necessarily distinct from the norm function.

In a Euclidean domain D, the sequence (s) terminates after a finite number of steps. We shall prove a more general statement (Theorem 32.14). Recall that $\{\xi\alpha \in D: \xi \in D\} = D\alpha$, where $\alpha \in D$, is the principal ideal generated by α (Example 30.6(h)).

32.11 Theorem: *If D is a Euclidean domain, then every ideal of D is a principal ideal.*

Proof: Let D be a Euclidean domain, and let d be the function of Definition 32.10. For any ideal A of D, we must find an α such that $A = D\alpha$. We argue as in Theorem 5.4.

When $A = \{0\}$, we clearly have $A = D0$, and the claim is true. Assume now $A \neq \{0\}$. Then $U = \{d(\alpha) \in \mathbb{N} \cup \{0\}: \alpha \in A, \alpha \neq 0\}$ is a nonempty subset of the set of nonnegative integers. Let m be the smallest integer in U. Then $m = d(\alpha)$ for some $\alpha \in A$, $\alpha \neq 0$; and $d(\alpha) \leqslant d(\beta)$ for all $\beta \in A$, $\beta \neq 0$.

We show that $A = D\alpha$. First we have $D\alpha \subseteq A$, because $\alpha \in A$ and A has the "absorbing" property. To prove $A \subseteq D\alpha$, take an arbitrary γ from A. There are $\kappa, \rho \in D$ such that

$$\gamma = \kappa\alpha + \rho, \quad \rho = 0 \text{ or } d(\rho) < d(\alpha),$$

provided $\gamma \neq 0$ (Definition 32.10). Now $\alpha \in A$, so $\kappa\alpha \in A$, and since $\gamma \in A$ as well, we see that $\rho \in A$. Here $d(\rho) < d(\alpha)$ is impossible, for then $d(\rho)$ in U would be less than m, which is the smallest number in U. Hence

necessarily $\rho = 0$ and $\gamma = \kappa\alpha \in D\alpha$. This shows $\gamma \in D\alpha$ for all $\gamma \in A$, provided $\gamma \neq 0$. Since $0 = 0\alpha \in D\alpha$ as well, we get $A \subseteq D\alpha$. Thus $A = D\alpha$. □

32.12 Definition: An integral domain D is called a *principal ideal domain* if every ideal of D is a principal ideal.

With this terminology, Theorem 32.11 can be reformulated as follows.

32.11 Theorem: *Every Euclidean domain is a principal ideal domain.*

In any integral domain, $\alpha \mid \beta$ if and only if $D\beta \subseteq D\alpha$, and $\alpha \approx \beta$ if and only if $D\alpha = D\beta$. Thus the sequence (s) gives rise to the chain

$$D\alpha = D\alpha_0 \subseteq D\alpha_1 \subseteq D\alpha_2 \subseteq \ldots$$

of principal ideals in D. The sequence (s) breaks down if and only if this chain of ideals breaks down. For principal ideal domains, this is always true.

32.13 Definition: Let D be an integral domain. D is said to *satisfy the ascending chain condition* (ACC) if, for every chain

$$A_0 \subseteq A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$$

of ideals in D, there is an index k such that $A_m = A_k$ for all $m \geqslant k$; or, what is the same, every chain

$$B_0 \subset B_1 \subset B_2 \subset B_3 \subset \ldots$$

of ideals in D consists of finitely many terms. An integral domain satisfying the ascending chain condition is also called a *noetherian* domain (in honor of Emmy Noether (1882-1935)).

32.14 Theorem: *Every principal ideal domain satisfies the ascending chain condition (is noetherian).*

Proof: Let D be a principal ideal domain and let

$$A_0 \subseteq A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$$

be a chain of ideals of D. We must show there is an integer k such that $A_m = A_k$ for all $m \geqslant k$. To this end, we put $B := \bigcup_{i=1}^{\infty} A_i$. We claim B is an ideal of D. Indeed, if $\alpha, \beta \in B$, then $\alpha \in A_j$, $\beta \in A_l$ for some indices j,l. Assuming $j \leqslant l$ without loss of generality, we have $A_j \subseteq A_l$. Since A_l is an ideal of D, we have $\alpha + \beta \in A_l$, so $\alpha + \beta \in B$. Also $-\alpha \in A_l$, so $\alpha \in B$. This shows that B is a subgroup of D under addition. Finally, if δ is an arbitrary element of D, then $\delta\alpha \in A_l$, since A_l is an ideal, so $\delta\alpha \in B$. Hence B is an ideal of D.

Since D is a principal ideal domain, $B = D\kappa$ for some $\kappa \in D$. As $\kappa = 1\kappa \in D\kappa = B = \bigcup_{i=1}^{\infty} A_i$, we see $\kappa \in A_k$ for some k. We claim $A_m = A_k$ for all $m \geqslant k$. We know that $A_k \subseteq A_m$ for all $m \geqslant k$ because each ideal in the chain is contained in the next one (the chain is ascending). On the other hand, for any $m \geqslant k$, we have $A_m \subseteq \bigcup_{i=1}^{\infty} A_i = B = D\kappa \subseteq A_k$, because $\kappa \in A_k$ and A_k is an ideal of D. Thus $A_m = A_k$ for all $m \geqslant k$ and D satisfies the ascending chain condition. □

Using Theorem 32.14, we shall prove the analog of Theorem 5.13 for any arbitrary principal ideal domain.

32.15 Theorem: *Let D be a principal ideal domain. Then every element of D that is neither zero nor a unit can be exressed as a product of finitely many irreducible elements of D.*

Proof: First we prove that every element α in D, which is neither zero nor a unit, has an irreducible divisor in D. Let $\alpha \in D$, $\alpha \neq 0$, $\alpha \neq$ unit. Arguing as on page 366, we get a sequence

$$\alpha = \alpha_0, \alpha_1, \alpha_2, \ldots \qquad (s)$$

of elements in D, where α_{i+1} is a proper divisor of α_i $(i = 0,1,2, \ldots)$. In particular, none of the α_i is a unit. This sequence gives rise to the ascending chain

$$D\alpha = D\alpha_0 \subset D\alpha_1 \subset D\alpha_2 \subset \ldots$$

of ideals of D (here $D\alpha_i \subset D\alpha_{i+1}$ because α_{i+1} is a *proper* divisor of α_i). Since D is noetherian (Theorem 32.14), this chain breaks off: the chain consists only of the ideals

$$D\alpha = D\alpha_0 \subset D\alpha_1 \subset D\alpha_2 \subset \ldots \subset D\alpha_k$$

say. Hence

$$\alpha = \alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$$

are the only elements in the sequence (s). We claim that α_k is irreducible in D. Otherwise, there would be proper divisors α_{k+1} and β_{k+1} of α_k with $\alpha_k = \alpha_{k+1}\beta_{k+1}$, and the sequence (s) would contain the term α_{k+1} after α_k, and would not terminate with the term α_k, a contradiction. Hence α_k is an irreducible divisor of α. We proved that every element α in D, which is neither zero nor a unit, has an irreducible divisor in D.

Let β be an arbitrary nonzero, nonunit element in D. We want to show that β can be written as a product of finitely many irreducible elements in D. By what we proved above, we know that β has an irreducible divisor, π_1 say. We put $\beta = \pi_1\beta_1$. Here $\beta_1 \neq 0$. If β_1 is not a unit, then β_1 has an irreducible divisor, π_2 say. We put $\beta_1 = \pi_2\beta_2$. Thus $\beta = \pi_1\pi_2\beta_2$. Here $\beta_2 \neq 0$. If β_2 is not a unit, then β_2 has an irreducible divisor, π_3 say. We put $\beta_2 = \pi_3\beta_3$. Thus $\beta = \pi_1\pi_2\pi_3\beta_3$. Continuing in this way, we get a sequence

$$\beta = \beta_0, \beta_1, \beta_2, \beta_3, \ldots$$

of elements in D inducing a chain

$$D\beta = D\beta_0 \subset D\beta_1 \subset D\beta_2 \subset D\beta_3 \subset \ldots$$

of ideals of D. By Theorem 32.14, this chain is finite, for example

$$D\beta = D\beta_0 \subset D\beta_1 \subset D\beta_2 \subset D\beta_3 \subset \ldots \subset D\beta_k$$

We claim β_k is a unit. Otherwise β_k would have an irreducible divisor π_{k+1} and, when we put $\beta_k = \pi_{k+1}\beta_{k+1}$, there would be, in the chain, an additional ideal $D\beta_{k+1}$ containing $D\beta_k$ properly, a contradiction. Thus β_k is a unit. Then

$$\beta = \pi_1\pi_2\pi_3\ldots\pi_{k-1}\pi_k\beta_k = \pi_1\pi_2\pi_3\ldots\pi_{k-1}(\pi_k\beta_k)$$

is a product of the irreducible elements $\pi_1, \pi_2, \pi_3, \ldots, \pi_{k-1}, \pi_k\beta_k$. □

Having established the analog of Theorem 5.13, we proceed to work out the counterpart of Euclid's lemma (Lemma 5.15). For this we need the notion of greatest common divisor.

32.16 Definition: Let D be an arbitrary integral domain and let $\alpha,\beta \in D$, not both zero. An element δ of D is called a *greatest common divisor of* α *and* β if

(i) $\delta|\alpha$ and $\delta|\beta$,
(ii) for all δ_1 in D, if $\delta_1|\alpha$ and $\delta_1|\beta$, then $\delta_1|\delta$.

Notice that any associate of δ above satisfies the same conditions and hence any associate of a greatest common divisor of α and β is also a greatest common divisor of α and β. It is seen easily that any two greatest common divisor of α and β, if α and β have a greatest common divisor at all, are associates. So a greatest common divisor of α and β is not uniquely determined and we have to say *a* greatest common divisor, not *the* greatest common divisor.

We let (α,β) stand for any greatest common divisor of α and β. Thus (α,β) is determined uniquely to within ambiguity among associate elements.

Although we defined a greatest common divisor of two elements in an integral domain, this does not mean, of course, that any two elements (not both zero) in that domain do have a greatest common divisor. Introducing a definition does not create the *definiendum*. Given two elements (not both zero) in an integral domain, we cannot assert that they have a greatest common divisor. As a matter of fact, in an arbitrary integral domain, not every pair of elements (not both zero) has a greatest common divisor. For the special class of principal ideal domains, however, the following theorem holds.

32.17 Theorem: *Let D be a principal ideal domain and let α,β be arbitrary elements in D, not both of them being zero. Then there is a greatest common divisor of α and β in D. Furthermore, if δ is a greatest common divisor of α and β, then there are $\xi,\eta \in D$ such that $\delta = \alpha\xi + \beta\eta$.*

Proof: (cf. Theorem 5.4) As in the proof of Theorem 5.4, we consider the set $A := \{\alpha\xi + \beta\eta : \xi,\eta \in D\}$. A is a nonempty subset of D. We claim that A is

an ideal of D. To prove this, let γ_1, γ_2 be arbitrary elements of A. Then $\gamma_1 = \alpha\xi_1 + \beta\eta_1$, $\gamma_2 = \alpha\xi_2 + \beta\eta_2$ for some $\xi_1, \xi_2, \eta_1, \eta_2 \in D$. Hence

$$\gamma_1 + \gamma_2 = (\alpha\xi_1 + \beta\eta_1) + (\alpha\xi_2 + \beta\eta_2) = \alpha(\xi_1 + \xi_2) + \beta(\eta_1 + \eta_2) \in D$$
$$-\gamma_1 = -(\alpha\xi_1 + \beta\eta_1) = \alpha(-\xi_1) + \beta(-\eta_1) \in D$$

and therefore A is a subgroup of D under addition. Also, for any $\kappa \in D$,

$$\kappa\gamma_1 = \kappa(\alpha\xi_1 + \beta\eta_1) = \alpha(\kappa\xi_1) + \beta(\kappa\eta_1) \in D$$

and thus A has the "absorbing" property as well. So A is an ideal of D.

Since α, β are not both equal to zero, $A \neq \{0\}$. Now D is a principal ideal domain, so $A = D\delta$ for some $\delta \in D$, and $A \neq \{0\}$ implies $\delta \neq 0$. Also, since $\delta = 1\delta \in D\delta = A$, there are $\xi_0, \eta_0 \in D$ with $\delta = \alpha\xi_0 + \beta\eta_0$. We prove now that δ is a greatest common divisor of α and β.

(i) $\alpha = \alpha 1 + \beta 0 \in A = D\delta$, so $\alpha = \kappa\delta = \delta\kappa$ for some $\kappa \in D$. Since we have $\delta \neq 0$, we can write $\delta | \alpha$. Likewise $\delta | \beta$.

(ii) If $\delta_1 \in D$ and $\delta_1 | \alpha$, $\delta_1 | \beta$, then $\delta_1 | \alpha\xi_0 + \beta\eta_0$, hence $\delta_1 | \delta$.

Thus δ is a greatest common divisor of α and β, and $\delta = \alpha\xi_0 + \beta\eta_0$ for some $\xi_0, \eta_0 \in D$. The proof is complete. □

In a principal ideal domain D, we see that $D\alpha + D\beta = D(\alpha,\beta)$ whenever $\alpha,\beta \in D$ are not both zero. Either from this remark, or better from Definition 32.16, it follows that $(\alpha,\beta) = (\beta,\alpha)$. When $(\alpha,\beta) \approx 1$, we say α is relatively prime to β, or α and β are relatively prime. In this case, there are ξ, η in D with $\alpha\xi + \beta\eta = 1$.

32.18 Lemma: *Let D be a principal ideal domain and $\alpha,\beta,\gamma,\pi \in D$.*
(1) *If $\gamma | \alpha\beta$ and $(\gamma,\alpha) \approx 1$, then $\gamma | \beta$.*
(2) *If π is irreducible and $\pi \nmid \alpha$, then $(\pi,\alpha) \approx 1$.*

Proof: (1) If $(\gamma,\alpha) \approx 1$, we have $\gamma\xi + \alpha\eta = 1$ for some $\xi, \eta \in D$. Hence $\gamma\beta\xi + \alpha\beta\eta = \beta$. Now $\gamma | \gamma\beta\xi$ and $\gamma | \alpha\beta$, so $\gamma | \gamma\beta\xi$ and $\gamma | \alpha\beta\eta$, so $\gamma | \gamma\beta\xi + \alpha\beta\eta$, so $\gamma | \beta$.

(2) Let π be irreducible. Then $\pi \neq 0$. So (π,α) exists by Theorem 32.17. Let δ be a greatest common divisor of π and α. Then $\delta | \pi$. Since π is irre-

ducible, either δ is associate to π, or δ is a unit. In the first case $\delta \approx \pi$, we get $\pi | \alpha$ (since $\delta | \alpha$) against our hypothesis $\pi \nmid \alpha$. Thus δ is a unit and $\delta \approx 1$, as claimed. □

32.19 Lemma: *Let D be a principal ideal domain and $\alpha, \beta, \pi \in D$. If π is irreducible and $\pi | \alpha\beta$, then $\pi | \alpha$ or $\pi | \beta$.*

Proof: If $\pi | \alpha$, the lemma is true. If $\pi \nmid \alpha$, then $(\pi, \alpha) \approx 1$ by Lemma 32.18(2) and so $\pi | \beta$ by Lemma 32.18(1), with π in place of γ. □

32.20 Definition: Let D be an arbitrary integral domain. If $\pi \in D$ is not zero or a unit, and if π has the property that

$$\text{for all } \alpha, \beta \in D, \qquad \pi | \alpha\beta \implies \pi | \alpha \text{ or } \pi | \beta,$$

then π is called a *prime* element in D.

In an arbitrary integral domain, all prime elements are irreducible. Indeed, let π be prime. Then π is not zero or a unit by definition. We show that π has no proper divisors. Suppose $\pi = \alpha\beta$. Then $\pi | \alpha\beta$ and therefore $\pi | \alpha$ or $\pi | \beta$. Without restricting generality, let us assume $\pi | \alpha$. But $\alpha | \pi$ as well, so $\pi \approx \alpha$ and β is a unit. Thus π admits no proper factorization and π is irreducible.

The converse of this remark is not true. That is to say, in an arbitrary integral domain, there may be irreducible elements which are not prime (see Ex. 13). Lemma 32.19 asserts that irreducible and prime elements coincide in a principal ideal domain. This is the basic reason why there turns out to be a unique factorization theorem in principal ideal domains.

32.21 Lemma: *Let D be an arbitrary ideal domain and $\alpha_1, \alpha_2, \ldots, \alpha_n, \pi \in D$. If π is prime and $\pi | \alpha_1\alpha_2 \ldots \alpha_n$, then $\pi | \alpha_1$ or $\pi | \alpha_2$ or ... or $\pi | \alpha_n$.*

Proof: Omitted.

32.22 Theorem: *Let D be a principal ideal domain. Every element of D, which is not zero or a unit, can be expressed as a product of irreducible elements of D in a unique way, apart from the order of the factors and the ambiguity among associate elements.*

Proof: (cf. Theorem 5.17.) Let $\alpha \in D$, $\alpha \neq 0$, $\alpha \neq$ unit. By Theorem 32.15, α can be expressed as a product of irreducible elements of D. We must show uniqueness. Given two decompositions

$$\pi_1\pi_2\ldots\pi_r = \alpha = \sigma_1\sigma_2\ldots\sigma_s$$

of α into irreducible elements, we must show $r = s$ and $\pi_1,\pi_2,\ldots,\pi_r$ are, in some order, associate to $\sigma_1,\sigma_2,\ldots,\sigma_s$. This will be proved by induction on r.

First assume $r = 1$. Then $\pi_1 = \alpha = \sigma_1\sigma_2\ldots\sigma_s$, so α is irreducible. This forces $s = 1$ and $\pi_1 = \alpha = \sigma_1$. This proves the theorem when $r = 1$.

Now assume $r \geqslant 2$ and that the theorem is proved for $r - 1$. This means, whenever we have an equation

$$\pi_1'\pi_2'\ldots\pi_{r-1}' = \sigma_1'\sigma_2'\ldots\sigma_t'$$

with irreducible π', σ', there holds $r - 1 = t$ and $\pi_1', \pi_2', \ldots, \pi_{r-1}'$ are, in some order, associates of $\sigma_1', \sigma_2', \ldots, \sigma_t'$.

We have $\pi_1\pi_2\ldots\pi_r = \alpha = \sigma_1\sigma_2\ldots\sigma_s$. So $\pi_r \mid \alpha$. So $\pi_r \mid \sigma_1\sigma_2\ldots\sigma_s$. Now π_r is prime by Lemma 32.19, and so $\pi_r \mid \sigma_j$ for some $j \in \{1,2,\ldots,s\}$. (Here we use the fact that irreducible elements are prime in a principal ideal domain. The conclusion $\pi_r \mid \sigma_j$ is *not* valid in an arbitrary integral domain.) Reordering the σ's if necessary, we may assume $\pi_r \mid \sigma_s$. Since σ_s is irreducible, σ_s has no proper divisors. Hence the divisor π_r of σ_s is either a unit or an associate of σ_s. But π_r is irreducible, so not a unit. Therefore π_r and σ_s are associate. So $\pi_r = \varepsilon\sigma_s$ for some unit $\varepsilon \in D$. Then we obtain

$$\pi_1\pi_2\ldots\pi_{r-1}(\varepsilon\sigma_s) = \alpha = \sigma_1\sigma_2\ldots\sigma_s$$
$$\pi_1\pi_2\ldots(\pi_{r-1}\varepsilon) = \sigma_1\sigma_2\ldots\sigma_{s-1}$$

and by induction, we get

$$r - 1 = s - 1,$$

$\pi_1, \pi_2, \ldots, (\pi_{r-1}\varepsilon)$ are, in some order, associates of $\sigma_1, \sigma_2, \ldots, \sigma_{s-1}$.

Hence $r = s$
and $\pi_1, \pi_2, \ldots, \pi_{r-1}$ are, in some order, associates of $\sigma_1, \sigma_2, \ldots, \sigma_{s-1}$, and π_r is associate to σ_s. This completes the proof. □

32.23 Definition: Let D be an integral domain. If every element of D, which is not zero or a unit, can be expressed as a product of finitely many irreducible elements of D in a unique way, apart from the order of the factors and the ambiguity among associate elements, then D is called a *unique factorization domain.*

With this definition, Theorem 32.22 reads as follows.

32.22 Theorem: *Every principal ideal domain is a unique factorization domain. In particular, every Euclidean domain is a unique factorization domain.* □

We generalize Lemma 32.19 to unique factorization domains.

32.24 Lemma: *Let D be a unique factorization domain. Then every irreducible element of D is prime.*

Proof: Let π be irreducible in D and $\pi | \alpha\beta$, where $\alpha, \beta \in D$. Thus there is a $\gamma \in D$ with $\pi\gamma = \alpha\beta$. Then

$$\alpha = \varepsilon\pi_1\pi_2\ldots\pi_r, \quad \beta = \varepsilon'\pi_1'\pi_2'\ldots\pi_s', \quad \gamma = \varepsilon''\pi_1''\pi_2''\ldots\pi_t'',$$

where $\varepsilon, \varepsilon', \varepsilon''$ are units and π_i, π_j', π_k'' are irreducible elements in D. From the uniqueness of the decomposition

$$\pi\varepsilon''\pi_1''\pi_2''\ldots\pi_t'' = \pi\gamma = \alpha\beta = \varepsilon\pi_1\pi_2\ldots\pi_r\varepsilon'\pi_1'\pi_2'\ldots\pi_s' = \varepsilon\varepsilon'\pi_1\pi_2\ldots\pi_r\pi_1'\pi_2'\ldots\pi_s'$$

we see that π must be associate to one of the irreducible elements π_j' or π_k''. Thus π divides α or β. So π is prime. □

There is the following generalization of Theorem 32.22. If D is an integral domain in which every nonzero, nonunit element can be written as a product of finitely many irreducible elements, and if every irreducible element in D is prime, then D is a unique factorization domain. The proof of Theorem 32.22 is valid in this more general case.

In a unique factorization domain D, any two elements α,β (not both zero) have a greatest common divisor. Clearly $\alpha \approx (\alpha,\beta)$ if $\beta = 0$ and $\beta \approx (\alpha,\beta)$ if $\alpha = 0$; and if $\alpha \neq 0 \neq \beta$, then $\alpha = \varepsilon \pi_1^{m_1}\pi_2^{m_2}\ldots\pi_r^{m_r}$ and $\beta = \varepsilon'\pi_1^{n_1}\pi_2^{n_2}\ldots\pi_r^{n_r}$ with suitable units ε,ε', irreducible elements $\pi_1, \pi_2, \ldots, \pi_r$ and nonnegative integers m_i, n_i, and it is easily seen that $\gamma = \pi_1^{k_1}\pi_2^{k_2}\ldots\pi_r^{k_r}$, where $k_i = \min\{m_i, n_i\}$, is a greatest common divisor of α and β. Thus (α,β) exists in any unique factorization domain, provided only α, β are not both equal to zero. However, in an arbitrary unique factorization domain, (α,β) cannot, in general, be expressed in the form $\alpha\xi + \beta\eta$.

There are unique factorization domains which are not principal ideal domains and there are principal ideal domains which are not Euclidean domains.

32.25 Theorem: *Let D be a principal ideal domain and let $\pi \in D$ be a nonzero, nonunit element of D. Then π is irreducible if and only if the factor ring $D/D\pi$ is a field.*

Proof: $D/D\pi$ is a commutative ring with identity $1 + D\pi$ (Example 30.9(c)). Suppose π is irreducible. We are to show that every nonzero element $\alpha + D\pi$ of $D/D\pi$ has an inverse in $D/D\pi$. Let $\alpha + D\pi$ be distinct from the zero element $0 + D\pi = D\pi$ of $D/D\pi$. This means $\alpha \notin D\pi$, so $\pi \nmid \alpha$. Since π is irreducible, we obtain $(\pi,\alpha) \approx 1$ from Lemma 32.18(2), and there are therefore β,γ in D such that $\pi\beta + \alpha\gamma = 1$. So $\alpha\gamma - 1 \in D\pi$ and

$$(\alpha + D\pi)(\gamma + D\pi) = 1 + D\pi.$$

Thus $\gamma + D\pi$ is an inverse of $\alpha + D\pi$. This proves that $D/D\pi$ is a field.

We now prove that, if π is not irreducible, then $D/D\pi$ is not a field. Indeed, if π is not irreducible, then $\pi = \alpha\beta$ for some $\alpha,\beta \in D$, where neither α nor β is a unit. Here $\pi \mid \alpha$ and, in view of this, $\alpha \mid \pi$ would imply

that $\pi \approx \alpha$; then β would be a unit, a contradiction. Hence $\pi \nmid \alpha$ and likewise $\pi \nmid \beta$. So $\alpha \notin D\pi$ and $\beta \notin D\pi$, so $\alpha + D\pi \neq 0 + D\pi \neq \beta + D\pi$, but

$$(\alpha + D\pi)(\beta + D\pi) = \alpha\beta + D\pi = \pi + D\pi = 0 + D\pi = \text{zero element of } D/D\pi.$$

Thus $\alpha + D\pi$ and $\beta + D\pi$ are zero divisors in $D/D\pi$ and $D/D\pi$ cannot be a field. □

Exercises

1. Let D be an integral domain and $\pi \in D$. Prove that π is a prime element of D if and only if $D\pi$ is a prime ideal of D (see §30, Ex. 11).

2. Let D be a principal ideal domain and $\pi \in D$. Prove that π is an irreducible element of D if and only if $D\pi$ is a maximal ideal of D (see §30, Ex. 12).

3. Show that $\mathbb{Z}[\sqrt{d}] := \{a + b\sqrt{d} \in \mathbb{R} : a,b \in \mathbb{Z}\}$ is a Euclidean domain when $d = 2,3,6$.

4. Show that $\mathbb{Z}[\sqrt{2}i] := \{a + b\sqrt{2}i \in \mathbb{C} : a,b \in \mathbb{Z}\}$ is a Euclidean domain.

5. Let $\theta = \frac{1+\sqrt{7}i}{2}$. Show that $\mathbb{Z}[\theta] := \{a + b\theta \in \mathbb{C} : a,b \in \mathbb{Z}\}$ is a Euclidean domain.

6. Let D be a Euclidean domain, with the function d as in Definition 32.10. Prove that $\varepsilon \in D$ is a unit if and only if $d(\varepsilon) = d(1)$.

7. Find the decomposition into irreducible elements of 2 in $\mathbb{Z}[i]$ and of 3 in $\mathbb{Z}[\omega]$.

8. Let $p \in \mathbb{N}$ be an odd prime number. Prove that (i) $p = p + 0i \in \mathbb{Z}[i]$ is prime in $\mathbb{Z}[i]$ in case $x^2 \equiv -1 \pmod{p}$ has no solution and (ii) $p \in \mathbb{Z}[i]$ is not prime in $\mathbb{Z}[i]$, and in fact $p = \alpha\bar{\alpha}$ with a suitable prime element α of $\mathbb{Z}[i]$, in case $x^2 \equiv -1 \pmod{p}$ has a solution.

9. Let $\alpha \in \mathbb{Z}[i]$. Show that $\mathbb{Z}[i]/\alpha\mathbb{Z}[i]$ has exactly $N(\alpha)$ elements.

10. Using the Euclidean algorithm, find a greatest common divisor of

$3 + 5i$ and $2 + 3i$; and of $14 + 23i$ and $11 + 44i$ in $\mathbb{Z}[i]$.

11. Prove: an integral domain is a principal ideal domain if and only if there is a function $d: D\backslash\{0\} \to \mathbb{N} \cup \{0\} \subseteq \mathbb{Z}$ satisfying

(i) $d(\alpha) \leqslant d(\beta)$ for any $\alpha,\beta \in D\backslash\{0\}$ with $\alpha \mid \beta$, and $d(\alpha) = d(\beta)$ if and only if $\alpha \approx \beta$;

(ii) for all $\alpha,\beta \in D\backslash\{0\}$ with $\alpha \nmid \beta$ and $\beta \nmid \alpha$, there are $\tau, \kappa, \rho \in D$ such that $\tau\alpha = \kappa\beta + \rho$ and $d(\rho) < \min\{d(\alpha),d(\beta)\}$.

12. Let $\lambda = \frac{1+\sqrt{19}i}{2}$. Show that $\mathbb{Z}[\lambda] := \{a + b\lambda \in \mathbb{C}: a,b \in \mathbb{Z}\}$ is a principal ideal domain, but not a Euclidean domain.

13. Prove that 2, 3, $1 + \sqrt{5}i$, $1 - \sqrt{5}i$ are irreducible in $\mathbb{Z}[\sqrt{5}i]$. Show that $2,3$ are not associate to $1 + \sqrt{5}i$, $1 - \sqrt{5}i$. Hence there are two essentially distinct decompositions

$$2 \cdot 3 = 6 = (1 + \sqrt{5}i)(1 - \sqrt{5}i)$$

of $6 \in \mathbb{Z}[\sqrt{5}i]$ and therefore $\mathbb{Z}[\sqrt{5}i]$ is not a unique factorization domain.

§33
Polynomial Rings

The reader is familiar with polynomials. In high school, it is taught that expressions like

$$x^2 + 2x + 5 \qquad x^3 + 2x^2 - 7x + 1$$

are polynomials. One learns how to add, subtract, multiply and divide two polynomials. Although one acquires a working knowledge about polynomials, a satisfactory definition of polynomials is hardly given. In this paragraph, we give a rigorous definition of polynomials.

Polynomials are treated in the calculus as functions. For example, $x^2 + 2x + 5$ is considered to be the function (defined on $\mathbb{R}$, say) that maps any $x \in \mathbb{R}$ to $x^2 + 2x + 5$. With this interpretation, a polynomial is a function and x is a generic element in its domain. The equality of two polynomials means then the equality of their domains and the equality of the function values at any element in their domain.

This is a perfectly sound approach, but it will prove convenient to treat polynomials differently in algebra. We propose to define the equality of two polynomials as the equality of their corresponding coefficients. This definition is motivated by the so-called comparison of coefficients. Note that this definition of equality does not involve x at all. Whatever x may be, it is not relevant to the definition of equality. Nor is it relevant to the addition and multiplication of two polynomials. So we may forget about x completely. We then deprive of a polynomial $a_0 + a_1x + \cdots + a_nx^n$ of the symbols x^r. What remains is a finite number of coefficients and "+" signs. The "+" signs can be thought of as connectives. Then a polynomial is essentially a finite number of coefficients. This leads to the following definition.

33.1 Definitions: Let R be a ring. A sequence

$$f = (a_0, a_1, a_2, \ldots)$$

of elements $a_0, a_1, a_2, \ldots$ in R, where only finitely many of them are distinct from the zero element of R, is called a *polynomial over* R.

The terms $a_0, a_1, a_2, \ldots$ are called the *coefficients* of the polynomial $f = (a_0, a_1, a_2, \ldots)$. The term a_0 will be referred to as the *constant term of* f.

Two polynomials $f = (a_0,a_1,a_2, \ldots)$ and $g = (b_0,b_1,b_2, \ldots)$ over R are declared *equal* when they are equal as sequences of course, that is to say, when $a_i = b_i$ for all $i = 0,1,2, \ldots$. In this case, we write $f = g$. Otherwise we put $f \neq g$.

If $f = (a_0, a_1, a_2, \ldots)$ is a polynomial over R, there is an index d such that $a_n = 0 \in R$ whenever $n > d$. If the coefficients $a_0, a_1, a_2, \ldots$ are not all equal to zero, there is an index d, uniquely determined by f, such that $a_d \neq 0$ and $a_n = 0$ for all $n > d$. This index d is called the *degree of* f. We write then $d = deg\ f$. If d is the degree of f, then a_d is said to be the *leading coefficient of* f. It is the last nonzero coefficient of f. If R happens to be a ring with identity 1 and if f is a polynomial over R with leading coefficient equal to 1, then f is called a *monic* polynomial.

A polynomial of degree one is called a *linear* polynomial, one of degree two is called a *quadratic* polynomial, one of degree three is called a *cubic* polynomial, one of degree four is called a *biquadratic* or *quartic* polynomial and one of degree five is called a *quintic* polynomial.

The polynomial $0^* = (0,0,0, \ldots)$ over R, whose terms are all equal to the zero element $0 \in R$ of R, is called the *zero polynomial* over R. The leading coefficient and the degree of the zero polynomial are *not* defined. The leading coefficient of any other polynomial *is* defined. The constant term of the zero polynomial *is* defined, and is $0 \in R$.

