An undergraduate course in

Abstract Algebra

Course notes for MATH3002 Rings and Fields

Robert Howlett

An undergraduate course in

Abstract Algebra

by

Robert Howlett

typesetting by TEX

Contents

Foreword...

The purpose of this book is to complement the lectures and thereby decrease, but not eliminate, the necessity of taking lecture notes. Reading the appropriate sections of the book before each lecture should enable you to understand the lecture as it is being given, provided you concentrate! This is particularly important in this course because, as theoretical machinery is developed, the lectures depend more and more heavily upon previous lectures, and students who fail to thoroughly learn the new concepts as they are introduced soon become completely lost.

> $\star\star\star$ *Proofs of the theorems are an important part of this course. You cannot expect to do third year Pure Mathematics without coming to grips with proofs. Mathematics is about proving theorems. You* **will** *be required to know proofs of theorems for the exam.* $\star\star\star$

It is the material dealt with in the lectures, not this book, which defines the syllabus of the course. The book is only intended to assist, and how much overlap there is with the course depends on the whim of the lecturer. There will certainly be things which are in the lectures and not in the book, and vice versa. The lecturer will probably dwell upon topics which are giving students trouble, and omit other topics. However, the book will still provide a reasonable guide to the course.

0
Prerequisites

Students will be assumed to be familiar with the material mentioned in this preliminary chapter. Anyone who is not should inform the lecturer forthwith.

§0a Concerning notation

When reading or writing mathematics you should always remember that the mathematical symbols which are used are simply abbreviations for words. Mechanically replacing the symbols by the words they represent should result in grammatically correct and complete sentences. The meanings of a few commonly used symbols are given in the following table.

Symbols	*To be read as*
$\{ \ldots \mid \ldots \}$	the set of all ... such that ...
$=$	is
$\in$	in *or* is in
$>$	greater than *or* is greater than

Thus for example the following sequence of symbols

$$\{ x \in X \mid x > a \} \neq \emptyset$$

is an abbreviated way of writing the sentence

The set of all x in X such that x is greater than a is not the empty set.

When reading mathematics you should mentally translate all symbols in this fashion. If you cannot do this and obtain meaningful sentences, seek help from your tutor. And make certain that, when you use mathematical symbols yourself, what you write can be translated into meaningful sentences.

§0b Concerning functions

The terminology we use in connection with functions could conceivably differ from that to which you are accustomed; so a list of definitions of the terms we use is provided here.

• The notation '$f\colon A \to B$' (read 'f, from A to B') means that f is a *function* with *domain* A and *codomain* B. In other words, f is a rule which assigns to every element a of the set A an element in the set B denoted by '$f(a)$'.

• A *map* is the same thing as a function. The term *mapping* is also used.

• A function $f\colon A \to B$ is said to be *injective* (or *one-to-one*) if and only if no two distinct elements of A yield the same element of B. In other words, f is injective if and only if for all a_1, $a_2 \in A$, if $f(a_1) = f(a_2)$ then $a_1 = a_2$.

• A function $f\colon A \to B$ is said to be *surjective* (or *onto*) if and only if for every element b of B there is an a in A such that $f(a) = b$.

• If a function is both injective and surjective we say that it is *bijective* (or a one-to-one correspondence).

• The *image* of a function $f\colon A \to B$ is the subset of B consisting of all elements obtained by applying f to elements of A. That is,

$$\operatorname{im} f = \{\, f(a) \mid a \in A \,\}.$$

An alternative notation is '$f(A)$' instead of '$\operatorname{im} f$'. Clearly, f is surjective if and only if $\operatorname{im} f = B$.

• The notation '$a \mapsto b$' means 'a maps to b'; in other words, the function involved assigns the element b to the element a. Thus '$a \mapsto b$ under f' means exactly the same as '$f(a) = b$'.

• If $f\colon A \to B$ is a function and C a subset of B then the *inverse image* or *preimage* of C is the subset of A

$$f^{-1}(C) = \{\, a \in A \mid f(a) \in C \,\}.$$

(The above line reads 'f inverse of C, which is the set of all a in A such that f of a is in C.' Alternatively, one could say 'The inverse image of C under f' instead of 'f inverse of C'.)

§0c Concerning vector spaces

Vector spaces enter into this course only briefly; the facts we use are set out in this section.

Associated with each vector space is a set of *scalars*. In the common and familiar examples this is $\mathbb{R}$, the set of all real numbers, but in general it can be any field. (Fields are defined in Chapter 2.)

Let V be a vector space over F. (That is, F is the associated field of scalars.) Elements of V can be added and multiplied by scalars:

$(*)$ If $v, w \in V$ and $\lambda \in F$ then $v + w, \lambda v \in V$.

These operations of addition and multiplication by scalars satisfy the following properties:

(i) $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$.

(ii) $u + v = v + u$ for all $u, v \in V$.

(iii) There exists an element $0 \in V$ such that $v + 0 = v$ for all $v \in V$.

(iv) For each $v \in V$ there exists an element $-v \in V$ such that $v + (-v) = 0$.

(v) $\lambda(\mu v) = (\lambda\mu)v$ for all $\lambda, \mu \in F$ and all $v \in V$.

(vi) $1v = v$ for all $v \in V$.

(vii) $\lambda(v + w) = \lambda v + \lambda w$ for all $\lambda \in F$ and all $v, w \in V$.

(viii) $(\lambda + \mu)v = \lambda v + \mu v$ for all $\lambda, \mu \in F$ and all $v \in V$.

The properties listed above are in fact the vector space axioms; thus in order to prove that a set V is a vector space over a field F one has only to check that $(*)$ and (i)–(viii) are satisfied.

Let V be a vector space over F and let $v_1, v_2, \ldots v_n \in V$. The elements $v_1, v_2, \ldots v_n$ are said to be *linearly independent* if the following statement is true:

$$\text{If } \lambda_1, \lambda_2, \ldots, \lambda_n \in F \text{ and } \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_n v_n = 0$$
$$\text{then } \lambda_1 = 0,\ \lambda_2 = 0,\ \ldots,\ \lambda_n = 0.$$

The elements $v_1, v_2, \ldots v_n$ are said to *span* the space V if the following statement is true:

$$\text{For every } v \in V \text{ there exist } \lambda_1, \lambda_2, \ldots, \lambda_n \in F$$
$$\text{such that } v = \lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_n v_n.$$

A *basis* of a vector space V is a finite subset of V whose elements are linearly independent and span V.

We can now state the only theorem of vector space theory which is used in this course.

0.1 THEOREM *If a vector space V has a basis then any two bases of V will have the same number of elements.*

Comment ▷▷▷
0.1.1 If V has a basis then the *dimension* of V is by definition the number of elements in a basis. ▷▷▷

§0d Some very obvious things about proofs

When trying to prove something, the logical structure of what you are trying to prove determines the logical structure of the proof. The following observations seem trivial, yet they are often ignored.

- To prove a statement of the form

$$\text{If } p \text{ then } q$$

your first line should be

$$\text{Assume that } p \text{ is true}$$

and your last line

$$\text{Therefore } q \text{ is true.}$$

- The statement

$$p \text{ if and only if } q$$

is logically equivalent to

$$\text{If } p \text{ then } q \text{ and if } q \text{ then } p,$$

and so the proof of such a statement involves first assuming p and proving q, then assuming q and proving p.

- To prove a statement of the form

$$\text{All } xxxx\text{'s are } yyyy\text{'s,}$$

the first line of your proof should be

$$\text{Let } a \text{ be an } xxxx$$

and the last line should be

$$\text{Therefore } a \text{ is a } yyyy.$$

(The second line could very well involve invoking the definition of '$xxxx$' or some theorem about $xxxx$'s to determine things about a; similarly the second to last line might correspond to the definition of '$yyyy$'.)

When trying to construct a proof it is sometimes useful to assume the opposite of the thing you are trying to prove, with a view to obtaining a contradiction. This technique is known as "indirect proof" (or "proof by contradiction"). The idea is that the conclusion c is a consequence of the hypotheses h_1, h_2, ..., if and only if the negation of c is incompatible with h_1, h_2, ... Hence we may assume the negation of c as an extra hypothesis, along with h_1, h_2 etc., and the task is then to show that the hypotheses contradict each other. Note, however, that although indirect proof is a legitimate method of proof in all situations, it is not a good policy to always use indirect proof as a matter of course. Most proofs are naturally expressed as direct proofs, and to recast them as indirect proofs may make them more complicated than necessary.

—Examples—

#1 Suppose that you wish to prove that a function $\lambda\colon X \to Y$ is injective. Consult the definition of injective. You are trying to prove the following statement:

$$\text{For all } x_1,\ x_2 \in X, \text{ if } \lambda(x_1) = \lambda(x_2) \text{ then } x_1 = x_2.$$

So the first two lines of your proof should be as follows:

$$\text{Let } x_1,\ x_2 \in X.$$
$$\text{Assume that } \lambda(x_1) = \lambda(x_2).$$

Then you will presumably consult the definition of the function λ to derive consequences of $\lambda(x_1) = \lambda(x_2)$, and eventually you will reach the final line

$$\text{Therefore } x_1 = x_2.$$

#2 Suppose you wish to prove that $\lambda\colon X \to Y$ is surjective. That is, you wish to prove

$$\text{For every } y \in Y \text{ there exists } x \in X \text{ with } \lambda(x) = y.$$

Your first line must be

$$\text{Let } y \text{ be an arbitrary element of } Y.$$

Somewhere in the middle of the proof you will have to somehow define an element x of the set X (the definition of x is bound to involve y in some way), and the last line of your proof has to be

$$\text{Therefore } \lambda(x) = y.$$

#3 Suppose that A and B are sets, and you wish to prove that $A \subseteq B$. (That is, A is a subset of or equal to B.) By definition the statement '$A \subseteq B$' is logically equivalent to

$$\text{All elements of } A \text{ are elements of } B.$$

So your first line should be

$$\text{Let } x \in A$$

and your last line should be

$$\text{Therefore } x \in B.$$

#4 Suppose that you wish to prove that $A = B$, where A and B are sets. The following statements are all logically equivalent to '$A = B$':

(i) For all x, $x \in A$ if and only if $x \in B$.

(ii) (For all x)$\big($(if $x \in A$ then $x \in B$) and (if $x \in B$ then $x \in A$)$\big)$.

(iii) All elements of A are elements of B and all elements of B are elements of A.

(iv) $A \subseteq B$ and $B \subseteq A$.

You must do two proofs of the general form given in #3 above.

1

Ruler and compass constructions

Abstract algebra is essentially a tool for other branches of mathematics. Many problems can be clarified and solved by identifying underlying structure and focussing attention on it to the exclusion of peripheral information which may only serve to confuse. Moreover, common underlying structures sometimes occur in widely varying contexts, and are more easily identifiable for having been previously studied in their own right. In this course we shall illustrate this idea by taking three classical geometrical problems, translating them into algebraic problems, and then using the techniques of modern abstract algebra to investigate them.

§1a Three problems

Geometrical problems arose very early in the history of civilization, presumably because of their relevance to architecture and surveying. The most basic and readily available geometrical tools are ruler and compass, for constructing straight lines and circles; thus it is natural to ask what geometrical problems can be solved with these tools.†

It is said that the citizens of Delos in ancient Greece, when in the grips of a plague, consulted an oracle for advice. They were told that a god was displeased with their cubical altar stone, which should be immediately replaced by one double the size. The Delians doubled the length, breadth and depth of their altar; however, this increased its volume eightfold, and the enraged god worsened the plague.

Although some historians dispute the authenticity of this story, the so-called "Delian problem"

† Note that the ruler is assumed to be unmarked; that is, it is not a measuring device but simply an instrument for ruling lines.

(1) Given a cube, construct another cube with double the volume

is one of the most celebrated problems of ancient mathematics. There are two other classical problems of similar stature:

(2) Construct a square with the same area as a given circle

(3) Trisect a given angle.

In this course we will investigate whether problems (1), (2) and (3) can be solved by ruler and compass constructions. It turns out that they cannot.

We should comment, however, that although the ancient mathematicians were unable to prove that these problems were insoluble by ruler and compass, they did solve them by using curves other than circles and straight lines.

§1b Some examples of constructions

Before trying to prove that some things cannot be done with ruler and compass, we need to investigate what **can** be done with those tools. Much of what follows may be familiar to you already.

#1 Given straight lines AB and AC intersecting at A the angle BAC can be bisected, as follows. Draw a circle centred at A, and let X, Y be the points where this circle meets AB, AC. Draw circles of equal radii centred at X and Y, and let T be a point of intersection of these circles. (The radius must be chosen large enough so that the circles intersect.) Then AT bisects the given angle BAC.

#2 Given lines AB and AC intersecting at A and a line PQ, the angle BAC can be copied at P, as follows. Draw congruent circles $\mathcal{C}_A$, $\mathcal{C}_P$ centred at A and P. Let $\mathcal{C}_A$ intersect AB at X and AC at Y, and let $\mathcal{C}_P$ intersect PQ at V. Draw a circle with centre V and radius equal to XY, and let T be a point of intersection of this circle and $\mathcal{C}_P$. Then the angle TPQ equals the angle BAC.

#3 Given a point A and a line PQ, one can draw a line through A parallel to PQ. Simply draw any line through A intersecting PQ at some point X, and then copy the angle AXQ at the point A.

#4 Given a line AB one can construct a point T such that the angle TAB equals $\frac{\pi}{3}$ radians (60 degrees). Simply choose T to be a point of intersection of the circle centred at A and passing through B and the circle centred at B and passing through A.

#5 Given line segments of lengths r, s and t one can construct a line segment of length rt/s, as follows. Draw distinct lines AP, AQ intersecting at A and draw circles $\mathcal{C}_r$, $\mathcal{C}_s$ and $\mathcal{C}_t$ of radii r, s and t centred at A. Let $\mathcal{C}_r$ intersect AP at B and let $\mathcal{C}_s$, $\mathcal{C}_t$ intersect AQ at X, Y. Draw a line through Y parallel to XB, and let C be the point at which it intersects AP. Then AC has the required length.

#6 There are simple constructions for angles equal to the sum and difference of two given angles, lengths equal to the sum and difference of two given lengths, and for a/n and na, where a is a given length and n a given positive integer. See the exercises at the end of the chapter.

#7 Given line segments of lengths a and b, where $a \geq b$, it is possible to construct a line segment of length $\sqrt{ab}$, as follows. First, construct line segments of lengths $r_1 = \frac{1}{2}(a+b)$ and $r_2 = \frac{1}{2}(a-b)$, and draw circles of radii r_1 and r_2 with the same centre O. Draw a line through O intersecting the smaller circle at P, and draw a line through P perpendicular to OP. (A right-angle can be constructed, for instance, by constructing an angle of $\frac{\pi}{3}$, bisecting it, and adding on another angle of $\frac{\pi}{3}$.) Let this perpendicular meet the large circle at Q. Then PQ has the required length.

Further ruler and compass constructions are dealt with in the exercises.

§1c Constructible numbers

Consider the Delian Problem once more: we are given a cube and wish to double its volume. We may as well choose our units of length so that the given cube has sides of length one. Then our problem is to construct a line segment of length $\sqrt[3]{2}$. The other problems can be stated similarly. A circle of unit radius has area π; to construct a square of this area one must construct a line segment of length $\sqrt{\pi}$. A right-angled triangle with unit hypotenuse and an angle θ has other sides $\cos\theta$ and $\sin\theta$; to trisect θ one must construct $\cos\left(\frac{\theta}{3}\right)$. So the problems become:

(1) Given a unit line segment, construct one of length $\sqrt{\pi}$.

(2) Given a unit line segment, construct one of length $\sqrt[3]{2}$.

(3) Given line segments of lengths 1 and $\cos\theta$, construct one of length $\cos\left(\frac{\theta}{3}\right)$.

To show that Problem 3 cannot be solved by ruler and compass, it will be sufficient to show that it cannot be done in the case $\theta = \frac{\pi}{3}$. In this case $\cos\theta = \frac{1}{2}$. Since a line segment of length $\frac{1}{2}$ can be constructed given a unit line segment, it suffices to show that given only a unit line segment it is not possible to construct one of length $\cos\left(\frac{\pi}{9}\right)$. In other words, an angle of 20 degrees cannot be constructed.

So, assume that we are given a line segment of length one. We first use this segment to define a coordinate system. Let one of the endpoints of the segment be the origin $(0,0)$ and the other endpoint the point $(1,0)$. After drawing a line through $(0,0)$ perpendicular to the x-axis we can find the position of the point $(0,1)$ by drawing a circle of centre $(0,0)$ and radius 1. We can now proceed to construct further points, lines and circles, in accordance with the following rules. We can construct

(a) a line if it passes through two previously constructed points,

(b) a circle if its centre is a previously constructed point and its radius the distance between two previously constructed points,

(c) a point if it is the point of intersection of two lines or circles or a circle and a line constructed in accordance with (a) and (b).

We now a define a number to be *constructible* if it is a coordinate of a constructible point. (Note that since lines perpendicular to the axes can be constructed, the point (a,b) can be constructed if and only if the points $(a,0)$ and $(0,b)$ can both be constructed. Furthermore, since a circle of radius a and centre O cuts the x-axis at $(a,0)$ and the y-axis at $(0,a)$, a number is constructible as an x-coordinate if and only if it is constructible as a y-coordinate.) Our aim will be to describe completely the set of constructible numbers and hence show that $\sqrt{\pi}$, $\sqrt[3]{2}$ and $\cos\left(\frac{\pi}{9}\right)$ are not constructible.

1.1 Theorem *If a and b are constructible numbers then so are $a+b$, $-a$, ab, a^{-1} (if $a \neq 0$) and $\sqrt{a}$ (if $a \geq 0$).*

Proof. Let a and b be constructible numbers. Then the points $(a,0)$ and $(0,b)$ can be constructed in accordance with rules (a), (b) and (c) above. The point $(a+b,0)$ is the point of intersection of the x-axis and a circle centre $(a,0)$ and radius the distance between $(0,0)$ and $(0,b)$; hence it is constructible by rule (b) above. So $a+b$ is constructible.

Draw the line joining $(0,b)$ and $(1,0)$. Using the process described in §1b the line through $(a,0)$ parallel to this can be constructed. It meets the y-axis at $(0,ab)$. Hence ab is constructible.