Notice that indexing begins with 0, not with 1. For example, $(1,0,2,5,0,0,0, \ldots)$ is a polynomial over $\mathbb{Z}$ of degree 3, not of degree 4. Its constant term is $1 \in \mathbb{Z}$, leading coefficient is $5 \in \mathbb{Z}$.

33.2 Definition: Let R be a ring and let

$$f = (a_0, a_1, a_2, \ldots) \quad \text{and} \quad g = (b_0, b_1, b_2, \ldots)$$

be two polynomials over R. Then the *sum of f and g*, denoted by $f + g$, is the sequence

$$f + g = (a_0 + b_0, a_1 + b_1, a_2 + b_2, \ldots)$$

obtained by termwise addition of the coefficients. The *product of f by g*, denoted by $f \cdot g$ or by fg, is the sequence

$$fg = (c_0, c_1, c_2, \ldots)$$

where the terms $c \in R$ are given by

$$\begin{aligned}
c_0 &= a_0 b_0 \\
c_1 &= a_0 b_1 + a_1 b_0 \\
c_2 &= a_0 b_2 + a_1 b_1 + a_2 b_0 \\
c_3 &= a_0 b_3 + a_1 b_2 + a_2 b_1 + a_3 b_0 \\
&\cdots\cdots\cdots\cdots \\
c_k &= a_0 b_k + a_1 b_{k-1} + a_2 b_{k-2} + \cdots + a_{k-2} b_2 + a_{k-1} b_1 + a_k b_0 \\
&\cdots\cdots\cdots\cdots
\end{aligned}$$

To find the k-th term c_k in fg, we multiply all a's with all b's in such a way that the sum of the indices is k, and add the results. We write $c_k = \sum_{i=0}^{k} a_i b_{k-i}$. The summation variable runs through different values for different k's (through 0,1,2,3 for $k = 3$, through 0,1,2,3,4,5 for $k = 5$, etc.). It will be convenient to write $c_k = \sum_{i+j=k} a_i b_j$, it being understood that i and j run through nonnegative integers in such a way that their sum is k.

33.3 Lemma: *Let R be a ring and let $f = (a_0, a_1, a_2, \ldots)$ and $g = (b_0, b_1, b_2, \ldots)$ be arbitrary polynomials over R. Let $0^* = (0,0,0, \ldots)$ be the zero polynomial over R.*

(1) $f + 0^* = f$ and $0^* + g = g$. Also $f0^* = 0^*$ and $0^* g = 0^*$.
(2) The sum $f + g$ is a polynomial over R. If $\deg f = m$ and $\deg g = n$, then

$deg(f+g) = max\{m,n\}$ *in case* $m \neq n$,

$deg(f+g) \leqslant m$ *in case* $m = n$ *and* $f+g \neq 0^*$.

(3) *The product* fg *is a polynomial over* R. *If* $deg\ f = m$ *and* $deg\ g = n$, *then*

$deg\ fg \leqslant m+n$ *in case* $fg \neq 0^*$,

$deg\ fg = m+n$ *in case* R *has no zero divisors.*

Proof: (1) The assertions $f + 0^* = f$ and $0^* + g = g$ are immediate from the definitions: $f + 0^* = (a_0,a_1,a_2,\ldots) + (0,0,0,\ldots) = (a_0 + 0,a_1 + 0,a_2 + 0,\ldots) = (a_0,a_1,a_2,\ldots) = f$ and similarly $0^* + g = g$. Also, the k-th coefficient of $f0^*$ is $a_00 + a_10 + a_20 + \cdots + a_k0 = 0 + 0 + 0 + \cdots + 0 = 0$ by Lemma 29.6, for any k. This proves $f0^* = 0^*$. Likewise $0^*g = 0^*$.

(2) We must show that $f + g$ has only finitely many terms distinct from 0. We proved it in part (1) when $f = 0^*$ or $g = 0^*$. Now we assume $f \neq 0^* \neq g$. Then f and g have degrees. Suppose *def* $f = m$ and *deg* $g = n$, so that $a_m \neq 0$, $a_r = 0$ for all $r > m$ and $b_n \neq 0$, $b_r = 0$ for all $r > n$.

If $m < n$, then $f + g = (a_0 + b_0, a_1 + b_1, \ldots, a_m + b_m, b_{m+1}, \ldots, b_n, 0, 0, 0, \ldots)$. So the n-th term in $f + g$ is $b_n \neq 0$, and the later terms are $a_r + b_r = 0 + 0 = 0$ for $r > n > m$. This shows that $f + g$ is a nonzero polynomial and $deg\ f + g = n = \max\{m,n\}$.

If $n < m$, then $f + g = (a_0 + b_0, a_1 + b_1, \ldots, a_n + b_n, a_{n+1}, \ldots, a_m, 0, 0, 0, \ldots)$. So the m-th term in $f + g$ is $a_m \neq 0$, and the later terms are $a_r + b_r = 0 + 0 = 0$ for $r > m > n$. This shows that $f + g$ is a nonzero polynomial and $deg\ f + g = m = \max\{m,n\}$. [Question: why cannot we combine the two cases $m < n$ and $n < m$ into a single one by assuming $m < n$ without loss of generality?]

If $m = n$, then $f + g = (a_0 + b_0, a_1 + b_1, \ldots, a_m + b_m, 0, 0, 0, \ldots)$. The r-th term in $f + g$ is $a_r + b_r = 0$ for all $r > m$. This shows that $f + g$ is a polynomial. Either it is the zero polynomial, or it is not the zero polynomial. In the latter case, the nonzero terms in $f + g$ have indices $\leqslant m$. In particular, the degree of $f + g$ is $\leqslant m$. (More exactly, $deg\ f + g = m$ if $a_m + b_m \neq 0$, and $deg\ f + g < m$ if $a_m + b_m = 0$.)

(3) To prove that the product fg is a polynomial over R, we must show that fg has only finitely many terms distinct from zero. We proved it in part (1) when $f = 0^*$ or $g = 0^*$. Now we assume $f \neq 0^* \neq g$. Then f and g

have degrees. Suppose $def\ f = m$ and $deg\ g = n$, so that $a_m \neq 0$, $a_r = 0$ for all $r > m$ and $b_n \neq 0$, $b_r = 0$ for all $r > n$.

The k-th term in $fg = (c_0, c_1, c_2, \ldots)$ is given by $c_k = \sum_{i+j=k} a_i b_j$. Suppose now $k > m + n$. If $i + j = k$, then either $i > m$ or $j > n$ (for $i \leq m$ and $j \leq n$ implies the contradiction $k = i + j \leq m + n < k$), so either $a_i = 0$ or $b_j = 0$ for each one of the summands $a_i b_j$ in $c_k = \sum_{i+j=k} a_i b_j$. So each summand is either $0b_j = 0$ or $a_i 0 = 0$ by Lemma 29.6 and $c_k = 0 + 0 + \cdots + 0 = 0$. This shows that $c_k = 0$ for all $k > m + n$. Hence fg has at most $m + n$ terms distinct from 0 and fg is a polynomial over R and $deg\ fg \leq m + n$ in case $fg \neq 0$*.

The $(m + n)$-th term c_{m+n} in fg is

$$\begin{aligned} c_{m+n} &= a_0 b_{m+n} + a_1 b_{m+n-1} + a_2 b_{m+n-2} + \cdots + a_{m-1} b_{n+1} \\ &\quad + a_m b_n \\ &\quad + a_{m+1} b_{n-1} + a_{m+2} b_{n-2} + \cdots + a_{m+n-1} b_1 + a_{m+n} b_0. \end{aligned}$$

Here the summands in the first line are 0 since $b_{m+n}, b_{m+n-1}, b_{m+n-2}, \ldots, b_{n+1}$ are 0 and the summands in the third line are 0 since $a_{m+1}, a_{m+2}, \ldots a_{m+n-1}$, a_{m+n} are 0. This gives $c_{m+n} = a_m b_n$. If R has no zero divisors, then $c_{m+n} = a_m b_n$ since $a_m \neq 0$ and $b_n \neq 0$. So $m + n$ is the greatest index k for which the k-th term in fg is distinct from 0. This proves that $deg\ fg = m + n$ in case R has no zero divisors. □

33.4 **Remark:** The last argument shows in fact that the leading coefficient of fg is the leading coefficient of f times the leading coefficient of g, provided R has no zero divisors.

33.5 **Theorem:** *Let R be a ring. The set of all polynomials over R is a ring with respect to the operations $+$ and $\cdot$ given in Definition* 33.2 (called the *addition* and *multiplication* of polynomials, respectively).

Proof: First of all, we must prove that $+$ makes the set of all polynomial over R into an abelian group. The closure property was shown in Lemma

33.3(2). The associativity and commutativity of addition of polynomials follow from the associativity and commutativity of addition in R. The zero polynomial 0^* is the zero element (Lemma 33.3(1)) and each polynomial $(a_0,a_1,a_2,\ldots)$ over R has an opposite $(-a_0,-a_1,-a_2,\ldots)$. The details are left to the reader.

Now the properties of multiplication in Definition 29.1. The closure of the set of all polynomial over R under multiplication was shown in Lemma 33.3(3). The associativity of multiplication is proved by observing that the m-th term in $(fg)h$, where

$$f = (a_0,a_1,a_2,\ldots), \qquad g = (b_0,b_1,b_2,\ldots), \qquad h = (c_0,c_1,c_2,\ldots)$$

are arbitrary polynomials over R, is given by

$$\sum_{k+l=m}(k\text{-th term in } fg)c_l = \sum_{k+l=m}\Big(\sum_{i+j=k}a_ib_j\Big)c_l = \sum_{i+j+l=m}(a_ib_j)c_l$$

and that the m-th term in $f(gh)$ is given by

$$\sum_{i+s=m}a_i(s\text{-th term in } gh) = \sum_{i+s=m}a_i\Big(\sum_{j+l=s}b_jc_l\Big) = \sum_{i+j+l=m}a_i(b_jc_l).$$

Here we used the distributivity in R. Since $(a_ib_j)c_l = a_i(b_jc_l)$, the m-th term in $(fg)h$ and $f(gh)$ are equal, and this for all m. So $(fg)h = f(gh)$ for all polynomials f,g,h over R and the multiplication is associative.

It remains to prove the distributivity laws. For any polynomials $f = (a_0,a_1,a_2,\ldots)$, $g = (b_0,b_1,b_2,\ldots)$, $h = (c_0,c_1,c_2,\ldots)$ over R, we have

$$f(g+h) = (a_0,a_1,a_2,\ldots)(b_0+c_0,b_1+c_1,b_2+c_2,\ldots)$$

$$= \text{polynomial whose } k\text{-th coefficient is } \sum_{i+j=k}a_i(b_j+c_j)$$

$$= \text{polynomial whose } k\text{-th coefficient is } \sum_{i+j=k}(a_ib_j+a_ic_j)$$

$$= \text{polynomial whose } k\text{-th coefficient is } \sum_{i+j=k}a_ib_j + \sum_{i+j=k}a_ic_j,$$

$$= \text{(polynomial whose } k\text{-th coefficient is } \sum_{i+j=k} a_i b_j \text{)}$$

$$+ \text{ (polynomial whose } k\text{-th coefficient is } \sum_{i+j=k} a_i c_j)$$

$$= fg + fh$$

and a similar argument proves $(f + g)h = fh + gh$ for all polynomials f,g,h over R. This completes the proof. □

The ring of all polynomials over R will be denoted by $R[x]$. When $f \in R[x]$, we say f is a *polynomial with coefficients in* R.

We now want to simplify our notation. A polynomial $f = (a_0,a_1,a_2, \ldots)$ over R, for which $a_n = 0$ whenever, say, $n > d$, can be written as

$$(a_0,0,0,0, \ldots) + (0,a_1,0,0, \ldots) + (0,0,a_2,0, \ldots) + \cdots + (0,0,\ldots,a_d,0, \ldots).$$

Each one of the polynomials above has at most one nonzero coefficient. A polynomial over R which has at most one nonzero coefficient will be called a *monomial over* R. We can write monomials over R more compactly as follows. If, for example, g is a monomial over R whose r-th coefficient is a (the possibility $a = 0$ is not excluded) and whose other coefficients are zero, then we can write $g = (0,0,\ldots,a,0, \ldots)$ shortly as (a,r). Here r denotes the index with the only the possibly nonzero element, and $a \in R$ is that possibly nonzero element in the r-th place. Then our f would be written as $(a_0,0) + (a_1,1) + (a_2,2) + \cdots + (a_d,d)$. The essential point is that a polynomial can be written as a sum of monomials, and a monomial is determined as soon as the index r and the possibly nonzero element a is given. We can choose other notations for monomials, of course, as long as they display the index r and the possibly nonzero element a. We prefer to write ax^r instead of (a,r) for the monomial $(0,0,\ldots,a,0, \ldots)$. In this notation, both the index r and the element a are displayed. It should be noted that x does *not* have a meaning by itself. It is like the comma in (a,r). In particular, x^r is *not* the r-th power of anything. r in ax^r is an index, a superscript showing where the element a sits in. With this notation, our f is written as

$$f = a_0x^0 + a_1x^1 + a_2x^2 + \cdots + a_dx^d.$$

The product of two monomials ax^r and bx^s is easily evaluated to be abx^{r+s}. The multiplication of two polynomials can be carried out in the familiar way by using this rule and the distributivity. The symbol x is a convenient device that simplifies computations. x will be called an *indeterminate (over R)*. This does not mean that x fails to be determined in some way. "Indeterminate" is just an odd name of a computational device. Finally, we agree to write a_0 for a_0x^0 and a_1x for a_1x^1. In particular, we write 0 for the zero polynomial 0*. This convention brings f to the form

$$a_0 + a_1x + a_2x^2 + \cdots + a_dx^d.$$

With the convention of writing a_0 for a_0x^0, we regard R as a subring of $R[x]$. In particular, we can multiply polynomials by elements of R in the natural way:

$$b(a_0 + a_1x + a_2x^2 + \cdots + a_dx^d) = ba_0 + ba_1x + ba_2x^2 + \cdots + ba_dx^d,$$
$$(a_0 + a_1x + a_2x^2 + \cdots + a_dx^d)b = a_0b + a_1bx + a_2bx^2 + \cdots + a_dbx^d.$$

If R is a ring with identity 1, then x can be interpreted in another way. The rule $ax^r \cdot bx^s = abx^{r+s}$ yields $1x^r \cdot 1x^s = 1x^{r+s}$. Let p denote the polynomial $1x = 1x^1 = (0,1,0,0, \ldots)$. We calculate that $p^2 = 1x^2$, $p^3 = 1x^3$, $p^4 = 1x^4$, etc. Our f can now be written as

$$f = a_0p + a_1p + a_2p^2 + \cdots + a_dp^d,$$

where this time the superscripts indicate the appropriate powers of $p = (0,1,0,0, \ldots)$, taken according to the definition of multiplication given in Definition 33.2. So any polynomial over R can be written as a sum of powers of p, and calculations are performed by using the distributivity. Since x obeys the same rules as a computational device as p does as a polynomial, we write the polynomial $p = 1x$ as x. Then x is the polynomial $(0,1,0,0, \ldots)$ in $R[x]$. We emphasize again that this interpretation of x as a polynomial is possible only when R has an identity. If R has no identity, then x is *not* a polynomial in $R[x]$.

The ring $R[x]$ is said to be constructed by *adjoining x to R*. When we want to examine several copies of $R[x]$ at the same time, we use different

letters to denote the indeterminates of the copies of $R[x]$. Thus we may have $R[x]$, $R[y]$, $R[z]$, etc.

Whenever convenient, we shall write $\sum_{i=0}^{d} a_i x^i$ for the polynomial

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d.$$

33.6 Lemma: *Let R be a ring.*

(1) *If R is commutative, then $R[x]$ is commutative.*

(2) *If R has an identity, then $R[x]$ has an identity.*

(3) *If R has no zero divisors, then $R[x]$ has no zero divisors.*

(4) *If R is an integral domain, then $R[x]$ is an integral domain.*

Proof: Let $f = \sum_{i=0}^{m} a_i x^i$ and $g = \sum_{j=0}^{n} b_j x^j$ be arbitrary polynomials in $R[x]$.

(1) If R is commutative, then $a_i b_j = b_j a_i$ for all $i = 0,1, \ldots, m$ and $j = 0,1, \ldots n$. We have then

$$fg = \sum_{k=0}^{m+n} \Big(\sum_{i+j=k} a_i b_j \Big) x^k = \sum_{k=0}^{n+m} \Big(\sum_{j+i=k} b_j a_i \Big) x^k = gf$$

and $R[x]$ is commutative.

(2) If R has an identity 1, then $1 = 1x^0 = (1,0,0,0, \ldots)$ is a polynomial in $R[x]$ and

$$f \cdot 1 = \Big(\sum_{i=0}^{m} a_i x^i \Big) 1 = \sum_{i=0}^{m} a_i 1 x^i = \sum_{i=0}^{m} a_i x^i = f,$$

$$1 \cdot f = 1 \Big(\sum_{i=0}^{m} a_i x^i \Big) = \sum_{i=0}^{m} 1 a_i x^i = \sum_{i=0}^{m} a_i x^i = f$$

for arbitrary $f \in R[x]$. Thus 1 is an identity element of $R[x]$.

(3) Assume now R has no zero divisors. Let us suppose also that $f \neq 0$ and $g \neq 0$. Without loss of generality, we may assume that a_m is the leading coefficient of f and that b_n is the leading coefficient of g. Then $a_m \neq 0$, $b_n \neq 0$. By remark 33.4, the leading coefficient of fg is $a_m b_n$ and $a_m b_n \neq 0$ since R has no zero divisors. Thus fg has a nonzero coefficient, namely the $(m + n)$-th coefficient and $fg \neq 0$. This shows that $R[x]$ has no zero divisors.

(4) An integral domain is a commutative ring with identity having no zero divisors, distinct from the null ring. Now if R is an integral domain, then R is not the null ring, and since $R \subseteq R[x]$, the polynomial ring $R[x]$ is not the null ring, either. The claim follows then immediately from (1),(2), and (3). □

33.7 Lemma: *Let R and S be two rings and let $\varphi: R \to S$ be a ring homomorphism. Then the mapping $\hat{\varphi}: R[x] \to S[x]$, defined by*

$$\Big(\sum_{i=0}^{m} a_i x^i\Big)\hat{\varphi} = \sum_{i=0}^{m} (a_i\varphi)x^i$$

is also a ring homomorphism. Furthermore, $Ker\ \hat{\varphi} = (Ker\ \varphi)[x]$ and $Im\ \hat{\varphi} = (Im\ \varphi)[x]$. (Note: $Ker\ \varphi$ and $Im\ \varphi$ are rings by Theorem 30.13, so $(Ker\ \varphi)[x]$ and $Im\ \hat{\varphi} = (Im\ \varphi)[x]$ are meaningful.)

Proof: Let $f = \sum_{i=0}^{m} a_i x^i$, $g = \sum_{j=0}^{n} b_j x^j$ be arbitrary polynomials in $R[x]$. We show that $\hat{\varphi}$ preserves addition. Here we may assume $m = n$, for we may add $0x^{m+1} + 0x^{m+2} + \cdots + 0x^n$ to f in case $m < n$ and $0x^{n+1} + 0x^{n+2} + \cdots + 0x^m$ to g in case $n < m$. We have

$$\begin{aligned}
(f+g)\hat{\varphi} &= \Big(\sum_{i=0}^{m} a_i x^i + \sum_{j=0}^{n} b_j x^j\Big)\hat{\varphi} \\
&= \Big(\sum_{i=0}^{m} a_i x^i + \sum_{i=0}^{m} b_i x^i\Big)\hat{\varphi} \\
&= \Big(\sum_{i=0}^{m} (a_i + b_i)x^i\Big)\hat{\varphi} \\
&= \sum_{i=0}^{m} [(a_i + b_i)\varphi]x^i \\
&= \sum_{i=0}^{m} (a_i\varphi + b_i\varphi)x^i \\
&= \sum_{i=0}^{m} (a_i\varphi)x^i + \sum_{i=0}^{m} (b_i\varphi)x^i \\
&= f\hat{\varphi} + g\hat{\varphi}
\end{aligned}$$

and so $\hat{\varphi}$ preserves addition. As for multiplication (here we do not have to assume $m = n$), we observe

$$
\begin{aligned}
(fg)\hat{\varphi} &= \Big[\sum_{k=0}^{m+n}\Big(\sum_{i+j=k} a_i b_j\Big)x^k\Big]\hat{\varphi} \\
&= \sum_{k=0}^{m+n}\Big[\Big(\sum_{i+j=k} a_i b_j\Big)\varphi\Big]x^k \\
&= \sum_{k=0}^{m+n}\Big(\sum_{i+j=k} (a_i b_j)\varphi\Big)x^k \\
&= \sum_{k=0}^{m+n}\Big(\sum_{i+j=k} a_i\varphi \cdot b_j\varphi\Big)x^k \\
&= \Big(\sum_{i=0}^{m}(a_i\varphi)x^i\Big)\Big(\sum_{j=0}^{n}(b_j\varphi)x^j\Big) \\
&= f\hat{\varphi}\cdot g\hat{\varphi}.
\end{aligned}
$$

Thus $\hat{\varphi}$ preserves multiplication as well. So $\hat{\varphi}$ is a ring homomorphism.

A polynomial $\sum_{i=0}^{m} a_i x^i$ belongs to the kernel of $\hat{\varphi}$ if and only if $\Big(\sum_{i=0}^{m} a_i x^i\Big)\hat{\varphi} = \sum_{i=0}^{m}(a_i\varphi)x^i$ is the zero polynomial in $S[x]$, so if and only if the coefficients $a_i\varphi$ are all equal to $0 \in S$ $(i = 0,1, \ldots,m)$, so if and only if $a_i \in Ker\ \varphi$ for all $i = 0,1, \ldots,m$, so if and only if $\sum_{i=0}^{m} a_i x^i \in (Ker\ \varphi)[x]$.

A polynomial $\sum_{i=0}^{m} c_i x^i \in S[x]$ belongs to the image of $\hat{\varphi}$ if and only if $\sum_{i=0}^{m} c_i x^i = \Big(\sum_{i=0}^{n} a_i x^i\Big)\hat{\varphi}$ for some $\sum_{i=0}^{n} a_i x^i \in R[x]$, so (assuming $m = n$ without loss of generality) if and only if, for each $i = 0,1, \ldots,m$, there is an $a_i \in R$ such that $c_i = a_i\varphi$, so if and only if $c_i \in Im\ \varphi$ for all $i = 0,1, \ldots,m$, and so if and only if $\sum_{i=0}^{m} c_i x^i \in (Im\ \varphi)[x]$. □

As an illustration of Lemma 33.7, we consider the natural homomorphism $\nu: \mathbb{Z} \to \mathbb{Z}_3$. Then the mapping $\hat{\nu}: \mathbb{Z}[x] \to \mathbb{Z}_3[x]$ is given by reducing the coefficients modulo 3. For example,

$$(5x^3 - 4x^2 + 2x + 1)\hat{v} = \bar{2}x^3 + \bar{2}x^2 + \bar{2}x + \bar{1}$$
$$(6x^4 - 3x^2 + x + 5)\hat{v} = \bar{1}x + \bar{2}.$$

The reader will easily verify that

$$(5x^3 - 4x^2 + 2x + 1)(6x^4 - 3x^2 + x + 5)$$
$$= 30x^7 - 24x^6 - 3x^5 + 23x^4 + 15x^3 - 21x^2 + 11x + 5,$$

whose image under $\hat{v}$ is

$$= \bar{30}x^7 - \bar{24}x^6 - \bar{3}x^5 + \bar{23}x^4 + \bar{15}x^3 - \bar{21}x^2 + \bar{11}x + \bar{5}$$
$$= \bar{2}x^4 + \bar{2}x + \bar{2}.$$

We have also $(\bar{2}x^3 + \bar{2}x^2 + \bar{2}x + \bar{1})(\bar{1}x + \bar{2}) = \bar{2}x^4 + \bar{2}x + \bar{2}$.

If $\varphi: R \to S$ is an isomorphism, then $Ker\ \varphi = 0$ and $Im\ \varphi = S$. This gives the following corollary to Lemma 33.7.

33.8 Theorem: *If R and S are isomorphic rings, then $R[x]$ and $S[x]$ are isomorphic.* □

Let R be a ring. Adjoining an indeterminate x to R, we get the ring $R[x]$. Now we can adjoin a new indeterminate y to $R[x]$ and get the ring $(R[x])[y] =: R[x][y]$. The elements of $R[x][y]$ are of the form $\sum_{i=0}^{m} f_i y^i$, where $f_i \in R[x]$. Similarly we can construct the ring $R[y][x] := (R[y])[x]$. We show that they are isomorphic.

33.9 Lemma: *Let R be a ring and let x,y be two indeterminates over R. Then $R[x][y] \cong R[y][x]$.*

Proof: We consider the mapping $T: R[x][y] \to R[y][x]$, given by

$$\sum_{i=0}^{m} \Big(\sum_{j=0}^{n} a_{ij} x^j \Big) y^i \longrightarrow \sum_{j=0}^{n} \Big(\sum_{i=0}^{m} a_{ij} y^i \Big) x^j,$$

which seems to be the only reasonable mapping from $R[x][y]$ to $R[y][x]$. It certainly preserves addition, for we have

$$\Big[\sum_{i=0}^{m} \Big(\sum_{j=0}^{n} a_{ij} x^j \Big) y^i + \sum_{i=0}^{r} \Big(\sum_{j=0}^{s} b_{ij} x^j \Big) y^i \Big] T$$

$$= \Big[\sum_{i=0}^{m}\Big(\sum_{j=0}^{n}a_{ij}x^j + \sum_{j=0}^{s}b_{ij}x^j\Big)y^i\Big]T \quad \text{(assuming } r = m \text{ without loss of generality)}$$

$$= \Big[\sum_{i=0}^{m}\Big(\sum_{j=0}^{n}(a_{ij} + b_{ij})x^j\Big)y^i\Big]T \quad \text{(assuming } s = n \text{ without loss of generality)}$$

$$= \sum_{j=0}^{n}\Big(\sum_{i=0}^{m}(a_{ij} + b_{ij})y^i\Big)x^j$$

$$= \sum_{j=0}^{n}\Big(\sum_{i=0}^{m}a_{ij}y^i + \sum_{i=0}^{m}b_{ij}y^i\Big)x^j$$

$$= \sum_{j=0}^{n}\Big(\sum_{i=0}^{m}a_{ij}y^i\Big)x^j + \sum_{j=0}^{n}\Big(\sum_{i=0}^{m}b_{ij}y^i\Big)x^j$$

$$= \Big[\sum_{i=0}^{m}\Big(\sum_{j=0}^{n}a_{ij}x^j\Big)y^i\Big]T + \Big[\sum_{i=0}^{m}\Big(\sum_{j=0}^{n}b_{ij}x^j\Big)y^i\Big]T$$

for all $\sum_{i=0}^{m}\Big(\sum_{j=0}^{n}a_{ij}x^j\Big)y^i$, $\sum_{i=0}^{r}\Big(\sum_{j=0}^{s}b_{ij}x^j\Big)y^i \in R[x][y]$.

Secondly T preserves multiplication of polynomials of the form $(ax^j)y^i$ (i.e., monomials over $R[x]$, whose eventually nonzero coefficients in $R[x]$ are themselves monomials over R; they will be referred to as *monomials in $R[x][y]$ over R*). We indeed have

$$\begin{aligned}
[(a_{ij}x^j)y^i \cdot (b_{rs}x^s)y^r]T &= [(a_{ij}x^j)(b_{rs}x^s)y^{i+r}]T \quad \text{(def. of multiplication in } R[x][y]) \\
&= [(a_{ij}b_{rs}x^{j+s})y^{i+r}]T \quad \text{(def. of multiplication in } R[x]) \\
&= (a_{ij}b_{rs}y^{i+r})x^{j+s} \\
&= [(a_{ij}y^i)(b_{rs}y^r)]x^{j+s} \\
&= (a_{ij}y^i)x^j \cdot (b_{rs}y^r)x^s \\
&= [(a_{ij}x^j)y^i]T \cdot [(b_{rs}x^s)y^r]T
\end{aligned}$$

for all monomials $(a_{ij}x^j)y^i$, $(b_{rs}x^s)y^r$ in $R[x][y]$.

Thirdly, T preserves multiplication of arbitrary polynomials. Any polynomial can be written as $p_1 + p_2 + \cdots + p_t$, where $p_1, p_2, \ldots, p_t$ are suitable monomials. Now for all polynomials $p_1 + p_2 + \cdots + p_t$, $q_1 + q_2 + \cdots + q_u$ in $R[x][y]$, where p's and q's are monomials, we have

$$[(p_1 + p_2 + \cdots + p_t)(q_1 + q_2 + \cdots + q_u)]T$$

$$= \Big(\sum_{i,j} p_i q_j\Big)T \qquad \text{(by distributivity)}$$

$$= \sum_{i,j} (p_i q_j)T \qquad \text{(since } T \text{ preserves addition)}$$

$$= \sum_{i,j} p_i T \cdot q_j T \qquad \text{(since } T \text{ preserves products of monomials)}$$

$$= (p_1 T + p_2 T + \cdots + p_t T)(q_1 T + q_2 T + \cdots + q_u T)$$

$$= (p_1 + p_2 + \cdots + p_t)T \cdot (q_1 + q_2 + \cdots + q_u)T,$$

so T preserves arbitrary products. Hence T is a ring homomorphism.

T is one-to-one, for if $\sum_{i=0}^{m} \Big(\sum_{j=0}^{n} a_{ij} x^j \Big) y^i \in R[x][y]$ is in the kernel of T, then its image $\sum_{j=0}^{n} \Big(\sum_{i=0}^{m} a_{ij} y^i \Big) x^j$ is the zero polynomial in $R[y][x]$, so all the coefficients $\sum_{i=0}^{m} a_{ij} y^i$ are equal to the zero polynomial in $R[y]$, so all elements a_{ij} of R are equal to the zero element in R, so all polynomials $\sum_{j=0}^{n} a_{ij} x^j$ are the zero polynomial in $R[x]$, so $\sum_{i=0}^{m} \Big(\sum_{j=0}^{n} a_{ij} x^j \Big) y^i$ is the zero polynomial in $R[x][y]$. Thus $Ker\, T$ consists of the zero polynomial and T is one-to-one.

Moreover, T is onto, for any polynomial $\sum_{j=0}^{n} \Big(\sum_{i=0}^{m} a_{ij} y^i \Big) x^j$ in $R[y][x]$ is the image of the polynomial $\sum_{i=0}^{m} \Big(\sum_{j=0}^{n} a_{ij} x^j \Big) y^i$ in $R[x][y]$.

Hence T is an isomorphism from $R[x][y]$ onto $R[y][x]$. □

In view of this result, we identify $R[x][y]$ and $R[y][x]$. To simplify the notation, we write $R[x,y]$ for $R[x][y]$. The elements of $R[x,y]$ are of the

form $\sum_{i,j} a_{ij}x^iy^j$, where $a_{ij} \in R$ and there are finitely many terms in the sum. Multiplication is carried out in the customary way, using distributivity and collecting terms. We have $R[x,y] = R[y,x]$.

We can of course adjoin a new inteterminate z to $R[x,y]$ and obtain $(R[x,y])[z] = (R[x][y])[z] =: R[x][y][z]$. We see

$$\begin{aligned}
R[x][y][z] &= (R[x][y])[z] && \text{(definition)}\\
&\cong (R[x][z])[y] && \text{(Lemma 33.9 with } R[x], z \text{ in place of } R, x)\\
&\cong (R[z][x])[y] && \text{(Lemma 33.9 and Theorem 33.8)}\\
&\cong (R[z][y])[x] && \text{(Lemma 33.9 with } R[z] \text{ in place of } R)\\
&\cong (R[y][z])[x]\\
&\cong (R[y][x])[z].
\end{aligned}$$

We regard these six rings as identical and write $R[x,y,z]$ for it. The notations "$R[x,y,z]$", "$R[x,z,y]$", "$R[z,x,y]$", "$R[z,y,x]$", "$R[y,z,x]$", "$R[y,x,z]$" will mean the same ring.