The proofs of the other parts are similarly based on constructions given in §1b, and are omitted. □

Comment ▷▷▷

1.1.1 Sometimes standard geometrical constructions include instructions like 'Draw an arbitrary line through A' or 'Draw any circle with centre B and radius large enough to intersect with PQ'. It is not immediately obvious that the rules (a), (b) and (c) given above are strong enough to permit constructions such as these. However, it follows from the above theorem that every element of the set $\mathbb{Q}$ of all rational numbers (numbers of the form n/m where n and m are integers) is constructible. When asked to draw an arbitrary line through A we may as well join A to a point with rational coordinates, and when asked to draw an arbitrary circle with centre B we may as well draw one that has rational radius. There will always be a rational number of suitable size, since rational numbers exist which are arbitrarily close to any given real number. So in fact anything that can be constructed with ruler and compass can be constructed by just following rules (a), (b) and (c). ◁◁◁

Obviously one cannot draw an infinite number of circles and/or lines; so the number of points obtained in any geometrical construction is finite. Suppose that $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the points occurring in a given construction, listed in the order in which they are constructed, with $\alpha_0 = (0,0)$ and $\alpha_1 = (1,0)$. Let the coordinates of α_i be (a_i, b_i) (for each i). Suppose that we now wish to construct another point. According to the rules we can draw a circle with centre α_i and radius equal to the distance between α_j and α_k (for any choice of i, j and k) and we can draw a straight line joining α_i and α_j (for any i and j). The points of intersection of such points and lines are the only points that we can construct at the next stage. (We can, of course, get further points by repeating the process.) The equation of such a circle is

$$(x - a_i)^2 + (y - b_i)^2 = (a_j - a_k)^2 + (b_j - b_k)^2 \tag{1}$$

and the equation of such a line is

$$(b_j - b_i)x - (a_j - a_i)y = a_i b_j - a_j b_i \,. \tag{2}$$

Hence the coordinates of the next point obtained can be found by solving simultaneously two equations each having one or other of the above forms.

1.2 THEOREM *Let $\alpha_1, \alpha_2, \ldots, \alpha_k$ be the points obtained in a ruler and compass construction, listed in the order obtained. For each n let S_n be the set of all real numbers obtainable from the coordinates of $\alpha_1, \alpha_2, \ldots, \alpha_n$ by finite sequences of operations of addition, subtraction, multiplication and division. Then there exists $a_n \in S_n$ such that the coordinates of α_{n+1} lie in the set $S_n(\sqrt{a_n}) = \{\, p + q\sqrt{a_n} \mid p, q \in S_n \,\}$.*

Proof. As described in the preamble, the coordinates of α_{n+1} are obtained by solving simultaneously two equations like (1) or (2). There are three cases to consider: both of the form (1), both of the form (2), and one of each. Let us deal with the last case first.

On expanding, (1) has the form

$$x^2 + y^2 + px + qy + r = 0 \tag{3}$$

with $p, q, r \in S_n$. Similarly, (2) has the form

$$sx + ty + u = 0 \tag{4}$$

with $s, t, u \in S_n$ and either $s \neq 0$ or $t \neq 0$. If $s \neq 0$, rewrite (4) as

$$x = -\frac{t}{s}y - \frac{u}{s} \tag{5}$$

and substitute into (3). This gives

$$p'y^2 + q'y + r' = 0, \tag{6}$$

where in fact

$$\begin{aligned} p' &= 1 + \frac{t^2}{s^2} \\ q' &= q - \frac{pt}{s} + \frac{2ut}{s^2}\,, \\ r' &= r - \frac{pu}{s} + \frac{u^2}{s^2} \end{aligned}$$

but all that concerns us is that $p', q', r' \in S_n$ and $p' \neq 0$. Let $a_n = (q')^2 - 4p'r'$, an element of S_n. Using the quadratic formula to solve (6) shows that the y-coordinate of α_{n+1} is in $S_n(\sqrt{a}_n)$ and then (5) shows that the x-coordinate is too.

In the case that both equations have the form (1) (so that α_{n+1} is the point of intersection of two circles) we must solve simultaneously equation (3) and a similar equation

$$x^2 + y^2 + lx + my + n = 0. \tag{7}$$

But subtracting (7) from (3) gives an equation of the form (4); so we can proceed as before.

In the remaining case both equations have the form (4), with coefficients in the set S_n. To solve them just involves operations of addition, subtraction, multiplication and division, and since by definition these operations cannot take us outside the set S_n it follows that the coordinates of α_{n+1} lie in $S_n = S_n(\sqrt{0})$. □

What Theorems 1.1 and 1.2 show is, essentially, that with ruler and compass one can add, subtract, multiply, divide and take square roots, and nothing else. To solve the Delian Problem one must construct the cube root of two. You may think that we have already settled the matter, since obviously it is impossible to find a cube root by taking square roots. Unfortunately, however, this is not obvious at all. How do you know, for instance, that the following formula is not correct?

$$\sqrt[3]{2} = \frac{a^3 + 4b^3}{3a^2b - 2ab^2 + 2b\sqrt{2b^4 + a^2b^2 - a^3b}}$$

where

$$a = 8 + 5\sqrt{10}$$

$$b = -10 + 6\sqrt{10} + \sqrt{225 - 40\sqrt{10}}.$$

Or if you can show that it is not, how do you know that there is not some far more complicated formula of the same kind which is correct? The algebraic machinery which will be developed in the subsequent chapters will show that there is not.

Exercises

1. Describe carefully how to perform the constructions mentioned in #6. (Hint: For a/n and na use #5.)

2. Given a line segment of length 1 unit, construct a line segment of each of the following lengths:

$$(i)\quad \sqrt{2} \qquad (ii)\quad \sqrt{3} \qquad (iii)\quad \sqrt{3}-\sqrt{2}$$
$$(iv)\quad \sqrt{2}\sqrt{3} \qquad (v)\quad \frac{\sqrt{3}}{\sqrt{2}} \qquad (vi)\quad \sqrt{\sqrt{2}+\sqrt{3}}.$$

Measure the line segments in each case to check the accuracy of your constructions.

3. Let $\theta = 2\pi/5$ and let $\alpha = e^{i\theta} = \cos\theta + \boldsymbol{i}\sin\theta$, where $\boldsymbol{i}$ is a complex square root of -1. Thus α is a complex fifth root of 1. Show that

$$\begin{aligned} x^4 + x^3 + x^2 + x + 1 &= (x-\alpha)(x-\alpha^{-1})(x-\alpha^2)(x-\alpha^{-2}) \\ &= (x^2 - 2(\cos\theta)x + 1)(x^2 - 2(\cos 2\theta)x + 1). \end{aligned}$$

Hence show that

$$\cos\theta + \cos 2\theta = -\frac{1}{2}$$
$$\cos\theta\cos 2\theta = -\frac{1}{4}$$

and solve to find $\cos\theta$.

4. Is it possible to construct an angle of $2\pi/5$?

5. Let OAB be an isosceles triangle with $OA = OB$ and the angle at O equal to $\pi/5$. Let the bisector of the angle OBA meet OA at the point P.

(*i*) Prove that the triangles OAB and BAP are similar.

(*ii*) Suppose that OA has length 1 and let AB have length x. Use the first part to prove that $x^2 + x - 1 = 0$.

(Hint: Prove that PA has length $1 - x$.)

(*iii*) Use the previous part to show that a regular decagon inscribed in a unit circle has sides of length $\frac{\sqrt{5}-1}{2}$, and hence devise a ruler and compass construction for a regular decagon.

6. Let $\theta = 2\pi/17$ and let $\omega = e^{i\theta} = \cos\theta + \boldsymbol{i}\sin\theta$, a complex 17^{th} root of 1. Prove that

$$\begin{aligned} &x^{16} + x^{15} + x^{14} + \cdots + x^2 + x + 1 \\ &= (x-\omega)(x-\omega^{-1})(x-\omega^2)(x-\omega^{-2})\ldots(x-\omega^8)(x-\omega^{-8}) \\ &= (x^2 - (2\cos\theta)x + 1)(x^2 - (2\cos 2\theta)x + 1)\ldots(x^2 - (2\cos 8\theta)x + 1). \end{aligned}$$

7. Let

$$\alpha_1 = \frac{-1+\sqrt{17}}{2}, \quad \alpha_2 = \frac{-1-\sqrt{17}}{2}$$

$$\beta_1 = \frac{1}{2}(\alpha_1+\sqrt{\alpha_1^2+4})$$

$$\beta_2 = \frac{1}{2}(\alpha_1-\sqrt{\alpha_1^2+4})$$

$$\beta_3 = \frac{1}{2}(\alpha_2+\sqrt{\alpha_2^2+4})$$

$$\beta_4 = \frac{1}{2}(\alpha_2-\sqrt{\alpha_2^2+4})$$

$$\gamma_1 = \frac{1}{2}(\beta_1 + \sqrt{\beta_1^2-4\beta_3}) \quad \gamma_5 = \frac{1}{2}(\beta_3 + \sqrt{\beta_3^2-4\beta_1})$$

$$\gamma_2 = \frac{1}{2}(\beta_1 - \sqrt{\beta_1^2-4\beta_3}) \quad \gamma_6 = \frac{1}{2}(\beta_3 - \sqrt{\beta_3^2-4\beta_1})$$

$$\gamma_3 = \frac{1}{2}(\beta_2 + \sqrt{\beta_2^2-4\beta_4}) \quad \gamma_7 = \frac{1}{2}(\beta_4 + \sqrt{\beta_4^2-4\beta_2})$$

$$\gamma_4 = \frac{1}{2}(\beta_2 - \sqrt{\beta_2^2-4\beta_4}) \quad \gamma_8 = \frac{1}{2}(\beta_4 - \sqrt{\beta_4^2-4\beta_2}).$$

(*i*) Check that $\gamma_1+\gamma_2 = \beta_1$ and $\gamma_1\gamma_2 = \beta_3$, and hence show that

$$(x^2-\gamma_1 x+1)(x^2-\gamma_2 x+1) = x^4-\beta_1 x^3+(2+\beta_3)x^2-\beta_1 x+1.$$

Similarly

$$(x^2-\gamma_3 x+1)(x^2-\gamma_4 x+1) = x^4-\beta_2 x^3+(2+\beta_4)x^2-\beta_2 x+1,$$
$$(x^2-\gamma_5 x+1)(x^2-\gamma_6 x+1) = x^4-\beta_3 x^3+(2+\beta_1)x^2-\beta_3 x+1,$$
$$(x^2-\gamma_7 x+1)(x^2-\gamma_8 x+1) = x^4-\beta_4 x^3+(2+\beta_2)x^2-\beta_4 x+1.$$

(*ii*) Check that

$$\begin{aligned}(x^4-\beta_1 x^3+(2+\beta_3)x^2-\beta_1 x+1)(x^4-\beta_2 x^3+(2+\beta_4)x^2-\beta_2 x+1)\\ = x^8+\left(\frac{1-\sqrt{17}}{2}\right)x^7+\left(\frac{5-\sqrt{17}}{2}\right)x^6+\left(\frac{7-\sqrt{17}}{2}\right)x^5+(2-\sqrt{17})x^4\\ +\left(\frac{7-\sqrt{17}}{2}\right)x^3+\left(\frac{5-\sqrt{17}}{2}\right)x^2+\left(\frac{1-\sqrt{17}}{2}\right)x+1.\end{aligned}$$

The product of the other two quartics appearing in part (*i*) is similar: just replace $-\sqrt{17}$ by $\sqrt{17}$.

(*iii*) Multiply the eighth degree polynomial appearing in part (*ii*) by its conjugate (obtained by replacing $-\sqrt{17}$ by $\sqrt{17}$) and show that the result is $x^{16}+x^{15}+x^{14}+\cdots+x^2+x+1$. Comparing with the previous exercise, deduce that the numbers γ_1, γ_2, ..., γ_8 are equal to $2\cos\theta$, $2\cos 2\theta$, ..., $2\cos 8\theta$ (not necessarily in that order), where $\theta = 2\pi/17$.

(*iv*) Use the previous parts to deduce that a regular seventeen sided polygon can be constructed with ruler and compass.

2
Introduction to rings

In this chapter we introduce the concepts which will be fundamental to the rest of the course, and which are necessary to adequately understand the set of constructible numbers.

§2a Operations on sets

If a and b are real numbers their sum $a + b$ and product ab are also real numbers. Addition and multiplication are examples of operations on the set of real numbers. Operations can be defined in many different ways on many different sets. For example, division of nonzero real numbers, addition and multiplication of 2×2 matrices over $\mathbb{R}$ (where $\mathbb{R}$ is the set of all real numbers) and so on. Let us state precisely what is meant by 'operation':

2.1 Definition An *operation* on a set S is a rule which assigns to each ordered pair of elements of S a uniquely determined element of S.

Thus, for example, addition is the rule which assigns to the ordered pair (a, b) of real numbers the real number $a + b$. Since the set of all ordered pairs of elements of S is usually denoted by '$S \times S$',

$$S \times S = \{\, (a, b) \mid a \in S \text{ and } b \in S \,\},$$

we could alternatively state Definition 2.1 as follows: an operation on S is a function $S \times S \longrightarrow S$. Addition on $\mathbb{R}$ is the function

$$\begin{aligned} \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{R} \\ (a, b) &\longmapsto a + b. \end{aligned}$$

In this course we will be concerned with many examples of sets equipped with two operations which have properties resembling addition and multiplication of numbers. The confusing thing is that sometimes some of the familiar

properties are not satisfied. For example, addition and multiplication can be defined on all the following sets:

$\mathbb{R}$	(real numbers)
$\mathbb{Z}$	(integers)
$\mathrm{Mat}(2,\mathbb{R})$	(2×2 matrices whose entries are real numbers)
$2\mathbb{Z}$	(even integers)
$\mathbb{R}[X]$	(polynomials in X with real coefficients—expressions like $2 + 5X + X^2$).

Each of these sets possesses a zero element; that is, an element 0 such that $\alpha + 0 = \alpha = 0 + \alpha$ for all elements α in the set. In four of the five examples the product of two nonzero elements is nonzero; however, this property fails for $\mathrm{Mat}(2,\mathbb{R})$. Similarly, in four of the examples there is an identity element; that is an element 1 such that $\alpha 1 = \alpha = 1\alpha$ for all α. However, this property fails for $2\mathbb{Z}$ (no **even** integer is an identity element). In $\mathbb{R}$ for any two nonzero elements a and b there is another element c such that $a = bc$. None of the other sets have this property. And the rule that $\alpha\beta = \beta\alpha$ for all α and β is not satisfied in $\mathrm{Mat}(2,\mathbb{R})$ but is in all the others.

We attempt to bring some order to this chaos by listing the most important properties, investigating which properties are consequences of which other properties, and classifying the various systems according to which properties hold.

§2b The basic definitions

2.2 Definition A *ring* is a set R together with two operations on R, called *addition* $\big((a,b) \mapsto a+b\big)$ and *multiplication* $\big((a,b) \mapsto ab\big)$, such that the following axioms are satisfied:

(i) $(a+b)+c = a+(b+c)$ for all $a, b, c \in R$.
(That is, addition is *associative*.)

(ii) There is an element $0 \in R$ such that $a + 0 = 0 + a = a$ for all $a \in R$.
(There is a *zero* element.)

(iii) For each $a \in R$ there is a $b \in R$ such that $a + b = b + a = 0$.
(Each element has a *negative*.)

(iv) $a + b = b + a$ for all $a, b \in R$. (Addition is *commutative*.)

(v) $(ab)c = a(bc)$ for all $a, b, c \in R$. (Multiplication is associative.)

(vi) $a(b+c) = ab + ac$ and $(a+b)c = ac + bc$ for all $a, b, c \in R$.
(Both *distributive* laws hold.)

Comment ▷▷▷
2.2.1 The five examples mentioned in §2a above are all rings. ▷▷▷

2.3 DEFINITION A *commutative ring* is a ring which satisfies $ab = ba$ for all elements a, b.

That is, a commutative ring is a ring which satisfies the commutative law for multiplication. (All rings satisfy the commutative law for addition.)

2.4 DEFINITION If R is a ring an element $e \in R$ is called an *identity* if $ea = ae = a$ for all $a \in R$.

Comment ▷▷▷
2.4.1 We will almost always use the symbol '1' rather than 'e' to denote an identity element. Not all rings have identity elements—for example $2\mathbb{Z}$ does not have one. ▷▷▷

2.5 DEFINITION If R is a commutative ring and $a \in R$ is such that $a \neq 0$ and $ab = 0$ for some nonzero $b \in R$ then a is called a *zero divisor*.

2.6 DEFINITION An *integral domain* is a commutative ring which has an identity element which is nonzero (that is, $1 \neq 0$) and no zero divisors.

Thus, in an integral domain the following property holds:

2.6.1 For all a, b, if $a \neq 0$ and $b \neq 0$ then $ab \neq 0$.

2.7 DEFINITION Let R be a ring with an identity and let $a \in R$. An element $b \in R$ is called a *multiplicative inverse* of a if $ab = ba = 1$.

Comment ▷▷▷
2.7.1 It can be proved—see the exercises at the end of the chapter—that if an element has a multiplicative inverse then the inverse is unique. That is, if b, $c \in R$ satisfy $ab = ba = 1$ and $ac = ca = 1$ then $b = c$. This fact means that there is no ambiguity in using the usual notation 'a^{-1}' for the inverse of an element a which has an inverse. (Compare with the remarks following Theorem 2.9 below.) ▷▷▷

2.8 DEFINITION A *field* is a commutative ring in which there is a nonzero identity element, and every nonzero element has a multiplicative inverse.

Thus, in a field the following property holds:

2.8.1 For all $a \neq 0$ there exists a b such that $ab = ba = 1$ (where 1 is the identity element).

We wish to investigate properties of the set of constructible numbers, which, as it happens, is a field. However, it turns out to be necessary to study other rings first before we can adequately describe and understand the relevant properties of constructible numbers. To familiarize ourselves with the various concepts we start by considering some examples.

#1 Fields

By definition a field satisfies all the ring axioms, and also

(i) multiplication is commutative,

(ii) there exists an identity element $1 \neq 0$,

(iii) all nonzero elements have multiplicative inverses.

The following are fields:

$\mathbb{R}$	(the set of all real numbers),
$\mathbb{Q}$	(the set of all rational numbers),
$\mathbb{C}$	(the set of all complex numbers),
$\mathbb{Q}[\sqrt{2}]$	(the set of all numbers of the form $a+b\sqrt{2}$ for $a, b \in \mathbb{Q}$),
Con	(the set of all real numbers which are constructible—that is, $\{ a \in \mathbb{R} \mid a \text{ is constructible} \}$).

(It is straightforward to check that $\mathbb{R}$ and $\mathbb{Q}$ satisfy the field axioms; a little more work is needed for the other two examples.)

#2 Integral domains

By definition an integral domain satisfies all the ring axioms, and also

(i) multiplication is commutative,

(ii) there exists an identity element $1 \neq 0$,

(iii) there are no zero divisors. (That is, there do not exist elements a, b such that $a \neq 0$ and $b \neq 0$ but $ab = 0$.)

The following are integral domains:

all the fields listed in #1 above,
$\mathbb{Z}$ (the set of all integers),
$\mathbb{R}[X]$ (the set of all polynomials in X with coefficients from the field $\mathbb{R}$),
$\mathbb{Z}[X]$ (the set of all polynomials in X with coefficients from the integral domain $\mathbb{Z}$).

We will prove in §2d below that all fields are integral domains.

#3 Other commutative rings

There exist commutative rings with no zero divisors which are not integral domains because they do not have identity elements. Examples of this are $2\mathbb{Z}$ (the set of all even integers), $3\mathbb{Z}$ (the set of all integers divisible by 3), and so on. In general, however, commutative rings are likely to have zero divisors. An example is:

$$\left\{ \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \;\middle|\; a,\, b \in \mathbb{R} \right\},$$

the set of all diagonal 2×2 matrices over $\mathbb{R}$. Further examples are provided by the rings $\mathbb{Z}_n$ (to be defined later).

#4 Other rings

$\mathrm{Mat}(2, \mathbb{R})$, $\mathrm{Mat}(3, \mathbb{R})$, ... (that is, square matrices of a given size over $\mathbb{R}$) are examples of noncommutative rings. In fact, as we will see in the next section, $\mathrm{Mat}(n, R)$ is a ring for any positive integer n and any ring R. Thus, for instance, $\mathrm{Mat}(5, 2\mathbb{Z})$ and $\mathrm{Mat}(4, \mathbb{R}[X])$ are examples of rings. One can also have matrices whose entries are matrices; thus $\mathrm{Mat}(2, \mathrm{Mat}(2, \mathbb{R}))$ is the ring of all 2×2 matrices whose entries are 2×2 matrices over $\mathbb{R}$.

Of the different kinds of rings listed above, fields are the simplest, since all of the usual properties of multiplication are satisfied. Next come integral domains, which only lack multiplicative inverses, then arbitrary commutative rings, and finally arbitrary rings, which are the most complicated since very few requirements are placed upon the multiplicative structure.

§2c Two ways of forming rings

Our principal theoretical objective in this course is to understand field extensions; that is, relationships that hold between a field and larger fields having the given field as a subset. (As, for instance, $\mathbb{Q} \subset \mathbf{Con} \subset \mathbb{R}$.) For this we

need to study various methods for constructing new rings from old ones, and this also increases our store of examples of rings. In this section we give two such methods of constructing rings.

#5 Direct sums

If R and S are rings we may define operations of addition and multiplication on the set $\mathcal{D}$ of all ordered pairs (r, s) with $r \in R$ and $s \in S$. We define

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2)$$
$$(r_1, s_1)(r_2, s_2) = (r_1 r_2, s_1 s_2).$$

(That is, the operations are defined "componentwise".) With these operations $\mathcal{D}$ is a ring, called the *direct sum* $R \dot{+} S$ of R and S.

To prove that the direct sum of R and S is a ring it is necessary to prove that Axioms (i)–(vi) of Definition 2.2 are satisfied. In each case the proof that $R \dot{+} S$ satisfies a given axiom simply involves using the same axiom for R and S. We prove only the first three axioms here, leaving the others as exercises.

Proof. (i) Let a, b, $c \in R \dot{+} S$. Then we have $a = (r_1, s_1)$, $b = (r_2, s_2)$ and $c = (r_3, s_3)$, for some r_1, r_2, $r_3 \in R$ and s_1, s_2, $s_3 \in S$, and so

$$\begin{aligned}
(a + b) + c &= ((r_1, s_1) + (r_2, s_2)) + (r_3, s_3) \\
&= (r_1 + r_2\,,\ s_1 + s_2) + (r_3, s_3) && \text{(by the definition of addition in } R \dot{+} S) \\
&= ((r_1 + r_2) + r_3\,,\ (s_1 + s_2) + s_3) && \text{(similarly)} \\
&= (r_1 + (r_2 + r_3)\,,\ s_1 + (s_2 + s_3)) && \text{(by Axiom (i) for } R \text{ and } S) \\
&= (r_1, s_1) + (r_2 + r_3\,,\ s_2 + s_3) \\
&= (r_1, s_1) + ((r_2, s_2) + (r_3, s_3)) \\
&= a + (b + c)
\end{aligned}$$

as required.

(ii) Let 0_R and 0_S be the zero elements of R and S, and let a be any element of $R \dot{+} S$. We have $a = (r, s)$ for some $r \in R$, $s \in S$. Now

$$a + (0_R, 0_S) = (r, s) + (0_R, 0_S) = (r + 0_R, s + 0_S) = (r, s) = a,$$

and similarly $(0_R, 0_S) + a = a$. Thus $(0_R, 0_S)$ is a zero element for $R \dot{+} S$.

(iii) Let a be an arbitrary element of $R \dot{+} S$. There exist $r \in R$ and $s \in S$ with $a = (r, s)$, and if we let $b = (-r, -s)$ then

$$a + b = (r, s) + (-r, -s) = \big((r + (-r))\,,\ (s + (-s))\big) = (0_R, 0_S).$$

Similarly $b + a = (0_R, 0_S)$, and since $(0_R, 0_S)$ is the zero element of $R \dot{+} S$ this shows that b is a negative of a. □

#6 Square matrices

Suppose that R is a ring and a_1, a_2, a_3, ... are elements of R. The axiom

$$(a+b)+c = a+(b+c) \qquad \text{for all } a,\, b,\, c \in R$$

shows that the expression $a+b+c$ is well defined. It makes no difference which way parentheses are inserted: the additions can be done in either order. The same obviously must apply for any number of terms; so there is no ambiguity if the parentheses are omitted. That is, the expression $a_1 + a_2 + \cdots + a_n$ is well defined. Moreover, the axiom

$$a+b = b+a \qquad \text{for all } a,\, b \in R$$

shows that the order of terms is immaterial. So there is no harm in using sigma notation: $a_1 + a_2 + \cdots + a_n = \sum_{i=1}^{n} a_i$.
(CAUTION: Rings are **not necessarily commutative**; so the same kind of thing does not apply for multiplication. In the expression $a_1 a_2 \ldots a_n$ the ordering of the factors must not be changed.)

The distributive laws $a(b+c) = ab + ac$ and $(a+b)c = ac + bc$ imply that

$$b(a_1 + a_2 + \cdots + a_n) = ba_1 + ba_2 + \cdots + ba_n$$

and

$$(a_1 + a_2 + \cdots + a_n)b = a_1 b + a_2 b + \cdots + a_n b\,,$$

or, in sigma notation,

$$b(\sum_{i=1}^{n} a_i) = \sum_{i=1}^{n} ba_i$$

and

$$(\sum_{i=1}^{n} a_i)b = \sum_{i=1}^{n} a_i b\,.$$

These formulae are known as the *generalized distributive laws.*

It is also worth noting that it is legitimate to interchange the order of summation in double sums:

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{m} a_{ij} \right) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} a_{ij} \right)$$

where the a_{ij} are ring elements (for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$). This follows from the commutative law for addition—ring axiom (iv).

If n is a positive integer let $\mathrm{Mat}(n, R)$ be the set of all $n \times n$ matrices with entries from R. If $A \in \mathrm{Mat}(n, R)$ and $i, j \in \{1, 2, \ldots, n\}$ let A_{ij} be the entry in the i^{th} row and j^{th} column of A. (Thus $A_{ij} \in R$ for each i, j.) Define addition and multiplication on $\mathrm{Mat}(n, R)$ by

$$(A + B)_{ij} = A_{ij} + B_{ij}$$
$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

That is, addition is defined componentwise, and for the product the $(i, j)^{\text{th}}$ entry of AB is obtained by multiplying the i^{th} row of A by the j^{th} column of B in the usual way.

It can be shown that, with these operations, $\mathrm{Mat}(n, R)$ is a ring. As with direct sums, the verification that $\mathrm{Mat}(n, R)$ satisfies a given axiom is, in most cases, a straightforward calculation based on the fact that R satisfies the same axiom. We will only do axioms (i) and (v) (which is harder than the others).

Proof. (i) Let A, B, $C \in \mathrm{Mat}(n, R)$. Then by the definition of addition in $\mathrm{Mat}(n, R)$ and the associative law for addition in R we have

$$\begin{aligned}((A + B) + C)_{ij} &= (A + B)_{ij} + C_{ij} \\ &= (A_{ij} + B_{ij}) + C_{ij} \\ &= A_{ij} + (B_{ij} + C_{ij}) \\ &= A_{ij} + (B + C)_{ij} \\ &= (A + (B + C))_{ij}\end{aligned}$$

and therefore $(A + B) + C = A + (B + C)$.

(ii) Let A, B, $C \in \mathrm{Mat}(n, R)$. Then

$$\begin{aligned}((AB)C)_{ij} &= \sum_{k=1}^{n} (AB)_{ik} C_{kj} \\ &= \sum_{k=1}^{n} \left(\sum_{h=1}^{n} A_{ih} B_{hk} \right) C_{kj}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{n} \left(\sum_{h=1}^{n} (A_{ih}B_{hk})\, C_{kj} \right) \\
&= \sum_{h=1}^{n} \left(\sum_{k=1}^{n} (A_{ih}B_{hk})\, C_{kj} \right) \\
&\qquad \text{(by interchanging the order of summation)} \\
&= \sum_{h=1}^{n} \left(\sum_{k=1}^{n} A_{ih}\, (B_{hk}C_{kj}) \right) \\
&\qquad \text{(by associativity of multiplication in } R\text{)} \\
&= \sum_{h=1}^{n} A_{ih} \left(\sum_{k=1}^{n} B_{hk}C_{kj} \right) \\
&= \sum_{h=1}^{n} A_{ih}(BC)_{hj} \\
&= (A(BC))_{ij}
\end{aligned}
$$

showing that $(AB)C = A(BC)$, as required. □

§2d Trivial properties of rings

Let R be any ring. By Axiom (ii) in Definition 2.2 we know that there is an element $0 \in R$ satisfying

($) $\qquad a + 0 = 0 + a = a \qquad$ for all $a \in R$.

Could there exist another element $z \in R$ with the same property—that is, satisfying

(£) $\qquad a + z = z + a = a \qquad$ for all $a \in R$?

The answer is no. If z satisfies (£) then $z = 0$. To see this observe that putting $a = z$ in ($) gives $z + 0 = z$, while putting $a = 0$ in (£) gives $z + 0 = 0$. Hence z must equal 0.

In the preceding paragraph we have proved that the zero element of any ring is unique. There are a number of other properties which are trivially true in the examples of rings familiar to us, and which are equally easily proved to hold in any ring; some of these are listed in the theorems in this section. Although they are trivial, it is necessary to prove that they are consequences of the ring axioms if we wish to claim that they are true in all rings.

2.9 THEOREM *In any ring R the zero element is unique, and each element has a unique negative.*

Proof. The first part has been proved above. For the second part, assume that $a \in R$ and that b, $c \in R$ are both negatives of a. Using Definition 2.2 we have

$$\begin{aligned} b &= b + 0 && \text{(by Axiom (ii))} \\ &= b + (a + c) && \text{(since } c \text{ is a negative of } a\text{)} \\ &= (b + a) + c && \text{(Axiom (i))} \\ &= 0 + c && \text{(since } b \text{ is a negative of } a\text{)} \\ &= c && \text{(Axiom (ii)).} \end{aligned}$$

Thus a cannot have two distinct negatives, which is what we had to prove. □

In view of the preceding theorem there is no ambiguity in denoting the negative of an element a by '$-a$', as usual. It is customary also to write '$x - y$' for '$x + (-y)$'.

2.10 THEOREM *Let R be any ring and $a, b, c \in R$.*

(i) *If $a + b = a + c$ then $b = c$.*
(ii) $-(a + b) = (-a) + (-b)$.
(iii) $-(-a) = a$.
(iv) $a0 = 0a = 0$.
(v) $a(-b) = -(ab) = (-a)b$.
(vi) $(-a)(-b) = ab$.
(vii) $a(b - c) = ab - ac$.

Proof. (i) Assume that $a + b = a + c$. Then we have

$$\begin{aligned} (-a) + (a + b) &= (-a) + (a + c) \\ ((-a) + a) + b &= ((-a) + a) + c && \text{(Axiom (i))} \\ 0 + b &= 0 + c && \text{(by definition of '}-a\text{')} \\ b &= c && \text{(Axiom (ii))} \end{aligned}$$

as required.

(ii) In view of the uniqueness of negatives it is sufficient to prove that $(-a) + (-b)$ is a negative of $a + b$; that is, it is sufficient to prove that

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

Furthermore, if we prove only the first of these equations, the other will follow as a consequence of Axiom (iv) in Definition 2.2. But use of the first four axioms readily gives

$$\begin{aligned}((-a)+(-b))+(a+b) &= (-a)+((-b)+(a+b))\\ &= (-a)+((-b)+(b+a))\\ &= (-a)+(((-b)+b)+a)\\ &= (-a)+(0+a)\\ &= (-a)+a\\ &= 0.\end{aligned}$$

(iii) By the definition, a negative of $-a$ is any element b which satisfies $(-a)+b=0=b+(-a)$. But these equations are satisfied if we put $b=a$; so it follows that a is a negative of $-a$. (In other words, the equations that say that x is a negative of y also say that y is a negative of x.) Since negatives are unique, we have proved that $-(-a)=a$.

(iv)
$$\begin{aligned}a0+0 &= a0 && \text{(Axiom (i))}\\ &= a(0+0) && \text{(Axiom (ii))}\\ &= a0+a0 && \text{(Axiom (vi)).}\end{aligned}$$

By part (i) above it follows that $a0=0$. The proof that $0a=0$ is similar.

The proofs of the other parts are left to the exercises. □

Our final result for this chapter gives a connection between two of the concepts we have introduced.

2.11 Theorem *Every field is an integral domain.*

Proof. Comparing the definitions of 'field' and 'integral domain' we see that this amounts to proving that there can be no zero divisors in a field. So, let F be a field and let $a\in F$ be a zero divisor. Then by definition $a\neq 0$ and there exists $b\in F$ with $b\neq 0$ and $ab=0$. But since F is a field all nonzero elements have multiplicative inverses; in particular there exists $c\in F$ such that $ca=1$. Thus

$$b=1b=(ca)b=c(ab)=c0=0$$

contradicting the fact that $b\neq 0$. So F can have no zero divisors. □

Exercises

1. Prove parts (v), (vi) and (vii) of Theorem 2.10.

2. Prove that Ring Axioms (iv), (v) and (vi) hold in the direct sum $R \dot{+} S$ of two rings R and S.

3. Prove that Ring Axioms (ii), (iii) and (iv) are satisfied in $\mathrm{Mat}(n, R)$.
(Hint: The zero element of $\mathrm{Mat}(n, R)$ is the matrix all of whose entries are zero, and the negative of a matrix A is the matrix whose entries are the negatives of the entries of A.)

4. Prove that Ring Axiom (vi) is satisfied in $\mathrm{Mat}(n, R)$.

5. Let A be any set and R any ring, and let $\mathcal{F}(A, R)$ be the set of all functions from A to R. Let addition and multiplication be defined on $\mathcal{F}(A, R)$ by the rules

$$(f + g)(a) = f(a) + g(a)$$
$$(fg)(a) = f(a)g(a)$$

for all f, $g \in \mathcal{F}(A, R)$ and $a \in A$. Prove that with these operations $\mathcal{F}(A, R)$ is a ring.

6. Suppose that e_1 and e_2 are both identity elements in the ring R. Prove that $e_1 = e_2$. (Hint: Consider $e_1 e_2$.)

7. Let R be a ring with an identity element 1. Prove that an element $a \in R$ can have at most one multiplicative inverse.

8. (*i*) Is the equation $a^2 - b^2 = (a - b)(a + b)$ valid in all rings?

(*ii*) Let R be a commutative ring and let x and y be elements of R having the property that $x^2 = 0$ and $y^2 = 0$. Prove that $(x + y)^3 = 0$.

(*iii*) Give an example of a (noncommutative) ring R having elements x and y such that $x^2 = 0$ and $y^2 = 0$ but $(x + y)^3 \neq 0$.
(Hint: The matrix $x = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ satisfies $x^2 = 0$.)

9. Suppose that R is a commutative ring and $a \in R$ is nonzero and not a zero divisor. Prove that if b, $c \in R$ satisfy $ab = ac$ then $b = c$.

10. By imitating the construction in #5, describe how to construct the direct sum $A \dot{+} B \dot{+} C$ of three rings A, B and C.

11. A ring R is said to be *Boolean* if $a^2 = a$ for all $a \in R$. Prove that if R is Boolean then $2a = 0$ for all $a \in R$. Prove also that all Boolean rings are commutative.

3
The integers

In this chapter we will investigate divisibility and factorization in the ring $\mathbb{Z}$. These properties will be used in the next chapter in the construction of the rings $\mathbb{Z}_n$, our first example of the important concept of a *quotient* of a ring. Our treatment of $\mathbb{Z}$ and its quotients will be mimicked later in our discussion of polynomial rings and their quotients, which are of central importance in field extension theory.

§3a Two basic properties of the integers

We begin with a property of the set $\mathbb{Z}^+$ of positive integers which is equivalent to the principle of mathematical induction. It should be regarded as an axiom.

3.1 Least Integer Principle *Every nonempty set of positive integers has a least element.*

Comment ▷▷▷

3.1.1 To convince yourself that the principle of mathematical induction follows from 3.1, it is worthwhile to try rewriting a simple proof by induction as a proof by the Least Integer Principle. The idea is this. Suppose we wish to prove that some statement $P(n)$ is true for all positive integers n. We check first that $P(1)$ is true. Now let S be the set of all positive integers for which $P(n)$ is not true; we aim to prove that S is empty. If it is not then by 3.1 it has a least element, k, and $k > 1$ since $1 \notin S$. Thus $k-1$ is positive and not in S; so $P(k-1)$ is true. If we can prove that $P(n)$ is true whenever $P(n-1)$ is true it will follow that $P(k)$ is true, contradicting $k \in S$, and thereby showing that $S = \emptyset$. Thus although we have appealed to 3.1 rather than induction, our task is the same: prove that $P(1)$ is true and prove that $P(n)$ is true whenever $P(n-1)$ is true.

As an illustration of this, let us rewrite the well known inductive proof that $\sum_{i=1}^{n} i^3 = (1/4)n^2(n+1)^2$ as a proof which appeals to 3.1 instead of induction. (We have no use for this formula in this course, but we will make use of the Least Integer Principle to prove other properties of $\mathbb{Z}$.)