More generally, if $x_1,x_2, \ldots ,x_n$ are indeterminates over a ring R, then $R[x_1,x_2, \ldots ,x_n]$ is defined to be the ring $R[x_1,x_2, \ldots ,x_{n-1}][x_n]$. It is isomorphic to each one of the $n!$ rings $R[x_{i_1},x_{i_2}, \ldots ,x_{i_n}]$, where $\begin{pmatrix} 1 & 2 & \ldots & n \\ i_1 & i_2 & \ldots & i_n \end{pmatrix}$ runs through the permutations in S_n. These $n!$ isomorphic rings will be considered identical. Elements of $R[x_1,x_2, \ldots ,x_n]$ are of the form

$$\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}\cdots\sum_{l=0}^{N_n} a_{ij\ldots l}x_1^{\,i}x_2^{\,j}\ldots x_n^{\,l}, \qquad a_{ij\ldots l} \in R.$$

The polynomials in $R[x_1,x_2, \ldots ,x_n]$ of the form $ax_1^{\,i}x_2^{\,j}\ldots x_n^{\,l}$ will be called *monomials over* R. It is customary to omit the indeterminates with exponent zero in a monomial. For example, $ax_1^{\,0}x_2^{\,2}x_3^{\,0}x_4^{\,3}$ in $R[x_1,x_2,x_3,x_4]$ is written $ax_2^{\,2}x_4^{\,3}$. An exponent is dropped when it is equal to 1. If R does not have an identity, the indeterminates $x_1,x_2, \ldots ,x_n$ are *not* elements of $R[x_1,x_2, \ldots ,x_n]$ and the expressions $x_1^{\,i}x_2^{\,j}\ldots x_n^{\,l}$ are *not* polynomials.

The *degree* of a nonzero monomial $ax_1^{\,i}x_2^{\,j}\ldots x_n^{\,l}$ is defined to be the nonnegative integer $i + j + \ldots + l$. The *total degree* of a polynomial $f =$

$\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}\cdots\sum_{l=0}^{N_n} a_{ij\ldots l}x_1^{\,i}x_2^{\,j}\ldots x_n^{\,l}$ is defined to be the maximum of the degrees of the monomials $a_{ij\ldots l}x_1^{\,i}x_2^{\,j}\ldots x_n^{\,l}$ with $a_{ij\ldots l}\neq 0$. The total degree of f will be denoted by *deg f*. The degree of f, considered as an element of $R[x_1,\ldots,x_{h-1},x_{h+1},\ldots,x_n][x_h]$ will be called the *degree of f in* x_h; this will be written $deg_h f$ $(h = 1,2,\ldots,n)$. The analog of Lemma 33.3 holds for polynomials in n indeterminates, both with the total degree and the degree in x_h in place of *deg f*.

We record a lemma that can be proved by induction on the number of indeterminates.

33.10 Lemma: *Let R be a ring and $x_1,x_2,\ldots,x_n$ indeterminates over R.*
(1) *If R is commutative, then $R[x_1,x_2,\ldots,x_n]$ is commutative.*
(2) *If R has an identity, then $R[x_1,x_2,\ldots,x_n]$ has an identity.*
(3) *If R has no zero divisors, then $R[x_1,x_2,\ldots,x_n]$ has no zero divisors.*
(4) *If R is an integral domain, then $R[x_1,x_2,\ldots,x_n]$ is an integral domain.*

Exercises

1. Evaluate: $(5x^2 - 3x + 1)(7x^3 + 6x - 1)$ in $\mathbb{Z}_8[x]$,

$(3x^3 + 4x + 1)(3x^2 + 7x + 2)$ in $\mathbb{Z}_9[x]$,

$\left[\begin{pmatrix}0&1\\1&0\end{pmatrix}x^4 + \begin{pmatrix}1&2\\0&0\end{pmatrix}x^2 + \begin{pmatrix}2&0\\1&1\end{pmatrix}\right]\left[\begin{pmatrix}1&0\\0&0\end{pmatrix}x^2 - \begin{pmatrix}1&1\\0&0\end{pmatrix}x + \begin{pmatrix}1&1\\1&1\end{pmatrix}\right]$ in $(Mat_2(\mathbb{Z}))[x]$,

$\left[\begin{pmatrix}0&1\\1&0\end{pmatrix}x^3 + \begin{pmatrix}1&3\\0&2\end{pmatrix}x^2 + \begin{pmatrix}1&0\\1&0\end{pmatrix}\right]\left[\begin{pmatrix}2&1\\0&3\end{pmatrix}x^2 + \begin{pmatrix}1&5\\2&0\end{pmatrix}x + \begin{pmatrix}2&4\\1&0\end{pmatrix}\right]$ in $(Mat_2(\mathbb{Z}_7))[x]$

(we dropped the bars for ease of notation).

2. Let R, R_1, R_2 be rings. Prove that

$(Mat_2(R))[x] \cong Mat_2R[x]$ and $(R_1 \oplus R_2)[x] \cong R_1[x] \oplus R_2[x]$

(see §29, Ex. 10).

3. Generalize Lemma 33.7 to polynomial rings in n indeterminates.

4. Let R be a commutative ring with identity and let $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a zero divisor in $R[x]$. Show that there exists a nonzero b in R such that $b\,a_n = ba_{n-1} = \cdots = b\,a_0 = 0$.

5. Let R be a ring and $f = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \cdots \sum_{l=0}^{N_n} a_{ij\ldots l} x_1^{\,i} x_2^{\,j} \ldots x_n^{\,l} \in R[x_1, x_2, \ldots, x_n]$. Prove that $deg_1 f$ is the largest i such that $a_{ij\ldots l} \neq 0$, $deg_2 f$ is the largest j such that $a_{ij\ldots l} \neq 0$, ..., $deg_n f$ is the largest l such that $a_{ij\ldots l} \neq 0$.

6. Extend Lemma 33.3 to polynomial rings in n indeterminates, both with total degree and the degree in x_h in place of the degree of f.

§34
Divisibility in Polynomial Domains

We learned in Lemma 33.6 that some properties of a ring R are transferred to the polynomial ring $R[x]$. In particular, if R is an integral domain, so is $R[x]$. In any integral domain, we have a theory of divisibility (§32). In this paragraph, we want to investigate the divisibility properties of polynomials. Lemma 33.6 suggests the questions: Is $R[x]$ a Euclidean domain if R is a Euclidean domain? Is $R[x]$ a principal ideal domain if R is a principal ideal domain? Is $R[x]$ a unique factorization domain if R is a unique factorization domain? The answer to the first two questions is 'no'. For example, $\mathbb{Z}[x]$ is not a principal ideal domain, let alone a Euclidean domain, although $\mathbb{Z}$ is Euclidean. On the other hand, the third question recieves an affirmative answer: if R is a unique factorization domain, so is $R[x]$. This will be proved as Theorem 34.13.

Let us recollect the basic definitions. Assume D is an integral domain. Then $D[x]$ is an integral domain (Lemma 33.6). A polynomial $f \in D[x]$ is said to be *divisible* by a nonzero polynomial $g \in D[x]$ if there is a polynomial h in $D[x]$ such that $f = gh$. We write then $g|f$. Notice that the coefficients of h are required to be in D. The notation $g|f$ does not merely mean that $f = gh$ for some arbitrary polynomial h. It means $f = gh$ for some polynomial h in $D[x]$.

When $f \neq 0$ and $f = gh$, we have $\deg f = \deg gh = \deg g + \deg h \geqslant \deg g$:

34.1 Lemma: *Let D be an integral domain. If $g, f \in D[x]$, $g \neq 0 \neq f$ and $g|f$, then,* $\deg g \leqslant \deg f$. □

A nonzero polynomial $e \in D[x]$ is a *unit of* $D[x]$ if $eh = 1$ for some $h \in D[x]$, or, equivalently, if $e|f$ for all $f \in D[x]$. In this case, Lemma 33.3 yields

$$0 = \deg 1 = \deg eh = \deg e + \deg h \geqslant 0 + 0 = 0,$$
$$\deg e = 0, \deg h = 0,$$

$$e \in D, \qquad h \in D,$$
$$eh = 1 \text{ holds in } D,$$
$$e \text{ is a unit in } D.$$

($e \neq 0 \neq h$, because $eh = 1 \neq 0$ and D is an integral domain.) So a unit in $D[x]$ is a unit in D: if a polynomial $e = \sum_{i=0}^{m} a_i x^i$ is a unit in $D[x]$, then $a_0 \in D$ is a unit in D and $a_1 = a_2 = \cdots = a_m = 0$. Conversely, if e is a unit in D so that $eh = 1$ for some $h \in D$, then of course $e,h \in D[x]$ and e is a unit in $D[x]$. We proved the following lemma.

34.2 **Lemma:** *Let D be an integral domain. Then $e \in D[x]$ is a unit in $D[x]$ if and only if $e \in D$ and e is a unit in D. In symbols, $D[x]^* = D^*$.* □

Thus any unit in $D[x]$ has degree 0 and the associates of a polynomial in $D[x]$ have the same degrees as the polynomial itself. Any proper divisor of $f \in D[x]$ is therefore of degree distinct from 0 and *deg f*.

A polynomial f in $D[x]\backslash\{0\}$ is irreducible if f is not a unit in $D[x]$ and if, in any factorization of f as $f = gh$ in $D[x]$, either g or h is a unit. This is Definition 32.7. We paraphrase this as follows: $f \in D[x]\backslash\{0\}$ is irreducible if $deg\ f > 0$ and if there are no polynomials g,h in $D[x]$ such that $f = gh$ and $0 < deg\ g, deg\ h < deg\ f$. The phrase "in $D[x]$" is important. Suppose $D \subseteq D_1$, where D_1 is another integral domain. Then $f \in D_1[x]$, too. Now it is possible that

there exist no $g,h \in D[x]$ such that $f = gh$, $0 < deg\ g < deg\ f$

and yet possibly

there exist some $g,h \in D_1[x]$ such that $f = gh$, $0 < deg\ g < deg\ f$.

Then f is irreducible in $D[x]$, but not in $D_1[x]$. This shows that irreducibility of f is not an intrinsic property of f. It is a property of f *relative to* the polynomial domain $D[x]$. For this reason, we have to mention the domain D whenever we speak about irreducible polynomials. We say f *is irreducible over* D when f is irreducible in $D[x]$. For example, $x^2 + 1 \in \mathbb{O}[x]$

is irreducible over $\mathbb{Q}$ since $x^2 + 1$ has no proper divisors in $\mathbb{Q}[x]$, but $x^2 + 1$ is reducible in $\mathbb{C}[x]$ since $x^2 + 1 = (x - i)(x + i)$, with $x - i, x + i \in \mathbb{C}[x]$ and $0 < 1 = deg(x - i) < 2 = deg\ (x^2 + 1)$.

We now compare the irreducibility of an element of D in D with its irreducibility in $D[x]$.

34.3 Lemma: *Let D be an integral domain and let a be any nonzero element of $D \subseteq D[x]$. Then a is irreducible in $D[x]$ if and only if a is irreducible in D.*

Proof: Suppose that a is irreducible in D. We prove that a is irreducible in $D[x]$. First we must show that a is not a unit in $D[x]$. Since a is irreducible in D, so not a unit in D, we have $a \notin D^\times = D[x]^\times$ (Lemma 34.2), so a is not a unit in $D[x]$. Secondly we must show that $a = bc$, where $b,c \in D[x]$, implies either b or c is a unit in $D[x]$. Indeed, if $a = bc$, then $0 = deg\ a = deg\ bc = deg\ b + deg\ c \geqslant 0$, so $deg\ b = 0 = deg\ c$. Then $a = bc$ is an equation in D. Since a is irreducible in D, either b or c is a unit in D, so, in view of Lemma 34.2, either b or c is a unit in $D[x]$. This proves that a is irreducible in $D[x]$.

Now the converse. We suppose that a is irreducible in $D[x]$ and show that a is irreducible in D. First we must show that a is not a unit in D. Since a is irreducible in $D[x]$, so not a unit in $D[x]$, we have $a \notin D[x]^\times = D^\times$ (Lemma 34.2), so a is not a unit in D. Secondly we must show that $a = bc$, where $b,c \in D$, implies either b or c is a unit in D. We read $a = bc$ as an equation in $D[x]$. Since a is irreducible in $D[x]$, either b or c is a unit in $D[x]$, so, in view of Lemma 34.2, either b or c is a unit in D. This proves that a is irreducible in D. □

We want to find the integral domains D such that $D[x]$ is a unique factorization domain. What conditions must be imposed on D? If $D[x]$ is to be a unique factorization domain, then each element of $D[x]\backslash\{0\}$ that is not a unit in $D[x]$, must be written as a product of irreducible elements of $D[x]$ in a unique way. In particular, each element of $D\backslash\{0\}$ that is not a unit in $D[x]$, must be written as a product of irreducible elements of $D[x]$ in a unique way. As any divisor in $D[x]$ of an element in D belongs to D

by degree considerations, the last statement means (Lemma 34.2, Lemma 34.3): each element of $D\backslash\{0\}$ that is not a unit in D, must be written as a product of irreducible elements of D in a unique way. Thus D must be a unique factorization domain. We shall prove conversely that $D[x]$ is a unique factorization domain whenever D is. The proof will make use of the polynomial ring $F[x]$, where F is the field of fractions of D (§31). $F[x]$ will turn out to be a Euclidean domain.

We show generally that $K[x]$ is a Euclidean domain if K is a field. In order to do that, let us remember, we must find a function $d:K[x]\backslash\{0\} \to \mathbb{N} \cup \{0\}$ such that $d(f) \leq d(fg)$ for all $f,g \in K[x]\backslash\{0\}$ and such that, for any nonzero polynomials f,g in $K[x]$, there are polynomials $q,r \in K[x]$ with $f = qg + r$ and $r = 0$ or $\deg r < \deg g$. The degree of polynomials will work as the function d. First we prove a slightly more general theorem.

34.4 **Theorem (Division algorithm):** *Let D be an integral domain, and let f,g be polynomials in $D[x]$. If the leading coefficient of g is a unit in D, then there are unique polynomials q,r in $D[x]$ such that*

$$f = qg + r, \qquad r = 0 \text{ or } \deg r < \deg g.$$

Proof: First we prove the existence of q and r. This is nothing but the long division of polynomials. Suppose we divide $f = x^5 - 2x^4 + 3x^3 + x^2 - x + 2$ by $g = x^2 + x + 1$. What do we do? We subtract x^3 times g from f:

$$\begin{array}{ll} x^5 - 2x^4 + 3x^3 + x^2 - x + 2 & \underline{x^2 + x + 1} \\ \underline{x^5 + x^4 + x^3} & x^3 \\ -3x^4 + 2x^3 + x^2 - x + 2 & \end{array}$$

and get the polynomial $f_1 = -3x^4 + 2x^3 + x^2 - x + 2$, whose degree is smaller than the degree of f. Then we subtract $-3x^2$ times g from f_1 and get a polynomial $f_2 = 5x^3 + 4x^2 - x + 2$, whose degree is smaller than the degree of f_1. We continue this process until we get a polynomial r whose degree is smaller than the degree of $g = x^2 + x + 1$:

$$
\begin{array}{rl|l}
x^5 - 2x^4 + 3x^3 + x^2 - x + 2 & & x^2 + x + 1 \\
\underline{x^5 + x^4 + x^3} & & \overline{x^3 - 3x^2 + 5x - 1} \\
-3x^4 + 2x^3 + x^2 - x + 2 & & \\
\underline{-3x^4 - 3x^3 - 3x^2} & & \\
5x^3 + 4x^2 - x + 2 & & \\
\underline{5x^3 + 5x^2 + 5x} & & \\
-x^2 - 6x + 2 & & \\
\underline{-x^2 - x - 1} & & \\
-5x + 3. & &
\end{array}
$$

Hence $f = (x^3 - 3x^2 + 5x - 1)g + (-5x + 3)$. In general, we have

$$
\begin{array}{rl|l}
f & & g \\
\underline{ax^m g} & & \overline{ax^m} \\
f_1 = f - ax^m g & &
\end{array}
$$

where a and $m \in \mathbb{N} \cup \{0\}$ are chosen appropriately, and $deg\ f_1 < deg\ f$. Then, by induction on the degree of f, we can divide f_1 (and hence f) by g and get a remainder r. This is essentially the proof.

Now let f,g be nonzero polynomials in $D[x]$ and suppose that the leading coefficient of g is a unit in D. We prove the existence of q and r by induction on $deg\ f$.

I. Induction begins at 0. Suppose $deg\ f = 0$. Then $f \in D\backslash\{0\}$. Since the leading coefficient of g is a unit in D by hypothesis, if $g \in D$, there is a $g^{-1} \in D$ such that $g^{-1}g = 1$, hence $fg^{-1} \in D$ and we can write

$$f = (fg^{-1})g + 0$$

If $g \in D[x]\backslash D$, then $deg\ g \geqslant 1$ and we can write

$$f = 0g + f.$$

This proves the existence of q and r with

$$q = fg^{-1}, \quad r = 0 \qquad \text{in case } g \in D,$$
$$q = 0, \quad r = f \qquad \text{in case } g \in D[x]\backslash D.$$

II. Now the inductive step. We use the principle of induction in the form 4.5. We assume that $deg\ f = n \geqslant 1$ and that, for any nonzero

polynomial h with $deg\ h < n$, there are polynomials q_1 and r_1 in $D[x]$ such that

$$h = q_1 g + r_1, \qquad r_1 = 0 \text{ or } deg\ r_1 < deg\ g.$$

In case $deg\ g > n$, we have

$$f = 0g + f \qquad deg\ f = n < deg\ g$$

and this proves the existence of q and r with

$$q = 0, \qquad r = f.$$

Having disposed of the case $deg\ g > n$, we assume now $deg\ g \leqslant n$. We subtract a suitable multiple of g from f to get a polynomial of degree smaller than n. If, say

$$f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0, \quad g = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_0,$$

b_m is a unit in D,

$b_m b_m^{-1} = 1$ for some $b_m^{-1} \in D$,

$m \leqslant n$,

then we put $f_1 := f - a_n b_m^{-1} x^{n-m} g$. Here either $f_1 = 0$ and the the existence of q and r is proved with $q = a_n b_m^{-1} x^{n-m}$, $r = 0$; or

$$f_1 = f - a_n b_m^{-1} x^{n-m} g$$
$$= (a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0) - a_n b_m^{-1} x^{n-m}(b_m x^m + b_{m-1}x^{m-1} + \cdots + b_0)$$

is a polynomial in $D[x]$ of degree $< n$. By the induction hypothesis, there are polynomials q_1, r_1 in $D[x]$ such that

$$f_1 = q_1 g + r_1, \qquad r_1 = 0 \text{ or } deg\ r_1 < deg\ g.$$

Hence

$$f = f_1 + a_n b_m^{-1} x^{n-m} g$$
$$= (q_1 g + r_1) + (a_n b_m^{-1} x^{n-m} g)$$
$$= (q_1 + a_n b_m^{-1} x^{n-m})g + r_1, \qquad r_1 = 0 \text{ or } deg\ r_1 < deg\ g$$

and this proves the existence of q and r with $q = q_1 + a_n b_m^{-1} x^{n-m}$, $r = r_1$ and completes the proof of the inductive step. The hypothesis that the leading coefficient of g be a unit has been used to construct the f_1 with $deg\ f_1 < deg\ f$.

The uniqueness of q and r. Suppose

$$f = qg + r = q'g + r'; \quad r = 0 \text{ or } deg\ r < deg\ g; \quad r' = 0 \text{ or } deg\ r' < deg\ g.$$

Then

$$(qg + r) - (q'g + r') = f - f = 0,$$
$$(q - q')g = r' - r$$

and the assumption $q - q' \neq 0$ leads to the contradiction

$$deg\ g \leqslant deg\ (q - q') + deg\ g = deg\ (q - q')g$$

$$= \deg (r' - r) \leq \max\{\deg r', \deg r\} < \deg g$$

by Lemma 33.3. This forces $q - q' = 0$, so $q = q'$, so $r = f - qg = f - q'g = r$. Thus q and r are uniquely determined. □

34.5 Theorem: *Let K be a field.*
(1) *For any nonzero polynomials f,g in $K[x]$, there are unique polynomials q and r in $K[x]$ such that*

$$f = qg + r, \qquad r = 0 \text{ or } \deg r < \deg g.$$

(2) *$K[x]$ is a Euclidean domain.*
(3) *$K[x]$ is a unique factorization domain.*

Proof: (1) Since $g \neq 0$, it has a leading coefficient, which is distinct from $0 \in K$. Then the leading coefficient of g is a unit in K (Example 32.6(b)). The assertion follows now from Theorem 34.4.

(2) We prove that $\deg: K[x]\backslash\{0\} \to \mathbb{N} \cup \{0\}$ satisfies the conditions in Definition 32.10. Certainly $\deg f$ is a nonnegative integer by definition and $\deg f \leq \deg fg$ for all $f,g \in K[x]\backslash\{0\}$ by Lemma 33.3. This proves the condition (i) in Definition 32.13. The condition (ii) is proved in part (1).

(3) This follows from Theorem 32.22. □

We record some consequences of Theorem 34.5.

34.6 Theorem: *Let K be a field. Any two polynomials f,g in $K[x]$, not both zero, have a greatest common divisor d in $K[x]$. If d is a greatest common divisor of f and g, then there are polynomials h and l in $K[x]$ such that $d = hf + lg$. Any two greatest common divisors of f and g are associate. In particular, there is one and only one monic greatest common divisor of f and g.* (This unique monic greatest common divisor of f and g is sometimes called *the* greatest common divisor of f and g). *Any irreducible polynomial in $K[x]$ is prime in $K[x]$ (Definition* 32.20). □

Theorem 34.5 is very satisfactory. If the underlying ring is a field, then the polynomial domain is a unique factorization domain. We turn our attention to polynomials with coefficients in a unique factorization domain. Let D be a unique factorization domain and let F be the field of fractions of D. We recall that the elements of F are fractions a/b of element $a,b \in D, b \neq 0$. We identify $a \in D$ with $a/1 \in F$ and thus regard D as a subring of F. In this way, $D[x] \subseteq F[x]$. (If you find this and the following discussion too abstract, you may just assume $D = \mathbb{Z}$ and $F = \mathbb{Q}$.)

Let $f \in D[x] \subseteq F[x]$. Now, a priori, f may be irreducible over D and not irreducible over F. See the comments preceding Lemma 34.3. In the case where D is a unique factorization domain and F is the field of fractions of D, it is in fact true that an irreducible polynomial in $D[x]$ is also irreducible in $F[x]$. After some preparation, this will be proved in Lemma 34.11. The hypothesis that D be a unique factorization domain is essential, for otherwise the following definition, which plays an important role in the proof of Lemma 34.11, does not make sense.

34.7 Definition: Let D be a unique factorization domain and let f be any nonzero polynomial in $D[x]$. A greatest common divisor of the coefficients of f is called a *content of* f.

Since greatest common divisors are uniquely determined to within ambiguity among associate elements, any two contents of f are associate. We write $C(f)$ for any content of f. Ignoring the distinction among associate elements, we sometimes call $C(f)$ *the* content of f by abuse of language.

The contents of $f = 2x^4 - 8x^2 + 2x + 6 \in \mathbb{Z}[x]$ and $g = 6x^2 - 9x + 18 \in \mathbb{Z}[x]$ are easily seen to be $C(f) = 2$ and $C(g) = 3$. The content of $fg =$ $12x^6 - 18x^5 - 12x^4 + 84x^3 - 126x^2 - 18x + 108$ is $C(fg) = 6 = 2 \cdot 3 = C(f)C(g)$. This is an example of a general phenomenon.

34.8 Lemma (Gauss' lemma): *Let D be a unique factorization domain and let f,g be arbitrary nonzero polynomials in $D[x]$. Then $C(fg) \approx C(f)C(g)$.*

Proof: First we remark that we cannot write $C(fg) = C(f)C(g)$, for contents are unique only up to associate elements.

f and g can be written as $f = C(f)f_1$ and $g = C(g)g_1$, where f_1 and g_1 are polynomials in $D[x]$ with $C(f_1) \approx 1$ and $C(g_1) \approx 1$. Similarly $fg = C(fg)h$, where $h \in D[x]$ and $C(h) \approx 1$. We have

$$C(f)f_1 C(g)g_1 = fg = C(fg)h$$
$$C(f)C(g)f_1 g_1 = C(fg)h.$$

Taking contents of both sides and observing $C(al) \approx aC(l)$ for $a \in D\backslash\{0\}$ and $l \in D[x]\backslash\{0\}$, we obtain

$$C(f)C(g)C(f_1 g_1) \approx C(fg)C(h)$$
$$C(f)C(g)C(f_1 g_1) \approx C(fg)$$

and the theorem will be proved if we can show $C(f_1 g_1) \approx 1$. Dropping the subscripts, we must prove:

$$\text{if } C(f) \approx 1 \text{ and } C(g) \approx 1, \text{ then } C(fg) \approx 1.$$

Suppose now $C(f) \approx 1$, $C(g) \approx 1$ and $C(fg)$ is not a unit. Then there is an *irreducible* element π in D with $\pi | C(fg)$. Since $C(f) \approx 1$ and $C(g) \approx 1$ by assumption, π cannot divide all the coefficients of

$$f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

nor of

$$g = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0,$$

say. Let a_h be the coefficient of f with the largest index that is not divisible by π and let b_k have a similar meaning for g. Then

$$\pi | a_n, \pi | a_{n-1}, \ldots, \pi | a_{h+1}, \pi \nmid a_h \tag{1}$$
$$\pi | b_m, \pi | b_{m-1}, \ldots, \pi | b_{k+1}, \pi \nmid b_k. \tag{2}$$

But π divides the coefficient

$$(\cdots + a_{h+2}b_{k-2} + a_{h+1}b_{k-1}) + a_h b_k + [a_{h-1}b_{k+1} + a_{h-2}b_{k+2} + \cdots]$$

of x^{h+k} in fg. Because of (1) and (2), π divides the expressions in ( ) and [ ]. So π divides $a_h b_k$ as well. Thus $\pi \nmid a_h$, $\pi \nmid b_m$ and $\pi | a_h b_k$, which tells us that π is not a prime element in D. On the other hand, D is a unique factorization domain and every irreducible element in D is prime (Lemma 32.24), hence π is prime. This is a contradiction. We conclude $C(fg) \approx 1$. □

34.9 Lemma: *Let D be a unique factorization domain and let F be the field of fractions of D. Let f,g be any nonzero polynomials in $D[x]$ with*

$C(f) \approx C(g)$. Then f and g are associate in $F[x]$ if and only if f and g are associate in $D[x]$.

Proof: By Lemma 34.2,

e is a unit in $D[x]$ $\Leftrightarrow$ e is a unit in D,

u is a unit in $F[x]$ $\Leftrightarrow$ u is a unit in F $\Leftrightarrow$ $u \in F\backslash\{0\}$.

If f and g are associate in $D[x]$, then $f = eg$ for some unit e in $D[x]$. Then e is a unit in D, so e is a nonzero element of D, so e is a nonzero element of F, so e is a unit in F, so e is a unit in $F[x]$, so f and g are associate in $F[x]$.

If f and g are associate in $F[x]$, then $f = ug$ for some unit u in $F[x]$. Thus $u \in F\backslash\{0\}$ and so $u = a/b$, where $a,b \in D\backslash\{0\}$. So $bf = ag$. Thus

$$bC(f) \approx C(bf) \approx C(ag) \approx aC(g) \approx aC(f)$$

and $b \approx a$ in D. So $a/b = u$ is a unit in D. Hence u is a unit in $D[x]$ and f is associate to g in $D[x]$. □

34.10 Lemma: *Let D be a unique factorization domain and let F be the field of fractions of D. Let f be a nonzero polynomial in $D[x]$ with $C(f) \approx 1$ and assume*

$$f = g_1 g_2 \cdots g_r,$$

where $g_1, g_2, \ldots, g_r$ are polynomials in $F[x]$. Then there are polynomials $h_1, h_2, \ldots, h_r$ in $D[x]$ such that g_i is associate to h_i in $F[x]$ and $C(h_i) \approx 1$ (for all $i = 1,2,\ldots,r$) and

$$f = h_1 h_2 \cdots h_r.$$

Proof: The coefficients of $g_1, g_2, \ldots, g_r$ are fractions of elements from D. We multiply each g_i by an appropriate element a_i in D, for example by the product of the "denominators" in the coefficients of g_i to get a polynomial $k_i \in D[x]$. Thus $a_i g_i = k_i \in D[x]$. We write $k_i = c_i h_i$, where $c_i \approx C(k_i) \in D$ and h_i is a polynomial in $D[x]$ with $C(h_i) \approx 1$. We have

$$a_1 a_2 \cdots a_r f = a_1 g_1 \cdot a_2 g_2 \cdots a_r g_r = k_1 k_2 \cdots k_r = c_1 c_2 \cdots c_r h_1 h_2 \cdots h_r$$

and, taking contents of both sides, and using Lemma 34.8 $r - 1$ times, we get

$$a_1 a_2 \cdots a_r C(f) = c_1 c_2 \cdots c_r \, C(h_1) C(h_2) \cdots C(h_r)$$

$$a_1 a_2 \cdots a_r \approx c_1 c_2 \cdots c_r.$$

Thus $e := c_1 c_2 \cdots c_r / a_1 a_2 \cdots a_r$ is a unit in D and

$$f = (eh_1)h_2 \ldots h_r.$$

Observe that $h_i = (a_i/c_i)g_i$ is associate to g_i in $F[x]$, because $a_i/c_i \in F\backslash\{0\}$ is a unit in $F[x]$. When we make a slight change of notation and write h_1 for eh_1, the proof is complete (eh_1 is also associate to g_1 in $F[x]$). □

34.11 Lemma: *Let D be a unique factorization domain and let F be the field of fractions of D. Let f be a nonzero polynomial in $D[x]$ with $C(f) \approx 1$. Then f is irreducible in $F[x]$ if and only if f is irreducible in $D[x]$.*

Proof: Assume first that f is irreducible in $F[x]$. Then f is not a unit in $F[x]$, hence $\deg f \geq 1$, hence f is not a unit in $D[x]$. Also, if $g,h \in D[x]$ and $f = gh$, we read this equation in $F[x]$ and conclude that either g or h is associate to f in $F[x]$. We know $1 \approx C(f) \approx C(gh) \approx C(g)C(h)$, so $C(g) \approx 1 \approx C(f)$ and $C(h) \approx 1 \approx C(f)$. Using Lemma 34.9, we deduce that either g or h is associate to f in $D[x]$. Thus f is not a unit in $D[x]$ and has no proper divisors in $D[x]$. This means f is irreducible in $D[x]$.

Conversely, assume that f is irreducible in $D[x]$. Then f is not a unit in $D[x]$ and so not a unit in D. This gives $\deg f \geq 1$, for otherwise $f \approx C(f) \approx 1$ would be a unit in D. So $\deg f \geq 1$ and f is not a unit in $F[x]$. We now want to show that f has no proper divisors in $F[x]$. Assume $f = g_1g_2$, where $g_1,g_2 \in F[x]$. By Lemma 34.10, $f = h_1h_2$, where $h_1,h_2 \in D[x]$, $C(h_1) \approx 1 \approx C(f)$, $C(h_2) \approx 1 \approx C(f)$ and g_1,g_2 are respectively associate to h_1,h_2 in $F[x]$. Since f is irreducible in $D[x]$, either h_1 or h_2 is associate to f in $D[x]$ and thus, by Lemma 34.9, either h_1 or h_2 is associate to f in $F[x]$, hence either g_1 or g_2 is associate to f in $F[x]$. Thus f has no proper divisors in $F[x]$ and f is irreducible in $F[x]$. □

We need one more lemma to prove that $D[x]$ is a unique factorization domain whenever D is. It comprises the main argument.

34.12 Lemma: *Let D be a unique factorization domain and let f be a nonzero polynomial in $D[x]$ such that $C(f) \approx 1$ and $\deg f \geq 1$. Then f can be written as a product of irreducible polynomials in a unique way.*

Proof: Let F be the field of fractions of D. We will use the fact that $F[x]$ is a unique factorization domain and the fact that irreducibility in $D[x]$ and in $F[x]$ coincide (Theorem 34.5, Lemma 34.11).

Consider f as a polynomial in $F[x]$. By Theorem 34.5,

$$f = g_1 g_2 \cdots g_r, \qquad g_1, g_2, \ldots, g_r \in F[x]$$

where $g_1, g_2, \ldots, g_r$ are irreducible in $F[x]$. According to Lemma 34.10,

$$f = h_1 h_2 \cdots h_r, \qquad h_1, h_2, \ldots, h_r \in D[x]$$

for some polynomials h_i in $D[x]$ with $C(h_i) \approx 1$ and h_i is associate to g_i in $F[x]$ $(i = 1,2, \ldots, r)$. Hence h_i is irreducible in $F[x]$ and, by Lemma 34.11, h_i is also irreducible in $D[x]$. We proved that f can be written as a product of irreducible polynomials in $D[x]$.