Let $S = \{\, n \in \mathbb{Z}^+ \mid \sum_{i=1}^{n} i^3 \neq (1/4)n^2(n+1)^2 \,\}$ (the set of all integers for which the given formula is false. Our aim is to prove that S is empty. Using indirect proof (see §0d), assume that $S \neq \emptyset$. By 3.1 it follows that S must have a least element. Let k be this least element. Since $\sum_{i=1}^{1} i^3 = 1^3 = (1/4)1^2(1+1)^2$ we see that $1 \notin S$, and so $k \neq 1$. Since $k \in \mathbb{Z}^+$ and $k \neq 1$, it follows that $k-1$ is also a positive integer. Since $k-1 < k$ and k is the least positive integer in S, it follows that $k-1 \notin S$. Thus $\sum_{i=1}^{k-1} i^3 = (1/4)(k-1)^2k^2$. Adding k^3 to both sides of this gives

$$\begin{aligned}\sum_{i=1}^{k} i^3 &= (1/4)(k-1)^2k^2 + k^3 \\ &= (1/4)((k^2-2k+1)k^2 + 4k^3) \\ &= (1/4)(k^4 - 2k^3 + k^2 + 4k^3) \\ &= (1/4)(k^4 + 2k^3 + k^2) \\ &= (1/4)k^2(k+1)^2,\end{aligned}$$

and by the definition of S we deduce that $k \notin S$. But this is a contradiction, since $k \in S$ by the definition of k. This contradiction completes the proof that S is empty, which is what we had to prove. ▷▷▷

The illustrative proof in 3.1.1 above is somewhat convoluted, and is more naturally expressed as a direct proof by induction, in the usual way. In other cases, however, the Least Integer Principle may be more natural and easier to use than induction. Indeed, our next proof is a case in point. We use 3.1 to prove the Division Theorem for integers:

3.2 Theorem *If $a \in \mathbb{Z}$ and $n \in \mathbb{Z}^+$ then there exist unique integers q and r with $a = qn + r$ and $0 \leq r < n$.*

Proof. Since r has to be $a - qn$, the theorem can be rephrased as follows:

(∗) If $a \in \mathbb{Z}$ and $n \in \mathbb{Z}^+$ then there exists a unique integer q such that $0 \leq a - qn < n$.

Given a and n, we first prove the existence of such a q.

Let $S = \{\, m \in \mathbb{Z}^+ \mid m = kn - a \text{ for some } k \in \mathbb{Z} \,\}$. If we put $k = a^2 + 1$ then

$$kn - a = (a^2 + 1)n - a \geq a^2 + 1 - a > (a - \frac{1}{2})^2 > 0$$

and so $kn - a \in S$. Hence $S \neq \emptyset$, and by 3.1 it follows that S has a least element.

Let $k_0 \in \mathbb{Z}$ be chosen so that $k_0 n - a$ is the least element of S. Then we have

$$k_0 n - a > 0. \tag{1}$$

Moreover, since $(k_0 - 1)n - a < k_0 n - a$ it follows that $(k_0 - 1)n - a \notin S$, and therefore

$$(k_0 - 1)n - a \leq 0. \tag{2}$$

Combining (1) and (2) yields

$$(k_0 - 1)n \leq a < k_0 n,$$

and, on subtracting $(k_0 - 1)n$ throughout,

$$0 \leq a - qn < n$$

where $q = k_0 - 1$. This establishes the existence part of $(*)$.

To prove the uniqueness assertion we must show that if q, q' are integers satisfying $0 \leq a - qn < n$ and $0 \leq a - q'n < n$ then $q = q'$.

Assume that q, q' are such integers. Then

$$(q - q')n = (a - q'n) - (a - qn) < n$$

since $a - q'n < n$ and $a - qn \geq 0$. Similarly,

$$(q' - q)n = (a - qn) - (a - q'n) < n.$$

So we obtain

$$-n < (q - q')n < n,$$

and, on dividing through by n,

$$-1 < q - q' < 1.$$

Since q and q' are integers it follows that $q = q'$, as required. □

Comments ▷▷▷

3.2.1 The integers q and r in 3.2 are called the *quotient* and *remainder* when a is divided by n.

3.2.2 If the remainder r in 3.2 is zero, so that $a = qn$, we write '$n|a$', which should be read as 'n is a factor of a', or 'n divides a' (short for 'n divides a exactly'), or 'n is a divisor of a', or 'a is a multiple of n'.

3.2.3 The integer 0 is divisible by all integers, since the equation $0 = 0n$ is valid for all n.

3.2.4 Observe that if a is an integer then the set of all divisors of $-a$ is the same as the set of all divisors of a. ▷▷▷

§3b The greatest common divisor of two integers

3.3 THEOREM *If a and b are integers which are not both zero then there is a unique positive integer d such that*

(a) $d|a$ and $d|b$,

(b) if c is an integer such that $c|a$ and $c|b$ then $c|d$.

Furthermore, there exist integers m and n such that $ma + nb = d$.

Comment ▷▷▷

3.3.1 Part (a) says that d is a common divisor of a and b, while part (b) says that any other common divisor of a and b is a divisor of d. So d is the greatest common divisor of a and b:

$$d \;=\; \gcd(a, b).$$

(A common notation is just $d = (a, b)$. Some authors prefer the nomenclature 'highest common factor' to 'greatest common divisor'.) ▷▷▷

We give two proofs of the existence of a d with the properties described in Theorem 3.3. The first, which is based on the *Euclidean Algorithm*, also provides a method for calculating d.

If a and b be integers which are not both zero, define $D = D(a, b)$ to be the set of all common divisors of a and b:

$$D(a, b) \;=\; \{\, c \in \mathbb{Z} \;\big|\; c|a \text{ and } c|b \,\}.$$

Our aim is to prove that there exists $d \in D$ such that d is divisible by all elements of D, and the strategy is to replace a and b by integers with smaller absolute value without changing D.

Observe first that by 3.2.4 we may replace a and b by $|a|$ and $|b|$ without changing the set of common divisors. So we may assume that $a \geq 0$ and $b \geq 0$. The next observation (Lemma 3.4 below) is that the set of common divisors is unchanged if a is replaced by another integer which differs from a by a multiple of b. The Euclidean Algorithm uses this fact repeatedly to replace a and b by smaller numbers until one of them becomes zero.

3.4 Lemma *If a, b, $m \in \mathbb{Z}$ then $D(a, b) = D(b, a + mb)$.*

Proof. Suppose that $c \in D(a, b)$. Then $a = rc$ and $b = sc$ for some integers r and s. Hence $a + mb = (r + ms)c$, and so c is a divisor of $a + mb$. Since c is also a divisor of b it follows that $c \in D(a + mb, b)$. This shows that $D(a, b) \subseteq D(a + mb, b)$.

Suppose, on the other hand, that $e \in D(a + mb, b)$. Then $a + mb = te$ and $b = ue$ for some integers t and u, and this gives $a = (t - mu)e$. Hence e is a divisor of a as well as of b, and so $e \in D(a, b)$. So $D(a+mb, b) \subseteq D(a, b)$. □

To find the greatest common divisor of two nonegative integers a and b, proceed as follows. Without loss of generality we may assume that $a \geq b$. If $b \neq 0$ let b' be the remainder on division of a by b, and define $a' = b$. Since $b' = a - qb$ for some q, Lemma 3.4 gives

$$D = D(a, b) = D(b, a - qb) = D(a', b').$$

So D is unchanged, and a' and b' are smaller than a and b were. If $b' \neq 0$ we can replace a by a' and b by b' and repeat the process. Eventually, after repeating the process often enough, the smaller of the two numbers will be zero. Let d be the other number. Then the set D, which is unchanged throughout, is equal to

$$D(d, 0) = \{\, c \in \mathbb{Z} \mid c|d \text{ and } c|0 \,\} = \{\, c \in \mathbb{Z} \mid c|d \,\}.$$

So the set of all common divisors of the two numbers a and b that we started with equals the set of all divisors of d. In particular, d itself is the greatest common divisor of a and b, divisible by every other common divisor.

We can conveniently paraphrase the above description of the algorithm

using terminology based on various computer programming languages:

EUCLIDEAN ALGORITHM.

```
while b ≠ 0 do
        [a, b] := [b, a − b ∗ (a div b)]
enddo
return a
```

The ":=" sign in this means "becomes". The pair of numbers a and b are replaced by b and $a - qb$ respectively, where q the quotient on dividing a by b, and this is repeated until $b = 0$. The program then returns the other number, which is the gcd of the initial two.

Using mathematical terminology of a more conventional kind, let $a_1 = a$ and $a_2 = b$, and (recursively) define a_{i+1} to be the remainder on dividing a_{i-1} by a_i, as long as $a_i \neq 0$. This results in the following formulae:

$$\begin{aligned} a_1 &= q_3 a_2 + a_3 & 0 < a_3 < a_2 \\ a_2 &= q_4 a_3 + a_4 & 0 < a_4 < a_3 \\ &\vdots & \vdots \\ a_{k-2} &= q_k a_{k-1} + a_k & 0 < a_k < a_{k-1} \\ a_{k-1} &= q_{k+1} a_k & a_{k+1} = 0. \end{aligned}$$

Since (by 3.2) the remainder on dividing a_{i-1} by a_i is a nonnegative integer less than a_i, the sequence $a_2, a_3, \ldots$ is a strictly decreasing sequence of nonnegative integers. Any such sequence must eventually reach 0. So the process must terminate. Now if $D = D(a_1, a_2)$ then repeated application of 3.4 gives

$$D = D(a_2, a_3) = D(a_3, a_4) = \cdots = D(a_k, a_{k+1}).$$

But $D(a_k, a_{k+1}) = D(a_k, 0)$ is just the set of all divisors of a_k, and so putting $d = a_k$ we conclude immediately that $d \in D$ and $c|d$ for all $c \in D$. That is, (a) and (b) of 3.3 are satisfied. Summarizing:

3.4.1 *The gcd of two integers is the last nonzero remainder obtained in the Euclidean Algorithm.*

The equations above also show how to express $\gcd(a_1, a_2)$ in the form $ma_1 + na_2$ with $m, n \in \mathbb{Z}$. The idea is that a_1 and a_2 are trivially expressed in this form, since we obtain

$$a_1 = m_1 a_1 + n_1 a_2$$
$$a_2 = m_2 a_1 + n_2 a_2$$

if we define $m_1 = 1, \quad n_1 = 0$ and $m_2 = 0, \quad n_2 = 1$, and the equations from the algorithm permit one to successively express a_3, a_4, ..., and eventually $a_k = \gcd(a_1, a_2)$, in the required form. Specifically, if we have expressions for a_{i-1} and a_i,

$$a_{i-1} = m_{i-1} a_1 + n_{i-1} a_2$$
$$a_i = m_i a_1 + n_i a_2,$$

then the equation

$$a_{i-1} = q_{i+1} a_i + a_{i+1}$$

(from the algorithm) gives

$$\begin{aligned} a_{i+1} &= a_{i-1} - q_{i+1} a_i \\ &= (m_{i-1} a_1 + n_{i-1} a_2) - q_{i+1}(m_i a_1 + n_i a_2) \\ &= m_{i+1} a_1 + n_{i+1} a_2, \end{aligned}$$

where

$$m_{i+1} = m_{i-1} - q_{i+1} m_i$$
$$n_{i+1} = n_{i-1} - q_{i+1} n_i.$$

Thus we have an expression for a_{i+1} and can repeat the process. This eventually yields an expression of the required kind for $a_k = \gcd(a_1, a_2)$.

Thus the full algorithm, for finding the gcd of a_0 and b_0 and expressing it in the form $ma_0 + nb_0$, is as follows:

Begin with $a = a_0$ and $b = b_0$.
    $m := 1$, $n := 0$, $m' := 0$, $n' := 1$
    while $b \neq 0$ do
        $q := a \operatorname{div} b$
        $[a, b] := [b, a - qb]$
        $[m, m'] := [m', m - qm']$
        $[n, n'] := [n', n - qn']$
    enddo
    return a, m, n

At the end of the process, a is the gcd of a_0 and b_0, and m and n satisfy $a = ma_0 + nb_0$.

We have now proved most of Theorem 3.3, but we have still to prove the uniqueness of the gcd. For this, assume that d_1 and d_2 are both gcd's of a and b; that is, (a) and (b) of 3.3 are satisfied with d replaced by d_1 and also with d replaced by d_2. Now since $d_1|a$ and $d_1|b$ (by (a) for d_1) it follows from (b) for d_2 that $d_1|d_2$. Hence $d_1 \leq d_2$. But exactly the same reasoning, using (a) for d_2 and (b) for d_1, gives $d_2 \leq d_1$. So $d_1 = d_2$.

—***Example***—

#1 Compute $\gcd(84, 133)$ and find integers m and n such that

$$\gcd(84, 133) = 84m + 133n.$$

⋙▷ The steps in the Euclidean Algorithm are:

(1) $$133 = 1 \times 84 + 49$$

(2) $$84 = 1 \times 49 + 35$$

(3) $$49 = 1 \times 35 + 14$$

(4) $$35 = 2 \times 14 + 7$$

(5) $$14 = 2 \times 7$$

The last nonzero remainder is 7, and so $7 = \gcd(84, 133)$.

Now (1) gives

(6) $$49 = 133 - 1 \times 84$$

and substituting this into (2) we obtain an expression for 35 as a combination of 84 and 133:

(7) $$\begin{aligned} 35 &= 84 - 1 \times 49 \\ &= 84 - (133 - 1 \times 84) \\ &= 2 \times 84 - 133. \end{aligned}$$

Substituting (6) and (7) into (3) gives an expression for 14:

(8) $$\begin{aligned} 14 &= 49 - 1 \times 35 \\ &= (133 - 1 \times 84) - (2 \times 84 - 133) \\ &= 2 \times 133 - 3 \times 84. \end{aligned}$$

Substitute (7) and (8) into (4):

(9)
$$\begin{aligned} 7 &= 35 - 2 \times 14 \\ &= (2 \times 84 - 133) - 2(2 \times 133 - 3 \times 84) \\ &= 8 \times 84 - 5 \times 133. \end{aligned}$$

So $m = 8, \quad n = -5$ is a solution. ⊲⋘

Now for the second proof of the existence of the gcd:

Proof. Let a and b be integers which are not both zero, and let

$$S = \{\, ma + nb \mid m, n \in \mathbb{Z} \,\} \cap \mathbb{Z}^+.$$

Since $a^2 + b^2 > 0$ we see that $a^2 + b^2 \in S$, and so, by 3.1, S has a least element. Let d be this least element. Then certainly there exist m, $n \in \mathbb{Z}$ with $d = ma + nb$.

Let d' be the remainder on dividing a by d. Then $0 \leq d' < d$, and, for some $q \in \mathbb{Z}$,

$$\begin{aligned} d' &= a - qd \\ &= a - q(ma + nb) \\ &= (1 - qm)a + (-qn)b. \end{aligned}$$

If $d' \neq 0$ this shows that $d' \in S$, contradicting the fact that d is the least element of S. Hence $d' = 0$; that is, $d|a$. The same reasoning with a replaced by b shows that $d|b$. □

§3c Factorization into primes

3.5 Definition A nonnegative integer is said to be *prime* if it is strictly greater than 1 and has no positive factors other than itself and 1.

3.6 Proposition *If $a|bc$ and $\gcd(a, b) = 1$ then $a|c$.*

Proof. Since $\gcd(a, b) = 1$ there exist (by Theorem 3.1) integers m and n such that $ma + nb = 1$. Multiplying through by c gives $c(ma) + c(nb) = c$, which can be rewritten as $(cm)a + (bc)n = c$ since multiplication in $\mathbb{Z}$ is commutative. Now both terms on the left hand side are multiples of a (since we are given that $a|bc$). Hence $a|c$. □

3.7 THEOREM *Each integer $n > 1$ can be expressed as a product of primes. (That is, $n = p_1 p_2 \dots p_r$ for some $r \geq 1$ and primes $p_1, p_2, \dots, p_r \in \mathbb{Z}^+$.)*

Proof. Assume that the above statement is false. By the Least Integer Principle there exists a least integer $n > 1$ which is not expressible as a product of primes. Then n itself is not prime (otherwise we could take $r = 1$ and $p_1 = n$); so $n = n_1 n_2$ with $1 < n_1 < n$ and $1 < n_2 < n$. By the minimality of n it follows that n_1 and n_2 are both expressible as products of primes. Hence so is n, contradiction. □

3.8 THEOREM *The product in 3.7 is unique up to the order of the factors. In other words, if $p_1, p_2, \dots, p_r$ and $q_1, q_2, \dots, q_s$ are positive prime integers such that $p_1 p_2 \dots p_r = q_1 q_2 \dots q_s$ then $r = s$, and, for some permutation σ of $\{1, 2, \dots, r\}$, we have $p_i = q_{\sigma(i)}$ for $i = 1, 2, \dots, r$.*

Proof. We have $p_1 | p_1 p_2 \dots p_r = q_1 (q_2 q_3 \dots q_s)$. Observe that since the only positive divisors of a prime are itself and 1, two distinct primes can have no positive common divisors other than 1. So if $p_1 \neq q_1$ then $\gcd(p_1, q_1) = 1$, and by 3.6 it follows that $p_1 | q_2 q_3 \dots q_s$. If $p_1 \neq q_2$ the same argument gives $p_1 | q_3 \dots q_s$. Thus if $p_1 \neq q_j$ for each j we get successively $p_1 | q_2 q_3 \dots q_s$, $p_1 | q_3 q_4 \dots q_s$, $p_1 | q_4 \dots q_s$, $\dots$, and eventually $p_1 | q_s$. But this is contrary to the assumption that $p_1 \neq q_j$ for each j. So $p_1 = q_{j_1}$ for some j_1. Now cancelling gives $p_2 p_3 \dots p_r = q_1 \dots q_{j_1 - 1} q_{j_1 + 1} \dots q_s$, and repeating the argument gives $p_2 = q_{j_2}$ for some $j_2 \neq j_1$. We can continue in this way cancelling factors until one side or the other is reduced to 1. But since 1 cannot equal a product of primes greater than 1 it follows that when all the factors have been cancelled from one side all the factors have been cancelled from the other side too. So we must have $r = s$ and $p_i = q_{j_i}$ for $i = 1, 2, \dots r$, where $j_1, j_2, \dots, j_r$ are all distinct—that is, $p_i = q_{\sigma(i)}$ where $\sigma(i) = j_i$ is a permutation of $\{1, 2, \dots, r\}$. □

—Example—

#2 Prove that $\sqrt{3}$ is irrational.

$\gg\!\!\rightarrow$ We prove first that if a and b are integers such that $a^2 - 3b^2 = 0$ then $a = b = 0$.

Suppose to the contrary that there exists a nontrivial integral solution to $a^2 - 3b^2 = 0$. Replacing a and b by their absolute values we may assume

that a and b are both in $\mathbb{Z}^+$. By Theorem 3.7 both a and b can be expressed as products of primes; say

$$a = p_1 p_2 \dots p_n \quad \text{and} \quad b = q_1 q_2 \dots q_m.$$

Now $a^2 = 3b^2$ gives

$$(**) \qquad p_1^2 p_2^2 \dots p_n^2 = 3q_1^2 q_2^2 \dots q_m^2.$$

But by Theorem 3.8 prime factorizations are unique up to ordering of the factors; hence the number of times 3 occurs on the left hand side of $(**)$ equals the number of times it occurs on the right hand side. But 3 occurs an even number of times on the left hand side (twice the number of i such that $p_i = 3$) and an odd number of times on the right hand side (one plus twice the number of j such that $q_j = 3$). This contradiction shows that no such a and b exist.

We can now see that $\sqrt{3}$ is irrational, for if $\sqrt{3} = a/b$ with $a,\ b \in \mathbb{Z}$ then $b \neq 0$ and $a^2 - 3b^2 = 0$. But the Lemma shows that this is impossible.

Of course, similar proofs apply for $\sqrt{2}$, $\sqrt{5}$, $\sqrt{6}$, $\sqrt[3]{2}$, and so on.

⊲⊣≪

Exercises

1. In each case compute the gcd of the given integers a and b and find integers m and n such that $\gcd(a, b) = ma + nb$:

(i) $a = 420,\ b = 2079$ (ii) $a = 1188,\ b = 4200.$

2. Prove that $\sqrt[3]{2}$ is irrational.

3. Show that if $a|b$ and $b|c$ then $a|c$.

4. Show that if $a|r$ and $b|s$ then $\gcd(a, b)|\gcd(r, s)$.

5. Let a, b and c be integers, and let $d = \gcd(a, \gcd(b, c))$. Show that $d = \gcd(\gcd(a, b), c)$. Show also that d is the largest positive integer which is a divisor of all of a, b and c, and that there exist integers l, m and n with $d = la + mb + nc$.

6. Let a, b be positive integers and M the set of positive integers which are multiples of both a and b. The least element of M is called the *least common multiple* of a and b and is denoted by $\text{lcm}(a, b)$.

(i) Show that if $a|c$ and $b|c$ then $ab|cd$, where $d = \gcd(a, b)$.

(ii) Show that $\text{lcm}(a, b) = ab/\gcd(a, b)$.

7. Let a, b and e be integers and let $d = \gcd(a, b)$. Prove that there exist integers m and n such that $ma + nb = e$ if and only if e is a multiple of d.

8. (i) Find an integral solution (m, n) to the equation

$$4641m + 2093n = 364.$$

(ii) Prove that there is no integral solution to the equation

$$91m + 63n = 6.$$

9. Let a and b be integers, and let $m = m_0$, $n = n_0$ be an integral solution to the equation

($) $$ma + nb = d$$

where $d = \gcd(a, b)$. Prove that for any $k \in \mathbb{Z}$

$$m = m_0 + k(b/d)$$
$$n = n_0 - k(a/d)$$

is another solution to ($). Prove also that every integral solution to ($) has this form.

10. Let m be a positive integer. Show that there exist unique integers $a_0, a_1, \ldots, a_r$ such that $a_r \neq 0$,

$$m = a_0 + 8a_1 + 8^2a_2 + \cdots + 8^r a_r$$

and $0 \leq a_i < 8$ for $i = 0, 1, \ldots, r$.

(Hint: Use the Division Theorem repeatedly.)

4

Quotients of the ring of integers

Our primary objective in this chapter is the construction of the rings $\mathbb{Z}_n$. We start with a preliminary section which will also be needed later in our discussion of quotient rings in general.

§4a Equivalence relations

Sometimes when considering elements of some set S it is convenient to lump together various elements of S if they are equivalent to one another, by some criterion of equivalence. For example, if S is the set of all cars in Sydney we may wish to regard two elements of S as equivalent if they are of the same make. (That is, all Holdens are equivalent, all Fords are equivalent, and so on.) Obviously the set of all equivalence classes will be much smaller than the set S itself. One can easily invent various other equivalence relations, but to justify the term 'equivalence' the following properties should hold:

(i) Every element should be equivalent to itself.

(ii) If a is equivalent to b then b should be equivalent to a.

(iii) If a is equivalent to b and b is equivalent to c then a should be equivalent to c.

4.1 Definition Let $\sim$ be a relation on a set S. That is, for every pair a, b of elements of S either $a \sim b$ (a is related to b) or $a \not\sim b$ (a is not related to b). Then $\sim$ is called an *equivalence relation* if the following hold for all a, b, $c \in S$:

(i) $a \sim a$. (*reflexive law*)

(ii) If $a \sim b$ then $b \sim a$. (*symmetric law*)

(iii) If $a \sim b$ and $b \sim c$ then $a \sim c$. (*transitive law*)

If $\sim$ is an equivalence relation on the set S and $a \in S$ define

$$\overline{a} = \{\, b \in S \mid b \sim a \,\}.$$

That is, $\overline{a}$ is the subset of S consisting of all elements equivalent to a. The subset $\overline{a}$ is called the *equivalence class* of a.

4.2 THEOREM *If $\sim$ is an equivalence relation on S and $a, b \in S$ then*
(i) $\overline{a} = \overline{b}$ if and only if $a \sim b$,
(ii) if $a \not\sim b$ then $\overline{a} \cap \overline{b} = \emptyset$.

Proof. (i) Suppose that $a \sim b$. By the Symmetric Law we also have that $b \sim a$. Now for $x \in S$ the Transitive Law gives us the following facts:

$$(*) \qquad \begin{array}{l} \text{If } x \sim a \text{ then } x \sim a \text{ and } a \sim b\text{; so } x \sim b. \\ \text{If } x \sim b \text{ then } x \sim b \text{ and } b \sim a\text{; so } x \sim a. \end{array}$$

By $(*)$ we see that $x \sim a$ if and only if $x \sim b$. Hence

$$\overline{a} = \{\, x \in S \mid x \sim a \,\} = \{\, x \in S \mid x \sim b \,\} = \overline{b}$$

Conversely, suppose that $\overline{a} = \overline{b}$. Since $a \sim a$ we have

$$a \in \{\, x \in S \mid x \sim a \,\} = \overline{a}.$$

Hence

$$a \in \overline{b} = \{\, x \in S \mid x \sim b \,\}.$$

So $a \sim b$.

(ii) Suppose that $a \not\sim b$ and $\overline{a} \cap \overline{b} \neq \emptyset$. Let $c \in \overline{a} \cap \overline{b}$. Then we have $c \in \overline{a}$ and $c \in \overline{b}$; so $c \sim a$ and $c \sim b$. By the Symmetric Law we deduce that $a \sim c$ and $c \sim b$, so that the Transitive Law gives $a \sim b$, contradicting our initial assumption. □

Comment ▷▷▷

4.2.1 Theorem 4.2 shows us that the equivalence classes form a partition of the set S—that is, every element of S lies in exactly one equivalence class. ▷▷▷

We now define

$$\overline{S} = \{\, \overline{a} \mid a \in S \,\}.$$

That is, $\overline{S}$ is the set of all equivalence classes of elements of S.

By 4.2, $\overline{a}$ and $\overline{b}$ are the same if $a \sim b$. So when we deal with equivalence classes we are, so to speak, amalgamating equivalent elements. Intuitively, we pretend that we cannot tell the difference between equivalent elements. Then all equivalent elements combine to be one single element (which, strictly speaking, is an equivalence class) of a set $\overline{S}$ which is smaller than our original set S.

4.3 DEFINITION The set $\bar{S}$ defined above is called the *quotient* of S by the equivalence relation $\sim$.

§4b Congruence relations on the integers

Throughout this section let n be a fixed positive integer.

4.4 DEFINITION Let $\equiv$ be the relation defined on $\mathbb{Z}$ by the rule

$$a \equiv b \text{ if and only if } a - b \text{ is a multiple of } n.$$

The relation $\equiv$ is called *congruence modulo* n.

Comment ▷▷▷

4.4.1 We usually write '$a \equiv b \pmod{n}$' rather than just '$a \equiv b$', unless there is no possible confusion. ▷▷▷

For example, $11 \equiv 3 \pmod{8}$, and $-6 \equiv 40 \pmod{23}$, and so on.

4.5 THEOREM *(i) Congruence modulo n is an equivalence relation on $\mathbb{Z}$.*
(ii) For every integer m there is exactly one integer r in the range $0 \leq r < n$ such that m is congruent to r modulo n.

Proof. (i) Let $a, b, c \in \mathbb{Z}$ be such that $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$. Then, by the definition, $n \mid (b - a)$ and $n \mid (c - b)$. That is,

$$b - a = nr \text{ and } c - b = ns$$

for some $r, s \in \mathbb{Z}$. Now

$$\begin{aligned} c - a &= (c - b) + (b - a) \\ &= ns + nr \\ &= n(s + r). \end{aligned}$$

So $n \mid (c - a)$, and we have proved that congruence modulo n is a transitive relation.

The proofs that congruence is reflexive and symmetric are left to the exercises.

(ii) By definition, $m \equiv r \pmod{n}$ if and only if $n \mid (m - r)$. That is, $m \equiv r \pmod{n}$ if and only if $m - r = qn$ for some $q \in \mathbb{Z}$. But by Theorem 3.2 the equation $m = qn + r$ has exactly one integral solution with $0 \leq r < n$. □

Comments ▷▷▷

4.5.1 By Theorem 4.5 the equivalence classes $\overline{0}$, $\overline{1}$, $\overline{2}$, ..., $\overline{n-1}$ are all the equivalence classes for the relation congruence modulo n, and these classes are all distinct from one another.

4.5.2 We will use the notation '$\mathbb{Z}_n$' rather than '$\overline{\mathbb{Z}}$' for the set of all equivalence classes. The equivalence classes for the congruence relation are usually called 'congruence classes'. ▷▷▷

§4c The ring of integers modulo n

Let n be a fixed positive integer. In the last section we defined the equivalence relation 'congruence modulo n' on $\mathbb{Z}$ and defined $\mathbb{Z}_n$ to be the set of all congruence classes. By Theorem 4.5

$$\mathbb{Z}_n = \{\overline{0}, \overline{1}, \overline{2}, \ldots, \overline{n-1}\}$$

(where by definition $\overline{r} = \{\, m \in \mathbb{Z} \mid m \equiv r \pmod{n} \,\}$). Intuitively, we have amalgamated into one object all integers which leave the same remainder on division by n. For example, if $n = 5$ we have

$$\mathbb{Z}_5 = \{\overline{0}, \overline{1}, \overline{2}, \overline{3}, \overline{4}\}$$

where

$$\begin{aligned}
\overline{0} &= \{\ldots, -5, 0, 5, 10, \ldots\} \\
\overline{1} &= \{\ldots, -4, 1, 6, 11, \ldots\} \\
\overline{2} &= \{\ldots, -3, 2, 7, 12, \ldots\} \\
\overline{3} &= \{\ldots, -2, 3, 8, 13, \ldots\} \\
\overline{4} &= \{\ldots, -1, 4, 9, 14, \ldots\}.
\end{aligned}$$

Intuitively, the integers ..., −6, −1, 4, 9, 14, 19, ... become a single object, as do ..., −2, 3, 8, ... and so on. That is, these numbers are all regarded as equal when working modulo 5. What this means strictly is that

$$\cdots = \overline{-6} = \overline{-1} = \overline{4} = \overline{9} = \cdots$$

and so on. In other words we have many different names for the same congruence class.

Observe that in the above example the sum of any element in set $\overline{3}$ and any element in the set $\overline{4}$ gives an element in the set $\overline{7} = \overline{2}$. Thus, for instance,

$-2 \in \overline{3}$, $19 \in \overline{4}$, and $-2 + 19 = 17 \in \overline{2}$. So it seems reasonable to define the sum of the sets $\overline{3}$ and $\overline{4}$ to be equal to the set $\overline{2}$. That is,

$$\overline{3} + \overline{4} \;=\; \overline{2}.$$

Similarly the product of any element in the set $\overline{2}$ and any element in the set $\overline{3}$ gives an element in the set $\overline{1}$. For example

$$2 \times 3 = 6 \in \overline{1}, \quad 7 \times (-2) = -14 \in \overline{1}, \quad (-8) \times 13 = -104 \in \overline{1}.$$

So we define

$$\overline{2} \times \overline{3} \;=\; \overline{1}.$$

This suggests the following general rule for addition and multiplication of congruence classes. To add (or multiply) two classes pick one element from each class and add (or multiply) the elements. The congruence class in which the answer lies is then defined to be the sum (or product) of the two given classes. We must show, however, that you get the same answer whichever elements you choose. Fortunately that is not hard to prove.

4.6 Lemma *Let n be a positive integer. If a, a', b, b' are integers such that $a \equiv a'$ (mod n) and $b \equiv b'$ (mod n) then*

$$a + b \equiv a' + b' \pmod{n} \qquad \textit{and} \qquad ab \equiv a'b' \pmod{n}.$$

Proof. Assume that $a \equiv a'$ and $b \equiv b'$. Then $n|a - a'$ and $n|b - b'$. That is,

$$a' = a + rn \text{ and } b' = b + sn \text{ for some } r, s \in \mathbb{Z}.$$

This gives

$$a' + b' = (a + rn) + (b + sn) = (a + b) + (r + s)n \equiv a + b \pmod{n}$$

and similarly

$$a'b' = ab + (rb + as + rsn)n \equiv ab \pmod{n}.$$

□

Comment ▷▷▷

4.6.1 In view of 4.2, an alternative formulation is this:

$$\text{If } \overline{a} = \overline{a'} \text{ and } \overline{b} = \overline{b'} \text{ then } \overline{a + b} = \overline{a' + b'} \text{ and } \overline{ab} = \overline{a'b'}.$$

▷▷▷

4.7 Theorem *Addition and multiplication can be defined on $\mathbb{Z}_n$ in such a way that $\overline{a} + \overline{b} = \overline{a + b}$ and $\overline{a}\,\overline{b} = \overline{ab}$ for all a, $b \in \mathbb{Z}$.*

Proof. Given α and β in $\mathbb{Z}_n$ there are uniquely determined integers r and s such that $0 \le r < n$, $0 \le s < n$ and $\overline{r} = \alpha$, $\overline{s} = \beta$. Define $\alpha + \beta = \overline{r + s}$

and $\alpha\beta = \overline{rs}$. We have now defined addition and multiplication on $\mathbb{Z}_n$; it remains to check that the formulae in the theorem statement are satisfied.

Let $a, b \in \mathbb{Z}$, and let $r, s \in \mathbb{Z}$ be such that $0 \leq r < n$, $0 \leq s < n$ and $\overline{r} = \overline{a}$, $\overline{s} = \overline{b}$. By the definitions just given, $\overline{a} + \overline{b} = \overline{r+s}$ and $\overline{a}\,\overline{b} = \overline{rs}$, and by 4.6.1 the required result follows. □

It is now trivial to verify that these operations of addition and multiplication on $\mathbb{Z}_n$ satisfy the ring axioms.

4.8 Theorem *The set $\mathbb{Z}_n$ forms a ring with respect to the operations of addition and multiplication as defined in 4.7.*

Proof. It is necessary to check all of the axioms (i)–(vi) of Definition 2.2. In each case one simply appeals to the same property of $\mathbb{Z}$ and the formulae in 4.7. We will do only the first three axioms, leaving the others as exercises.

(i) Let $\alpha, \beta, \gamma \in \mathbb{Z}_n$. Then there exist $a, b, c \in \mathbb{Z}$ with $\overline{a} = \alpha$, $\overline{b} = \beta$, $\overline{c} = \gamma$, and we have

$$\begin{aligned}
(\alpha+\beta)+\gamma &= (\overline{a}+\overline{b})+\overline{c} \\
&= \overline{a+b}+\overline{c} && \text{(by the definition of addition in } \mathbb{Z}_n) \\
&= \overline{(a+b)+c} && \text{(similarly)} \\
&= \overline{a+(b+c)} && \text{(since addition in } \mathbb{Z} \text{ is associative)} \\
&= \overline{a}+\overline{b+c} && \text{(by the definition of addition in } \mathbb{Z}_n) \\
&= \overline{a}+(\overline{b}+\overline{c}) && \text{(similarly)} \\
&= \alpha+(\beta+\gamma).
\end{aligned}$$

(ii) We will prove that $\overline{0}$ is a zero element in $\mathbb{Z}_n$. Let α be any element of $\mathbb{Z}_n$, and choose any a in $\mathbb{Z}$ with $\overline{a} = \alpha$. Then

$$\alpha + \overline{0} = \overline{a} + \overline{0} = \overline{a+0} = \overline{a} = \alpha.$$

Similarly, $\overline{0} + \alpha = \alpha$.

(iii) Let $\alpha \in \mathbb{Z}_n$ and let $a \in \mathbb{Z}$ with $\overline{a} = \alpha$. Then

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

and similarly $\overline{-a} + \alpha = \overline{0}$. Since $\overline{0}$ is the zero element of $\mathbb{Z}_n$ this shows that $\overline{-a}$ is a negative of α. □

§4d Properties of the ring of integers modulo n

Since multiplication in $\mathbb{Z}$ is commutative we have in $\mathbb{Z}_n$ that

$$\overline{a_1}\,\overline{a_2} = \overline{a_1 a_2} = \overline{a_2 a_1} = \overline{a_2}\,\overline{a_1}$$

for all a_1, $a_2 \in \mathbb{Z}$. Since every element of $\mathbb{Z}_n$ is of the form $\overline{a}$ for some $a \in \mathbb{Z}$ this shows that multiplication in $\mathbb{Z}_n$ is commutative. We also have

$$\overline{1}\,\overline{a} = \overline{1a} = \overline{a1} = \overline{a}\,\overline{1}$$

for all $a \in \mathbb{Z}$, and it follows that $\overline{1}$ is a multiplicative identity for $\mathbb{Z}_n$. Thus we have proved

4.9 THEOREM *The ring $\mathbb{Z}_n$ is commutative and has an identity element.*

The ring $\mathbb{Z}$ itself, as well as being commutative and having an identity element, has the property that there are no zero divisors. Hence $\mathbb{Z}$ is an integral domain. However the ring $\mathbb{Z}_n$ usually has got zero divisors. For instance in $\mathbb{Z}_6$ we have $\overline{3} \neq \overline{0}$ and $\overline{2} \neq \overline{0}$ but $\overline{3} \times \overline{2} = \overline{6} = \overline{0}$. Hence we see that $\mathbb{Z}_n$ is not generally an integral domain. A little thought shows that it is not possible to find zero divisors in $\mathbb{Z}_n$ in this way if n is prime. In fact:

4.10 THEOREM *The ring $\mathbb{Z}_n$ is an integral domain if and only if n is prime.*

Proof. If n is not prime then there exist r, $s \in \mathbb{Z}$ with $1 < r < n$, $1 < s < n$ and $rs = n$. This gives

$$\overline{r}\,\overline{s} = \overline{rs} = \overline{n} = \overline{0}$$

although $\overline{r} \neq 0$ and $\overline{s} \neq 0$. Since $\overline{0}$ is the zero element of $\mathbb{Z}_n$ it follows that $\mathbb{Z}_n$ has zero divisors and is therefore not an integral domain.

On the other hand, suppose that n is prime and suppose that $\overline{r}\,\overline{s} = 0$. Then $\overline{rs} = 0$; that is, $rs \equiv 0 \pmod{n}$. So n is a divisor of rs. But if r and s are expressed as products of primes and these two expressions are multiplied together, the result is an expression for rs as a product of primes. By Theorem 3.8 we deduce that the prime divisors of rs are precisely the prime divisors of r together with the prime divisors of s. Since n is a prime it follows that n must be either a prime divisor of r or a prime divisor of s. Hence either $r \equiv 0 \pmod{n}$ or $s \equiv 0 \pmod{n}$; that is, either $\overline{r} = \overline{0}$ or $\overline{s} = \overline{0}$. Thus we have proved that it is impossible for the product of two nonzero elements of $\mathbb{Z}_n$ to be zero. So $\mathbb{Z}_n$ has no zero divisors. Since also $\mathbb{Z}_n$ is commutative and has an identity element (Theorem 4.9), $\mathbb{Z}_n$ is an integral domain. □

So if p is prime $\mathbb{Z}_p$ is an integral domain. However even more is true—in fact $\mathbb{Z}_p$ is a field in this case. To prove this one must show that each element of $\mathbb{Z}_p$ has a multiplicative inverse. (Observe, for instance, that this is true in $\mathbb{Z}_5$. Since $\overline{2} \times \overline{3} = \overline{6} = \overline{1}$ it can be seen that $\overline{2}$ and $\overline{3}$ are multiplicative inverses of each other, while similar calculations show that $\overline{1}$ and $\overline{4}$ are each their own inverse.)