Now uniqueness (up to the order of factors and ambiguity among associate polynomials). Let $f \in D[x]$ with $C(f) \approx 1$ and $\deg f \geqslant 1$, and let

$$f = p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s \qquad p_i, q_j \in D[x] \tag{1}$$

be two representations of f as a product of irreducible polynomials $p_1, p_2, \ldots, p_r, q_1, q_2, \ldots, q_s$ in $D[x]$. Taking contents and using Lemma 34.8, we get

$$C(p_1)C(p_2)\cdots C(p_r) \approx C(f) \approx 1 \approx C(q_1)C(q_2)\cdots C(q_s)$$

so that $C(p_i)$ and $C(q_s)$ are units in D. By Lemma 34.11, the polynomials p_i, q_τ are irreducible in $F[x]$. Since $F[x]$ is a unique factorization domain, we deduce from (1) that $r = s$ and, eventually after reindexing the polynomials, p_i is associate to q_i in $F[x]$. Since $C(p_i) \approx C(q_i)$, Lemma 34.9 tells us that p_i is associate to q_i in $D[x]$ $(i = 1,2, \ldots, r)$. This completes the proof. □

34.13 Theorem: *If D is a unique factorization domain, then $D[x]$ is a unique factorization domain.*

Proof: Given any nonzero polynomial f in $D[x]$ which is not a unit in $D[x]$, we have to show that f can be written as a product of irreducible polynomials in $D[x]$, and that this representation is unique up to the order of factors and ambiguity between associate polynomials.

Now let $f \in D[x], f \neq 0, f \neq$ unit in $D[x]$. If $\deg f = 0$, then $f \in D$ and, since D is a unique factorization domain, f can be written as a product of irreducible elements $p_1, p_2, \ldots, p_r$ of D. These elements are uniquely determined, and they are irreducible also in $D[x]$ (Lemma 34.3). So f can be written as a product of irreducible elements in a unique way if $\deg f = 0$.

Suppose next $\deg f \geqslant 1$. We write $f = cf_1$, where $c \approx C(f) \in D$ and $f_1 \in D[x]$ with $C(f_1) \approx 1$, $\deg f_1 \geqslant 1$. Here c and f_1 are uniquely determined up to a unit in D. Now $c \in D$ can be written as a product of irreducible elements in D, which are also irreducible in $D[x]$:

$$c = a_1 a_2 \ldots a_r \qquad a_i \text{ are irreducible in } D[x],$$

and a_i are uniquely determined. By Lemma 34.12, f_1 can be written as a product of irreducible polynomials in $D[x]$:

$$f_1 = q_1 q_2 \ldots q_s \qquad q_j \text{ are irreducible in } D[x]$$

and q_i are uniquely determined. Hence

$$f = a_1 a_2 \ldots a_r q_1 q_2 \ldots q_s$$

is a product of the irreducible polynomials a_i, q_j in $D[x]$, which are unique up to the order of factors and ambiguity between associate elements. □

By repeated application of Theorem 34.13, we get

34.14 Theorem: *If D is a unique factorization domain, then $D[x_1, x_2, \ldots, x_n]$ is a unique factorization domain.* □

In particular,

34.15 Theorem: *If K is a field, then $K[x_1, x_2, \ldots, x_n]$ is a unique factorization domain.* □

Exercises

1. Prove that $x^4 + 1 \in \mathbb{Z}[x]$ is irreducible over $\mathbb{Z}$ by comparing the coefficients of both sides in a hypothetical factorization $x^2 + 1 = fg$ and deriving a contradiction from it. Investigate the cases *deg* $f = 1$, *deg* $g = 3$ and *deg* $f = 2 =$ *deg* g separately.

2. Do Ex. 1 for $x^4 + 2$ and $x^4 + 3 \in \mathbb{Z}[x]$.

3. Show that $x^4 + 4$ is reducible over $\mathbb{Z}$.

4. Show that $x^4 + \overline{1} \in \mathbb{Z}_2[x]$ is reducible over $\mathbb{Z}_2$.

5. Show that $x^4 + 1 \in (\mathbb{Z}[\sqrt{2}])[x]$ is reducible over $\mathbb{Z}[\sqrt{2}]$ (see §32, Ex. 3).

6. Find a content of

(a) $65x^4 + 26x^2 - 9x + 143$	$\in \mathbb{Z}[x]$
(b) $(5 + i)x^3 + (-1 + 5i)x + (-4 + 7i)$	$\in (\mathbb{Z}[i])[x]$
(c) $(1 + \omega)x^4 + (-1 + 2\omega)x^3 + (1 - 2\omega)x^2 + 3x + (2 + 3\omega)$	$\in (\mathbb{Z}[\omega])[x]$
(d) $8x^4 + 24x^3 - 32x^2 - 48x + 56$	$\in \mathbb{Q}[x]$
(e) $\overline{3}x^2 + \overline{5}x + \overline{7}$	$\in \mathbb{Z}_{97}[x]$.

7. Let D be a unique factorization domain and let F be the field of fractions of D. Let $f \in D[x]$ be a nonzero polynomial whose leading coefficient is a unit in D. Suppose that $g, h \in F[x]$ and $f = gh$. Prove that then $g \in D[x]$ and $h \in D[x]$.

8. Let D be a unique factorization domain and let $f, g \in D[x] \setminus D$. Prove that a greatest common divisor of f and g has degree $\geqslant 1$ if and only if there are polynomials h, k in $D[x]$ satisfying *deg* $h <$ *deg* g and *deg* $k <$ *deg* f such that $fh = gk$.

§35
Substitution and Differentiation

In this paragraph, we study the divisibility of polynomials by those of the first degree. We prove the familiar remainder theorem. Roots of polynomials are introduced and multiple roots are examined.

Everything in this paragraph is based on the substitution homomorphism which we now define.

35.1 Definition: Let R be a ring and let $f = \sum_{i=0}^{m} a_i x^i$ be an arbitrary polynomial in $R[x]$. Let S be a ring containing R. For any $s \in S$, the element $\sum_{i=0}^{m} a_i s^i$ of S is called the *value of f at s*. The value of f at s is said to be obtained by *substituting s for x* or by *evaluating f at s*. The value $\sum_{i=0}^{m} a_i s^i$ of f at s will be denoted by $f(s)$.

In many cases, S is taken to be R, and then $f(s) \in S$. In fact, we may always assume $S = R$ by taking f as a polynomial in $S[x]$. However, if $R \subset S$ and $s \in S \backslash R$, then $f(s)$ need not belong to R.

35.2 Examples: (a) Let $g = 4x^2 + 6x + 8 \in E[x]$, where E is the ring of even integers; so $E \subseteq \mathbb{Z}$. Now $1 \in \mathbb{Z}$ and $g(1) = 4 \cdot 1^2 + 6 \cdot 1 + 8 = 18 \in \mathbb{Z}$.

(b) Let $h = 3x^3 + 4x^2 + x - 1 \in \mathbb{Z}[x]$. Here $\mathbb{Q}$ is a ring that contains $\mathbb{Z}$ and $\frac{2}{5} \in \mathbb{Q}$. We have $h(\frac{2}{5}) = 3(\frac{2}{5})^3 + 4(\frac{2}{5})^2 + (\frac{2}{5}) - 1 = \frac{29}{125} \in \mathbb{Q}$.

(c) Let $f = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}x^2 + \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}x + \begin{pmatrix} -1 & 0 \\ 2 & 0 \end{pmatrix} \in (Mat_2(\mathbb{Z}))[x]$. Now $Mat_2(\mathbb{Z})$ is a ring containing $Mat_2(\mathbb{Z})$ and $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in Mat_2(\mathbb{Z})$. Then $f\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\right)$

$$= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2 + \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 2 \\ 4 & 0 \end{pmatrix} \in Mat_2(\mathbb{Z}).$$

(d) Let R be a ring with identity and $f = \sum_{i=0}^{m} a_i x^i \in R[x]$. Then $R \subseteq R[x]$ and $x \in R[x]$. The value of f at $x \in R[x]$ is $f(x) = \sum_{i=0}^{m} a_i x^i = f$, so $f(x) = f \in R[x]$. From now on, the notations f and $f(x)$ for a polynomial in $R[x]$ will be used interchangably.

(e) Again let R be a ring with identity and $f = \sum_{i=0}^{m} a_i x^i \in R[x]$ be a polynomial with coefficients in R. Let y be an indeterminate distinct from x. Then R is contained in $R[y]$ and $y \in R[y]$. The value of f at y is $f(y) = \sum_{i=0}^{m} a_i y^i \in R[y]$.

(f) Let $p = x^3 - x + 1 \in \mathbb{O}[x]$. Now $\mathbb{O} \subseteq \mathbb{O}[x]$, $x + 1 \in \mathbb{O}[x]$ and $p(x + 1) = (x + 1)^3 - (x + 1) + 1 = x^3 + 3x^2 + 2x + 1 \in \mathbb{O}[x]$. Similarly $x^2 \in \mathbb{O}[x]$ and $p(x^2) = (x^2)^3 - (x^2) + 1 = x^6 - x^2 + 1 \in \mathbb{O}[x]$.

(g) Let R be a ring. For any $f \in R[x]$, the value of f at $g \in R[x]$ can be found as in the last example, and it is a polynomial $f(g(x))$ in $R[x]$.

(h) Let $f = \bar{3}x^2 - \bar{3}x + \bar{2} \in \mathbb{Z}_{12}[x]$. The value $f(1)$ of f at $1 \in \mathbb{Z}$ is not defined, for $\mathbb{Z}$ does not contain $\mathbb{Z}_{12}$.

(i) Let $q = x^2 + x + 2$ and $r = x^3 + x + 3 \in \mathbb{Z}[x]$. We put $t = qr = x^5 + x^4 + 3x^3 + 4x^2 + 5x + 6 \in \mathbb{Z}[x]$. One checks easily that $q(2) = 8$, $r(2) = 13$, $t(2) = 104$. Notice $t(2) = 8 \cdot 13 = q(2) \cdot r(2)$. This is explained in the next lemma.

35.3 Lemma: *Let R be a ring, S a ring that contains R, and s an element of S. If S is commutative, then the mapping*

$$\begin{aligned} T_s: R[x] &\to S \\ f &\to f(s) \end{aligned}$$

is a ring homomorphism (called the *substitution* or *evaluation homomorphism*).

Proof: For any $f = \sum_{i=0}^{m} a_i x^i$, $g = \sum_{j=0}^{n} b_j x^j$ in $R[x]$, we have

$$(f+g)T_s = \Big(\sum_{i=0}^{m} a_i x^i + \sum_{j=0}^{n} b_j x^j\Big)T_s$$

$$= \Big(\sum_{i=0}^{m} (a_i + b_i)x^i\Big)T_s \qquad \text{(assuming } n = m \text{ without loss of generality)}$$

$$= \sum_{i=0}^{m} (a_i + b_i)s^i$$

$$= \sum_{i=0}^{m} a_i s^i + \sum_{i=0}^{m} b_i s^i$$

$$= f(s) + g(s)$$

$$= fT_s + gT_s,$$

and further

$$(fg)T_s = \Big[\sum_{k=0}^{m+n}\Big(\sum_{i+j=k} a_i b_j\Big)x^k\Big]T_s = \sum_{k=0}^{m+n}\Big(\sum_{i+j=k} a_i b_j\Big)s^k,$$

$$(fT_s)(gT_s) = \Big(\sum_{i=0}^{m} a_i x^i\Big)T_s \cdot \Big(\sum_{j=0}^{n} b_j x^j\Big)T_s = \Big(\sum_{i=0}^{m} a_i s^i\Big)\cdot\Big(\sum_{j=0}^{n} b_j s^j\Big)$$

$$= \sum_{i,j} a_i s^i b_j s^j$$

$$= \sum_{i,j} a_i b_j s^{i+j} \qquad \text{(using commutativity of } S)$$

$$= \sum_{k=0}^{m+n}\Big(\sum_{i+j=k} a_i b_j\Big)s^k$$

$$= (fg)T_s.$$

Hence T_s preserves sums and products, and is therefore a ring homomorphism. □

In the proof of Lemma 35.3, the commutativity of S is used in a crucial way. If S is not commutative, then T_s is *not* a homomorphism. For example,

$$Ix^2 - \begin{pmatrix}1&0\\0&1\end{pmatrix} = \left[Ix + \begin{pmatrix}0&1\\1&0\end{pmatrix}\right]\left[Ix - \begin{pmatrix}0&1\\1&0\end{pmatrix}\right] \quad \text{in } (Mat_2(\mathbb{Z}))[x]$$

but substituting $\begin{pmatrix}1&1\\0&0\end{pmatrix}$ for x does not preserve sums and products:

$$\begin{pmatrix}1&1\\0&0\end{pmatrix}^2 - \begin{pmatrix}1&0\\0&1\end{pmatrix} \neq \left[\begin{pmatrix}1&1\\0&0\end{pmatrix} + \begin{pmatrix}0&1\\1&0\end{pmatrix}\right]\left[\begin{pmatrix}1&1\\0&0\end{pmatrix} - \begin{pmatrix}0&1\\1&0\end{pmatrix}\right].$$

The substitution homomorphism is closely related to the division algorithm in an integral domain.

35.4 Theorem (Remainder theorem): *Let D be an integral domain, $f \in D[x]$ and $a \in D$. There is a unique polynomial q in $D[x]$ such that*

$$f(x) = q(x)(x - a) + f(a).$$

Proof: We divide f by $(x - a)$. This is possible by Theorem 34.4, because the leading coefficient of $x - a$ is a unit in D (in fact = 1). Thus there are unique polynomials q and r such that

$$f(x) = q(x)(x - a) + r(x) \qquad r = 0 \text{ or } \deg r < \deg (x - a) = 1.$$

So r is an element of D (zero or not). To find r, we substitute a for x; since substitution is a homomorphism by Lemma 35.3, we get

$$f(a) = q(a)(a - a) + r(a)$$
$$f(a) = r.$$

This completes the proof. □

35.5 Definition: Let R be a ring, S a commutative ring that contains R and let f be a polynomial in $R[x]$. An element a of S is called a *root* or *zero of* f if $f(a) = 0$.

35.6 Theorem (Factor theorem): *Let D be an integral domain, and let f be an arbitrary polynomial in $D[x]$. Let E be an integral domain containing D and let $a \in E$. Then a is a root of f if and only if $(x - a) | f$ in $E[x]$.*

Proof: By the remainder theorem (with E in place of D), there is a polynomial q in $E[x]$ such that $f(x) = q(x)(x - a) + f(a)$. If a is a root of f, then $f(a) = 0$, so $f(x) = q(x)(x - a)$ and $(x - a) | f(x)$ in $E[x]$. Conversely, if $(x - a) | f(x)$ in $E[x]$, then $(x - a) | [f(x) - q(x)(x - a)]$ in $E[x]$, so $(x - a) | f(a)$ in $E[x]$. Thus $f(x) = u(x)(x - a)$ for some $u(x) \in E[x]$. Substituting a for x, we get $f(a) = u(a)(a - a) = 0$. So a is a root of f. □

The factor theorem puts an upper bound to the number of roots of polynomials over integral domains, in particular of those over fields.

35.7 Theorem: *Let D be an integral domain, f a nonzero polynomial in $D[x]$ and let E be an integral domain containing D. Then there are at most deg f distinct roots of f in E.*

Proof: We make induction on the degree of f. Polynomials of degree 0 are just the nonzero elements of D, and they have no roots in E (zero roots). So the theorem is true when $deg\ f = 0$. Assume now $deg = 1$, so that $f = cx + d$, where $c, d \in D$ and $c \neq 0$. If f had more then one roots in E, say if a_1, a_2 were roots of f in E and $a_1 \neq a_2$, we would get

$$\begin{aligned} ca_1 + d = f(a_1) = 0 = f(a_2) &= ca_2 + d \\ ca_1 &= ca_2 \\ c(a_1 - a_2) &\equiv 0 \quad c \neq 0 \\ a_1 - a_2 &= 0, \end{aligned}$$

contrary to $a_1 \neq a_2$. Thus $cx + d$ has either no roots in E or one and only one root in E, and the theorem is proved when $deg\ f = 1$.

Suppose now $n \geqslant 2$, $deg\ f = n$ and that, for all integral domains D', any polynomial of degree $n - 1$ in $D'[x]$ has at most $n - 1$ distinct roots in any integral domain E' that contains D'. If f has no roots in E, the theorem is true. If f has a root a_0 in E, we have

$$f(x) = q(x)(x - a_0) \qquad \text{for some } q(x) \in E[x]$$

by the factor theorem. Here $q(x)$ is of degree $n-1$ by Lemma 33.3(3). By our induction hypothesis, $q(x)$ has at most $n-1$ distinct roots in E. Now let A be the set of all distinct roots of $q(x)$ in E (possibly $A = \varnothing$) so that $|A| \leq n-1$.

If $b \in E$ is any root of f, then $f(b) = 0$, so $q(b)(b - a_0) = 0$, so $q(b) = 0$ or $b = a_0$, so $b \in A$ or $b = a_0$. Hence $B \subseteq A \cup \{a_0\}$, where B is the set of all distinct roots of $f(x)$ in E. Thus $|B| \leq |A| + 1 \leq (n-1) + 1 = n$ and f has at most $n = \deg f$ distinct roots in E. This completes the proof. □

Theorem 35.7 may be false if the underlying ring is not commutative or if it has zero divisors. For example, $x^2 + \bar{1} \in H[x]$ of degree two over the noncommutative ring H of Ex. 9 in §29 has infinitely many roots in H. Also, the polynomial $x^2 - \bar{1} = \bar{1}x^2 - \bar{1}$ over $\mathbb{Z}_8$, which has zero divisors, possesses four distinct roots $\bar{1}, \bar{3}, \bar{5}, \bar{7}$ in $\mathbb{Z}_8$.

We give two applications of Theorem 35.7. In these applications, the underlying integral domain is a field.

35.8 Theorem (Lagrange's interpolation formula): *Let K be a field and $a_0, a_1, \ldots, a_n$ be distinct elements of K. Let $b_0, b_1, \ldots, b_n$ be arbitrary elements of K (not necessarily distinct). Then there is a unique polynomial in $K[x]$ such that $f(a_0) = b_0, f(a_1) = b_1, \ldots, f(a_n) = b_n$ and such that $\deg f \leq n$ (one less than the number of a's or b's) or $f = 0$. This polynomial is given explicitly by the formula*

$$f = \sum_{i=0}^{n} \frac{(x - a_0)\ldots(x - a_{i-1})(x - a_{i+1})\ldots(x - a_0)}{(a_i - a_0)\ldots(a_i - a_{i-1})(a_i - a_{i+1})\ldots(a_i - a_0)} b_i.$$

Proof: The i-th summand $f_i :=$

$$\frac{(x - a_0)\ldots(x - a_{i-1})(x - a_{i+1})\ldots(x - a_0)}{(a_i - a_0)\ldots(a_i - a_{i-1})(a_i - a_{i+1})\ldots(a_i - a_0)} b_i$$

in the formula is $0 \in K[x]$ (when $b_i = 0$) or a polynomial in $K[x]$ of degree n (when $b_i \neq 0$). Here $f_i(a_i) = b_i$ and $f_i(a_j) = 0$ for $i \neq j$. So $f := f_1 + f_2 + \ldots + f_n$ is either the zero polynomial or a polynomial of degree at most n such that $f(a_i) = f_1(a_i) + f_2(a_i) + \cdots + f_n(a_i) = 0 + \cdots + f_i(a_i) + 0 + \cdots + 0 = b_i$ for all

$i = 1,2, \ldots, n$. This proves the existence of a polynomial with the properties stated in the theorem, namely the one given explicitly above.

The uniqueness of f follows from Theorem 35.7. If g is a polynomial in $K[x]$ with $deg\ g \leqslant n$, and if $g(a_1) = b_1, g(a_2) = b_2, \ldots, g(a_n) = b_n$, then the polynomial $h = f - g$ has at least $n + 1$ roots $a_0, a_1, \ldots, a_n$ in K, and, if $h \neq 0$, then h has degree at most equal to n (Lemma 33.3(2)). This is not compatible with Theorem 35.7, so $h = 0$ and $g = f$. Therefore f is the unique polynomial satisfying the conditions above. □

The formula for f is easy to remember. We have $f = f_1 + f_2 + \cdots + f_n$, where $f_i(a_i) = b_i$ and $f_i(a_j) = 0$ for $i \neq j$. The second condition leads to $f_i = (x - a_0)\ldots(x - a_{i-1})(x - a_{i+1})\ldots(x - a_0)c_i$ for some $c_i \in K$, and c_i must as in the formula if $f_i(a_i)$ is to be equal to b_i.

35.9 Theorem (Wilson's theorem): *If $p \in \mathbb{N}$ is a prime number, then* $(p - 1)! + 1 \equiv 0 \pmod{p}$.

Proof (Lagrange): Fermat's theorem (Theorem 12.6) states that $a^{p-1} \equiv 1 \pmod{p}$ for any integer a with $(a,p) = 1$. We can write this as

$$\bar{a}^{p-1} - \bar{1} = \bar{0} \quad \text{in } \mathbb{Z}_p \quad \text{if } \bar{a} \neq \bar{0}.$$

Thus the polynomial $f = x^{p-1} - \bar{1} = \bar{1}x^{p-1} - \bar{1} \in \mathbb{Z}_p[x]$ has $p - 1$ distinct roots in $\mathbb{Z}_p$, namely $\bar{1}, \bar{2}, \ldots, \overline{p-1}$. The polynomial

$$g = (x - \bar{1})(x - \bar{2})\ldots(x - \overline{p-1})$$

has the same roots. Hence the polynomial

$$h = f - g = (\bar{1}x^{p-1} - \bar{1}) - (x - \bar{1})(x - \bar{2})\ldots(x - \overline{p-1}) = (x^{p-1} - \bar{1}) - (x^{p-1} + \cdots)$$

over $\mathbb{Z}_p$ has at least $p - 1$ roots $\bar{1}, \bar{2}, \ldots, \overline{p-1}$ in $\mathbb{Z}_p$. If h were not the zero polynomial in $\mathbb{Z}_p[x]$, its degree would be less than $p - 1$. This contradicts Theorem 35.7. So h is the zero polynomial in $\mathbb{Z}_p[x]$: each coefficient of h is equal to $\bar{0} \in \mathbb{Z}_p$. In particular,

$$\begin{aligned} \bar{0} &= \text{coefficient of } x^0 \text{ in } h \\ &= (\text{coefficient of } x^0 \text{ in } f) - (\text{coefficient of } x^0 \text{ in } g) \\ &= (-\bar{1}) - ((-\bar{1})(-\bar{2})\ldots(-\overline{p-1})) \end{aligned}$$

$$= -1 - (-1)^{p-1}(p-1)!$$
$$= -((p-1)! + 1) \text{ in } \mathbb{Z}_p$$

provided p is odd. Hence $(p-1)! + 1 \equiv 0 \pmod{p}$ when p is an odd prime number. But this congruence holds also when $p = 2$. This completes the proof. □

The next theorem will be familiar to the reader in the case of $D = \mathbb{Z}$, $F = \mathbb{Q}$ under the name of "rational root theorem".

35.10 Theorem: *Let D be a unique factorization domain and let F be the field of fractions of D. Let $f = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in D[x]$ be an arbitrary polynomial in $D[x]$. If $a = \frac{b}{c} \in F$ is a root of f, where $b,c \in D$ and $(b,c) \approx 1$, then*

$$c \mid a_n \quad \text{and} \quad b \mid a_0 \qquad \text{in } D.$$

In particular, if the leading coefficient of f is a unit in D, then any root of f in F is actually in D.

Proof: By hypothesis, $a = \frac{b}{c}$ is a root of f so that

$$a_n \frac{b^n}{c^n} + a_{n-1}\frac{b^{n-1}}{c^{n-1}} + \cdots + a_1 \frac{b}{c} + a_0 = 0.$$

Multiplying both sides by c^n, we obtain

$$a_n b^n + (a_{n-1}b^{n-1}c + \cdots + a_1 bc^{n-1} + a_0 c^n) = 0,$$
$$[a_n b^n + a_{n-1}b^{n-1}c + \cdots + a_1 bc^{n-1}] + a_0 c^n = 0.$$

c divides the expression in ( ), so $c \mid a_n b^n$. As $(b,c) \approx 1$, we have $(b^n,c) \approx 1$. From $(b^n,c) \approx 1$ and $c \mid a_n b^n$, we conclude $c \mid a_n$. Likewise, b divides the expression in [ ], so $b \mid a_0 c^n$. As $(b,c) \approx 1$, we have $(b,c^n) \approx 1$. From $(b,c^n) \approx 1$ and $b \mid a_0 c^n$, we conclude $b \mid a_0$. In particular, if a_n is a unit in D, then c is also a unit in D since $c \mid a_n$, so there is a $c^{-1} \in D$ such that $cc^{-1} = 1$ and the root $a = \frac{b}{c} = \frac{bc^{-1}}{cc^{-1}} = \frac{bc^{-1}}{1} = bc^{-1} \in D$. □

35.11 Example: As an illustration of Theorem 35.10, we prove that the real number $\sqrt{2}$ is irrational. Let $f(x) = x^2 - 2 \in \mathbb{Z}[x]$. Since $\mathbb{Z}$ is a unique factorization domain and the leading coefficient of f is a unit in $\mathbb{Z}$ (f is in fact a monic polynomial), any root of f in $\mathbb{Q}$ must be actually in $\mathbb{Z}$ by Theorem 35.10. But

$$f(0) = -2 \neq 0; \quad f(\mp 1) = -1 \neq 0;$$

$$f(\mp m) = m^2 - 2 \geqslant 2, \text{ so } f(\mp m) \neq 0 \text{ for } m \geqslant 2;$$

so f has no integer roots, and consequently no rational roots, as claimed.

Next we discuss the multiplicity of roots. Let D be an integral domain and f a nonzero polynomial in $D[x]$. If $a \in D$ is a root of f, then we have $f(x) = (x - a)q_1(x)$ for some $q_1(x) \in D[x]$ by the factor theorem (Theorem 35.6). Either a is not a root of $q_1(x)$, or we have $q_1(x) = (x - a)q_2(x)$ and therefore $f(x) = (x - a)^2 q_2(x)$ for some $q_2(x) \in D[x]$. In the latter case, either a is not a root of $q_2(x)$, or we have $q_2(x) = (x - a)q_3(x)$ and therefore $f(x) = (x - a)^3 q_3(x)$ for some $q_3(x) \in D[x]$. We repeat this argument. Since the degrees of $q_1(x), q_2(x), q_3(x), \ldots$ get smaller and smaller, we will reach a polynomial $q_m(x)$ with

$$f(x) = (x - a)^m q_m(x), \quad q_m(a) \neq 0.$$

35.12 Definition: Let D be an integral domain and f a nonzero polynomial in $D[x]$. Suppose $a \in D$ and $f(a) = 0$. The uniquely determined integer $m \geqslant 1$ such that

$$f(x) = (x - a)^m q_m(x), \quad q_m(x) \in D[x], \quad q_m(a) \neq 0,$$

that is, the uniquely determined integer $m \geqslant 1$ such that

$$(x - a)^m | f(x), \quad (x - a)^{m+1} \nmid f(x) \quad \text{in } D[x]$$

is called the *multiplicity* of the root a of f. The root a of f is called a *simple* root when $m = 1$ and a *multiple* root when $m > 1$.

This definition makes sense also when a is a root of f in E, where E is an integral domain containing D: we need only regard f as a polynomial over E and use the definition with E in place of D. When E_1 and E_2 are two integral domains containing D and a root a of f is both in E_1 and E_2, we have, say,

$$\begin{aligned} f(x) &= (x-a)^{m_1}q_1(x), \quad & q_1(x) &\in E_1[x], & q_1(a) &\neq 0, \\ f(x) &= (x-a)^{m_2}q_2(x), \quad & q_2(x) &\in E_2[x], & q_2(a) &\neq 0, \\ f(x) &= (x-a)^{m_0}q_0(x), \quad & q_0(x) &\in (E_1 \cap E_2)[x], & q_0(a) &\neq 0, \end{aligned}$$

as the equations defining the multiplicity of a as a root in $E_1, E_2, E_1 \cap E_2$. Then

$$(x-a)^{m_1}q_1(x) = (x-a)^{m_0}q_0(x) \qquad \text{in } E_1[x]$$

and the assumption $m_1 > m_0$ or $m_1 < m_0$ leads to the contradiction

$$\begin{gathered} (x-a)^{m_1-m_0}q_1(x) = q_0(x) \quad \text{or} \quad q_1(x) = (x-a)^{m_0-m_1}q_0(x), \\ 0 = q_0(a) \quad \text{or} \quad q_1(a) = 0. \end{gathered}$$

Hence $m_1 = m_0$. Likewise $m_2 = m_0$ and therefore $m_1 = m_2$: the multiplicity of a root of $f \in D[x]$ is independent of the integral domain to which the root belongs.

In order to find out whether a polynomial has multiple roots, we take derivatives.

In analysis, the derivative of a real-valued function u of a real variable x is defined by

$$u'(x) = \lim_{h \to 0} \frac{u(x+h) - u(x)}{h}.$$

This definition cannot be extended to polynomials over a ring. For one thing, polynomials are not functions. Second, what should $\dfrac{u(x+h) - u(x)}{h}$ mean in a ring? Third, we did not define limits in a ring. In fact, in many rings, a reasonable limit process cannot be introduced at all. But we know from analysis that the derivative of the

function $x \to \sum_{k=0}^{m} a_k x^k$ is the function $x \to \sum_{k=1}^{m} k\, a_k x^{k-1}$. This suggests the following definition.

35.13 Definition: Let R be an arbitrary ring and let $f = \sum_{k=0}^{m} a_k x^k$ be an arbitrary polynomial in $R[x]$. The *derivative of* f is defined as the polynomial

$$f' = f'(x) = \sum_{k=1}^{m} k\, a_k x^{k-1} = \sum_{k=0}^{m-1} (k+1) a_{k+1} x^k \in R[x].$$

$k a_k$ means of course $a_k + a_k + \cdots + a_k$ in R (k times). This definition has nothing to do with limits. Taking the derivative of a polynomial is called differentiation.

35.14 Examples: (a) Let $f(x) = x^4 - 3x^2 + x + 10 \in \mathbb{Z}[x]$. Then $f'(x) = 4x^3 - 6x + 1 \in \mathbb{Z}[x]$.

(b) Let $g(x) = \frac{1}{3}x^5 + \frac{1}{7}x^4 + \frac{2}{5}x^3 + \frac{4}{3}x - 3 \in \mathbb{Q}[x]$. Then $g'(x) = \frac{5}{3}x^4 + \frac{4}{7}x^3 + \frac{6}{5}x^2 + \frac{4}{3} \in \mathbb{Q}[x]$.

(c) Let $h(x) = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}x^3 + \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}x^2 + \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}x + \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \in (Mat_2(\mathbb{Z}))[x]$. Then

$$h'(x) = 3\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}x^2 + 2\begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}x + 1\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}$$

$$= \begin{pmatrix} 3 & 6 \\ 9 & 12 \end{pmatrix}x^2 + \begin{pmatrix} 0 & 2 \\ -2 & 2 \end{pmatrix}x + \begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix} \in (Mat_2(\mathbb{Z}))[x].$$

(d) Let $k(x) = \bar{2}x^4 + \bar{4}x^2 + \bar{3}x + \bar{5} \in \mathbb{Z}_8[x]$. Then
$k'(x) = 4\cdot\bar{2}x^3 + 2\cdot\bar{4} + 1\cdot\bar{3} = \bar{3} \in \mathbb{Z}_8[x]$.

(e) Let $l(x) = x^{125} + x^{25} + \bar{2}x^5 + \bar{3} \in \mathbb{Z}_5[x]$. Then
$l'(x) = 125\cdot\bar{1}x^{124} + 25\cdot\bar{1}x^{24} + 5\cdot\bar{2}x^4 = 0 \in \mathbb{Z}_5[x]$.

The familiar rules of differentiation hold in any polynomial ring.