We prove a slightly more general statement, namely:

4.11 THEOREM *Suppose that D is an integral domain which has only a finite number of elements. Then D is a field.*

Proof. Recall that an integral domain is a commutative ring with identity with no zero divisors, while a field is a commutative ring with identity for which all nonzero elements have multiplicative inverses. So it suffices to prove that if $a \in D$ and $a \neq 0$ then there exists $b \in D$ with $ab = 1$.

Let $a \in D$ with $a \neq 0$. Define a mapping $\lambda\colon D \to D$ by

$$\lambda(b) = ab \qquad \text{for all } b \in D.$$

We will show that λ is injective. That is, we show that $\lambda(b)$ and $\lambda(c)$ can only be equal if $b = c$.

Suppose that $\lambda(b) = \lambda(c)$. Then $ab = ac$, and so $ab - ac = 0$. By the distributive law we obtain $a(b - c) = 0$. But D has no zero divisors (since it is an integral domain), and since $a \neq 0$ it follows that $b - c = 0$; that is, $b = c$. Hence λ is injective, as claimed above.

Now suppose that $b_1, b_2, \ldots, b_k$ are all the distinct elements of D. Then $\lambda(b_1), \lambda(b_2), \ldots, \lambda(b_k)$ are all distinct (since $\lambda(b_i) = \lambda(b_j)$ would imply that $b_i = b_j$). But D has only k elements altogether; so each element of D is equal to some $\lambda(b_i)$. In particular the identity element 1 equals $\lambda(b_i)$ for some i. That is, $1 = \lambda(b) = ab$ for some $b \in D$. So a has a multiplicative inverse, namely b. This argument applies to any nonzero element a of D, and so it follows that D is a field. □

Note also that even if n is not prime an element $\overline{a}$ of $\mathbb{Z}_n$ will have a multiplicative inverse if $\gcd(a, n) = 1$. This is so since by Theorem 3.3 there will exist integers r and s with $ra + sn = 1$, giving

$$ra = 1 - sn \equiv 1 \pmod{n}$$

so that $\overline{r}\,\overline{a} = \overline{1}$, and $\overline{r}$ is the inverse of $\overline{a}$.

—***Example***—

#1 Calculate the inverse of the element $\overline{24}$ in $\mathbb{Z}_{1001}$

⋙▷ Apply the Euclidean Algorithm to find integers r and s which satisfy $24r + 1001s = 1$:

$$\begin{aligned} 1001 &= 41 \times 24 + 17 \\ 24 &= 1 \times 17 + 7 \\ 17 &= 2 \times 7 + 3 \\ 7 &= 2 \times 3 + 1 \end{aligned}$$

Substituting back gives

$$\begin{aligned} 1 = 7 - 2 \times 3 &= 7 - 2(17 - 2 \times 7) = 5(24 - 17) - 2 \times 17 \\ &= 5 \times 24 - 7(1001 - 41 \times 24) = 292 \times 24 - 7 \times 1001. \end{aligned}$$

Thus $\overline{292}$ is the required inverse. ◁⋘

Exercises

1. Prove that congruence modulo n is reflexive and symmetric.

2. Prove that $\mathbb{Z}_n$ satisfies Axioms (iv), (v) and (vi) of Definition 2.2.

3. Use mathematical induction to show that $6^n \equiv 6 \pmod{10}$ for every positive integer n.

4. Let m be an odd positive integer. Show that

(i) $m^2 \equiv m \pmod{2m}$

(ii) $m^2 \equiv 1 \pmod{4}$.

5. Find all solutions of the congruence $54x \equiv 13 \pmod{37}$.

6. Find the remainder when 19^{15} is divided by 36.

7. Find all the zero divisors in the rings $\mathbb{Z}_8$ and $\mathbb{Z}_2 \dot{+} \mathbb{Z}_2 \dot{+} \mathbb{Z}_2$.

8. Show that an integer is divisible by four if and only if the sum of the units digit and twice the tens digit is divisible by four.

9. Suppose that m is a positive integer, and let C be one of the congruence classes modulo m. Prove that if $a \in C$ and $b \in C$ then $\gcd(a, m) = \gcd(b, m)$.

10. Determine if the elements $\overline{13}$ and $\overline{14}$ have inverses in the ring $\mathbb{Z}_{22}$, and find the inverses if they exist.

11. Prove that if $m|n$ and $a \equiv b \pmod{n}$ then $a \equiv b \pmod{m}$.

12. Let $X = \{-1, 0, 1\}$ and define the relation $\sim$ on $\mathbb{Z}^+$ by the rule

$$a \sim b \text{ if } a - b \in X.$$

Show that $\sim$ satisfies two of the properties of an equivalence relation, and give an example to indicate why the third property in not satisfied.

13. Define the relation $\sim$ on $\mathbb{R}$ by the rule

$$x \sim y \text{ if and only if } y - x \in \mathbb{Z}.$$

Prove that $\sim$ is an equivalence relation and describe the eqivalence classes.

14. For the given set and relation below, determine which define equivalence relations, giving proofs or counterexamples:

(*i*) S is the set of all people living in Australia; $a \sim b$ if a lives within 100 km of b.

(*ii*) S is the set of all integers; $a \sim b$ if $a \geq b$.

(*iii*) S is the set of all subsets of a finite set T; $a \sim b$ if a and b have the same number of elements.

5
Some Ring Theory

In this chapter we introduce some of the concepts which are needed to study abstract rings, and prove the first theorems of the subject.

§5a Subrings and subfields

5.1 DEFINITION (i) A subset S of a ring R is called a *subring* if S is itself a ring with respect to the operations of R.

(ii) A subset S of a field F is called a *subfield* if S is itself a field with respect to the operations of F.

For example, the ring $\mathbb{Z}$ is a subring of the field $\mathbb{R}$, but not a subfield. The rational numbers, $\mathbb{Q}$, form a subfield of $\mathbb{R}$, which is in turn a subfield of $\mathbb{C}$. The even integers, $2\mathbb{Z}$, form a subring of $\mathbb{Z}$.

5.2 THEOREM *Let R be a ring and S a subset of R such that*

(i) S is nonempty,
(ii) S is closed under multiplication (that is, $ab \in S$ for all $a, b \in S$),
(iii) S is closed under addition ($a + b \in S$ for all $a, b \in S$),
(iv) S is closed under forming negatives ($-a \in S$ for all $a \in S$).

Then S is a subring of R.

Conversely, any subring of R has these four properties.

Proof. Assume first that S is a subring of R. We must prove that the four properties above are satisfied.

Since S is a ring it must have a zero element. So $S \neq \emptyset$, and the first of the properties holds. Note also that if z is the zero of S and 0 the zero of R then $z + z = z$ (by the defining property of the zero of S) and $z + 0 = z$

(by the defining property of the zero of R), so that by Theorem 2.10 (i) we must have $z = 0$.

Now let a and b be arbitrary elements of S. Since the operations of addition and multiplication in R define operations on S we must have that $a+b$ and ab are elements of S. So properties (ii) and (iii) hold. Furthermore, since a must have a negative in S there must exist $x \in S$ such that

$$a + x = z = x + a.$$

But since $z = 0$ these equations also say that x is a negative of a in R. By Theorem 2.9 it follows that $x = -a$, and we have proved that $-a \in S$, as required.

For the converse we must assume that S satisfies properties (i)–(iv) and prove that it satisfies Definition 2.2. Observe that properties (ii) and (iii) guarantee that the sum and product in R of two elements of S are actually elements of S; hence the operations of R do give rise to operations on S. It remains to prove that Axioms (i)–(vi) of Definition 2.2 are satisfied in S. In each case the proof uses the fact that since R is a ring the same axiom is satisfied in R. The hardest part is to prove that the zero element of R is actually in S; so let us do this first.

We are given that S is nonempty; hence there exists at least one element $s \in S$. By property (iv) it follows that $-s \in S$, and so by property (iii)

$$0 = s + (-s) \in S.$$

Let a, b, $c \in S$. By Axioms (i), (iv), (v) and (vi) in R we have

$$\begin{aligned}(a + b) + c &= a + (b + c)\\ a + b &= b + a\\ (ab)c &= a(bc)\\ a(b + c) &= ab + ac\\ (a + b)c &= ac + bc\end{aligned}$$

and so it follows that Axioms (i), (iv), (v) and (vi) are satisfied in S.

We proved above that $0 \in S$. Now if a is any element of S we have (by Axiom (ii) in R) that $a+0 = a = 0+a$, and therefore 0 is a zero element for S. Moreover by property (iv) we have that $-a \in S$; thus each element of S has a negative in S. So S satisfies Axioms (ii) and (iii). □

Comments ▷▷▷

5.2.1 In the above proof we have also shown that every subring contains the zero element of the ring.

5.2.2 The point of proving theorems is that the work which goes into proving them never has to be repeated. One has simply to check that the hypotheses of the theorem are satisfied to be able to assert that its conclusion is satisfied, without repeating the steps of the proof. In particular, if we have to prove that something is a ring we can usually contrive to use a theorem (such as the above) in whose proof the tedium of checking the axioms one by one has already been dealt with. ▷▷▷

—Examples—

#1 Prove that $S = \{\, a + b\sqrt{3} \mid a, b \in \mathbb{Z} \,\}$ is a subring of $\mathbb{R}$.

≫▷ By Theorem 5.2 it is sufficient to check that S is nonempty and satisfies the three closure properties.

It is obvious that S is nonempty—for example $0 = 0 + 0\sqrt{3} \in S$.

Let α, $\beta \in S$. We must show that $\alpha\beta$, $\alpha + \beta$ and $-\alpha$ are all in S. By definition of S we have $\alpha = a + b\sqrt{3}$ and $\beta = c + d\sqrt{3}$ for some a, b, c, $d \in \mathbb{Z}$. Thus

$$\begin{aligned} \alpha + \beta &= (a + c) + (b + d)\sqrt{3} \\ \alpha\beta &= (ac + 3bd) + (ad + bc)\sqrt{3} \\ -\alpha &= (-a) + (-b)\sqrt{3}. \end{aligned}$$

In each case the right hand side has the form (integer)+(integer)$\sqrt{3}$, and so $\alpha + \beta$, $\alpha\beta$ and $-\alpha$ are all in S, as required. ◁≪

#2 Prove that

$$S = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \;\middle|\; a, b, c \in \mathbb{Z} \right\}.$$

is a subring of $\mathrm{Mat}(2, \mathbb{Z})$.

≫▷ $S \neq \emptyset$ since $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in S$.

Let $\alpha = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$ and $\beta = \begin{pmatrix} d & e \\ 0 & f \end{pmatrix}$ be arbitrary elements of S. Then

$$\alpha + \beta = \begin{pmatrix} a+d & b+e \\ 0 & c+f \end{pmatrix} \in S,$$

$$\alpha\beta = \begin{pmatrix} ad & ae+bf \\ 0 & cf \end{pmatrix} \in S,$$

$$-\alpha = \begin{pmatrix} -a & -b \\ 0 & -c \end{pmatrix} \in S.$$

Hence the closure properties hold. ⋘◁

#3 Let $S = \left\{ \begin{pmatrix} 0 & a \\ b & c \end{pmatrix} \;\middle|\; a,\, b,\, c \in \mathbb{Z} \right\}$. Prove that S is not a subring of $\mathrm{Mat}(2, \mathbb{Z})$.

⋙▷ The multiplication operation on $\mathrm{Mat}(2, \mathbb{Z})$ does not yield an operation on S, since the product of two elements of S need not be in S. For example,

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in S \qquad \text{and} \qquad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in S,$$

but

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \notin S.$$

⋘◁

#4 Let $R = \mathbb{Z}_8$ and let $S \subseteq R$ be given by

$$S = \{\bar{0}, \bar{2}, \bar{4}, \bar{6}\}.$$

Prove that S is a subring of $\mathbb{Z}_8$.

⋙▷ Observe that $S = \{\, \overline{2r} \mid r \in \mathbb{Z} \,\}$. Obviously $S \neq \emptyset$. Since

$$\overline{2r} + \overline{2s} = \overline{2r + 2s} \in S,$$

and

$$\overline{2r}\,\overline{2s} = \overline{4rs} = \overline{2(2rs)} \in S,$$

and

$$-\left(\overline{2r}\right) = \overline{2(-r)} \in S,$$

the required closure properties hold. ⋘◁

#5 The subset $\{\bar{1}, \bar{3}, \bar{5}, \bar{7}\}$ of $\mathbb{Z}_8$ is not a subring. It is closed under multiplication but not addition.

5.3 THEOREM *Let F be a field and S a subset of F such that*

(i) $0 \in S$ *and* $1 \in S$,

(ii) if $a \in S$ *and* $b \in S$ *then* $a + b \in S$ *and* $ab \in S$,

(iii) if $a \in S$ *then* $-a \in S$,

(iv) if $a \in S$ *and* $a \neq 0$ *then* $a^{-1} \in S$.

Then S is a subfield of F.

Conversely, any subfield of F satisfies these properties.

Proof. Since F is a field it is certainly a ring. Suppose that S is a subset of F satisfying the properties above. Then S satisfies properties (i)–(iv) of Theorem 5.2, and so by Theorem 5.2 it follows that S is a subring of F. Hence S is a ring.

We are given that the identity element of F is in the subset S; hence S has an identity element. The identity element is nonzero, since by Definition 2.8 applied to F we know that $1 \neq 0$. Furthermore, $ab = ba$ for all $a, b \in S$, since by Definition 2.8 the same is true for all $a, b \in F$. Finally, if a is a nonzero element of S then property (iv) above guarantee that a has an inverse in S. We have now checked that all the requirements of Definition 2.8 are satisfied; thus S is a field, and therefore a subfield of F.

The proof of the converse (that any subfield has the listed properties) is similar to the first part of the proof of Theorem 5.2, and is left to the exercises. □

—***Example***—

#6 $S = \{\, a + b\sqrt{3} \mid a, b \in \mathbb{Q} \,\}$ is a subfield of $\mathbb{R}$.

The hardest part of the proof is to show that S is closed under forming inverses of its nonzero elements—part (iv) of 5.3.

Suppose that $a, b \in \mathbb{Q}$ and that $a + b\sqrt{3} \neq 0$. Then $a^2 - 3b^2 \neq 0$ (by §3c#2), and we find that

$$(a + b\sqrt{3})^{-1} = \frac{a}{a^2 - 3b^2} + \frac{-b}{a^2 - 3b^2}\sqrt{3}.$$

Since the coefficients on the right hand side are rational numbers, the result follows.

---

§5b Homomorphisms

5.4 Definition Let R and S be rings. A mapping $\theta: R \to S$ is called a *ring homomorphism* if

$$\theta(a+b) = \theta(a) + \theta(b)$$

and

$$\theta(ab) = \theta(a)\theta(b)$$

for all a and b in R.

Definition 5.4 is sometimes expressed as follows:

> A ring homomorphism from R to S is a mapping which *preserves addition and multiplication.*

Some elementary properties of homomorphisms are listed in the next theorem.

5.5 Theorem *If $\theta: R \to S$ is a ring homomorphism then*

(i) $\theta(0_R) = 0_S$ *(where '0_R' means 'zero element of R' and '0_S' means 'zero element of S'),*

(ii) $\theta(-a) = -\theta(a)$ *for all $a \in R$,*

(iii) $\theta(a-b) = \theta(a) - \theta(b)$ *for all $a, b \in R$,*

(iv) $\theta(a_1 a_2 \dots a_n) = \theta(a_1)\theta(a_2)\dots\theta(a_n)$ *for all $a_1, a_2, \dots, a_n \in R$,*

(v) $\theta(a_1 + a_2 + \cdots + a_n) = \theta(a_1) + \theta(a_2) + \cdots + \theta(a_n)$ *for all $a_i \in R$.*

Proof. (i) Using the defining properties of 0_R and 0_S and the fact that θ preserves addition we have

$$\theta(0_R) + \theta(0_R) = \theta(0_R + 0_R) = \theta(0_R) = \theta(0_R) + 0_S$$

and by Theorem 2.10 (i) it follows that $\theta(0_R) = 0_S$.

(ii) Let $a \in R$. Since θ preserves addition

$$\theta(a) + \theta(-a) = \theta(a + (-a)) = \theta(0_R) = 0_S$$

by part (i). Applying the commutative law for addition we obtain that $\theta(-a) + \theta(a) = 0_S$ also. By Theorem 2.9 (uniqueness of negatives) it follows that $\theta(-a) = -\theta(a)$.

(iii) Let $a, b \in R$. We have

$$\theta(a - b) = \theta(a + (-b)) = \theta(a) + \theta(-b) = \theta(a) + (-\theta(b)) = \theta(a) - \theta(b).$$

(iv) If $n = 1$ this is immediate. Proceding by induction we assume that $n > 1$ and the statement holds with $n - 1$ in place of n. Let $a_1, a_2, \ldots, a_n \in R$. Using the fact that θ preserves multiplication and the induction hypothesis we obtain

$$\begin{aligned}\theta(a_1 a_2 \ldots a_n) &= \theta\big((a_1 a_2 \ldots a_{n-1})a_n\big) \\ &= \theta(a_1 a_2 \ldots a_{n-1})\theta(a_n) \\ &= \big(\theta(a_1)\theta(a_2)\ldots\theta(a_{n-1})\big)\theta(a_n) \\ &= \theta(a_1)\theta(a_2)\ldots\theta(a_n)\end{aligned}$$

as required.

(v) The proof of this is similar to the proof of (iv) and is omitted. □

Comment ▷▷▷

5.5.1 If R has an identity element 1 then it is not necessarily true that $\theta(1)$ is an identity element of S. If θ is surjective (onto), however, then $\theta(1)$ must be an identity. See the exercises at the end of the chapter.

5.5.2 If S is a subring of the ring R and $\theta: R \to T$ is a homomorphism then the *restriction* of θ to S—the mapping $\phi: S \to T$ given by $\phi(s) = \theta(s)$ for all $s \in S$—is clearly a homomorphism. ▷▷▷

—Examples—

#7 Define $\theta: \mathbb{Z} \to \mathbb{Z}_n$ by $\theta(a) = \overline{a}$ for all $a \in \mathbb{Z}$. (That is, θ takes any integer to its congruence class modulo n.) By definition of addition and multiplication of congruence classes

$$\begin{aligned}\theta(a + b) &= \overline{a + b} = \overline{a} + \overline{b} = \theta(a) + \theta(b) \\ \theta(ab) &= \overline{ab} = \overline{a}\,\overline{b} = \theta(a)\theta(b)\end{aligned}$$

So θ is a homomorphism. (This is an example of a homomorphism from a ring to a quotient ring of itself. Whenever a quotient ring can be formed such a homomorphism exists.)