35.15 Lemma: *Let R be a ring, $c \in R$, and let $f,g \in R[x]$. Then*
$$(f+g)' = f' + g', \quad (cf)' = cf', \quad (fg)' = f'g + fg'.$$

Proof: Let $f = \sum_{k=0}^{m} a_k x^k$ and $g = \sum_{j=0}^{n} b_j x^j$. We have

$$(f+g)' = \left(\sum_{k=0}^{m} a_k x^k + \sum_{j=0}^{n} b_j x^j\right)'$$

$$= \left(\sum_{k=0}^{m} a_k x^k + \sum_{k=0}^{m} b_k x^k\right)' \text{(assuming } n = m \text{ without loss of generality)}$$

$$= \left(\sum_{k=0}^{m} (a_k + b_k) x^k\right)'$$

$$= \sum_{k=1}^{m} k(a_k + b_k) x^{k-1}$$

$$= \sum_{k=1}^{m} (k\,a_k + k\,b_k) x^{k-1}$$

$$= \sum_{k=1}^{m} k\,a_k x^{k-1} + \sum_{k=1}^{m} k\,b_k x^{k-1}$$

$$= f' + g',$$

$$(cf)' = \left(c \sum_{k=0}^{m} a_k x^k\right)' = \left(\sum_{k=0}^{m} c\,a_k x^k\right)' = \sum_{k=1}^{m} k\,c\,a_k x^{k-1} = c \sum_{k=1}^{m} k\,a_k x^{k-1} = cf'$$

Next we find $(fg)'$ and $f'g + fg'$. We have

$$(fg)' = \left[\left(\sum_{k=0}^{m} a_k x^k\right)\left(\sum_{j=0}^{n} b_j x^j\right)\right]'$$

$$= \left[\sum_{s=0}^{m+n} \left(\sum_{k+j=s} a_k b_j\right) x^s\right]'$$

$$= \sum_{s=1}^{m+n} s \left(\sum_{k+j=s} a_k b_j\right) x^{s-1}, \tag{1}$$

$$
\begin{aligned}
f'g + fg' &= \Big(\sum_{k=0}^{m} a_k x^k\Big)'\Big(\sum_{j=0}^{n} b_j x^j\Big) + \Big(\sum_{k=0}^{m} a_k x^k\Big)\Big(\sum_{j=0}^{n} b_j x^j\Big)' \\
&= \Big(\sum_{k=1}^{m} k\, a_k x^{k-1}\Big)\Big(\sum_{j=0}^{n} b_j x^j\Big) + \Big(\sum_{k=0}^{m} a_k x^k\Big)\Big(\sum_{j=1}^{n} j\, b_j x^{j-1}\Big) \\
&= \sum_{s=1}^{m+n}\Big(\sum_{k+j=s} k a_k b_j\Big)x^{s-1} + \sum_{s=1}^{m+n}\Big(\sum_{k+j=s} j a_k b_j\Big)x^{s-1} \\
&= \sum_{s=1}^{m+n}\Big(\sum_{k+j=s} k a_k b_j + j a_k b_j\Big)x^{s-1} \\
&= \sum_{s=1}^{m+n} s\Big(\sum_{k+j=s} a_k b_j\Big)x^{s-1}. \qquad (2)
\end{aligned}
$$

From (1) and (2), we conclude $(fg)' = f'g + fg'$. This completes the proof. □

35.16 Lemma: *Let R be a ring and let $f_1, f_2, \ldots f_n, f, g \in R[x]$.*

(1) $(f_1 + f_2 + \cdots + f_n)' = f_1' + f_2' + \cdots + f_n'$.

(2) $(f_1 f_2 \ldots f_n)' = f_1' f_2 \ldots f_n + f_1 f_2' \ldots f_n + \cdots + f_1 f_2 \ldots f_n'$.

(3) $(g^n)' = n g^{n-1} g'$.

(4) $[f(g(x))]' = f'(g(x))g'(x)$.

Proof: (1) and (2) follow from Lemma 35.16 by induction on n. (3) is a special case of (2), with $f_1 = f_2 = \cdots = f_n = g$. We now prove (4). Let $f = \sum_{k=0}^{m} a_k x^k$. Then $f(g(x)) = \sum_{k=0}^{m} a_k g^k \in R[x]$ and, by (1) and (3), the derivative of $f(g(x))$ is

$$
\begin{aligned}
\Big(\sum_{k=0}^{m} a_k g^k\Big)' &= \sum_{k=0}^{m} a_k (g^k)' = \sum_{k=1}^{m} a_k (g^k)' = \sum_{k=1}^{m} k\, a_k g^{k-1} g' \\
&= \Big(\sum_{k=1}^{m} k\, a_k g^{k-1}\Big) g' = f'(g) g'. \qquad \square
\end{aligned}
$$

We are now in a position to determine which roots are multiple roots.

35.17 Theorem: *Let D be an integral domain, and E an integral domain that contains D. Let $c \in E$ and let f be a nonzero polynomial in $D[x]$. Then c is a multiple root of f if and only if c is a root of both f and f'.*

Proof: Suppose c is a multiple root of f. Then it is a root of f. We wish to show that c is a root of f' as well. We have $f(x) = (x - c)^2 g(x)$ for some $g(x) \in E[x]$. Differentiating and substituting c for x, we obtain

$$f'(x) = 2(x - c)g(x) + (x - c)^2 g'(x)$$
$$f'(c) = 2(c - c)g(c) + (c - c)^2 g'(c) = 0$$

and c is indeed a root of f'.

Conversely, suppose c is a root of f and f'. We write $f(x) = (x - c)h(x)$, where $h(x) \in E[x]$. We want to show that c is a root of h. Since

$$f'(x) = h(x) + (x - c)h'(x)$$
$$f'(c) = h(c) + (c - c)h'(c)$$
$$0 = h(c) + 0,$$

$h(c) = 0$ and c is a multiple root of f. □

35.18 Theorem: *Let K be a field and E an integral domain that contains K. Let $f(x)$, $g(x)$ be arbitrary nonzero polynomials in $K[x]$.*
(1) If f and g are relatively prime, then f and g have no common root in E.
(2) If f and f' are relatively prime, then f has no multiple roots in E.
(3) If f is irreducible in $K[x]$, then either f and g are relatively prime or $f|g$ in $K[x]$.
(4) If f is irreducible in $K[x]$ and deg $f >$ deg g, then f and g have no common root in E.
(5) If f is irreducible in $K[x]$ and $f' \neq 0$, then there is no root of f in E which is a multiple root.
(6) If f is irreducible in $K[x]$ and if f has a root in E which is not a multiple root of f, then $f' \neq 0$.

Proof: (1) Suppose f and g are relatively prime in $K[x]$. By Theorem 34.6, there are polynomials h, l in $K[x]$ such that

$$1 = h(x)f(x) + l(x)g(x),$$

where 1 is the identity element of K. If f and g had a root $c \in E$ in common, we would have

$$1 = h(c)f(c) + l(c)g(c) = h(c)0 + l(c)0 = 0 + 0 = 0,$$

a contradiction. So f and g have no common root in E.

(2) Assume f and f' are relatively prime. If f has no root in E, then certainly f has no multiple root in E. Now we suppose f has a root c in E and prove that c is not a multiple root of f. Indeed, since f and f' are relatively prime, f and f' have no common root by part (1), so $f'(c) \neq 0$ and c is not a multiple root of f by Theorem 35.17.

(3) Suppose f is irreducible in $K[x]$ and let $d \in K[x]$ be a greatest common divisor of f and g. Since $d|f$ and f is irreducible, d is either a unit in $K[x]$ or an associate of f. In the first case, f and g are relatively prime, in the second case, $f \approx d$ and $d|g$ yields $f|g$.

(4) Suppose f is irreducible in $K[x]$ and $deg\ g < deg\ f$, then f cannot divide g, so f and g are relatively prime by part (3). By part (1), f and g have no common root in E.

(5) Suppose f is irreducible in $K[x]$ and $f' \neq 0$. Then $deg\ f' < deg\ f$. Since f is irreducible, f and f' have no common root in E by part (4). Now if f has no root in E, then f has certainly no multiple root in E. If f has a root c in E, then c is not a root of f', so c is not a multiple root of f by Theorem 35.17. In any case, f has no multiple root in E.

(6) Suppose f is irreducible in $K[x]$ and suppose $c \in E$ is a simple root of f in E. If we had $f' = 0$, we would have $f(c) = 0$ and $f'(c) = 0$ and c would be a multiple root of f by Theorem 35.17, a contradiction. Thus, if there are roots in E and if they are all simple, then $f' \neq 0$. □

We finish this paragraph with a brief discussion of successive substitutions.

35.19 Definition: Let R be a ring and let

$$f = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \cdots \sum_{k=0}^{N_{n-1}} \sum_{l=0}^{N_n} a_{ij\ldots kl} x_1^i x_2^j \ldots x_{n-1}^k x_n^l$$

be a polynomial in $R[x_1, x_2, \ldots, x_{n-1}, x_n]$. Let S be a ring that contains R and let $c_1, c_2, \ldots, c_{n-1}, c_n$ be elements of S. The element

$$\sum_{i=0}^{N_1} \sum_{j=0}^{N_2} \cdots \sum_{k=0}^{N_{n-1}} \sum_{l=0}^{N_n} a_{ij\ldots kl} c_1^i c_2^j \ldots c_{n-1}^k c_n^l$$

of S is called the *value of* f *at* $(c_1,c_2,\ldots,c_{n-1},c_n)$. It will be denoted by $f(c_1,c_2,\ldots,c_{n-1},c_n)$.

With the foregoing notation, $f = \sum_{i=0}^{N_1}\Big(\sum_{j=0}^{N_2}\cdots\sum_{k=0}^{N_{n-1}}\sum_{l=0}^{N_n} a_{ij\ldots kl}x_1^{\,i}x_2^{\,j}\ldots x_{n-1}^{k}\Big)x_n^{\,l}$ is a polynomial in $R[x_1,x_2,\ldots,x_{n-1}][x_n]$. Substituting c_n for x_n in the sense of Definition 35.1 (with $S[x_1,x_2,\ldots,x_{n-1}]$, $R[x_1,x_2,\ldots,x_{n-1}]$, x_n, c_n in place of S, R, x, c, respectively), we get an element of $S[x_1,x_2,\ldots,x_{n-1}]$, namely

$$\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}\cdots\sum_{k=0}^{N_{n-1}}\Big(\sum_{l=0}^{N_n} c_n^{\,l}a_{ij\ldots kl}\Big)x_1^{\,i}x_2^{\,j}\ldots x_{n-1}^{k} \in S[x_1,x_2,\ldots,x_{n-2}][x_{n-1}].$$

Substituting c_{n-1} for x_{n-1} in this polynomial over $S[x_1,x_2,\ldots,x_{n-2}]$, we get a polynomial in $S[x_1,x_2,\ldots,x_{n-2}]$, namely

$$\sum_{i=0}^{N_1}\sum_{j=0}^{N_2}\cdots\sum_{j'=0}^{N_{n-2}}\Big(\sum_{k=0}^{N_{n-1}}\sum_{l=0}^{N_n} c_{n-1}^{\,k}c_n^{\,l}a_{ij\ldots j'kl}\Big)x_1^{\,i}x_2^{\,j}\ldots x_{n-2}^{j'}.$$

We continue in this way. If S is commutative, we obtain $f(c_1,c_2,\ldots,c_{n-1},c_n)$ after n substitutions. Thus

$$f(c_1,c_2,\ldots,c_{n-1},c_n) = fT_{c_n}T_{c_{n-1}}\ldots T_{c_2}T_{c_1},$$

where $T_{c_n}: R[x_1,x_2,\ldots,x_{n-1},x_n] \longrightarrow S[x_1,x_2,\ldots,x_{n-1}]$

$T_{c_h}: S[x_1,x_2,\ldots,x_{h-1},x_h] \longrightarrow S[x_1,x_2,\ldots,x_{h-1}]$ $\quad (h = 2,\ldots,n-1)$

and $T_{c_1}: S[x_1] \longrightarrow S$

are the substitution homomorphisms in the sense of Definition 35.1. Since the composition of homomorphisms is a homomorphism (Theorem 30.12), we obtain the following lemma.

35.20 Lemma: *Let R be a ring, S a ring that contains R, and $c_1,c_2,\ldots,c_{n-1},c_n$ elements of S. If S is commutative, then the mapping*

$$T_{(c_1,c_2,\ldots,c_{n-1},c_n)}: R[x_1,x_2,\ldots,x_{n-1},x_n] \longrightarrow S$$
$$f \longrightarrow f(c_1,c_2,\ldots,c_{n-1},c_n)$$

is a ring homomorphism (called the *evaluation* or *substitution* homomorphism). □

Exercises

1. Let $f = x^3 + ax^2 + bx + c \in \mathbb{Z}[x]$. Prove that f is reducible over $\mathbb{Q}$ if and only if f has an integer root.

2. Find a polynomial $f \in \mathbb{Q}[x]$ with $\deg f \leqslant 4$ satisfying
$$f(-2) = 9, \quad f(-1) = -2, \quad f(0) = 1, \quad f(1) = 4, \quad f(2) = 25.$$

3. Let p be a prime number of the form $4k + 1$. Using Wilson's theorem, show that $\overline{\left(\frac{p-1}{2}\right)!}$ is a root of $x^2 + 1 \in \mathbb{Z}_p[x]$.

4. Let R be a ring and $f = \sum_{i,j,k} a_{ijk} x^i y^j z^k \in R[x,y,z]$. The derivative of f, when f is regarded as a polynomial in $R[y,z][x]$, is called the *derivative of f with respect to x* and is written $\frac{\partial f}{\partial x}$. Thus $\frac{\partial f}{\partial x} = \sum_{\substack{i,j,k \\ i \geqslant 1}} i\, a_{ijk} x^{i-1} y^j z^k$. The derivatives with respect to y and z are defined similarly. f is said to be *homogeneous of degree m* if $i + j + k = m$ for all i,j,k with $a_{ijk} \neq 0$. Prove the following assertions.

(a) Let t be an indeterminate over $R[x,y,z]$. If $f(x,y,z) \in R[x,y,z]$ is a homogeneous polynomial of degree m, then
$$f(tx,ty,tz) = t^m f(x,y,z) \in R[x,y,z,t]. \qquad (*)$$

(b) Let t be an indeterminate over $R[x,y,z]$ and $f(x,y,z) \in R[x,y,z]$. If $(*)$ holds in $R[x,y,z,t]$, then $f(x,y,z)$ is a homogeneous polynomial of degree m.

(c) If $f(x,y,z) \in R[x,y,z]$ is a homogeneous polynomial of degree m, then
$$f(rx,ry,rz) = r^m f(x,y,z)$$
for all $r \in R$.

(d) If $f(x,y,z) \in \mathbb{Q}[x,y,z]$ and $f(rx,ry,rz) = r^m f(x,y,z)$ for all $r \in \mathbb{Q}$, then $f(x,y,z)$ is a homogeneous polynomial of degree m.

(e) Find a polynomial $f(x,y,z) \in \mathbb{Z}_5[x,y,z]$ such that
$$f(rx,ry,rz) = r^m f(x,y,z) \text{ for all } r \in \mathbb{Z}_5$$

and which is not homogeneous of degree m.

(f) If $f(x,y,z) \in R[x,y,z]$ is homogeneous of degree m, then

$$x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y} + z\frac{\partial f}{\partial y} = mf.$$

5. Let R be a ring and $f \in R[x]$. The derivative of f' is called the *second derivative of* f, and is written as f'' or as $f^{(2)}$. More generally, the $(n+1)$-*st derivative of* f is defined recursively as the derivative of the n-th derivative $f^{(n)}$ of f, and is written as $f^{(n+1)}$. Thus $f^{(n+1)} = (f^{(n)})'$. We write $f^{(1)}$ for f' and $f^{(0)} = f$. Prove that, for any $f,g \in R[x]$, any $c \in R$, any $n \in \mathbb{N}$

$$(f+g)^{(n)} = f^{(n)} + g^{(n)}, \qquad (cf)^{(n)} = cf^{(n)},$$

$$(fg)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} f^{(n-k)}g^{(k)}.$$

6. Let K be a field, f a nonzero polynomial of degree n in $K[x]$ and assume that $(n!)1_K \neq 0$, where 1_K is the identity of K. Show that

$$f(x+y) = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} y^k$$

in $K[x,y]$, where, of course, $\frac{f^{(k)}(x)}{k!}$ means $[(k!)1_K]^{-1}f(x)$.

7. Let p be a prime number and $f \in \mathbb{Z}_p[x]$. Show that $f' = 0$ if and only if $f(x) = g(x^p)$ for some $g \in \mathbb{Z}_p[x]$.

8. Let K be a field. We put $M = Mat_2(K)$ for brevity. Let us recall that the determinant of $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M$ is $ad - bc$ and that $A \in M$ is a unit in M if and only if *det* A is a unit in K.

Let $A(x), B(x) \in M[x]$ be nonzero polynomials and assume that the leading coefficient of $B(x)$ has a nonzero determinant. Show that there are uniquely determined polynomials $Q(x), R(x), Q\dagger(x), R\dagger(x)$ in $M[x]$ such that $A(x) = Q(x)B(x) + R(x)$, $R(x) = 0$ or $deg\ R(x) < deg\ B(x)$. and $A(x) = B(x)Q\dagger(x) + R\dagger(x)$, $R\dagger(x) = 0$ or $deg\ R\dagger(x) < deg\ B(x)$. $Q(x)$ and $R(x)$ are called the *right quotient* and *right remainder*, $Q\dagger(x)$ and $R\dagger(x)$ are called the *left quotient* and *left remainder* when $A(x)$ is divided by $B(x)$.

If $F(x) = F_n x^n + F_{n-1}x^{n-1} + \cdots + F_1 x + F_0 \in M[x]$ and $A \in M$, then

$$F(A) := F_n A^n + F_{n-1}A^{n-1} + \cdots + F_1 A + F_0 \in M$$

is called the *right value of* $F(x)$ *at* A and

$$F\dagger(A) := A^n F_n + A^{n-1}F_{n-1} + \cdots + AF_1 + F_0 \in M$$

is called the *left value of* $F(x)$ *at* A. Prove that the right (resp. left) remainder of $F(x) \in M[x]$, when $F(x)$ is divided by $Ix - A$, is equal to $F(A)$ (resp. $F\dagger(A)$).

9. Let R be a ring and $D_i : R[x] \to R[x]$ be functions ($i = 1,2$) such that

$$D_i(f+g) = D_i f + D_i g \quad D_i(cf) = c\,D_i f, \qquad D_i(fg) = (D_i f)g + fD_i(g)$$

for all $f,g \in R[x]$. Define $D : R[x] \to R[x]$ by

$$Df = D_1(D_2 f) - D_2(D_1 f).$$

Prove that

$$D(f+g) = Df + Dg \quad D(cf) = c\,Df, \qquad D(fg) = (Df)g + fD(g)$$

for all $f,g \in R[x]$.

§36
Fields of Rational Functions

The reader might have missed the familiar quotient rule $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$ in Lemma 35.15. It was missing because $\frac{f}{g}$ is not a polynomial. We now introduce these quotients $\frac{f}{g}$.

36.1 Definition: Let D be an integral domain and $x, x_1, x_2, \ldots, x_n$ indeterminates over D. Then $D[x]$ and $D[x_1, x_2, \ldots, x_n]$ are integral domains (Lemma 33.6, Lemma 33.10). An element in the the field of fractions of $D[x]$ is called a *rational function (in x) over D*. The field of fractions of $D[x]$ will be called the *field of rational functions over D (in x)* and will be denoted by $D(x)$. An element in the the field of fractions of $D[x_1, x_2, \ldots, x_n]$ is called a *rational function (in $x_1, x_2, \ldots, x_n$) over D*. The field of fractions of $D[x_1, x_2, \ldots, x_n]$ will be called the *field of rational functions over D (in $x_1, x_2, \ldots, x_n$)* and will be denoted by $D(x_1, x_2, \ldots, x_n)$.

Thus a rational function over D is a fraction $\frac{f}{g}$ of two polynomials over D, with $g \neq 0$. Two rational functions $\frac{f_1}{g_1}$ and $\frac{f_2}{g_2}$ are equal if and only if the polynomials $f_1 g_2$ and $g_1 f_2$ are equal. Two rational functions $\frac{f_1}{g_1}$ and $\frac{f_2}{g_2}$ are added and multiplied according to the rules

$$\frac{f_1}{g_1} + \frac{f_2}{g_2} = \frac{f_1 g_2 + g_1 f_2}{g_1 g_2}, \qquad \frac{f_1}{g_1}\frac{f_2}{g_2} = \frac{f_1 f_2}{g_1 g_2}.$$

Here g_1 and g_2 are distinct from the zero polynomial over D.

This terminology is unfortunate and misleading, because a rational function is *not* a function in the sense of Definition 3.1. A rational function is *not* a function of the 'rational' kind, whatever that might mean. The technical term we defined is *rational function*, a term

consisting of two words "rational" and "function". The meaning of the words "rational" and "function" do not play any role in Definition 36.1. A rational function is a fraction of polynomials over D. The reader should exercise caution about this point. One should not conclude that

$$\frac{x^2 - 1}{x - 1} \quad \text{and} \quad \frac{x + 1}{1} \qquad \text{in } \mathbb{C}(x)$$

are different rational functions, on grounds that that their domains are different, since the domain of the first one does not contain 1, whereas 1 is in the domain of the second one. Neither of them has a domain, for neither of them is a function. And these rational functions are equal because the polynomials $(x^2 - 1)1$ and $(x - 1)(x + 1)$ in $\mathbb{C}[x]$ are equal.

36.2 Lemma: *Let D be an integral domain and F the field of fractions of D. Let x be an indeterminate over D. Then $D(x) = F(x)$.*

Proof: F consists of the fractions $\frac{a}{b}$, where $a,b \in D$ and $b \neq 0$; and $D(x)$ consists of the fractions

$$\frac{a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0}{b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0},$$

where $a_n,a_{n-1},\ldots,a_1,a_0,b_m,b_{m-1},\ldots,b_1,b_0 \in D$ and the denominator is distinct from the zero polynomial in $D[x]$. Finally, $F(x)$ consists of the fractions

$$\frac{c_n x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0}{d_m x^m + d_{m-1}x^{m-1} + \cdots + d_1 x + d_0},$$

where $c_n,c_{n-1},\ldots,c_1,c_0,d_m,d_{m-1},\ldots,d_1,d_0 \in F$ and the denominator is distinct from the zero polynomial in $F[x]$.

An element of D is identified with the fraction $\frac{a}{1}$ in F (Theorem 31.5), whence $D \subseteq F$. Thus $D[x] \subseteq F[x]$ as sets. Note that two elements $\frac{f(x)}{g(x)}$ and $\frac{p(x)}{q(x)}$ of $D(x)$ are equal in $D(x)$ if and only if $f(x)q(x) = g(x)p(x)$ in $D[x]$, and this holds if and only if $f(x)q(x) = g(x)p(x)$ in $F[x]$, so if and only if $\frac{f(x)}{g(x)}$

and $\frac{p(x)}{q(x)}$ are equal in $F(x)$. Thus every element of $D(x)$ is in $F(x)$ and equality in $D(x)$ coincides with equality in $F(x)$. So $D(x) \subseteq F(x)$.

Next we show $F(x) \subseteq D(x)$. Let $\frac{p(x)}{q(x)} \in F(x)$, with $p(x), q(x) \in F[x]$, $q(x) \neq 0$. Then $p(x) = \sum_{i=0}^{n} \frac{a_i}{b_i} x^i$, $q(x) = \sum_{j=0}^{m} \frac{c_j}{d_j} x^j$, where $a_i, b_i, c_j, d_j \in D$, $b_i \neq 0$, $d_j \neq 0$ for all i,j and not all of c_j are equal to $0 \in D$. We put $b = b_0 b_1 \ldots b_{n-1} b_n$ and $d = d_0 d_1 \ldots d_{m-1} d_m$. Then $dbp(x)$ and $dbq(x)$ are polynomials in $D[x]$, and hence $\frac{p(x)}{q(x)} = \frac{dbp(x)}{dbq(x)} \in D(x)$. So $F(x) \subseteq D(x)$. This proves $D(x) = F(x)$. □

As an illustration of Lemma 36.2, observe that $\frac{\frac{2}{3}x^2 - \frac{1}{7}x + \frac{1}{4}}{\frac{2}{5}x^2 + \frac{1}{3}x - \frac{1}{2}} \in \mathbb{Q}(x)$ is equal to the rational function $\frac{5(56x^2 - 12x + 21)}{14(12x^2 + 10x - 15)}$ in $\mathbb{Z}(x)$.

36.3 Remark: Let D be an integral domain and F the field of fractions of D. Then

$$\begin{aligned} D(x_1,x_2,\ldots,x_n) &= \text{field of fractions of } D[x_1,x_2,\ldots,x_n] \\ &= \text{field of fractions of } D[x_1,x_2,\ldots,x_{n-1}][x_n] \\ &= D[x_1,x_2,\ldots,x_{n-1}](x_n) \\ &= D(x_1,x_2,\ldots,x_{n-1})(x_n) \end{aligned}$$

by Lemma 36.2, with $D[x_1,x_2,\ldots,x_{n-1}]$, $D(x_1,x_2,\ldots,x_{n-1})$, x_n in place of D, F, x, respectively.

Also, we have $D(x_1,x_2,\ldots,x_n) = F(x_1,x_2,\ldots,x_n)$, for this is true when $n = 1$ (Lemma 36.2) and, when it is true for $n = k$, so that $D(x_1,x_2,\ldots,x_k) = F(x_1,x_2,\ldots,x_k)$, it is also true for $n = k + 1$:

$$\begin{aligned} D(x_1,x_2,\ldots,x_k,x_{k+1}) &= D(x_1,x_2,\ldots,x_k)(x_{k+1}) \\ &= F(x_1,x_2,\ldots,x_k)(x_{k+1}) \\ &= F(x_1,x_2,\ldots,x_k,x_{k+1}), \end{aligned}$$

the last equation by the remark above, with F in place of D and $k + 1$ in place of n.

In the remainder of this paragraph, we discuss partial fraction expansions of ratinonal functions.

36.4 Lemma: *Let K be a field and let $f(x)$ be a nonzero polynomial in $K[x]$. Let $q(x)$, $r(x)$ be two nonzero, relatively prime polynomials of positive degree in $K[x]$. Suppose $\deg f(x) < \deg q(x)r(x)$ and suppose that $f(x)$ is relatively prime to $q(x)r(x)$. Then there are uniquely determined nonzero polynomials $a(x)$, $b(x)$ in $K[x]$ such that*

$$a(x)r(x) + b(x)q(x) = f(x), \quad \deg a(x) < \deg q(x), \quad \deg b(x) < \deg r(x).$$

Proof: We first prove the existence of $a(x)$ and $b(x)$. Since $q(x)$, $r(x)$ are relatively prime, there are polynomials $h(x)$, $k(x)$ in $K[x]$ with

$$h(x)r(x) + k(x)q(x) = 1.$$

Multiplying both sides of this equation by $f(x)$ and putting $A(x) = f(x)h(x)$, $B(x) = f(x)k(x)$, we obtain

$$A(x)r(x) + B(x)q(x) = f(x).$$

We now divide $A(x)$ by $q(x)$ and $B(x)$ by $r(x)$:

$$A(x) = s(x)q(x) + a(x), \quad a(x) = 0 \text{ or } \deg a(x) < \deg q(x),$$
$$B(x) = u(x)r(x) + b(x), \quad b(x) = 0 \text{ or } \deg b(x) < \deg r(x).$$

Thus
$$\begin{aligned} a(x)r(x) + b(x)q(x) &= (A(x) - s(x)q(x))r(x) + (B(x) - u(x)r(x))q(x) \\ &= (A(x)r(x) + B(x)q(x)) - (s(x) + u(x))q(x)r(x) \\ &= f(x) - (s(x) + u(x))q(x)r(x). \end{aligned}$$

We claim $s(x) + u(x)$ is the zero polynomial in $K[x]$. Otherwise, we would have
$$\deg (s(x) + u(x)) \geqslant 0,$$
$$\deg (s(x) + u(x))q(x)r(x) \geqslant \deg q(x)r(x),$$
and since by hypothesis $\deg f(x) < \deg q(x)r(x)$,
$$\deg f(x) - (s(x) + u(x))q(x)r(x) \geqslant \deg q(x)r(x),$$

so that $a(x)r(x) + b(x)q(x) \neq 0$; in particular, both $a(x)$ and $b(x)$ cannot be zero. Assume, without loss of generality, that $a(x) \neq 0$ in case one of $a(x)$, $b(x)$ is zero and that $\deg a(x)r(x) \geqslant \deg b(x)q(x)$ in case neither of them is zero. Then we get the contradiction

$$\begin{aligned} deg\ [f(x) - (s(x) + u(x))q(x)r(x)] &= deg\ (a(x)r(x) + b(x)q(x)) \\ &\leqslant deg\ a(x)r(x) \\ &= deg\ a(x) + deg\ r(x) \\ &< deg\ q(x) + deg\ r(x) \\ &= deg\ q(x)r(x). \end{aligned}$$

Thus $s(x) + u(x)$, and consequently $(s(x) + u(x))q(x)r(x)$ is the zero polynomial in $K[x]$. This gives $a(x)r(x) + b(x)q(x) = f(x)$. It remains to show that $a(x)$ and $b(x)$ are distinct from the the zero polynomial in $K[x]$. Both of them cannot be 0, for then $f(x)$ would be also 0, which it is not by hypothesis. If one of them is 0, say if $a(x) = 0$, then $b(x) \neq 0$ and $f(x) = b(x)q(x)$ would not be relatively prime to $q(x)r(x)$ (because $q(x)$ is of positive degree, so not a unit in $K[x]$), against the hypothesis. This proves the existence of $a(x)$, $b(x)$.

It remains to show the uniqueness of $a(x)$ and $b(x)$. If we have also $a_1(x)r(x) + b_1(x)q(x) = f(x)$, $deg\ a_1(x) < deg\ q(x)$, $deg\ b_1(x) < deg\ r(x)$, we obtain $0 = f(x) - f(x) = (a_1(x)r(x) + b_1(x)q(x)) - (a(x)r(x) + b(x)q(x))$

$$= (a(x) - a_1(x))r(x) - (b_1(x) - b(x))q(x),$$

so
$$(a(x) - a_1(x))r(x) = (b_1(x) - b(x))q(x). \qquad (*)$$

Hence

$r(x) \mid (b_1(x) - b(x))q(x)$ in $K[x]$

$r(x) \mid b_1(x) - b(x)$ in $K[x]$ as $r(x)$ and are $q(x)$ relatively prime.

Now $b(x) \neq b_1(x)$ implies $b(x) - b_1(x) \neq 0$ and this gives

$$deg\ r(x) \leqslant deg\ (b_1(x) - b(x)) \leqslant max\{deg\ b_1(x), deg\ b(x)\} < deg\ r(x),$$

a contradiction. Thus $b(x) = b_1(x)$ and we get then $a(x) = a_1(x)$ from (*). So $a(x)$ and $b(x)$ are uniquely determined. □

36.5 Lemma: *Let K be a field and let $\frac{f(x)}{g(x)}$ be a nonzero rational function in $K(x)$, with $deg\ f(x) < deg\ g(x)$. Suppose that $f(x)$ and $g(x)$ are both monic and that $f(x)$ is relatively prime to $g(x)$. Assume $g(x) = q(x)r(x)$, where $q(x)r(x)$ are two relatively prime polynomials of positive degree in $K[x]$. Then there are uniquely determined nonzero polynomials $a(x)$, $b(x)$ in $K[x]$ such that*

$$\frac{f(x)}{g(x)} = \frac{f(x)}{q(x)r(x)} = \frac{a(x)}{q(x)} + \frac{b(x)}{r(x)}$$

and $$deg\ a(x) < deg\ q(x), \quad deg\ b(x) < deg\ r(x).$$

Proof: If $\frac{f(x)}{g(x)}$ is a nonzero rational function in $K(x)$, then $f(x)$ is a nonzero polynomial in $K[x]$, and $f(x)$ is relatively prime to $g(x) = q(x)r(x)$. As $f(x)$ and $g(x)$ are monic, these conditions determine $f(x)$ and $g(x)$ uniquely. The polynomials $q(x)$, $r(x)$ are relatively prime and *deg* $f(x)$ is smaller than *deg* $q(x)r(x)$. So the hypotheses of Lemma 36.4 are satisfied and therefore there are uniquely determined nonzero polynomials $a(x),b(x)$ in $K[x]$ such that

$$f(x) = a(x)r(x) + b(x)q(x),$$

and $$deg\ a(x) < deg\ q(x), \quad deg\ b(x) < deg\ r(x).$$

Dividing both sides of the equation above by $g(x) = q(x)r(x)$, we see that there are uniquely determined nonzero polynomials $a(x),b(x)$ in $K[x]$ such that

$$\frac{f(x)}{g(x)} = \frac{f(x)}{q(x)r(x)} = \frac{a(x)}{q(x)} + \frac{b(x)}{r(x)}$$

and $$deg\ a(x) < deg\ q(x), \quad deg\ b(x) < deg\ r(x).$$ □

By induction on m, we obtain the following lemma.

36.6 Lemma: *Let K be a field and let $\frac{f(x)}{g(x)}$ be a nonzero rational function in $K(x)$, with deg $f(x) <$ deg $g(x)$. Suppose that $f(x)$ and $g(x)$ are both monic and that $f(x)$ is relatively prime to $g(x)$. Assume $g(x) = q_1(x)q_2(x)\ldots q_m(x)$, where $q_1(x), q_2(x), \ldots, q_m(x)$ are pairwise relatively prime monic polynomials of positive degree in $K[x]$. Then there are uniquely determined nonzero polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ in $K[x]$ such that*

$$\frac{f(x)}{g(x)} = \frac{f(x)}{q_1(x)q_2(x)\ldots q_m(x)} = \frac{a_1(x)}{q_1(x)} + \frac{a_2(x)}{q_2(x)} + \cdots + \frac{a_m(x)}{q_m(x)}$$

and $$deg\ a_i(x) < deg\ q_i(x) \quad \text{for all } i = 1,2,\ldots,m.$$

36.7 Lemma: *Let K be a field and x an indeterminate over K.* Let $g(x)$ *be a polynomial in $K[x]$ of degree ≥ 1. Then, for any $f(x) \in K[x]$, there are uniquely determined polynomials $r_0(x), r_1(x), r_2(x), \ldots, r_n(x)$ such that*

$$f(x) = r_0(x) + r_1(x)g(x) + r_2(x)g(x)^2 + \cdots + r_n(x)g(x)^n$$

and

$$r_i(x) = 0 \quad \text{or} \quad \deg\ r_i(x) < \deg\ g(x) \quad \text{for all } i = 1,2,\ldots,n.$$

Proof: From $\deg\ g \geqslant 1$, we know that $g \neq 0$. So we may divide f by g and obtain $f = q_0 g + r_0$, where $q_0, r_0 \in K[x]$, with $r_0 = 0$ or $\deg\ r_0 < \deg\ g$. Here q_0 and r_0 are uniquely determined by f and g (Theorem 34.4) and we have $f = r_0 + q_0 g$. If $q_0 = 0$, we are done (with $n = 0$). Otherwise, since $f = q_0 g + r_0$, $\deg\ g \geqslant 1$ and $r_0 = 0$ or $\deg\ r_0 < \deg\ g$, we have $\deg\ q_0 < \deg\ f$ (Lemma 33.3). We now divide q_0 by g and obtain $q_0 = q_1 g + r_1$, where q_1, $r_1 \in K[x]$, with $r_1 = 0$ or $\deg\ r_1 < \deg\ g$. Here q_1 and r_1 are uniquely determined by q_0 and g (hence by f and g) and $f = r_0 + r_1 g + q_1 g^2$. If $q_1 = 0$, we are done. Otherwise, $\deg\ q_1 < \deg\ q_0$. We then divide q_1 by g and obtain $q_1 = q_2 g + r_2$, where $q_2, r_2 \in K[x]$, with $r_2 = 0$ or $\deg\ r_2 < \deg\ g$. Here q_2 and r_2 are uniquely determined by q_1 and g (hence by f and g) and $f = r_0 + r_1 g + r_2 g^2 + q_2 g^3$. If $q_2 = 0$, we are done. Otherwise, we have $\deg\ q_2 < \deg\ q_1$. We continue this process. As the degrees of $q_0, q_1, q_2, \ldots$ get smaller and smaller, this process cannot go on indefinitely. Sooner or later, we will meet a q_n equal to $0 \in K[x]$. Then, with uniquely determined $r_0, r_1, r_2, \ldots r_n$, we have $f = r_0 + r_1 g + r_2 g^2 + \cdots + r_n g^n$, where $r_i(x) = 0$ or $\deg\ r_i < \deg\ g$ for all $i = 1,2,\ldots,n$. □

In the situation of Lemma 36.7, the unique expression

$$f = r_0 + r_1 g + r_2 g^2 + \cdots + r_n g^n$$

of $f(x)$, where $r_i(x) = 0$ or $\deg\ r_i < \deg\ g$ for all $i = 1,2,\ldots,n$, is called the *g-adic expansion of f*.

36.8 Theorem: *Let K be a field and $\dfrac{p(x)}{q(x)}$ a nonzero rational function in $K(x)$, where $p(x), q(x) \in K[x]$ are relatively prime in $K[x]$. Let u be the leading coefficient of $q(x)$ and let $q(x) = u g_1(x)^{m_1} g_2(x)^{m_2} \ldots g_t(x)^{m_t}$ be the decomposition of $q(x)$ into polynomials irreducible over K, where $g_i(x)$ are monic. Then there are uniquely determined polynomials $G(x)$, $a_1^{(1)}(x), a_2^{(1)}(x), \ldots, a_{m_1}^{(1)}(x), a_1^{(2)}(x), a_2^{(2)}(x), \ldots, a_{m_2}^{(2)}(x), \ldots, a_1^{(t)}(x), a_2^{(t)}(x), \ldots, a_{m_t}^{(t)}(x)$ in $K[x]$ such that*

$$\frac{p(x)}{q(x)} = G(x) + \frac{a_1^{(1)}(x)}{g_1^{1}(x)} + \frac{a_2^{(1)}(x)}{g_1^{2}(x)} + \cdots + \frac{a_{m_1}^{(1)}(x)}{g_1^{m_1}(x)}$$
$$+ \frac{a_1^{(2)}(x)}{g_2^{1}(x)} + \frac{a_2^{(2)}(x)}{g_2^{2}(x)} + \cdots + \frac{a_{m_2}^{(2)}(x)}{g_2^{m_2}(x)}$$
$$+ \cdots\cdots$$
$$+ \frac{a_1^{(t)}(x)}{g_t^{1}(x)} + \frac{a_2^{(t)}(x)}{g_t^{2}(x)} + \cdots + \frac{a_{m_t}^{(t)}(x)}{g_t^{m_t}(x)}$$

and $\deg a_i^{(k)}(x) \leqslant \deg g_k(x)$ *or* $a_i^{(k)}(x) = 0$ *for all* i *and* k.

Proof: We divide $p(x)$ by $q(x)$ and find unique polynomials $G(x)$, $H(x)$ in $K[x]$ with $p(x) = q(x)G(x) + H(x)$, $\deg H(x) < \deg q(x)$ or $H(x) = 0$. In the latter case, everything is proved ($a_i^{(k)}(x) = 0$ for all i and k). If $H(x) \neq 0$, let v be the leading coefficient of $H(x)$ and put $c = v/u$. Then $H(x)$ and $q(x)$ are relatively prime (since $p(x)$ and $q(x)$ are). We have $H(x) = vh(x)$, where $h(x)$ is monic, relatively prime to $q(x)$ and

$$\frac{p(x)}{q(x)} = G(x) + c\frac{h(x)}{q(x)}$$

with $\deg h(x) < \deg q(x)$. We may use Lemma 36.6 and get uniquely determined nonzero polynomials $b_1(x), b_2(x), \ldots, b_t(x)$ in $K[x]$ such that

$$\frac{h(x)}{q(x)} = \frac{b_1(x)}{g_1(x)^{m_1}} + \frac{b_2(x)}{g_2(x)^{m_2}} + \cdots + \frac{b_t(x)}{g_t(x)^{m_t}}$$

and $\deg b_k(x) < \deg g_k(x)^{m_k}$ for all $k = 1,2,\ldots,t$. We put $f_k(x) = cb_k(x)$. Then

$$\frac{p(x)}{q(x)} = G(x) + \frac{f_1(x)}{g_1(x)^{m_1}} + \frac{f_2(x)}{g_2(x)^{m_2}} + \cdots + \frac{f_t(x)}{g_t(x)^{m_t}}$$

and, since c is uniquely determined by $p(x)$ and $q(x)$, the polynomials $f_k(x)$ are also uniquely determined. Since

$$\deg f_k(x) = \deg b_k(x) < \deg g_k(x)^{m_k},$$

in the $g_k(x)$-adic expansion

$$f_k(x) = r_0(x) + r_1(x)g_k(x) + r_2(x)g_k(x)^2 + \cdots + r_n(x)g_k(x)^n$$

of $f_k(x)$, the polynomials $r_s(x) = 0$ for $s \geqslant m_k$. So let

$$f_k(x) = a_1^{(k)}(x)g_k(x)^{m_k-1} + a_2^{(k)}(x)g_k(x)^{m_k-2} + \cdots + a_{m_k-1}^{(k)}(x)g_k(x) + a_{m_k}^{(k)}(x)$$

be the $g_k(x)$-adic expansion of $f_k(x)$. The polynomials $a_1^{(k)}, a_2^{(k)}, \ldots, a_{m_k}^{(k)}$ in $K[x]$ are uniquely determined and $deg\ a_i^{(k)} < deg\ g_k(x)$ or $a_i^{(k)} = 0$ for all $i = 1,2,\ldots,m_k$. Hence, for all $k = 1,2,\ldots,t$, there holds

$$\frac{f_k(x)}{g_k(x)^{m_k}} = \frac{a_1^{(k)}(x)}{g_k^{1}(x)} + \frac{a_2^{(k)}(x)}{g_k^{2}(x)} + \cdots + \frac{a_{m_k}^{(k)}(x)}{g_k^{m_k}(x)}$$

and this completes the proof.

The equation

$$\begin{aligned}\frac{p(x)}{q(x)} = G(x) &+ \frac{a_1^{(1)}(x)}{g_1^{1}(x)} + \frac{a_2^{(1)}(x)}{g_1^{2}(x)} + \cdots + \frac{a_{m_1}^{(1)}(x)}{g_1^{m_1}(x)} \\ &+ \frac{a_1^{(2)}(x)}{g_2^{1}(x)} + \frac{a_2^{(2)}(x)}{g_2^{2}(x)} + \cdots + \frac{a_{m_2}^{(2)}(x)}{g_2^{m_2}(x)} \\ &+ \cdots\cdots \\ &+ \frac{a_1^{(t)}(x)}{g_t^{1}(x)} + \frac{a_2^{(t)}(x)}{g_t^{2}(x)} + \cdots + \frac{a_{m_t}^{(t)}(x)}{g_t^{m_t}(x)}\end{aligned}$$

in Theorem 36.8 is known as the *expansion of* $\frac{p(x)}{q(x)}$ *in partial fractions.*

Exercises

1. Let K be a field. For any nonzero rational function $\frac{f}{g}$ in $K(x)$, we define the *degree of* $\frac{f}{g}$, denoted by $deg\ \frac{f}{g}$, by $deg\ \frac{f}{g} = deg\ f - deg\ g$. Prove that the degree of a rational function is well defined. Can you extend the degree assertions in Lemma 33.3 to rational functions?

2. Let K be a field. For any rational function $\frac{f}{g}$ in $K(x)$, we define the *derivative of* $\frac{f}{g}$, denoted by $\left(\frac{f}{g}\right)'$, by declaring

$$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}.$$

Prove that differentiation is well defined, i.e., prove that $\frac{f}{g} = \frac{a}{b}$ implies $(\frac{f}{g})' = (\frac{a}{b})'$.

3. Extend Lemma 35.15 and Lemma 35.16 to derivatives of rational functions in one indeterminate over a field.

4. Expand $\frac{2x^3 + 3x^2 + 8x + 6}{(x^3 + 3x + 3)(x^2 + 2x + 3)} \in \mathbb{Q}(x)$ and

$$\frac{4x^3 + 3x^2 + x + 2}{x^5 + 4x^4 + 4x^3 + 2x + 2} \in \mathbb{Z}_5(x)$$

in partial fractions.

5. Let K be a field and let $a_1, a_2, \ldots, a_m$ be pairwise distinct elements in K. Put $g(x) = (x - a_1)(x - a_2)\ldots(x - a_m)$ and let $f(x)$ be a nonzero polynomial in $K[x]$ with $deg\ f(x) < m$. Show that

$$\frac{f(x)}{g(x)} = \sum_{i=1}^{m} \frac{f(a_i)/g'(a_i)}{x - a_i}.$$

§37
Irreducibility Criteria

In this paragraph, we develope some sufficient conditions for a polynomial to be irreducible. In general, given a specific polynomial, it is extremely difficult to determine whether it is irreducible. This is not surprising when we remember that it is also exceedingly difficult to determine whether a given specific integer is prime.

We start with Eisenstein's criterion, which is very simple to use.

37.1 Lemma (Eisenstein's criterion): *Let D be a unique factorization domain and let*

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

be a nonzero polynomial in $D[x]$ with $C(f) \approx 1$. If there is a prime (irreducible) element p in D such that

$$p \nmid a_n,$$
$$p \mid a_{n-1}, \ldots\ldots, p \mid a_1, p \mid a_0,$$
$$p^2 \nmid a_0,$$

then f is irreducible over D.

Proof: Suppose, by way of contradiction, that $f(x)$ is reducible over D. Then its proper factors must have degrees > 0, because $C(f) \approx 1$. Assume $f(x) = g(x)h(x)$, where

$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0 \qquad (b_m \neq 0,\ m \geqslant 1)$$
$$h(x) = c_k x^k + c_{k-1}x^{k-1} + \cdots + c_1 x + c_0 \qquad (c_k \neq 0,\ k \geqslant 1)$$

are polynomials in $D[x]$.

Then $a_0 = b_0 c_0$. Since $p \mid a_0$ and so $p \mid b_0 c_0$ by hypothesis and p is prime, we see $p \mid b_0$ or $p \mid c_0$. Here both $p \mid b_0$ and $p \mid c_0$ cannot be simultaneously true, for then we would have $p^2 \mid b_0 c_0$, so $p^2 \mid a_0$, against our hypothesis. Thus one and only one of $p \mid b_0$, $p \mid c_0$ is true. Let us assume, without loss of generality, that $p \mid b_0$ and $p \nmid c_0$.

Also $a_n = b_m c_k$. Since $p \nmid a_n$ and so $p \nmid b_m c_k$ by hypothesis, we have $p \nmid b_m$. Thus $p \mid b_0$ and $p \nmid b_m$. Let r be the smallest index for which the coefficient b_r in $g(x)$ is not divisible by p, so that

$$p \mid b_0,\ p \mid b_1,\ \ldots,\ p \mid b_{r-1},\ p \nmid b_r \qquad (*)$$

(possibly $r = 1$ or $r = m$).

Now $a_r = (b_0 c_r + b_1 c_{r-1} + \ldots + b_{r-1} c_1) + b_r c_0$, and $r \leqslant m < m + k = n$. So $p \mid a_r$ by hypothesis and p divides the expression in ( ) by (*), so $p \mid b_r c_0$. Then, since p is prime, this forces $p \mid b_r$ or $p \mid c_0$, whereas $p \nmid b_r$ and $p \nmid c_0$. This contradiction completes the proof. □

37.2 Examples: (a) $x^5 + 5x + 5 \in \mathbb{Z}[x]$ is irreducible over $\mathbb{Z}$, because its content is 1 and

$$5 \nmid 1,$$
$$5 \mid 0,\ 5 \mid 0,\ 5 \mid 0,\ 5 \mid 5,\ 5 \mid 5,$$
$$5^2 \nmid 5.$$

(b) Let $D = \mathbb{Z}[i]$ and $f(x) = 3x^3 + 2x^2 + (4 - 2i)x + (1 + i) \in D[x]$. Then D is a unique factorization domain and $C(f) \approx 1$. Moreover $1 + i \in D$ is a prime element in D and

$$1 + i \nmid 3$$
$$1 + i \mid 2,\quad 1 + i \mid 4 - 2i,\quad 1 + i \mid 1 + i,$$
$$(1 + i)^2 \nmid 1 + i.$$

Hence $f(x)$ is irreducible over D.

(c) Let D be a unique factorization domain and $g(x,y) = x^n - y \in (D[y])[x]$. The content of g is $1 \in D[y]$, since g is in fact a monic polynomial. Also, y is irreducible in $D[y]$ and

$$y \nmid 1$$
$$y \mid 0,\ y \mid 0,\ \ldots,\ y \mid 0,\ y \mid -y,$$
$$y^2 \nmid -y,$$

hence $g(x,y) = x^n + 0x^{n-1} + 0x^{n-2} + \cdots + 0x - y \in (D[y])[x]$ is irreducible over $D[y]$.

(d) Let $p \in \mathbb{N}$ be a prime number and $\Phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1 \in \mathbb{Z}[x]$. The polynomial $\Phi_p(x)$ is known as the p-th cyclotomic polynomial. We show that $\Phi_p(x)$ is irreducible over $\mathbb{Z}$. Eisenstein's criterion is not directly applicable, but we observe that

$$(x - 1)\Phi_p(x) = x^p - 1,$$

and, when we substitute $x + 1$ for x in both sides of this equation, we get

$$x\Phi_p(x+1) = (x+1)^p - 1 = \sum_{k=0}^{p-1} \binom{p}{k} x^{p-k}$$

by the binomial theorem (Theorem 29.16), so

$$\Phi_p(x+1) = x^{p-1} + \binom{p}{1}x^{p-2} + \binom{p}{2}x^{p-3} + \cdots + \binom{p}{p-1}$$

and we will try to apply Eisenstein's criterion to this polynomial. We note $p|p!$, so $p|(p-k)!k!\binom{p}{k}$. Since p is relatively prime to $(p-k)!\,k!$ when $1 \leqslant k \leqslant p-1$, Theorem 5.12 gives $p|\binom{p}{k}$ for $k = 1,2,\ldots,p-1$. So

$$p \nmid 1,$$

$$p|\binom{p}{1}, \qquad p|\binom{p}{2}, \ldots, \; p|\binom{p}{p-1},$$

$$p^2 \nmid \binom{p}{p-1},$$

and the content of $\Phi_p(x+1) = 1$. Hence $\Phi_p(x+1)$ is irreducible over $\mathbb{Z}$.

This implies that $\Phi_p(x)$ is also irreducible over $\mathbb{Z}$, since $\Phi_p(x)$ is clearly not a unit in $\mathbb{Z}[x]$ and any factorization $\Phi_p(x) = f(x)g(x)$ of $\Phi_p(x)$ into nonunit polynomials $f(x), g(x) \in \mathbb{Z}[x]$ would give a factorization $\Phi_p(x+1) = f(x+1)g(x+1) = f_1(x)g_1(x)$ of $\Phi_p(x+1)$ into nonunit polynomials $f_1(x)$, $g_1(x)$ in $\mathbb{Z}[x]$, contrary to the irreducibility of $\Phi_p(x+1)$ over $\mathbb{Z}$.

The argument in the last example can be generalized.

37.3 Lemma: *Let D be an integral domain, α a unit in D and let β be an arbitrary element of D.*

(1) *The mapping $T: D[x] \to D[x]$ is a ring isomorphism such that $\gamma T = \gamma$*

$$f(x) \to f(\alpha x + \beta)$$

for all $\gamma \in D$.

(2) *$\deg f(\alpha x + \beta) = \deg f(x)$ for any $f(x) \in D[x]\backslash\{0\}$ (that is, T preserves degrees of polynomials).*

(3) *$f(x)$ is irreducible over D if and only if $f(\alpha x + \beta)$ is irreducible over D.*

(4) *If, in addition, D is a unique factorization domain, then $C(f(x)) \approx C(f(\alpha x + \beta))$ for any $f(x) \in D[x]\backslash\{0\}$ (that is, T preserves contents of polynomials).*

Proof: (1) The mapping $T: f(x) \to f(\alpha x + \beta)$ is just the substitution homomorphism $T_{\alpha x+\beta}$ (Lemma 35.3 with D, $D[x]$, $\alpha x + \beta$ in place of R, S, s, respectively). We are to show that T is one-to-one and onto. To this end, we need only find an inverse of T (Theorem 3.17(2)). This is quite easy. We are tempted to substitute $(x - \beta)/\alpha$ for x. This idea is correct, but we must formulate it properly. Since α is a unit in D, there is an inverse α^{-1} of α in D, and we put $S: D[x] \longrightarrow D[x]$. Then we have

$$f(x) \to f(\alpha^{-1}(x - \beta))$$

$$f(x)TS = f(\alpha x + \beta)S = f\big(\alpha(\alpha^{-1}(x - \beta)) + \beta\big) = f(x)$$
$$f(x)ST = f(\alpha^{-1}(x - \beta))T = f\big(\alpha^{-1}((\alpha x + \beta) - \beta)\big) = f(x)$$

for all $f(x) \in D[x]$. Hence $TS = \iota_{D[x]} = ST$ and T is therefore an isomorphism. Finally, polynomials of degree 0 and the polynomial $0 \in D[x]$ are not effected by the substitution $x \to \alpha x + \beta$ and so $\gamma T = \gamma$ for all $\gamma \in D$.

(2) For any $f(x) \in D[x]\backslash\{0\}$, if $deg\ f = n$ and

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with $a_n \neq 0$, we have

$$\begin{aligned} f(\alpha x + \beta) &= a_n(\alpha x + \beta)^n + a_{n-1}(\alpha x + \beta)^{n-1} + \cdots + a_1(\alpha x + \beta) + a_0 \\ &= a_n \alpha^n x^n + \text{terms of lower degree,} \end{aligned}$$

with $a_n \alpha^n \neq 0$ as the leading coefficient. So $deg\ f(\alpha x + \beta) = n$, as claimed.

(3) If $f(x) \in D[x]\backslash\{0\}$ is not irreducible over D, then either $f(x)$ is a unit in $D[x]$, hence $f(x) \in D$ is a unit in D and $f(\alpha x + \beta) = f(x)$ (by part (1)) is also a unit in D and in $D[x]$; or $f(x) = g(x)h(x)$ for some polynomials $g(x)$, $h(x)$ in $D[x]$ with $1 \leqslant deg\ g(x) < deg\ f(x)$, and then $f(\alpha x + \beta) = g(\alpha x + \beta)h(\alpha x + \beta)$ with $g(\alpha x + \beta)$, $h(\alpha x + \beta) \in D[x]$ and $1 \leqslant deg\ g(x) = deg\ g(\alpha x + \beta) = deg\ g(x) < deg\ f(x) = deg\ f(\alpha x + \beta)$ (by part (2)), and thus $f(\alpha x + \beta)$ has a proper divisor. In either case, $f(\alpha x + \beta)$ is not irreducible over D.

Repeating the same argument for the substitution $x \to \alpha^{-1}(x - \beta)$, we conclude: if $f(\alpha x + \beta)$ is not irreducible over D, then $f(x)$ is not irreducible over D.

(4) Suppose now that D is a unique factorization domain, that $f(x) =$

$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, and that $C(f(x)) \approx \gamma$. Then

$$f(\alpha x + \beta) = \binom{n}{0} a_n \alpha^n x^n + \left(\binom{n}{1} a_n \alpha^{n-1}\beta + \binom{n}{0} a_{n-1}\alpha^{n-1} \right) x^{n-1}$$

$$+ \left(\binom{n}{2} a_n \alpha^{n-2}\beta^2 + \binom{n}{1} a_{n-1}\alpha^{n-1}\beta + \binom{n}{0} a_{n-2}\alpha^{n-2} \right) x^{n-2} + \cdots .$$

A content δ of $f(\alpha x + \beta)$ divides $\binom{n}{0} a_n \alpha^n$, hence $\delta | a_n$ (α and α^n is a unit); and δ divides the coefficient of x^{n-1}, hence $\delta | \binom{n}{0} a_{n-1}\alpha^{n-1}$, hence $\delta | a_{n-1}$; and δ divides the coefficient of x^{n-2}, hence $\delta | \binom{n}{0} a_{n-2}\alpha^{n-2}$, hence $\delta | a_{n-2}$; etc. Proceeding in this way, we see that δ divides all the coefficients of $f(x)$. Since $\gamma \approx C(f(x))$, we obtain $\delta | \gamma$. The same argument with $f(\alpha x + \beta)$, $f(x)$, T^{-1} in place of $f(x)$, $f(\alpha x + \beta)$, T shows that $\gamma | \delta$. Thus $\delta \approx \gamma$, as was to be proved. □

When $C(f(x)) \approx 1$ but the divisibility conditions in Eisenstein's criterion are not satisfied, we might attempt to find a unit α and an element β so that $f(\alpha x + \beta)$ will satisfy the divisibility conditions. If we succeed in finding such α, β, then $f(\alpha x + \beta)$ will be irreducible by Eisenstein's criterion (as $C(f(\alpha x + \beta)) \approx 1$ by Lemma 37.3(4)) and $f(x)$ will be irreducible, too (by Lemma 37.3(3)). This is what we did in Example 37.2(d).

Eisenstein's criterion is a sufficient condition for irreducibility. It is not necesary, even if we extend it using Lemma 37.3(3). That is to say, $f(x)$ may be irreducible and yet, for all units α in D and for all elements β in D, the polynomial $f(\alpha x + \beta)$ may fail to satisfy the divisibility conditions in Eisenstein's criterion. In fact, a closer study of its proof reveals that we are essentially reading the polynomials mod Dp, i.e., we are taking the images of polynomials in $D[x]$ under the mapping $\hat{v}: D[x] \to (D/Dp)[x]$ (see Lemma 33.7).

37.4 Lemma: *Let D be an integral domain and let K be a field. Let*

$\varphi: D \to K$ be a ring homomorphism and let $\hat{\varphi}: D \to K$ be the homomorphism of Lemma 33.7.
(1) If $f \in D[x]$ and $f = gh$ with $g,h \in D[x]$, then $f\hat{\varphi} = g\hat{\varphi}h\hat{\varphi}$.
(2) If $f \in D[x]\backslash D$, deg f = deg $f\hat{\varphi}$ and $f\hat{\varphi}$ is irreducible in $K[x]$, then f has no divisors g in $D[x]$ such that $0 <$ deg $g <$ deg f.

Proof: (1) This follows from the fact that $\hat{\varphi}$ is a homomorphism.

(2) Suppose, on the contrary, that $f = gh$ in $D[x]$, with $0 <$ deg $g <$ deg f. Then $f\hat{\varphi} = g\hat{\varphi}h\hat{\varphi}$ by (1). Since $f\hat{\varphi}$ is irreducible in $K[x]$, $f\hat{\varphi} \neq 0$, so $g\hat{\varphi} \neq 0 \neq h\hat{\varphi}$ and either deg $g\hat{\varphi} = 0$ or deg $h\hat{\varphi} = 0$. We get then

$$\begin{aligned} \text{deg } f\hat{\varphi} &= \text{deg } g\hat{\varphi}h\hat{\varphi} = \text{deg } g\hat{\varphi} + \text{deg } h\hat{\varphi} \\ &\leqslant \text{deg } g + \text{deg } h\hat{\varphi} \leqslant \text{deg } g + \text{deg } h \\ &\leqslant \text{deg } g + \text{deg } h = \text{deg } gh = \text{deg } f = \text{deg } f\hat{\varphi}, \end{aligned}$$

which forces deg $g\hat{\varphi}$ = deg g and deg $h\hat{\varphi}$ = deg h. Thus either deg $g = 0$ or deg $h = 0$, and so either $0 =$ deg g or deg $g =$ deg f, against our hypothesis $0 <$ deg $g <$ deg f. □

In Lemma 37.4, we relaxed the hypothesis on $C(f)$ that was imposed in Eisenstein's criterion. We pay for it, of course. Notice we did *not* claim that f is irreducible over D. We claimed only that f has no proper factor of positive degree less than deg f. Here f may have proper divisors, but any factorization of f in $D[x]$ has the form $f = \alpha f_1$, where $\alpha \in D$ and deg f_1 = deg f.

37.5 Examples: (a) Let $q(x) = x^3 + x + \overline{1} = \overline{1}x^3 + \overline{1}x + \overline{1} \in \mathbb{Z}_2[x]$. If $q(x)$ were reducible in $\mathbb{Z}_2[x]$, it would have a factor of degree $\leqslant 3/2$, so a factor of degree 1. So $q(x)$ would have a root in $\mathbb{Z}_2 = \{\overline{0},\overline{1}\}$ by the factor theorem (Theorem 35.6). But $q(\overline{0}) = \overline{1} \neq \overline{0}$ and $q(\overline{1}) = \overline{1} \neq \overline{0}$, so $q(x)$ is irreducible in $\mathbb{Z}_2[x]$.

Let $f(x) = x^3 + 2x^2 + x + 7 \in \mathbb{Z}[x]$. Under the mapping $\hat{\nu}: \mathbb{Z}[x] \to \mathbb{Z}_2[x]$, where $\nu: \mathbb{Z} \to \mathbb{Z}_2$ is the natural homomorphism, we have

$$f\hat{\nu} = \overline{1}x^3 + \overline{2}x^2 + \overline{1}x + \overline{7} = x^3 + x + \overline{1} = q(x) \in \mathbb{Z}_2[x],$$

and so $f\hat{\nu}$ is irreducible over $\mathbb{Z}_2$. By Lemma 37.4(2), f has no polynomial divisors of degree 1, nor of degree 2. Since f does not have any divisors of degree 0 either ($C(f) \approx 1$), f is irreducible over $\mathbb{Z}$.

(b) Lemma 37.4 can be useful even if $f\hat{\nu}$ is not irreducible. The factorization of $f\hat{\varphi}$ in $K[x]$ gives us information about possible factors of f in $D[x]$ and restricts their number drastically.

As an illustration, consider $f(x) = x^5 + 5x^4 + 4x^3 + 16x^2 + 8x + 1 \in \mathbb{Z}[x]$. Under $\hat{\nu}: \mathbb{Z}[x] \to \mathbb{Z}_3[x]$, where $\nu: \mathbb{Z} \to \mathbb{Z}_3$ is the natural homomorphism, we have (we drop the bars for ease of notation)

$$\begin{aligned} f\hat{\nu} &= x^5 + 2x^4 + x^3 + x^2 + 2x + 1 \in \mathbb{Z}_3[x] \\ &= (x^2 + 2x + 1)(x^3 + 1) \\ &= (x + 1)^2(x + 1)(x^2 - x + 1) \\ &= (x + 1)^2(x + 1)(x^2 + 2x + 1) \\ &= (x + 1)^5, \end{aligned}$$

so any monic factor g of f in $\mathbb{Z}[x]$ with $1 \leqslant deg\ g \leqslant 2$ satisfies

$$g\hat{\nu} = x + 1 \in \mathbb{Z}_3[x] \quad \text{or} \quad g\hat{\nu} = (x + 1)^2 \in \mathbb{Z}_3[x]$$

($\mathbb{Z}_3[x]$ is a unique factorization domain).

Does $f \in \mathbb{Z}[x]$ have a divisor of degree one? If it had, it would have a rational root, and that root would be 1 or -1 by Theorem 35.10. Since $f(1) = 35 \neq 0$ and $f(-1) = 9 \neq 0$, f has no rational root, and f has no divisor of degree one.

Does $f \in \mathbb{Z}[x]$ have a divisor of degree two? If f has a monic divisor $g = g(x) = ax^2 + bx + c \in \mathbb{Z}[x]$ of degree two, then $g\hat{\nu} = x^2 + 2x + 1 \in \mathbb{Z}_3[x]$, and so $a \equiv 1$, $b \equiv 2$, $c \equiv 1 \pmod 3$. Besides, a divides the leading coefficient of f, and c divides the constant term in f: thus $a|1$ and $c|1$. So $a = \mp 1$ and $c = \mp 1$. Without restricting generality, we may assume $a = 1$. The possible monic factors of f of second degree are therefore to be found among

$$g_m(x) = x^2 + (3m + 2)x + 1, \qquad h_m(x) = x^2 + (3m + 2)x - 1 \qquad (m \in \mathbb{Z}).$$

We check if any g_m or h_m divides f. Supposing $g_m(x)|f(x)$ in $\mathbb{Z}[x]$, we get

$$\begin{aligned} & g_m(1)|f(1) && \text{in } \mathbb{Z} \\ & 3m + 4 \mid 35 \\ & 3m + 4 \in \{1,5,7,35,-1,-5,-7,-35\} \\ & 3m + 4 = 1,7,-5,-35 \end{aligned}$$

$$3m + 2 = -1, 5, -7, -37$$

$$g_m(x) = x^2 - x + 1 \quad \text{or} \quad x^2 + 5x + 1 \quad \text{or} \quad x^2 + -7x + 1 \quad \text{or} \quad x^2 + -37x + 1.$$

Testing these four polynomials in turn, we find $x^2 - x + 1$ does not divide $f(x)$, and $x^2 + 5x + 1$ divides $f(x)$; in fact $f(x) = (x^2 + 5x + 1)(x^3 + 3x + 1)$. [If none of the four polynomials divided $f(x)$, we would repeat the argument with h_m. In this way, we would find a divisor of $f(x)$ or we would show that $f(x)$ is irreducible.]