#8 Let

$$\theta(a+b\boldsymbol{i}) = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

define a map $\theta\colon \mathbb{C} \to \mathrm{Mat}(2,\mathbb{R})$. Since we have†

$$\begin{aligned} \theta\big((a+b\boldsymbol{i})+(c+d\boldsymbol{i})\big) &= \theta\big((a+c)+(b+d)\boldsymbol{i}\big) \\ &= \begin{pmatrix} a+c & b+d \\ -(b+d) & a+c \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ -b & a \end{pmatrix} + \begin{pmatrix} c & d \\ -d & c \end{pmatrix} \\ &= \theta(a+b\boldsymbol{i}) + \theta(c+d\boldsymbol{i}), \end{aligned}$$

and

$$\begin{aligned} \theta\big((a+b\boldsymbol{i})(c+d\boldsymbol{i})\big) &= \theta\big((ac-bd)+(ad+bc)\boldsymbol{i}\big) \\ &= \begin{pmatrix} ac-bd & ad+bc \\ -(ad+bc) & ac-bd \end{pmatrix} \\ &= \begin{pmatrix} a & b \\ -b & a \end{pmatrix}\begin{pmatrix} c & d \\ -d & c \end{pmatrix} \\ &= \theta(a+b\boldsymbol{i})\theta(c+d\boldsymbol{i}) \end{aligned}$$

it follows that θ is a homomorphism.

If a homomorphism $\theta\colon R \to S$ is bijective (one-to-one and onto) then it sets up a one-to-one correspondence between elements of R and elements of S such that the sum of the elements of S corresponding to two given elements a, $b \in R$ is the element corresponding to $a+b$, and similarly their product is the element of S corresponding to ab. Thus the rings R and S are essentially the same as one another—although they have different elements, R and S have the same underlying additive and multiplicative structure.

---

† Here and subsequently the boldface letter $\boldsymbol{i}$ denotes a complex square root of -1

5.6 DEFINITION A ring homomorphism which is bijective is called an *isomorphism*. If there exists an isomorphism $\theta: R \to S$ then R and S are said to be *isomorphic*, and we write '$R \cong S$'.

—***Examples***—

#9 The map

$$\theta: a + b\boldsymbol{i} \longmapsto \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$$

provides an isomorphism from $\mathbb{C}$ to the subring S of $\mathrm{Mat}(2, \mathbb{R})$ consisting of all matrices of the form $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$.

To prove this one must prove first that S is a subring of $\mathrm{Mat}(2, \mathbb{R})$, then prove that θ preserves addition and multiplication and is bijective. We omit the proof that S is a subring since it is similar to the proofs in #1 and #2 above.

Let $\alpha, \beta \in \mathbb{C}$ with $\theta(\alpha) = \theta(\beta)$. We have $\alpha = a + b\boldsymbol{i}$, $\beta = c + d\boldsymbol{i}$ for some $a, b, c, d \in \mathbb{R}$, and since

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = \theta(a + b\boldsymbol{i}) = \theta(c + d\boldsymbol{i}) = \begin{pmatrix} c & d \\ -d & c \end{pmatrix}$$

we see that $a = c$ and $b = d$, whence $\alpha = \beta$. Thus θ is injective.

Let A be an arbitrary element of S. Then $A = \begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ for some $a, b \in \mathbb{R}$, and so $A = \theta(a + b\boldsymbol{i})$. Hence θ is surjective. This completes the proof, since we have already seen in #8 that θ preserves addition and multiplication.

Note that the fact that θ preserves addition and multiplication, together with the fact that $S = \operatorname{im}\theta$, can be used to prove that S is closed under addition and multiplication, as follows. Let $A, B \in S$. Then $A = \theta(\alpha)$, $B = \theta(\beta)$ for some $\alpha, \beta \in \mathbb{C}$, and so

$$A + B = \theta(\alpha) + \theta(\beta) = \theta(\alpha + \beta) \in \operatorname{im}\theta = S$$
$$AB = \theta(\alpha)\theta(\beta) = \theta(\alpha\beta) \in \operatorname{im}\theta = S$$

#10 If R is any ring the direct sum $R \dot{+} R$ is isomorphic to the subring of $\mathrm{Mat}(2, R)$ consisting of all diagonal matrices. The map

$$(a, b) \longmapsto \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

is an isomorphism.

Let ψ be the given map. By the definition of addition and multiplication in $R \dot{+} R$ we have

$$(a, b) + (c, d) = (a + c, b + d)$$
$$(a, b)(c, d) = (ac, bd)$$

so that

$$\psi((a, b) + (c, d)) = \psi(a + c, b + d) = \begin{pmatrix} a + c & 0 \\ 0 & b + d \end{pmatrix}$$
$$= \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} + \begin{pmatrix} c & 0 \\ 0 & d \end{pmatrix} = \psi(a, b) + \psi(c, d),$$

and $\psi((a, b)(c, d)) = \psi(ac, bd) = \psi(a, b)\psi(c, d)$ similarly.

We have shown that ψ preserves addition and multiplication. It is also necessary to show that ψ is bijective and that the set of all diagonal matrices is a subring of $\mathrm{Mat}(2, R)$. These proofs are straightforward and are omitted.

#11 Prove that if F is a field and R a ring is isomorphic to F then R is also a field.

$\gg\!\!\rightarrow$ We must show that R is commutative and has a nonzero identity element, and that all nonzero elements of R have inverses.

Let $\phi\colon F \to R$ be an isomorphism, and let $x, y \in R$. Since ϕ is surjective there exist $a, b \in F$ with $\phi(a) = x$ and $\phi(b) = y$. Multiplication in F is commutative (since F is a field); so $ab = ba$, and

$$xy = \phi(a)\phi(b) = \phi(ab) = \phi(ba) = \phi(b)\phi(a) = yx$$

showing that R is commutative.

By Exercise 2 at the end of the chapter, $\phi(1)$ is an identity for R. Since $\phi(0) = 0$ and $1 \neq 0$ the fact that ϕ is injective shows that $\phi(1)$ is nonzero.

Suppose that x is a nonzero element of R. Then $x = \phi(a)$ for some $a \in F$, and since

$$\phi(0) = 0 \neq x = \phi(a)$$

injectivity of ϕ gives $a \neq 0$. Since F is a field it follows that a has an inverse, and

$$x\phi(a^{-1}) = \phi(a^{-1})x = \phi(a^{-1})\phi(a) = \phi(a^{-1}a) = \phi(1).$$

Thus $\phi(a^{-1})$ is an inverse for x; so all nonzero elements of R have inverses. ◁⊢≪

---

§5c Ideals

5.7 Definition A subring I of a ring R is called an *ideal* of R if $ar \in I$ and $ra \in I$ for all $a \in I$ and $r \in R$.

Comment ▷▷▷

5.7.1 If I is an ideal in R then multiplying an element of I by any element of R and must give an element of I. Note that this is a more stringent requirement than closure under multiplication, which merely says that the product of two elements of I lies in I. An ideal must be closed under multiplication by arbitrary elements of the ring. ▷▷▷

—***Example***—

#12 Let $R = \mathbb{Z}$ and $I = 2\mathbb{Z}$. Then I is nonempty ($0 \in 2\mathbb{Z}$), closed under addition (the sum of two even integers is even), closed under multiplication (the product of two even integers is even), and closed under forming negatives (the negative of an even integer is even). So I is a subring of R. To observe that in fact it is an ideal it remains to show that I is closed under multiplication by arbitrary elements of R—that is, show that the product of an even integer and an arbitrary integer gives an even integer. But this is obvious.

Note that in the above example it was not really necessary to prove closure under multiplication separately since it follows from closure under multiplication by ring elements. This observation yields the following proposition:

5.8 Proposition *A subset I of a ring R is an ideal if and only if the following all hold:*

(i) *I is nonempty.*
(ii) *For all x and y, if $x \in I$ and $y \in I$ then $x + y \in I$.*
(iii) *For all x, if $x \in I$ then $-x \in I$.*
(iv) *For all x and y if $x \in I$ and $y \in R$ then $xy \in I$ and $yx \in I$.*

Proof. Suppose first that I is an ideal of R. Then I is a subring of R, and by Theorem 5.2 properties (i), (ii) and (iii) above all hold. Property (iv) holds too since it is explicitly assumed in the definition of an ideal.

Conversely, assume that I satisfies properties (i)–(iv). As remarked above it follows from property (iv) that I is closed under multiplication; thus all the requirements of Theorem 5.2 are satisfied, and it follows that I is a subring of R. This together with property (iv) shows that I is an ideal. □

—***Example***—

#13 Let

$$R = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \;\middle|\; a, b, c \in \mathbb{Z} \right\}$$

and

$$I = \left\{ \begin{pmatrix} 0 & d \\ 0 & 0 \end{pmatrix} \;\middle|\; d \in \mathbb{Z} \right\}.$$

Prove that R is a subring of $\mathrm{Mat}(2, \mathbb{Z})$ and I is an ideal of R.

≫▷ That R is a subring of $\mathrm{Mat}(2, \mathbb{Z})$ was proved in #2. Clearly I is nonempty—for instance, $\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \in I$. Let x and y be arbitrary elements of I and r an arbitrary element of R. Then for some integers a, b, c, d, e,

$$x = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix} \qquad y = \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \qquad r = \begin{pmatrix} c & d \\ 0 & e \end{pmatrix}$$

giving

$$x + y = \begin{pmatrix} 0 & a+b \\ 0 & 0 \end{pmatrix} \qquad rx = \begin{pmatrix} 0 & ac \\ 0 & 0 \end{pmatrix}$$

$$-x = \begin{pmatrix} 0 & -a \\ 0 & 0 \end{pmatrix} \qquad xr = \begin{pmatrix} 0 & ae \\ 0 & 0 \end{pmatrix}$$

and since these are all in I it follows that I is an ideal. ◁≪

§5d The characteristic of a ring

Let R be any ring. If $a \in R$ we define

$$\begin{aligned} 1a &= a \\ 2a &= a + a \\ 3a &= a + a + a \end{aligned}$$

and so on. In general, if m is any positive integer,

$$ma = \underbrace{a + a + \cdots + a}_{m \text{ terms}}.$$

If m is a negative integer we define

$$ma = -\big((-m)a\big)$$

observing that $(-m)a$ has already been defined since $-m$ is positive. And for the case $m = 0$ we define $0a = 0$. We have now defined ma whenever $m \in \mathbb{Z}$ and $a \in R$. This is a method of multiplying ring elements by integers, and is not to be confused with the multiplication operation within R itself. (But—fortunately—the value $0a$ is the same whether 0 is interpreted as an integer or the zero of the ring, and the same applies to $1a$ if R has an identity element.)

Similarly we define $a^m = aa \ldots a$ (m factors) if $m \in \mathbb{Z}^+$; if R has an identity element 1 we define $a^0 = 1$ for all $a \in R$; if m is negative and $a \in R$ has an inverse we define $a^m = (a^{-1})^{-m}$. The following should be clear:

5.9 Proposition *Let R be any ring, $a \in R$ and $m, n \in \mathbb{Z}$. Then*

(i) $m(na) = (mn)a$ *and* $(m + n)a = ma + na$,

(ii) $(a^m)^n = a^{mn}$ *and* $a^{m+n} = a^m a^n$.

(If either m or n is negative the second part is only applicable if a has an inverse; similarly, if either m or n is zero it is only applicable if R has an identity.)

5.10 Definition Let R be a ring. If there is a positive integer n such that $na = 0$ for all $a \in R$ then the least such n is called the *characteristic* of R. If there is no such n then R is said to have characteristic 0.

—Examples—

#14 The characteristic of $\mathbb{Z}_2$ is 2, the characteristic of $\mathbb{Z}_3$ is 3, and so on.

#15 Let S be the subring of $\mathbb{Z}_8$ given by $S = \{\bar{0}, \bar{2}, \bar{4}, \bar{6}\}$. Observe that

$$\bar{2} \neq \bar{0}, \quad \bar{2} + \bar{2} = \bar{4} \neq \bar{0}, \quad \bar{2} + \bar{2} + \bar{2} = \bar{6} \neq \bar{0}.$$

So the characteristic of S is not 1, 2, or 3. But

$$\begin{aligned} \bar{2} + \bar{2} + \bar{2} + \bar{2} &= \bar{8} = \bar{0} \\ \bar{4} + \bar{4} + \bar{4} + \bar{4} &= \overline{16} = \bar{0} \\ \bar{6} + \bar{6} + \bar{6} + \bar{6} &= \overline{24} = \bar{0} \\ \bar{0} + \bar{0} + \bar{0} + \bar{0} &= \bar{0}. \end{aligned}$$

So S has characteristic 4. (This shows that the characteristic of a subring can be less than the characteristic of the ring, since $\mathbb{Z}_8$ has characteristic 8.)

If a ring R has an identity element, 1, then $na = 0$ for all $a \in R$ if and only if $n1 = 0$. From this we can deduce the following proposition:

5.11 Proposition *If R is a ring with identity element 1 then the characteristic of R is the least positive integer n such that $n1 = 0$, or zero if there is no such n.*

Proof. Define

$$H = \{\, m \in \mathbb{Z}^+ \mid m1 = 0 \,\}$$

and

$$K = \{\, m \in \mathbb{Z}^+ \mid ma = 0 \text{ for all } a \in R \,\}.$$

We prove that $H = K$.

Let $m \in H$. Then $m1 = 0$, and so for all $a \in R$ we have

$$ma = \underbrace{a + a + \cdots + a}_{m \text{ terms}} = a(\underbrace{1 + 1 + \cdots + 1}_{m \text{ terms}}) = a(m1) = a0 = 0.$$

Hence $m \in K$.

Conversely, if $m \in K$ then $ma = 0$ for all $a \in R$, and, in particular, $m1 = 0$, whence $m \in H$. Thus $m \in K$ if and only if $m \in H$, and so $H = K$, as claimed.

By Definition 5.10 the characteristic of R is the least element of K, or zero if $K = \emptyset$. Since $H = K$ this shows that the characteristic of R is the least element of H, or zero if $H = \emptyset$, and this is precisely the assertion of Proposition 5.11. □

5.12 THEOREM *Let R be a ring with identity $1 \neq 0$ and let S be the subset of R given by $S = \{ n1 \mid n \in \mathbb{Z} \}$. Then S is a subring of R and*

(i) if R has characteristic 0 then $S \cong \mathbb{Z}$,

(ii) if R has characteristic $m \neq 0$ then $S \cong \mathbb{Z}_m$.

Proof. S is nonempty since, for instance, $1 \in S$. Let $a, b \in S$. Then $a = n1$, $b = m1$ for some $n, m \in \mathbb{Z}$, and, by Proposition 5.9,

$$a + b = n1 + m1 = (n + m)1 \in S$$
$$ab = (n1)(m1) = (nm)1 \in S$$
$$-a = -(n1) = (-n)1 \in S$$

so that by Theorem 5.2 it follows that S is a subring of R.

(i) Assume that R has characteristic 0.

Define a function $\psi: \mathbb{Z} \to S$ by

$$\psi(r) = r1 \qquad \text{for all } r \in \mathbb{Z}.$$

We have

$$\psi(r + s) = (r + s)1 = r1 + s1 = \psi(r) + \psi(s)$$
$$\psi(rs) = (rs)1 = (r1)(s1) = \psi(r)\psi(s)$$

so that ψ is a homomorphism. To show that $S \cong \mathbb{Z}$ it remains to show that ψ is bijective.

Let $a \in S$. Then, for some $n \in \mathbb{Z}$, $a = n1 = \psi(n)$. Thus ψ is surjective.

Let $r, s \in \mathbb{Z}$ be such that $\psi(r) = \psi(s)$. Then $r1 = s1$, and hence $(r - s)1 = 0$. But since R has characteristic 0 we know by Proposition 5.11 that there is no positive integer n such that $n1 = 0$. So $r - s \leq 0$, and by the same reasoning $s - r \leq 0$. Hence $r = s$, and we have shown that ψ is injective.

(ii) Suppose that the characteristic of R is $m > 0$. Then m is the least positive integer which annihilates 1 (in the sense that $m1 = 0$). We show first that if $r, s \in \mathbb{Z}$ then $r1 = s1$ if and only if $r \equiv s \pmod{m}$.

If $r \equiv s \pmod{m}$ then $r - s = tm$ for some $t \in \mathbb{Z}$, and

$$r1 = (s + tm)1 = s1 + (tm)1 = s1 + t(m1) = s1 + t0 = s1 + 0 = s1.$$

Suppose, conversely, that $r1 = s1$. Then $(r-s)1 = 0$, and, by what we have just proved, $k1 = 0$ for all $k \in \mathbb{Z}$ with $k \equiv (r-s) \pmod{m}$. By Theorem 4.5 (ii) we may choose such a k with $0 \leq k < m$. But since there is no positive integer less than m which annihilates 1 it follows that $k = 0$. Thus $(r-s) \equiv 0$, and therefore $r \equiv s$.

We now define a function $\theta\colon \mathbb{Z}_m \to S$ by $\theta(\overline{r}) = r1$ for all $r \in \mathbb{Z}$; this is unambiguous since if $\overline{r} = \overline{s}$ then $r \equiv s$ and hence $r1 = s1$, and it defines θ on all elements of $\mathbb{Z}_m$ since every element of $\mathbb{Z}_m$ is of the form $\overline{r}$.

By the definitions of addition and multiplication in $\mathbb{Z}_m$ we find that

$$\theta(\overline{r} + \overline{s}) = \theta(\overline{r+s}) = (r+s)1 = r1 + s1 = \theta(\overline{r}) + \theta(\overline{s})$$
$$\theta(\overline{r}\,\overline{s}) = \theta(\overline{rs}) = (rs)1 = (r1)(s1) = \theta(\overline{r})\theta(\overline{s})$$

so that θ is a homomorphism. Since every element of S is of the form $r1 = \theta(\overline{r})$ for some $r \in \mathbb{Z}$ it follows that θ is surjective. If $\theta(\overline{r}) = \theta(\overline{s})$ then $r1 = s1$ so that, as shown above, $r \equiv s$, and $\overline{r} = \overline{s}$. Thus θ is injective also, and $S \cong \mathbb{Z}_m$. □

—*Example*—

#16 In this example we will be dealing simultaneously with rings $\mathbb{Z}_2$, $\mathbb{Z}_3$ and $\mathbb{Z}_6$. To avoid confusion we will use the following notation (for $a \in \mathbb{Z}$):

'$\hat{a}$' denotes the congruence class of a modulo 2,
'$\tilde{a}$' denotes the congruence class of a modulo 3,
'$\bar{a}$' denotes the congruence class of a modulo 6.