(c) Lemma 37.4 gives a very elegant proof of Eisenstein's criterion. in case the underlying ring is a principal ideal domain. Suppose D is a principal ideal domain and

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

is a nonzero polynomial in $D[x]$ with $C(f) \approx 1$ and p is a prime element D such that

$$p \nmid a_n,$$
$$p \mid a_{n-1}, \ldots\ldots, p \mid a_1, p \mid a_0,$$
$$p^2 \nmid a_0.$$

Since p is irreducible, the factor ring D/Dp is a field (Theorem 32.25). We can use Lemma 37.4 with the natural homomorphism $\nu: D \to D/Dp$. The divisibility conditions on the coefficients of f imply

$$f\hat{\nu} = (a_n \nu)x^n, \qquad a_n \nu \in D/Dp, \qquad a_n \nu \neq 0.$$

If f had a proper factorization $f = gh$ in $D[x]$, where $0 < deg\ g < n$, we would get

$$g\hat{\nu}h\hat{\nu} = f\hat{\nu} = (a_n \nu)x^n$$

hence $g\hat{\nu} = b\nu x^r$, $h\hat{\nu} = c\nu x^s$ with $0 < r < n$, $0 < s < n$ and $b\nu c\nu = a_n \nu$. Then the constant terms of g and h would be divisible by p, and p^2 would be divide their product a_0, contrary to the hypothesis. Hence f is irreducible over D.

The idea (that $g_m(x) | f(x) \Rightarrow g_m(1) | f(1)$) in Example 37.5(b) has been exploited by L. Kronecker (1823-1891). Let D be an integral domain and let $f(x)$ be an arbitrary nonzero polynomial in $D[x]$. To find out whether f is irreducible over D, one must check whether $g | f$ or $g \nmid f$ holds for *all* polynomials g with $deg\ g < deg\ f$. If D happens to be finite (and thus a field; Theorem 31.1), there are finitely many g's with $deg\ g < deg\ f$; and the

question whether f is irreducible over D can be decided by checking $g|f$ for these the finitely many g's. If D is not finite, this argument does not work, and we must, so it seems, check if $g|f$ for infinitely many polynomials $g \in D[x]$. Kronecker showed that, if D is a unique factorization domain which possesses a finite number of units and if we have a method for finding the irreducible factors of any given nonzero element of D, then, to find out whether a given nonzero polynomial is irreducible or not, we need check $g|f$ for only a finite number of polynomials g in $D[x]$.

His idea is that, if $g(x)|f(x)$ in $D[x]$, then $g(a)|f(a)$ in D for any $a \in D$, and that a polynomial g is determined uniquely if its values are known at more than *deg* g elements of D (Lagrange's interpolation formula).

Let D be an infinite unique factorization domain. Asume there are finitely many units in D, and assume that there is a method for finding the irreducible factors of any given nonzero element of D. Let f be a nonzero polynomial in $D[x]$ of degree n. If $n = 0$, then $f \in D$ and we can find the irreducible factors of f in D by assumption. If $n = 1$, then $f = cf_1$, where $c \approx C(f)$ and f_1 is an irreducible polynomial in $D[x]$. The irreducible factors of $c \in D$ can be found by assumption, and thus the irreducible factors of f, too, can be found effectively. If $n \geqslant 2$ and f is reducible, there is a factor $g \in D[x]$ of f with *deg* $g \leqslant n/2$ (Lemma 33.3(3)). We put $m := [n/2]$. We take $m + 1$ distinct elements $a_0, a_1, a_2, \ldots, a_m$ from D and evaluate $f(a_0), f(a_1), f(a_2), \ldots f(a_m) \in D$. If any $f(a_i)$ happens to be $0 \in D$, then $x - a_i$ is a factor of f (Theorem 35.6). Therefore we may assume that $f(a_0), f(a_1), f(a_2), \ldots f(a_m)$ are all distinct from zero. Each one of them has finitely many divisors in D, because D is a unique factorization domain and D has finitely many units. There is asumed to be a method of finding these divisors. Let N_i be the number of factors of $f(a_i)$. A factor g of $f \in D[x]$ with *deg* $g \leqslant m$ satisfies one of the $N_0 N_1 N_2 \ldots N_m$ systems of equations

$$g(a_0) = c_0, \quad g(a_1) = c_1, \quad g(a_2) = c_2, \quad \ldots, \quad g(a_m) = c_m, \qquad (\dagger)$$

where $c_0, c_1, c_2, \ldots, c_m$ run independently over the divisors of the elements $f(a_0), f(a_1), f(a_2), \ldots f(a_m)$, respectively. For each one of these $N_0 N_1 N_2 \ldots N_m$ choices of $c_0, c_1, c_2, \ldots, c_m$, we build the unique polynomial g satisfying (†). This is done by Lagrange's interpolation formula; but this formula requires that the underlying ring be in fact a field. Thus Lagrange's interpolation formula gives us a list of $N_0 N_1 N_2 \ldots N_m$ polynomials g in $F[x]$,

where F is the field of fractions of D, one for each choice $c_0, c_1, c_2, \ldots, c_m$ of the divisors of $f(a_0), f(a_1), f(a_2), \ldots f(a_m)$.

From this list of polynomials, we delete those which are not in $D[x]$. If any polynomial g remains, we divide f by g in $F[x]$. Then $f = qg + r$, with $q, r \in F[x]$. If $r \neq 0$ or $r = 0$ but $q \notin D[x]$, we delete g from our list. We delete g from our list also the the polynomials which are units in D. If any polynomial g survives, it is a factor of f. Otherwise, f is irreducible over D.

When a proper divisor g of f is found in this way, the same procedure can be applied to g and f/g. Repeating this process, we can find all irreducible factors of f.

$\mathbb{Z}$ satisfies the conditions imposed on D in Kronecker's method. Thus the irreducibility of a polynomial in $\mathbb{Z}[x]$ can be determined effectively. This in turn implies that the irreducibility of a polynomial in $\mathbb{Z}[x][y]$ can be determined effctively. By repeated application of Kronecker's method, we can always decide whether a given polynomial in $\mathbb{Z}[x_1, x_2, \ldots, x_n]$ is irreducible or reducible. The same holds for polynomials in the rings $\mathbb{Z}[i][x_1, x_2, \ldots, x_n]$ and $\mathbb{Z}[\omega][x_1, x_2, \ldots, x_n]$.

Kronecker's method is very long and very cumbersome in any specific case. However, it is important philosophically, because it assures that the irreducibility or reducibility of a polynomial can be determined effectively in a finite number of steps.

Exercises

1. Using Eisenstein's criterion, show that the following polynomials are irreducible over the rings indicated:

$x^4 - 6x^3 + 24x^2 - 30x + 14$ over $\mathbb{Z}$,

$x^4 + 6x^3 - 42x^2 + 57x + 78$ over $\mathbb{Z}$,

$3x^5 + (21 - i)x^4 + (14 - 5i)x^3 + (-10 + 11i)$ over $\mathbb{Z}[i]$,

$x^5 - 7x^4 + (3 + 2\omega)x^3 + (2 - \omega)x + (1 - 4\omega)$ over $\mathbb{Z}[\omega]$.

2. Let $f = x^6 - 2x^5 + 3x^4 - 2x^3 + 3x^2 - 2x + 2 \in \mathbb{Z}[x]$. Either prove that f is irreducible over $\mathbb{Z}$ or find all irreducible factors of f in $\mathbb{Z}[x]$.

3. Do Ex. 2 for the polynomials $x^4 - 2x^3 - 2x^2 + 15x + 30$ and $x^5 + 8x^4 + 25x^3 + 39x^2 + 30x + 7$ in $\mathbb{Z}[x]$.

§38
Symmetric Polynomials

Let D be an integral domain and let $f(x_1,x_2,\ldots,x_m) \in D[x_1,x_2,\ldots,x_m]$. For any permutation $\sigma = \begin{pmatrix} 1 & 2 & \ldots & m \\ i_1 & i_2 & \ldots & i_m \end{pmatrix}$ in S_m, the value of f at $(x_{i_1},x_{i_2},\ldots,x_{i_m})$ is a polynomial $f(x_{i_1},x_{i_2},\ldots,x_{i_m})$ in $D[x_1,x_2,\ldots,x_m]$, which we can shortly denote by f^σ (Definition 35.19). For example, if $f(x,y,z) = x^2 + y^2 - xz$ in $\mathbb{Z}[x,y,z]$, then $f(z,x,y) = z^2 + x^2 - zy$; and if $g(x,y) = x^2 - xy + y^3$ in $\mathbb{Z}[x,y]$, then $g(y,x) = y^2 - yx + x^3 \in \mathbb{Z}[x,y]$. In general, $f(x_{i_1},x_{i_2},\ldots,x_{i_m})$ will be a polynomial distinct from $f(x_1,x_2,\ldots,x_m)$.

38.1 Definition: Let D be an integral domain and let $f(x_1,x_2,\ldots,x_m)$ be a polynomial in $D[x_1,x_2,\ldots,x_m]$. If $f(x_{i_1},x_{i_2},\ldots,x_{i_m}) = f(x_1,x_2,\ldots,x_m)$ for all permutations $\sigma = \begin{pmatrix} 1 & 2 & \ldots & m \\ i_1 & i_2 & \ldots & i_m \end{pmatrix}$ in S_m, then $f = f(x_1,x_2,\ldots,x_m)$ is called a *symmetric polynomial* in $D[x_1,x_2,\ldots,x_m]$. We also say that $f(x_1,x_2,\ldots,x_m)$ is *symmetric* in the indeterminates $x_1,x_2,\ldots,x_m$.

The polynomials $x + y$, xy, $x^2 + y^2$, $x^3 + y^3$ are symmetric polynomials in $D[x,y]$. Also, the polynomials $x^2 + y^2 + z^2$ and $xy + yz + zx$ are symmetric polynomials in $D[x,y,z]$.

The sum, difference and product of symmetric polynomials are symmetric polynomials. Indeed, if $f(x_1,x_2,\ldots,x_m)$ and $g(x_1,x_2,\ldots,x_m)$ are symmetric polynomials in $D[x_1,x_2,\ldots,x_m]$, and if

$$h(x_1,x_2,\ldots,x_m) = f(x_1,x_2,\ldots,x_m) + g(x_1,x_2,\ldots,x_m)$$

is their sum, then, for any permutation $\begin{pmatrix} 1 & 2 & \ldots & m \\ i_1 & i_2 & \ldots & i_m \end{pmatrix}$ in S_m, we have

$$\begin{aligned} h(x_{i_1},x_{i_2},\ldots,x_{i_m}) &= f(x_{i_1},x_{i_2},\ldots,x_{i_m}) + g(x_{i_1},x_{i_2},\ldots,x_{i_m}) \\ &= f(x_1,x_2,\ldots,x_m) + g(x_1,x_2,\ldots,x_m) \\ &= h(x_1,x_2,\ldots,x_m), \end{aligned}$$

and so $h(x_1,x_2,\ldots,x_m)$ is a symmetric polynomial. The same argument works also when $h = f - g$ and $h = fg$. This proves

38.2 Lemma: *Let D be an integral domain. The symmetric polynomials in $D[x_1,x_2,\ldots,x_m]$ form a subring of $D[x_1,x_2,\ldots,x_m]$.* □

We introduce a new indeterminate t and consider the polynomial

$$f(t) = (t - x_1)(t - x_2)\ldots(t - x_m) \text{ in } D[x_1,x_2,\ldots,x_m][t].$$

We see that $x_1,x_2,\ldots,x_m \in D[x_1,x_2,\ldots,x_m]$ are the roots of $f(t)$. We have

$$f(t) = t^m - \sigma_1(x_1,x_2,\ldots,x_m)t^{m-1} + \sigma_2(x_1,x_2,\ldots,x_m)t^{m-2} - + \cdots + (-1)^m\sigma_m(x_1,x_2,\ldots,x_m)$$

for some $\sigma_1,\sigma_2,\ldots,\sigma_m$ in $D[x_1,x_2,\ldots,x_m]$. Since

$$f(t) = (t - x_{i_1})(t - x_{i_2})\ldots(t - x_{i_m})$$
$$= t^m - \sigma_1(x_{i_1},x_{i_2},\ldots,x_{i_m})t^{m-1} + \sigma_2(x_{i_1},x_{i_2},\ldots,x_{i_m})t^{m-2} - + \cdots + (-1)^m\sigma_m(x_{i_1},x_{i_2},\ldots,x_{i_m})$$

for any permutation $\begin{pmatrix} 1 & 2 & \ldots & m \\ i_1 & i_2 & \ldots & i_m \end{pmatrix}$ in S_m, we have

$$\sigma_j(x_{i_1},x_{i_2},\ldots,x_{i_m}) = \sigma_j(x_1,x_2,\ldots,x_m) \quad \text{for all } j = 1,2,\ldots,m.$$

Thus $\sigma_1,\sigma_2,\ldots,\sigma_m$ are symmetric polynomials in $D[x_1,x_2,\ldots,x_m]$.

38.3 Definition: Let D be an integral domain and let

$$(t - x_1)(t - x_2)\ldots(t - x_m)$$
$$= t^m - \sigma_1(x_1,x_2,\ldots,x_m)t^{m-1} + \sigma_2(x_1,x_2,\ldots,x_m)t^{m-2} - + \cdots + (-1)^m\sigma_m(x_1,x_2,\ldots,x_m).$$

The symmetric polynomials $\sigma_1,\sigma_2,\ldots,\sigma_m$ are called the *elementary symmetric polynomials* in $D[x_1,x_2,\ldots,x_m]$.

By routine computation, we find the elementary symmetric polynomials in explicitly. For example,

$\sigma_1 = x + y,$ $\sigma_2 = xy$ in $D[x,y]$

$\sigma_1 = x + y + z, \qquad \sigma_2 = xy + yz + zx, \qquad \sigma_3 = xyz \quad \text{in } D[x,y,z]$

$\sigma_1 = x + y + z + u, \qquad \sigma_2 = xy + xz + xu + yz + yu + zu,$
$\sigma_3 = xyz + xyu + xzu + yzu, \quad \sigma_4 = xyzu \qquad \text{in } D[x,y,z,u]$

are the elementary symmetric polynomials.

Notice that $(t - x_1)(t - x_2)\ldots(t - x_m)$, when multiplied out, is a sum of certain terms $a_1a_2\ldots a_m$, where each a_i is either t or one of $-x_1, -x_2, \ldots, -x_m$. The term $(-1)^j\sigma_j(x_1,x_2,\ldots,x_m)t^{m-j}$ is the sum of those $a_1a_2\ldots a_m$'s for which exactly $m - j$ of the a's are equal to t. Hence $(-1)^j\sigma_j(x_1,x_2,\ldots,x_m)$ is the sum of all products $b_1b_2\ldots b_j$, where $b_1,b_2,\ldots,b_j$ run independently over the set $\{-x_1, -x_2, \ldots, -x_m\}$. In other words, $\sigma_j(x_1,x_2,\ldots,x_m)$ is the sum of all $\binom{m}{j}$ products of $x_1,x_2,\ldots,x_m$, taken j at a time. Thus

$$\sigma_1 = \sum x_i$$

$$\sigma_2 = \sum x_ix_j$$

$$\sigma_3 = \sum x_ix_jx_k$$

$$\cdots\cdots\cdots\cdots$$

$$\sigma_m = x_1x_2\ldots x_m.$$

Note that "σ_j" stands for many polynomials: σ_j in $D[x_1,x_2,\ldots,x_m]$ is distinct from σ_j in $D[x_1,x_2,\ldots,x_n]$ when $m \neq n$. This ambiguity in notation will not cause any confusion if we pay attention to the number of indeterminates. When confusion is likely, we write $\sigma_j(x_1,x_2,\ldots,x_m)$ instead of σ_j.

Now $\sigma_1,\sigma_2,\ldots,\sigma_m$ are symmetric polynomials in $D[x_1,x_2,\ldots,x_m]$, and, by reapeated application of Lemma 38.2, we conclude that $g(\sigma_1,\sigma_2,\ldots,\sigma_m)$ is also a symmetric polynomial, where g is any polynomial in m indeterminates. Hence the set $\{g(\sigma_1,\sigma_2,\ldots,\sigma_m) : g \in D[u_1,u_2,\ldots,u_m]\}$ consist only of symmetric polynomials. We will prove conversely that every symmetric polynomial is in this set (the subring of symmetric polynomials in $D[x_1,x_2,\ldots,x_m]$ is the subring of $D[x_1,x_2,\ldots,x_m]$ generated by $\sigma_1,\sigma_2,\ldots,\sigma_m$).

38.4 Theorem (Fundamendal theorem on symmetric polynomials): *Let D be an integral domain and $f(x_1,x_2,\ldots,x_m)$ a symmetric poly-*

nomial in $D[x_1,x_2,\ldots,x_m]$. *Then there is a unique polynomial* $g(u_1,u_2,\ldots,u_m)$ *in* $D[u_1,u_2,\ldots,u_m]$ *such that* f *is the value of* g *at* $(\sigma_1,\sigma_2,\ldots,\sigma_m)$:

$$f(x_1,x_2,\ldots,x_m) = g(\sigma_1,\sigma_2,\ldots,\sigma_m) \in D[x_1,x_2,\ldots,x_m].$$

Loosely speaking, every symmetric polynomial is a polynomial in the elementary symmetric polynomials $\sigma_1,\sigma_2,\ldots,\sigma_m$. We introduced new indeterminates $u_1,u_2,\ldots,u_m$ in order to distinguish clearly between g and $g(\sigma_1,\sigma_2,\ldots,\sigma_m)$.

For example, $f(x,y) = x^2 + y^2 \in \mathbb{Z}[x,y]$ is a symmetric polynomial, and we have $x^2 + y^2 = (x + y)^2 - 2xy = \sigma_1^2 - 2\sigma_2$. Hence $f(x,y) = g(\sigma_1,\sigma_2)$, where $g(u,v) = u^2 - 2v \in \mathbb{Z}[u,v]$. Likewise, if $f(x,y,z)$ is the symmetric polynomial $x^2y + xy^2 + x^2z + xz^2 + y^2z + yz^2$ in $\mathbb{Z}[x,y,z]$, we have $f(x,y,z) = (x + y + z)(xy + yz + zx) - 3xyz = \sigma_1\sigma_2 - 3\sigma_3$. Thus $f(x,y,z) = g(\sigma_1,\sigma_2,\sigma_3)$, where $g(u,v,w) = uv - 3w \in \mathbb{Z}[u,v,w]$.

The proof of the fundemental theorem requires some preparation. First we need an ordering of m-tuples. Given any two m-tuples $(r_1,r_2,\ldots,r_m)$, $(s_1,s_2,\ldots,s_m)$ of nonnegative integers, we will say $(r_1,r_2,\ldots,r_m)$ is *higher than* $(s_1,s_2,\ldots,s_m)$, or $(s_1,s_2,\ldots,s_m)$ is *lower than* $(r_1,r_2,\ldots,r_m)$ when $r_1 > s_1$. If $r_1 = s_1$, we will say $(r_1,r_2,\ldots,r_m)$ is *higher than* $(s_1,s_2,\ldots,s_m)$, or $(s_1,s_2,\ldots,s_m)$ is *lower than* $(r_1,r_2,\ldots,r_m)$ when $r_2 > s_2$. If $r_1 = s_1$ and $r_2 = s_2$, we will compare r_3 and s_3, etc. This is very much like the ordering of words alphabetically, and will be referred to as the *alphabetical* or *lexigographical* ordering of m-tuples. Stated differently, $(r_1,r_2,\ldots,r_m)$ is higher than $(s_1,s_2,\ldots,s_m)$ if and only if the first nonzero difference among

$$r_1 - s_1, \quad r_2 - s_2, \quad \ldots, \quad r_m - s_m$$

is positive. Clearly, if $(r_1,r_2,\ldots,r_m)$ is higher than $(s_1,s_2,\ldots,s_m)$ and $(s_1,s_2,\ldots,s_m)$ is higher than $(t_1,t_2,\ldots,t_m)$, then $(r_1,r_2,\ldots,r_m)$ is higher than $(t_1,t_2,\ldots,t_m)$.

Now let f be a polynomial in $D[x_1,x_2,\ldots,x_m]$. So f is a sum of monomials $ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$, where $a \in D$ and $(k_1,k_2,\ldots,k_m)$ is an m-tuple of nonnegative integers. Here there may be several monomials $ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$, $bx_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$, $cx_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$, etc. with the same exponent system $(k_1,k_2,\ldots,k_m)$. In this case, we collect these monomials into a single one

$(a+b+c+\cdots)x_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$. We assume this has been done for each of the exponent systems, so that each m-tuple $(k_1,k_2,\ldots,k_m)$ occurs as an exponent system of a monomial at most once. If, after this collection process, a monomial $ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$ occuring in f has a nonzero coefficient $a \in D$, we will say that *a appears in f.*

Let us now assume $f \neq 0$. We order the monomials appearing in f by the alphabetical ordering of their exponent systems. First we write the monomial appearing in f whose exponent system is highest (i.e., higher than the exponent systems of all other monomials appearing in f). Among the remaining monomials appearing in f, we find the one with the highest exponent system and write it in the second place. Among the remaining monomials appearing in f, the one with the highest exponent system will be written it in the third place, and so on. In this ordering of monomials, the one that is written in the first place, that is to say, the one with the highest exponent system will be called the *leading monomial* of the nonzero polynomial $f \in D[x_1,x_2,\ldots,x_m]$. Note that the coefficients of monomials play no role in this ordering. Only the exponent systems are relevant.

For instance, $f(x,y,z) = xz^5 + z^7 + 2x^3 + 5x^2y + 100x^2y^2 - x^2y^2z \in \mathbb{Z}[x,y,z]$ will be written as $2x^3 - x^2y^2z + 100x^2y^2 + 5x^2y + xz^5 + z^7$when we order the monomials in the described manner. The leading monomial of $f(x,y,z)$ is $2x^3$.

38.5 Lemma: *Let D be an integral domain and $f,g \in D[x_1,x_2,\ldots,x_m]\backslash\{0\}$. If $ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$ is the leading monomial of f and $bx_1^{n_1}x_2^{n_2}\ldots x_m^{n_m}$ is the leading monomial of g, then $abx_1^{k_1+n_1}x_2^{k_2+n_2}\ldots x_m^{k_m+n_m}$ is the leading monomial of fg.*

Proof: By hypothesis, $a \neq 0$, $b \neq 0$, so $ab \neq 0$. Now $fg \neq 0$ and fg is the sum of all products $(cx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m})(dx_1^{s_1}x_2^{s_2}\ldots x_m^{s_m})$, where $cx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m}$ and $dx_1^{s_1}x_2^{s_2}\ldots x_m^{s_m}$ run through all monomials appearing in f and g, respectively. We contend that, among all these products, the highest exponent system is $(k_1+n_1,k_2+n_2,\ldots,k_m+n_m)$, and that this exponent

system arises only from the product $(ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m})(bx_1^{n_1}x_2^{n_2}\ldots x_m^{n_m})$. This will imply

$$fg = abx_1^{k_1+n_1}x_2^{k_2+n_2}\ldots x_m^{k_m+n}$$
$$+ [\text{a sum of monomials, each with an exponent system lower than } (k_1+n_1, k_2+n_2, \ldots, k_m+n_m)],$$

and, since $ab \neq 0$, the leading monomial of fg will be equal to $abx_1^{k_1+n_1}x_2^{k_2+n_2}\ldots x_m^{k_m+n_m}$.

To prove our contention, let $cx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m}$ be a monomial appearing in f and let $dx_1^{s_1}x_2^{s_2}\ldots x_m^{s_m}$ be one appearing in g, but assume that either $cx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m}$ is distinct from $ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$ or $dx_1^{s_1}x_2^{s_2}\ldots x_m^{s_m}$ is distinct from $bx_1^{n_1}x_2^{n_2}\ldots x_m^{n_m}$. We are to show that the exponent system $(r_1+s_1, r_2+s_2, \ldots, r_m+s_m)$ is lower than $(k_1+n_1, k_2+n_2, \ldots, k_m+n_m)$. Now $(r_1, r_2, \ldots, r_m)$ is lower than $(k_1, k_2, \ldots, k_m)$ or equal to it, and $(s_1, s_2, \ldots, s_m)$ is lower than $(n_1, n_2, \ldots, n_m)$ or equal to it, but the case of simultaneous equality is excluded. Hence the first nonzero integer in

$$k_1 - r_1,\ k_2 - r_2,\ \ldots,\ k_m - r_m$$

is positive, or $(k_1, k_2, \ldots, k_m) = (r_1, r_2, \ldots, r_m)$, and the first nonzero integer in

$$n_1 - s_1,\ n_2 - s_2,\ \ldots,\ n_m - s_m$$

is positive, or $(n_1, n_2, \ldots, n_m) = (s_1, s_2, \ldots, s_m)$. Since simultaneous equality is excluded, there are nonzero integers in

$$(k_1 - r_1) + (n_1 - s_1),\ (k_2 - r_2) + (n_2 - s_2),\ \ldots,\ (k_m - r_m) + (n_m - s_m)$$

and the first of them, being a sum of two positive integers or a sum of a positive integer and zero, is certainly positive. This means that $(k_1+n_1, k_2+n_2, \ldots, k_m+n_m)$ is higher than $(r_1+s_1, r_2+s_2, \ldots, r_m+s_m)$.

Since the exponent system $(k_1+n_1, k_2+n_2, \ldots, k_m+n_m)$ does arise from the product $(ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m})(bx_1^{n_1}x_2^{n_2}\ldots x_m^{n_m})$, it is indeed the highest exponent system of all the products $(cx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m})(dx_1^{s_1}x_2^{s_2}\ldots x_m^{s_m})$ where $cx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m}$ and $dx_1^{s_1}x_2^{s_2}\ldots x_m^{s_m}$ run through all monomials appearing in f and g, respectively. This proves our contention, and also the lemma. □

By induction, we obtain

38.6 Lemma: *Let D be an integral domain and $f_1, f_2, \ldots, f_t$ be nonzero polynomials in $D[x_1, x_2, \ldots, x_m]$. Then the leading monomial of $f_1 f_2 \ldots f_t$ is the product of the leading monomials of $f_1, f_2, \ldots, f_t$.* □

38.7 Lemma: *Let D be an integral domain, $a \in D\backslash\{0\}$, and let $\sigma_1, \sigma_2, \ldots, \sigma_m$ be the elementary symmetric polynomials in $D[x_1, x_2, \ldots, x_m]$.*
If $k_1 \geqslant k_2 \geqslant k_3 \geqslant \ldots \geqslant k_m \geqslant 0$ are integers, then the leading monomial of $a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$ is $a{x_1}^{k_1}{x_2}^{k_2}\ldots{x_{m-1}}^{k_{m-1}}{x_m}^{k_m}$.

Proof: The leading monomials of $\sigma_1, \sigma_2, \sigma_3, \sigma_4, \ldots, \sigma_m$ are respectively x_1, x_1x_2, $x_1x_2x_3$, $x_1x_2x_3x_4$, …, $x_1x_2\ldots x_m$, because σ_j is a sum of $\binom{m}{j}$ monomials, each of which is a product of j indeterminates from $x_1, x_2, \ldots, x_m$. In view of Lemma 38.6, the leading monomial of $a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$ is

$$a(x_1)^{k_1-k_2}(x_1x_2)^{k_2-k_3}(x_1x_2x_3)^{k_3-k_4}\ldots(x_1x_2x_3\ldots x_{m-1})^{k_{m-1}-k_m}(x_1x_2x_3\ldots x_{m-1}x_m)^{k_m}$$
$$= a{x_1}^{k_1}{x_2}^{k_2}\ldots{x_{m-1}}^{k_{m-1}}{x_m}^{k_m}.$$ □

We need one more lemma for the proof of the fundamental theorem.

38.8 Lemma: *Let D be an integral domain and let $f(x_1, x_2, \ldots, x_m)$ be a nonzero symmetric polynomial in $D[x_1, x_2, \ldots, x_m]$. Let $a{x_1}^{k_1}{x_2}^{k_2}\ldots{x_m}^{k_m}$ be the leading monomial of f (here $a \in D$, $a \neq 0$ and $k_1, k_2, \ldots, k_m$ are nonnegative integers).*
(1) *We have $k_1 \geqslant k_2 \geqslant \cdots \geqslant k_{m-1} \geqslant k_m$.*
(2) *If $b{x_1}^{r_1}{x_2}^{r_2}\ldots{x_m}^{r_m}$ is a monomial appearing in f, then*
$$k_1 \geqslant r_1,\ k_1 \geqslant r_2,\ \ldots,\ k_1 \geqslant r_m.$$

Proof: Let σ be any permutation in S_m and let τ be the inverse of σ.
(1) As $a{x_1}^{k_1}{x_2}^{k_2}\ldots{x_m}^{k_m}$ appears in $f(x_1, x_2, \ldots, x_m)$,
$a{x_{1\tau}}^{k_1}{x_{2\tau}}^{k_2}\ldots{x_{m\tau}}^{k_m}$ appears in $f(x_{1\tau}, x_{2\tau}, \ldots, x_{m\tau}) = f^\tau = f = f(x_1, x_2, \ldots, x_m)$,
$a{x_1}^{k_{1\sigma}}{x_2}^{k_{2\sigma}}\ldots{x_m}^{k_{m\sigma}}$ appears in $f(x_1, x_2, \ldots, x_m)$,
and, since $a{x_1}^{k_1}{x_2}^{k_2}\ldots{x_m}^{k_m}$ is the leading monomial of f, we obtain:

for all $\sigma \in S_m$, $(k_1, k_2, \ldots, k_m)$ is higher than or equal to $(k_{1\sigma}, k_{2\sigma}, \ldots, k_{m\sigma})$.

Using this with $\sigma = (12) \in S_m$, we see $(k_1,k_2,\ldots,k_m)$ is higher than or equal to $(k_2,k_1,\ldots,k_m)$, so $k_1 \geqslant k_2$. And $\sigma = (23)$ yields that $(k_1,k_2,k_3\ldots,k_m)$ is higher than or equal to $(k_1,k_3,k_2,\ldots,k_m)$, so $k_2 \geqslant k_3$. In like manner, when we choose $\sigma = (34), \ldots, (m-1,m) \in S_m$, we get $k_3 \geqslant k_4, \ldots, k_{m-1} \geqslant k_m$. This proves (1).

(2) As $bx_1^{r_1}x_2^{r_2}\ldots x_m^{r_m}$ appears in $f(x_1,x_2,\ldots,x_m)$,
$bx_{1\tau}^{r_1}x_{2\tau}^{r_2}\ldots x_{m\tau}^{r_m}$ appears in $f(x_{1\tau},x_{2\tau},\ldots,x_{m\tau}) = f^\tau = f = f(x_1,x_2,\ldots,x_m)$,
$bx_1^{r_{1\sigma}}x_2^{r_{2\sigma}}\ldots x_m^{r_{m\sigma}}$ appears in $f(x_1,x_2,\ldots,x_m)$,
and so:

for all $\sigma \in S_m$, $(k_1,k_2,\ldots,k_m)$ is higher than or equal to $(r_{1\sigma},r_{2\sigma},\ldots,r_{m\sigma})$.

Thus $k_1 \geqslant r_{1\sigma}$ for all $\sigma \in S_m$. Here 1σ assumes all values $1,2,\ldots,m$ as σ runs through S_m, and hence $k_1 \geqslant r_1, k_1 \geqslant r_2, \ldots, k_1 \geqslant r_m$. □

Proof of the fundamental theorem: Throughout the proof, the number m of the indeterminates will be fixed. We make induction on the exponent system of the leading monomial of the symmetric polynomial. This will be explained shortly.

Let f be a nonzero symmetric polynomial in $D[x_1,x_2,\ldots,x_m]$ and let $ax_1^{k_1}x_2^{k_2}\ldots x_m^{k_m}$ be its leading monomial.

First we claim: if $(k_1,k_2,\ldots,k_m) = (0,0,\ldots,0)$, then there is a polynomial g in m indeterminates $u_1,u_2,\ldots,u_m$ over D such that $f(x_1,x_2,\ldots,x_m)$ is equal to $g(\sigma_1,\sigma_2,\ldots,\sigma_m)$. This is very easy to prove. Indeed, if $(k_1,k_2,\ldots,k_m) = (0,0,\ldots,0)$, then, by Lemma 38.2(2), the exponent system of any monomial appearing in f is $(0,0,\ldots,0)$, so f is the constant polynomial a in $D[x_1,x_2,\ldots,x_m]$. Then of course $f(x_1,x_2,\ldots,x_m) = g(\sigma_1,\sigma_2,\ldots,\sigma_m)$, where g is the constant polynomial a in $D[u_1,u_2,\ldots,u_m]$.

Now suppose that $(k_1,k_2,\ldots,k_m)$ is higher than $(0,0,\ldots,0)$ and that, for any nonzero symmetric polynomial $f_1 \in D[x_1,x_2,\ldots,x_m]$ whose leading monomial has a lower exponent system than $(k_1,k_2,\ldots,k_m)$, there is a polynomial g_1 in $D[u_1,u_2,\ldots,u_m]$ such that $f_1(x_1,x_2,\ldots,x_m) = g_1(\sigma_1,\sigma_2,\ldots,\sigma_m)$. Under this assumption, we will prove the existence of a polynomial g in $D[u_1,u_2,\ldots,u_m]$ with $f(x_1,x_2,\ldots,x_m) = g(\sigma_1,\sigma_2,\ldots,\sigma_m)$. This will establish the fundamental theorem because $(0,0,\ldots,0)$ is the lowest possible exponent

system and the theorem has been proved in this case above. Moreover, as there are only a finite number of m-tuples lower than $(k_1,k_2,\ldots,k_m)$, the method of proof can be used effectively to find the polynomial g explicitly in concrete cases. [Basically, we write the m-tuples $L_1,L_2,L_3,\ldots$ in alphabetical order and prove that (1) the theorem is true for all nonzero symmetric polynomials whose leading monomials have the exponent system $L_1 = (0,0,\ldots,0)$ and that (2) for any $s > 1$, if the theorem is true for all nonzero symmetric polynomials whose leading monomials have exponent systems equal to one of $L_1,L_2,\ldots L_{s-1}$, then the theorem is also true for all nonzero symmetric polynomials whose leading monomials have the exponent system L_s. Once the leading monomial of a symmetric polynomial is given, there can be only a finite number of exponent systems of monomials appearing in that symmetric polynomial (Lemma 38.8(2).]

By Lemma 38.8(1), the integers $k_1 - k_2, k_2 - k_3, \ldots, k_m - k_{m-1}, k_m$ are nonnegative and, by Lemma 38.7, the polynomial $a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$ has the same leading monomial as f. Let $f_1 = f - a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$. Thus f_1 is a symmetric polynomial in $D[x_1,x_2,\ldots,x_m]$. If $f_1 = 0$, then $f = a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$ and $f = g(\sigma_1,\sigma_2,\ldots,\sigma_m)$, where $g = au_1^{k_1-k_2}u_2^{k_2-k_3}\ldots u_{m-1}^{k_{m-1}-k_m}u_m^{k_m} \in D[u_1,u_2,\ldots,u_m]$, and the proof is completed in this case. If $f_1 \neq 0$, then f_1 has a leading monomial. The exponent system of this leading monomial of f_1 is the exponent system of a monomial appearing in f or in $a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$ (or in both). This exponent system is distinct from $(k_1,k_2,\ldots,k_m)$. Since it arises from a monomial appearing in f or in $a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$, it is lower than the common exponent system $(k_1,k_2,\ldots,k_m)$ of the leading monomials of f and $a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m}$. By hypothesis, there is a polynomial g_1 in $D[u_1,u_2,\ldots,u_m]$ such that $f_1(x_1,x_2,\ldots,x_m) = g_1(\sigma_1,\sigma_2,\ldots,\sigma_m)$. Hence

$$\begin{aligned} f &= f_1 + a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m} \\ &= g_1(\sigma_1,\sigma_2,\ldots,\sigma_m) + a\sigma_1^{k_1-k_2}\sigma_2^{k_2-k_3}\ldots\sigma_{m-1}^{k_{m-1}-k_m}\sigma_m^{k_m} \end{aligned}$$

and there is a polynomial g in $D[u_1,u_2,\ldots,u_m]$, namely

$$g_1 + au_1^{k_1-k_2}u_2^{k_2-k_3}\ldots u_{m-1}^{k_{m-1}-k_m}u_m^{k_m},$$

such that $f(x_1,x_2,\ldots,x_m) = g(\sigma_1,\sigma_2,\ldots,\sigma_m)$.

This completes the proof of the existence of g. It remains to show the uniqueness of g. Suppose now f is a nonzero symmetric polynomial in $D[x_1,x_2,\ldots,x_m]$ and assume that $g,h \in D[u_1,u_2,\ldots,u_m]$ with $g(\sigma_1,\sigma_2,\ldots,\sigma_m) =$

$f(x_1,x_2,\ldots,x_m) = h(\sigma_1,\sigma_2,\ldots,\sigma_m)$. If g were distinct from h, then $g - h \neq 0$ would have a leading monomial which we may write in the form $u_1^{s_1-s_2}u_2^{s_2-s_3}\ldots u_{m-1}^{s_{m-1}-s_m}u_m^{s_m}$, where $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_{m-1} \geqslant s_m$. Then $0 = f - f = g(\sigma_1,\sigma_2,\ldots,\sigma_m) - h(\sigma_1,\sigma_2,\ldots,\sigma_m)$ in $D[x_1,x_2,\ldots,x_m]$ would have a leading monomial $bx_1^{s_1}x_2^{s_2}\ldots x_{m-1}^{s_{m-1}}x_m^{s_m}$, a contradiction. Hence $g = h$, as was to be proved. □

The fundamental theorem can be summarized by saying that the substitution mapping

$$T: D[u_1,u_2,\ldots,u_m] \longrightarrow S$$
$$g(u_1,u_2,\ldots,u_m) \to g(\sigma_1,\sigma_2,\ldots,\sigma_m)$$

is a ring isomorphism, where S is the subring of $D[x_1,x_2,\ldots,x_m]$ consisting of the symmetric polynomials in $D[x_1,x_2,\ldots,x_m]$.

38.9 Examples: (a) We express the polynomial

$$f(x,y,z) = 5xyz + x^2y + xy^2 + xz^2 + yz^2 + y^2z + x^2z \in \mathbb{Z}[x,y,z]$$

in terms of $\sigma_1,\sigma_2,\sigma_3$.

We first arrange the monomials appearing in f in the alphabetical order of their exponent systems:

$$f(x,y,z) = x^2y + x^2z + xy^2 + 5xyz + xz^2 + y^2z + yz^2.$$

The leading monomial of f is $1x^2y^1z^0$. We therefore subtract $1\sigma_1^{2-1}\sigma_2^{1-0}\sigma_3^{0}$ from f and get

$$f - \sigma_1\sigma_2 = (x^2y + x^2z + xy^2 + 5xyz + xz^2 + y^2z + yz^2) - (x + y + z)(xy + yz + zx)$$
$$= 2xyz.$$

The leading monomial of $f - \sigma_1\sigma_2$ is $2x^1y^1z^1$. So we subtract $2\sigma_1^{1-1}\sigma_2^{1-1}\sigma_3^{1}$ from $f - \sigma_1\sigma_2$ and get

$$(f - \sigma_1\sigma_2) - 2\sigma_3 = 2xyz - 2xyz = 0.$$

Hence $f(x,y,z) = \sigma_1\sigma_2 + 2\sigma_3$.

(b) We express

$$f(x,y,z,w) = x^3 + y^3 + z^3 + w^3 \in \mathbb{Z}[x,y,z]$$

in terms of $\sigma_1,\sigma_2,\sigma_3,\sigma_4$. The monomials are in alphabetical order, and the leading monomial of f is $1x^3y^0z^0w^0$. So we subtract $1\sigma_1^{3-0}\sigma_2^{0-0}\sigma_3^{0-0}\sigma_4^{0}$ from f and get

$$f - \sigma_1^3 = (x^3 + y^3 + z^3 + w^3) - (x + y + z + w)^3$$
$$= \cdots\cdots$$
$$= -3x^2y - 3xy^2 - 3x^2z - 3xz^2 - 3x^2w - 3xw^2 - 3y^2z - 3yz^2$$

$$-3y^2w - 3yw^2 - 3z^2w - 3zw^2 - 6xyz - 6xyw - 6xzw - 6yzw.$$

The leading monomial of $f - \sigma_1^3$ is $-3x^2y = -3x^2y^1z^0w^0$. We therefore subtract $-3\sigma_1^{2-1}\sigma_2^{1-0}\sigma_3^{0-0}\sigma_4^0$ from $f - \sigma_1^3$ and get

$$\begin{aligned}(f - \sigma_1^3) - (-3\sigma_1\sigma_2) &= (f - \sigma_1^3) + 3(x + y + z + w)(xy + xz + xw + yz + yw + zw)\\ &= \cdots\cdots\\ &= 3xyz + 3xyw + 3xzw + 3yzw\\ &= 3\sigma_3.\end{aligned}$$

Hence $f(x,y,z,w) = \sigma_1^3 - 3\sigma_1\sigma_2 + 3\sigma_3$.

We now derive formulas connecting the sum of the k-th powers of $x_1, x_2, \ldots, x_m$ with the elementary symmetric polynomials. These formulas are due to I. Newton (1642-1727).

38.10 Theorem (Newton): *Let D be an integral domain and $x_1, x_2, \ldots, x_m$ indeterminates over D. For $k = 1,2,3, \ldots$, we put $s_k = x_1^k + x_2^k + \ldots + x_m^k$, so that $s_k \in D[x_1, x_2, \ldots, x_m]$. Then*

$$0 = s_1 - \sigma_1$$

$$0 = s_2 - \sigma_1 s_1 + 2\sigma_2$$

$$0 = s_3 - \sigma_1 s_2 + \sigma_2 s_1 - 3\sigma_3$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$0 = s_{m-1} - \sigma_1 s_{m-2} + \sigma_2 s_{m-3} + \cdots + (-1)^{m-2}\sigma_{m-2}s_1 + (-1)^{m-1}(m-1)\sigma_{m-1}$$

and

$$0 = s_m - \sigma_1 s_{m-1} + \sigma_2 s_{m-2} + \cdots + (-1)^{m-2}\sigma_{m-2}s_2 + (-1)^{m-1}\sigma_{m-1}s_1 + (-1)^m m\sigma_m$$

$$0 = s_{m+1} - \sigma_1 s_m + \sigma_2 s_{m-1} + \cdots + (-1)^{m-2}\sigma_{m-2}s_3 + (-1)^{m-1}\sigma_{m-1}s_2 + (-1)^m\sigma_m s_1$$

$$0 = s_{m+2} - \sigma_1 s_{m+1} + \sigma_2 s_m + \cdots + (-1)^{m-2}\sigma_{m-2}s_4 + (-1)^{m-1}\sigma_{m-1}s_3 + (-1)^m\sigma_m s_2$$

$$0 = s_{m+3} - \sigma_1 s_{m+2} + \sigma_2 s_{m+1} + \cdots + (-1)^{m-2}\sigma_{m-2}s_5 + (-1)^{m-1}\sigma_{m-1}s_4 + (-1)^m\sigma_m s_3$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

Proof: We make use of the polynomial $f(t) = (t - x_1)(t - x_2)\ldots(t - x_m)$. We know that

$$f(t) = t^m - \sigma_1 t^{m-1} + \sigma_2 t^{m-2} - \cdots + (-1)^{m-1}\sigma_{m-1} t + (-1)^m \sigma_m$$

and that $x_1, x_2, \ldots, x_m$ are the roots of $f(t) \in D[x_1, x_2, \ldots, x_m]$. Hence

$$0 = x_i^m - \sigma_1 x_i^{m-1} + \sigma_2 x_i^{m-2} - \cdots + (-1)^{m-1}\sigma_{m-1} x_i + (-1)^m \sigma_m$$

for all $i = 1,2, \ldots, m$. Multiplying both sides of this equation by x_i^j, where $j = 0,1,2,3, \ldots$, we get

$$0 = x_i^{m+j} - \sigma_1 x_i^{m+j-1} + \sigma_2 x_i^{m+j-2} - \cdots + (-1)^{m-1}\sigma_{m-1} x_i^{j+1} + (-1)^m \sigma_m x_i^j$$

for all $i = 1,2, \ldots, m$. Adding these m equations side by side, we obtain $0 =$

$$\sum_{i=1}^{m} x_i^{m+j} - \sigma_1 \sum_{i=1}^{m} x_i^{m+j-1} + \sigma_2 \sum_{i=1}^{m} x_i^{m+j-2} - + \cdots + (-1)^{m-1}\sigma_{m-1} \sum_{i=1}^{m} x_i^{j+1} + (-1)^m \sigma_m \sum_{i=1}^{m} x_i^j$$

i.e.,

$$0 = s_{m+j} - \sigma_1 s_{m+j-1} + \sigma_2 s_{m+j-2} + \cdots + (-1)^{m-2}\sigma_{m-2} s_{j+2} + (-1)^{m-1}\sigma_{m-1} s_{j+1} + (-1)^m \sigma_m s_j.$$

This establishes all the equations except the first $m - 1$ of them. The first $m - 1$ equations will be established by a similar reasoning. This time we make use of the derivative of $f(t)$. By Lemma 35.16(2), we have

$$f'(t) = (t-x_2)(t-x_3)\ldots(t-x_m) + (t-x_1)(t-x_3)\ldots(t-x_m) + \cdots + (t-x_1)(t-x_2)\ldots(t-x_{m-1})$$

$$= \frac{f(t)}{t - x_1} + \frac{f(t)}{t - x_2} + \cdots + \frac{f(t)}{t - x_m}.$$

For $i = 1,2, \ldots, m$, we put

$$\frac{f(t)}{t - x_i} = q_{m-1}^{(i)} t^{m-1} + q_{m-2}^{(i)} t^{m-2} + \cdots + q_1^{(i)} t + q_0^{(i)}.$$

Hence $mt^{m-1} - (m-1)\sigma_1 t^{m-2} + (m-2)\sigma_2 t^{m-3} - + \cdots + (-1)^{m-1}\sigma_{m-1} = f'(t)$

$$= \sum_{i=1}^{m} \frac{f(t)}{t - x_i} = \sum_{i=1}^{m} \left(q_{m-1}^{(i)} t^{m-1} + q_{m-2}^{(i)} t^{m-2} + \cdots + q_1^{(i)} t + q_0^{(i)} \right)$$

$$= \left(\sum_{i=1}^{m} q_{m-1}^{(i)} \right) t^{m-1} + \left(\sum_{i=1}^{m} q_{m-2}^{(i)} \right) t^{m-2} + \cdots + \left(\sum_{i=1}^{m} q_1^{(i)} \right) t + \left(\sum_{i=1}^{m} q_0^{(i)} \right),$$

so that

$$m = \sum_{i=1}^{m} q_{m-1}^{(i)}, \qquad -(m-1)\sigma_1 = \sum_{i=1}^{m} q_{m-2}^{(i)}, \qquad \ldots, \quad (-1)^{m-2}2\sigma_{m-2} = \sum_{i=1}^{m} q_1^{(i)},$$

$$(-1)^{m-1}\sigma_{m-1} = \sum_{i=1}^{m} q_0^{(i)}. \qquad (*)$$

On the other hand, $t^m - \sigma_1 t^{m-1} + \sigma_2 t^{m-2} - + \cdots + (-1)^{m-1}\sigma_{m-1}t + (-1)^m\sigma_m$

$$= f(t) = (t - x_i)(q_{m-1}^{(i)}t^{m-1} + q_{m-2}^{(i)}t^{m-2} + \cdots + q_1^{(i)}t + q_0^{(i)})$$

$$= q_{m-1}^{(i)}t^m + q_{m-2}^{(i)}t^{m-1} + q_{m-3}^{(i)}t^{m-2} + \cdots + q_1^{(i)}t^2 + q_0^{(i)}t$$

$$- q_{m-1}^{(i)}x_i t^{m-1} - q_{m-2}^{(i)}x_i t^{m-2} - \cdots - q_2^{(i)}x_i t^2 - q_1^{(i)}x_i t - q_0^{(i)}x_i.$$

Comparing the coefficients of powers of t on both sides, we get

$$\begin{aligned}
1 &= q_{m-1}^{(i)} \\
-\sigma_1 &= q_{m-2}^{(i)} - q_{m-1}^{(i)}x_i \\
+\sigma_2 &= q_{m-3}^{(i)} - q_{m-2}^{(i)}x_i \\
-\sigma_3 &= q_{m-4}^{(i)} - q_{m-3}^{(i)}x_i \\
&\cdots\cdots\cdots\cdots \\
(-1)^{m-1}\sigma_{m-1} &= q_0^{(i)} - q_1^{(i)}x_i \\
(-1)^m\sigma_m &= \quad - q_0^{(i)}x_i,
\end{aligned}$$

which may be written

$$\begin{aligned}
q_{m-1}^{(i)} &= 1 \\
q_{m-2}^{(i)} &= -\sigma_1 + q_{m-1}^{(i)}x_i \\
q_{m-3}^{(i)} &= +\sigma_2 + q_{m-2}^{(i)}x_i \\
q_{m-4}^{(i)} &= -\sigma_3 + q_{m-3}^{(i)}x_i \\
&\cdots\cdots\cdots\cdots \\
q_0^{(i)} &= (-1)^{m-1}\sigma_{m-1} + q_1^{(i)}x_i \\
0 &= (-1)^m\sigma_m + q_0^{(i)}x_i.
\end{aligned}$$

So, for each $i = 1,2, \ldots,m$,

$$q_{m-2}^{(i)} = -\sigma_1 + x_i \qquad (1)$$

$$q_{m-3}^{(i)} = +\sigma_2 + (-\sigma_1 + x_i)x_i = \sigma_2 - \sigma_1 x_i + x_i^2 \qquad (2)$$

$$q_{m-4}^{(i)} = -\sigma_3 + (\sigma_2 - \sigma_1 x_i + x_i^2)x_i = -\sigma_3 + \sigma_2 x_i - \sigma_1 x_i^2 + x_i^3 \qquad (3)$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$q_0^{(i)} = (-1)^{m-1}\sigma_{m-1} + ((-1)^{m-2}\sigma_{m-2} + (-1)^{m-3}\sigma_{m-3}x_i + \cdots + (-1)\sigma_1 x_i^{m-3} + x_i^{m-2})x_i$$

$$= (-1)^{m-1}\sigma_{m-1} + (-1)^{m-2}\sigma_{m-2}x_i + (-1)^{m-3}\sigma_{m-3}x_i^2 + \cdots + (-1)\sigma_1 x_i^{m-2} + x_i^m. \quad (m-1)$$

We add the m equations (1) together, also the m equations (2), the m equations (3),..., the m equations $(m - 1)$ (for $i = 1,2, \ldots,m$). Using (*), we obtain

$$-(m-1)\sigma_1 = -m\sigma_1 + s_1$$

$$+(m-2)\sigma_2 = +m\sigma_2 - \sigma_1 s_1 + s_2$$

$$-(m-3)\sigma_3 = -m\sigma_3 + \sigma_2 s_1 - \sigma_1 s_2 + s_3$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

$$(-1)^{m-1}\sigma_{m-1} = (-1)^{m-1} m\sigma_{m-1} + (-1)^{m-2}\sigma_{m-2}s_1 + (-1)^{m-3}\sigma_{m-3}s_2 + \cdots + (-1)\sigma_1 s_{m-2} + s_{m-1},$$

which are equivalent to

$$\begin{aligned}
&s_1 - \sigma_1 = 0\\
&s_2 - \sigma_1 s_1 + 2\sigma_2 = 0\\
&s_3 - \sigma_1 s_2 + \sigma_2 s_1 - 3\sigma_3 = 0\\
&\qquad\cdots\cdots\cdots\cdots\cdots\\
&s_{m-1} - \sigma_1 s_{m-2} + \sigma_2 s_{m-3} + - \cdots + (-1)^{m-2}\sigma_{m-2}s_1 + (-1)^{m-1}(m-1)\sigma_{m-1} = 0.
\end{aligned}$$

This completes the proof. □

Newton's formulas express s_k recursively in terms of $s_1, s_2, \ldots, s_{k-1}$ and of $\sigma_1, \sigma_2, \ldots, \sigma_m$. We can eliminate $s_1, s_2, \ldots, s_{k-1}$ and write s_k solely in terms of $\sigma_1, \sigma_2, \ldots, \sigma_m$. For instance:

$$\begin{aligned}
s_1 &= \sigma_1\\
s_2 &= \sigma_1 s_1 - 2\sigma_2 = \sigma_1\sigma_1 - 2\sigma_2 = {\sigma_1}^2 - 2\sigma_2\\
s_3 &= \sigma_1 s_2 - \sigma_2 s_1 + 3\sigma_3 = \sigma_1({\sigma_1}^2 - 2\sigma_2) - \sigma_2\sigma_1 + 3\sigma_3 = {\sigma_1}^3 - 3\sigma_1\sigma_2 + 3\sigma_3\\
s_4 &= \sigma_1 s_3 - \sigma_2 s_2 + \sigma_3 s_1 - 4\sigma_4 = \sigma_1({\sigma_1}^3 - 3\sigma_1\sigma_2 + 3\sigma_3) - \sigma_2({\sigma_1}^2 - 2\sigma_2) + \sigma_3\sigma_1 - 4\sigma_4\\
&= {\sigma_1}^4 - 4{\sigma_1}^2\sigma_2 + 4\sigma_1\sigma_3 + 2{\sigma_2}^2 - 4\sigma_4
\end{aligned}$$

(here σ_j should be replaced by 0 when $j > m$).

Now let D, E be integral domains and $D \subseteq E$. Let

$$p(t) = c_0 t^m + c_1 t^{m-1} + \cdots + c_{m-1}t + c_m$$

be a nonzero polynomial of degree m in $D[t]$, and assume that there are exactly m roots $a_1, a_2, \ldots, a_m$ of p in E (counted with multiplicities). Then

$$p(t) = c_0(t - a_1)(t - a_2)\ldots(t - a_m) \quad \text{in } E[t].$$

Hence $p(t) \in E[t]$ is obtained from

$$c_0 f(t) = c_0(t - x_1)(t - x_2)\ldots(t - x_m) \quad \in D[x_1, x_2, \ldots, x_m][t]$$

by substituting a_i for x_i $(i = 1,2, \ldots, m)$. Now $c_0 f(t) =$ $c_0(t^m - \sigma_1(x_1, x_2, \ldots, x_m)t^{m-1} + \sigma_2(x_1, x_2, \ldots, x_m)t^{m-2} - + \ldots + (-1)^m\sigma_m(x_1, x_2, \ldots, x_m))$ and, since substitution is a homomorphism (Lemma 35.20), we have

$$p(t) =$$
$$c_0(t^m - \sigma_1(a_1,a_2,\ldots,a_m)t^{m-1} + \sigma_2(a_1,a_2,\ldots,a_m)t^{m-2} - + \cdots + (-1)^m\sigma_m(a_1,a_2,\ldots,a_m)).$$

Therefore
$$\begin{aligned} c_1 &= -c_0\sigma_1(a_1,a_2,\ldots,a_m) \\ c_2 &= +c_0\sigma_2(a_1,a_2,\ldots,a_m) \\ c_3 &= -c_0\sigma_3(a_1,a_2,\ldots,a_m) \\ &\cdots\cdots\cdots\cdots \\ c_{m-1} &= (-1)^{m-1}\sigma_{m-1}(a_1,a_2,\ldots,a_m) \\ c_m &= (-1)^m\sigma_m(a_1,a_2,\ldots,a_m); \end{aligned}$$

in words: the values of the elementary symmetric polynomials at the roots of a polynomial are equal to the coefficients of that polynomial, except for a factor $\mp c_0$, where c_0 is the leading coefficient of the polynomial. The equations above tell us that (i) $\sigma_i(a_1,a_2,\ldots,a_m)$ belong to D if c_0 is a unit in D; (ii) $\sigma_i(a_1,a_2,\ldots,a_m)$ belong to the field of fractions of D in any case; (iii) in particular, $\sigma_i(a_1,a_2,\ldots,a_m)$ belong to D if D is a field. Moreover, if $h(x_1,x_2,\ldots,x_m) \in D[x_1,x_2,\ldots,x_m]$ is a symmetric polynomial, then $h(x_1,x_2,\ldots,x_m) = g(\sigma_1,\sigma_2,\ldots,\sigma_m)$ for some polynomial in m indeterminates over D, and substitution yields

$$h(a_1,a_2,\ldots,a_m) = g(\sigma_1(a_1,a_2,\ldots,a_m), \sigma_2(a_1,a_2,\ldots,a_m), \ldots, \sigma_m(a_1,a_2,\ldots,a_m))$$

so that (i) $h(a_1,a_2,\ldots,a_m)$ belongs to D if c_0 is a unit in D; (ii) $h(a_1,a_2,\ldots,a_m)$ belongs to the field of fractions of D in any case; (iii) $h(a_1,a_2,\ldots,a_m)$ belongs to D if D is a field. We summarize this discussion in the next theorem.

38.11 Theorem: *Let D be an integral domain and let $p(t) = c_0t^m + c_1t^{m-1} + \cdots + c_{m-1}t + c_m$ a polynomial over D. Assume that $p(t)$ has exactly m roots $a_1,a_2,\ldots,a_m$ in an integral domain containing D.*

(1) $c_i = (-1)^i\sigma_m(a_1,a_2,\ldots,a_m)$ *for* $i = 1,2,\ldots,m$.
(2) *If h is any symmetric polynomial in m indeterminates over D, then $h(a_1,a_2,\ldots,a_m)$, which is an element of the integral domain containing the roots of $p(t)$, is in fact an element of the field of fractions of D.*
(3) *If, in addition, the leading coefficient of $p(t)$ is a unit in D, then $h(a_1,a_2,\ldots,a_m)$ belongs to D.*
(4) *If, in particular, D is a field, then $h(a_1,a_2,\ldots,a_m)$ belongs to D.* □

It is true that any nonzero polynomial of degree m over an integral domain D has exactly m roots in some integral domain containing D. This will be proved later (Theorem 53.6). In the following examples, we will assume that the polynomials have as many roots as their degrees in some integral domain.

Examples: (a) Let us evaluate $a^2b^2 + a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 + c^2d^2$, where a,b,c,d are the roots of $t^4 - t^2 + 1 \in \mathbb{Z}[t]$. To this end, we express the symmetric polynomial $x^2y^2 + x^2z^2 + x^2u^2 + y^2z^2 + y^2u^2 + z^2u^2$ in terms of $\sigma_1,\sigma_2,\sigma_3,\sigma_4$. Subtracting $1\sigma_1^{2-2}\sigma_2^{2-0}\sigma_3^{0-0}\sigma_4^0$ from this polynomial, we get

$$(x^2y^2 + \cdots + z^2u^2) - \sigma_2{}^2 = -2x^2yz - \cdots - 6xyzu,$$

and $$(x^2y^2 + \cdots + z^2u^2) - \sigma_2{}^2 - (-2\sigma_1^{2-1}\sigma_2^{1-1}\sigma_3^{1-0}\sigma_4^0) = \cdots = 0,$$

so $$x^2y^2 + x^2z^2 + x^2u^2 + y^2z^2 + y^2u^2 + z^2u^2 = \sigma_2{}^2 - \sigma_1\sigma_3.$$

Then $$a^2b^2 + a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 + c^2d^2$$
$$= (ab + ac + ad + bc + bd + cd)^2 - 2(a + b + c + d)(abc + abd + acd + bcd).$$

Here a,b,c,d are the roots of $t^4 - t^2 + 1$, so

$$a + b + c + d = -0, \qquad ab + ac + ad + bc + bd + cd = +(-1),$$
$$abc + abd + acd + bcd = -0, \quad abcd = +1$$

and therefore $a^2b^2 + a^2c^2 + a^2d^2 + b^2c^2 + b^2d^2 + c^2d^2 = (-1)^2 - 2(0)(0) = 1$.

(b) We find a polynomial of degree three in $\mathbb{Z}[t]$ whose roots are the cubes of the roots of $t^3 + 2t^2 + 3t + 4 \in \mathbb{Z}[t]$. Let us denote the roots of this polynomial by a,b,c, so that $a + b + c = -2$, $ab + ac + bc = 3$, $abc = 4$. We put $$t^3 + q_1t^2 + q_2t + q_3 = (t - a^3)(t - b^3)(t - c^3).$$

From Theorem 38.11, we know that

$$q_1 = a^3 + b^3 + c^3, \qquad q_2 = a^3b^3 + a^3c^3 + b^3c^3, \qquad q_3 = a^3b^3c^3.$$

Since $s_3 = \sigma_1{}^3 - 3\sigma_1\sigma_2 + 3\sigma_3$, we conclude

$$\begin{aligned} q_1 &= a^3 + b^3 + c^3 \\ &= (a + b + c)^3 - 3(a + b + c)(ab + ac + bc) + 3(abc) \\ &= (-2)^3 - 3(-2)(3) + 3(-4) \\ &= -2. \end{aligned}$$

We find easily that $x^3y^3 + x^3z^3 + y^3z^3 = \sigma_2{}^3 - 3\sigma_1\sigma_2\sigma_3 + 3\sigma_3{}^2$; hence

$$\begin{aligned} q_2 &= a^3b^3 + a^3c^3 + b^3c^3 \\ &= (ab + ac + bc)^3 - 3(a + b + c)(ab + ac + bc)(abc) + 3(abc)^2 \\ &= (3)^3 - 3(-2)(3)(-4) + 3(-4)^2 \\ &= 3. \end{aligned}$$

Finally, $q_3 = a^3b^3c^3$
$= (abc)^3$
$= (-4)^3$
$= -64.$

Thus

$$t^3 - (-2)t^2 + (3)t - (-64) = t^3 + 2t^2 + 3t + 64 \in \mathbb{Z}[x]$$

is a polynomial whose roots are the cubes of the roots of $t^3 + 2t^2 + 3t + 4$.

Exercises

1. Express the following symmetric polynomials over $\mathbb{Z}$ in terms of the elementary symmetric polynomials:
   (a) $x^3y^2 + x^2y^3 + x^3z^2 + x^2z^3 + y^3z^2 + y^2z^3$;
   (b) $x^2y^2 + x^2z^2 + x^2u^2 + y^2z^2 + y^2u^2 + z^2u^2$;
   (c) $x^5 + y^5 + x^5 + x^4y + y^4x + x^4z + z^4x + y^4z + z^4y$.

2. Find a polynomial over $\mathbb{Z}$ whose roots are the
   (a) squares of the roots of $t^3 + 5t^2 + 7t + 1 \in \mathbb{Z}[t]$;
   (b) squares of the roots of $t^5 + 5t^4 - 6t^3 + t^2 - 7t - 4 \in \mathbb{Z}[t]$;
   (c) cubes of the roots of $t^4 - 3t^3 + 2t^2 + 2 \in \mathbb{Z}[t]$.

3. Let K be a field. A rational function $\dfrac{f(x_1,x_2,\ldots,x_m)}{g(x_1,x_2,\ldots,x_m)}$ in $K[x_1,x_2,\ldots,x_m]$ is said to be a *symmetic rational function over* K if

$$\frac{f(x_{1\sigma},x_{2\sigma},\ldots,x_{m\sigma})}{g(x_{1\sigma},x_{2\sigma},\ldots,x_{m\sigma})} = \frac{f(x_1,x_2,\ldots,x_m)}{g(x_1,x_2,\ldots,x_m)}$$

for all $\sigma \in S_n$. Prove that a symmetic rational function over K can be expressed as a fraction of two symmetic polynomials over K. Conclude that any symmetic rational function over K can be written as

$$\frac{p(\sigma_1,\sigma_2,\ldots,\sigma_m)}{q(\sigma_1,\sigma_2,\ldots,\sigma_m)}$$

with suitable polynomials p,q in $K[u_1,u_2,\ldots,u_m]$. (Loosely speaking, any symmetic rational function is a rational function of the elementary symmetric polynomials.)

4. Express the following rational functions over $\mathbb{Z}$ in terms of the elementary symmetric polynomials:
   (a) $\dfrac{x}{y}+\dfrac{y}{x}+\dfrac{x}{z}+\dfrac{z}{x}+\dfrac{y}{z}+\dfrac{z}{y}$;

(b) $\dfrac{x^2}{yz}+\dfrac{y^2}{xz}+\dfrac{z^2}{xy}$;

(c) $\dfrac{1}{1-x}+\dfrac{1}{1-y}+\dfrac{1}{1-z}$.

5. Prove: for any symmetric polynomial $f(x_1,x_2,\ldots,x_m)$ over $\mathbb{Z}$, there is a polynomial $h(u_1,u_2,\ldots,u_m)$ in $\mathbb{Q}[u_1,u_2,\ldots,u_m]$ such that $f(x_1,x_2,\ldots,x_m) = h(s_1,s_2,\ldots,s_m)$, where s_j are the power sums of x_i.

6. Write the symmetric polynomials in Ex. 1 as polynomials in s_j over $\mathbb{Q}$.