(Obviously it would not be suitable to use '$\bar{a}$' for all of these three!)
We have that

$$\mathbb{Z}_2 = \{\hat{0}, \hat{1}\}$$
$$\mathbb{Z}_3 = \{\tilde{0}, \tilde{1}, \tilde{2}\}$$
$$\mathbb{Z}_6 = \{\bar{0}, \bar{1}, \bar{2}, \bar{3}, \bar{4}, \bar{5}\}.$$

Consider now the ring $\mathbb{Z}_2 \dot{+} \mathbb{Z}_3$, which consists of all pairs (a, b) with $a \in \mathbb{Z}_2$ and $b \in \mathbb{Z}_3$. It has six elements, since there are two choices for a and three choices for b. The element $(\hat{1}, \tilde{1})$ is easily seen to be an identity element (since $\hat{1}$ is an identity for $\mathbb{Z}_2$ and $\tilde{1}$ is an identity element for $\mathbb{Z}_3$) and to be

nonzero. For any $n \in \mathbb{Z}^+$ we have

$$\begin{aligned} n(\hat{1}, \tilde{1}) &= \underbrace{(\hat{1}, \tilde{1}) + \cdots + (\hat{1}, \tilde{1})}_{n \text{ terms}} \\ &= (n\hat{1}, n\tilde{1}) \\ &= (\hat{n}, \tilde{n}) \end{aligned}$$

and this equals the zero element $(\hat{0}, \tilde{0})$ if and only if $\hat{n} = \hat{0}$ and $\tilde{n} = \tilde{0}$. But $\hat{n} = \hat{0}$ if and only if n is even, and $\tilde{n} = \tilde{0}$ if and only if n is divisible by three. So $n(\hat{1}, \tilde{1}) = (\hat{0}, \tilde{0})$ if and only if n is divisible by six. Hence the characteristic of $\mathbb{Z}_2 \dot{+} \mathbb{Z}_3$ is six.

By Theorem 5.12 the set $\{\, n(\hat{1}, \tilde{1}) \mid n \in \mathbb{Z} \,\}$ is a subring of $\mathbb{Z}_2 \dot{+} \mathbb{Z}_3$ isomorphic to $\mathbb{Z}_6$. However, $\mathbb{Z}_2 \dot{+} \mathbb{Z}_3$ has only six elements altogether, and so this subring must equal the whole of $\mathbb{Z}_2 \dot{+} \mathbb{Z}_3$. So we have shown that

$$\mathbb{Z}_2 \dot{+} \mathbb{Z}_3 \cong \mathbb{Z}_6.$$

The proof of Theorem 5.12 can be used to write down an explicit isomorphism $\phi\colon \mathbb{Z}_6 \to \mathbb{Z}_2 \dot{+} \mathbb{Z}_3$. We obtain:

$$\begin{aligned} \bar{1} &\mapsto (\hat{1}, \tilde{1}) \\ \bar{2} &\mapsto (\hat{1}, \tilde{1}) + (\hat{1}, \tilde{1}) = (\hat{0}, \tilde{2}) \\ \bar{3} &\mapsto (\hat{1}, \tilde{1}) + (\hat{1}, \tilde{1}) + (\hat{1}, \tilde{1}) = (\hat{1}, \tilde{0}) \\ \bar{4} &\mapsto 4(\hat{1}, \tilde{1}) = (\hat{4}, \tilde{4}) = (\hat{0}, \tilde{1}) \\ \bar{5} &\mapsto 5(\hat{1}, \tilde{1}) = (\hat{5}, \tilde{5}) = (\hat{1}, \tilde{2}) \\ \bar{6} &\mapsto 6(\hat{1}, \tilde{1}) = (\hat{6}, \tilde{6}) = (\hat{0}, \tilde{0}). \end{aligned}$$

(Note that $\bar{6} = \bar{0}$.)

Exercises

1. Complete the proof of Theorem 5.3.

2. Let R and S be rings and let $\theta\colon R \to S$ be a surjective ring homomorphism. Let R have an identity element 1. Prove that $\theta(1)$ is an identity element in S.

3. In the ring $\mathbb{Z}$ find:

(i) the smallest subring containing 7,

(ii) the smallest subring containing 5 and 7.

4. Prove that **Con** (the constructible numbers) is a subfield of $\mathbb{R}$.
(Hint: Use Theorems 5.3 and 1.1.)

5. Suppose K is a subfield of $\mathbb{C}$ (the complex numbers) which contains $\mathbb{R}$ (the real numbers). Show that either $K = \mathbb{R}$ or $K = \mathbb{C}$.
(Hint: Assume that K contains $\mathbb{R}$ and some complex number not in $\mathbb{R}$, and show that K contains all complex numbers.)

6. Let $\mathbb{C}$ be the field of complex numbers and let $\theta\colon \mathbb{C} \to \mathbb{C}$ be defined by the formula

$$\theta(a + ib) = a - ib \qquad \text{for all } a, b \in \mathbb{R}.$$

Show that θ is a ring isomorphism from $\mathbb{C}$ to $\mathbb{C}$.

7. Prove that isomorphic rings have the same characteristic.

8. Let D be an integral domain with nonzero characteristic m. Prove that m is a prime. (Hint: If $m = rs$ then $(r1)(s1) = 0$.)

9. Let R and S be rings, and let $\theta\colon R \to S$ be an isomorphism. Prove that $a \in R$ is a zero divisor if and only if $\theta(a) \in S$ is a zero divisor. Deduce that isomorphic rings have the same number of zero divisors.

10. Prove that $\mathbb{Z}_{12}$ and $\mathbb{Z}_6 \dot{+} \mathbb{Z}_2$ are not isomorphic.

11. Define

$$\mathbb{Q}[\sqrt[3]{2}] = \{\, a + b\sqrt[3]{2} + c\left(\sqrt[3]{2}\right)^2 \mid a, b, c \in \mathbb{Q} \,\}.$$

Prove that $\mathbb{Q}[\sqrt[3]{2}]$ is a subring of $\mathbb{R}$, and prove also that it is an integral domain.

12. Let $\mathbb{Q}_2 = \left\{ \frac{m}{2^k} \mid m \in \mathbb{Z} \text{ and } k \in \mathbb{Z}^+ \right\}$, the set of all rational numbers with denominator a power of 2. Show that $\mathbb{Q}_2$ is a subring of $\mathbb{Q}$ but not a subfield.

13. Let

$$R = \left\{ \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & f & g \end{pmatrix} \;\middle|\; a, b, c, d, e, f, g \in \mathbb{Z} \right\}.$$

Is R a ring with respect to the operations of addition and multiplication defined as usual for matrices?

Let $\theta\colon R \to \mathrm{Mat}(2, \mathbb{Z})$ be the mapping defined by

$$\theta \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & f & g \end{pmatrix} = \begin{pmatrix} d & e \\ f & g \end{pmatrix}.$$

(*i*) Is θ injective?

(*ii*) Is θ surjective?

(*iii*) Is θ a homomorphism?

(*iv*) Define a relation $\sim$ on R by

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & f & g \end{pmatrix} \sim \begin{pmatrix} p & q & r \\ 0 & s & t \\ 0 & u & v \end{pmatrix} \text{ if and only if } \begin{pmatrix} d & e \\ f & g \end{pmatrix} = \begin{pmatrix} s & t \\ u & v \end{pmatrix}.$$

Is $\sim$ an equivalence relation?

(*v*) Define $\approx$ on R by

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & f & g \end{pmatrix} \approx \begin{pmatrix} p & q & r \\ 0 & s & t \\ 0 & u & v \end{pmatrix} \text{ if and only if } b = q \text{ and } c = r.$$

Is $\approx$ an equivalence relation?

(*vi*) Suppose that X, Y, Z, $W \in R$ are such that $X \sim Y$ and $Z \sim W$. Is it true that $X + Z \sim Y + W$? Is it true that $XZ \sim YW$?

(*vii*) The same as (*vi*) with $\sim$ replaced by $\approx$.

(*viii*) Let $\overline{R}$ be the set of all the equivalence classes into which R is partitioned by the relation $\sim$. Show that there is a one to one correspondence between $\overline{R}$ and $\mathrm{Mat}(2, \mathbb{Z})$.

6
Polynomials

Given a ring R it is possible to form new rings containing R as a subring by "adjoining" new elements to R. The simplest example of this is the ring of polynomials in X with coefficients from R, and it is necessary to study polynomial rings before dealing with the general problem of adjoining elements. The study of geometrical constructions leads naturally to this problem. For, suppose that $\alpha_1, \alpha_2, \ldots, \alpha_n$ are the points obtained in a geometrical construction, and let S_i be the smallest subring of $\mathbb{R}$ containing the coordinates of $\alpha_1, \alpha_2, \ldots, \alpha_i$. Then S_{i+1} can be thought of as obtained from S_i by adjoining the coordinates of α_{i+1}.

§6a Definitions

6.1 DEFINITION Let R be a commutative ring which has a nonzero identity element 1. A *polynomial* in the *indeterminate* X over R is an expression of the form

$$(*) \qquad a_0 + a_1X + \cdots + a_nX^n$$

where n is a positive integer and $a_0, a_1, a_2, \ldots, a_n \in R$. We call a_i the i^{th} *coefficient* of the polynomial.

Comments ▷▷▷

6.1.1 It is possible to define polynomials over arbitrary rings, but in this course we will only talk about polynomials over commutative rings with 1.

6.1.2 The coefficients of the polynomial are elements of the ring R, but X is not. In fact X, X^2, X^3, ... are nothing more than symbols written alongside the coefficients to enable us to see which is the 0^{th}, which the 1^{st}, which the 2^{nd}, and so on. Indeed, in some treatments of the topic the symbols

$X, X^2, \ldots$ are not used in the definition, and a polynomial is defined to be a sequence of ring elements $(a_0, a_1, \ldots, a_n, 0, 0, \ldots)$. So a polynomial is nothing more than its coefficients. Accordingly, to say that two polynomials p and q are equal is to say that for each i the i^{th} coefficient of p is equal to the i^{th} coefficient of q.

6.1.3 The terms in the expression $(*)$ above may be written in any order, and if $a_i = 0$ the corresponding term may be omitted. Similarly we may omit unnecessary coefficients equal to 1 (writing 'X' instead of '$1X$', and so on). Thus if

$$p = 2 + X^3 - 5X$$

then the 0^{th} coefficient of p is 2, the 1^{st} coefficient is -5, the 2^{nd} is 0, the 3^{rd} is 1. It is also convenient to say that the 4^{th}, 5^{th}, ... coefficients are zero (rather than saying that they do not exist). Thus a polynomial always has an infinite sequence of coefficients, one for each nonnegative integer, but all the coefficients beyond some point must be zero. ▷▷▷

6.2 Definition The polynomial all of whose coefficients are zero is called the *zero polynomial*.

6.3 Definition If p is a polynomial the largest i for which the i^{th} coefficient is nonzero is called the *degree* of p, and this coefficient is called the *leading* coefficient of p.

So if $p = a_0 + a_1X + \cdots + a_nX^n$ with $a_n \neq 0$ then a_n is the leading coefficient and $\deg(p) = n$. Note that we do not define the degree of the zero polynomial. In some treatments the zero polynomial is said to have degree $-\infty$. It would **not** be suitable to define the degree of the zero polynomial to be zero.

If p is a polynomial in the indeterminate X we often write '$p(X)$' instead of just 'p' to remind ourselves that p is a polynomial or to remind ourselves that the indeterminate is X.

Notation. The set of all polynomials over R in the indeterminate X is denoted by '$R[X]$'.

§6b Addition and multiplication of polynomials

6.4 DEFINITION Let R be a commutative ring with 1 and let $a, b \in R[X]$. Let a_i, b_i be the i^{th} coefficients of a, b (for $i = 0, 1, 2, \dots$). Define $a+b$ to be the polynomial with i^{th} coefficient $a_i + b_i$, and define ab to be the polynomial with i^{th} coefficient $a_i b_0 + a_{i-1} b_1 + \cdots + a_0 b_i$ (for $i = 0, 1, 2, \dots$).

By Definition 6.4, if

$$a = a_0 + a_1 X + \cdots + a_n X^n$$
$$b = b_0 + b_1 X + \cdots + b_m X^m$$

then

$$a + b = (a_0 + b_0) + (a_1 + b_1)X + (a_2 + b_2)X^2 + \cdots$$
$$ab = a_0 b_0 + (a_1 b_0 + a_0 b_1)X + (a_2 b_0 + a_1 b_1 + a_0 b_2)X^2 + \cdots.$$

Note that the formula for ab is obtained by multiplying out the expressions for a and b and collecting like terms in the usual way. (In particular, therefore, the i^{th} coefficient is zero for all i sufficiently large.)

6.5 THEOREM *If R is a commutative ring with 1 and X is an indeterminate then $R[X]$ is a commutative ring with 1.*

Proof. We must check the axioms in Definition 2.2 and the commutative law for multiplication, and find an identity element. It will be convenient to use the same notation as in the definition above: if p is a polynomial, then p_i is the i^{th} coefficient of p.

Let $a, b, c \in R[X]$. Then for all i

$$\begin{aligned}((a + b) + c)_i = (a + b)_i + c_i &= (a_i + b_i) + c_i \\ &= a_i + (b_i + c_i) = a_i + (b + c)_i = (a + (b + c))_i\end{aligned}$$

and so $(a + b) + c = a + (b + c)$. The proof that $a + b = b + a$ is similar.

The i^{th} coefficient of ab is the sum of all terms $a_r b_s$ with $r + s = i$; that is,

$$(ab)_i = \sum_{r+s=i} a_r b_s$$

in a convenient notation. We find that

$$\begin{aligned}((ab)c)_i &= \sum_{r+s=i} (ab)_r c_s \\ &= \sum_{r+s=i} \left(\sum_{u+v=r} a_u b_v \right) c_s \\ &= \sum_{u+v+s=i} (a_u b_v) c_s \\ &= \sum_{u+v+s=i} a_u (b_v c_s) \\ &= \sum_{u+t=i} a_u \left(\sum_{v+s=t} b_v c_s \right) \\ &= \sum_{u+t=i} a_u (bc)_t \\ &= (a(bc))_i\end{aligned}$$

and therefore $(ab)c = a(bc)$. Similarly

$$\begin{aligned}(a(b+c))_i &= \sum_{r+s=i} a_r (b+c)_s = \sum_{r+s=i} a_r (b_s + c_s) \\ &= \sum_{r+s=i} a_r b_s + a_r c_s = \sum_{r+s=i} a_r b_s + \sum_{r+s=i} a_r c_s = (ab)_i + (ac)_i\end{aligned}$$

Similar proofs also apply for the other distributive law and the commutativity of multiplication.

If we define z to be the polynomial for which $z_i = 0$ for all i then it is readily checked that $(a+z)_i = a_i = (z+a)_i$ for all $a \in R[X]$, so that z is a zero element for $R[X]$. It is also easily seen that $-a$ defined by $(-a)_i = -(a_i)$ for all i satisfies $a + (-a) = z = (-a) + a$; so each $a \in R[X]$ has a negative. Finally, define e to be the polynomial for which the 0^{th} coefficient is the identity element of R and all the other coefficients are equal to zero. That is, $e = 1 + 0X + 0X^2 + \cdots$. Then for all $a \in R[X]$,

$$(ae)_i = a_i e_0 + a_{i-1} e_1 + \cdots + a_0 e_i = a_i$$

since $e_0 = 1$ and $e_j = 0$ for $j \neq 0$. Thus $ae = a$, and since also $ea = a$ (similarly, or by commutativity of multiplication) it follows that e is an identity. □

Comment ▷▷▷

6.5.1 In accordance with the rules described in 6.1.3 above, the zero and identity polynomials are usually denoted by '0' and '1', but for the proof of Theorem 6.5 we needed to distinguish them from the zero and identity of R. See also §6c below. ▷▷▷

6.6 THEOREM *(i) Let R be an integral domain and a and b nonzero polynomials over R. Then the leading coefficient of ab is the product of the leading coefficients of a and b, and* $\deg(ab) = \deg(a) + \deg(b)$.

(ii) If R is an integral domain then so is $R[X]$.

Proof. (i) Let n be the degree of a and m the degree of b. If $r+s > n+m$ then necessarily either $r > n$ or $s > m$, and so either $a_r = 0$ or $b_s = 0$. Hence if $i > n + m$ then

$$(ab)_i = \sum_{r+s=i} a_r b_s = 0.$$

Similarly

$$(ab)_{n+m} = \sum_{r+s=n+m} a_r b_s = a_n b_m$$

since all other terms have either $r > n$ or $s > m$. Since $a_n \neq 0$ and $b_m \neq 0$ and R is an integral domain it follows that $(ab)_{n+m} \neq 0$. Thus $n + m$ is the largest value of i for which $(ab)_i \neq 0$, and so

$$\deg(ab) = n + m = \deg(a) + \deg(b).$$

Moreover, the leading coefficient of ab is $(ab)_{n+m}$, and, as we have seen, it is equal to $a_n b_m$, the product of the leading coefficients of a and b.

(ii) We have already proved that $R[X]$ is a commutative ring with 1; so it remains to prove that it has no zero divisors. But the first part of this proof shows that if the polynomials a and b each have a nonzero coefficient then so too does ab. That is, if $a \neq 0$ and $b \neq 0$ then $ab \neq 0$, as required. □

§6c Constant polynomials

Let R be a commutative ring with 1. For each $a \in R$ there is a polynomial for which the 0^{th} coefficient is a and all the other coefficients are zero. Using the notation described in 6.1.3 this polynomial would be denoted by 'a';

our notation does not distinguish between elements of R and these so-called *constant* polynomials. However, for the purposes of the next theorem we need a notation which does distinguish; so, temporarily, we will denote the constant polynomial a by '$c(a)$'. (That is, $c(a) = a + 0X + 0X^2 + \cdots$.)

6.7 THEOREM *The set $S = \{\, c(a) \mid a \in R \,\}$ of constant polynomials is a subring of $R[X]$ isomorphic to R, and the function $c\colon R \to S$ defined by $a \mapsto c(a)$ is an isomorphism.*

Proof. As in the proof of 6.5, denote the i^{th} coefficient of p by p_i. Then for all $a \in R$ we have $c(a)_0 = a$ and $c(a)_i = 0$ for all $i > 0$.

Let $a, b \in R$ with $c(a) = c(b)$. Then $a = c(a)_0 = c(b)_0 = b$. Thus c is injective. Since every constant polynomial is of the form $c(a)$ for some $a \in R$, c is surjective also. Furthermore, c preserves addition and multiplication, since

$$c(a+b)_0 = a + b = c(a)_0 + c(b)_0 = (c(a) + c(b))_0$$
$$c(ab)_0 = ab = c(a)_0 c(b)_0 = \sum_{r+s=0} c(a)_r c(b)_s = (c(a)c(b))_0$$

and for $i > 0$

$$c(a+b)_i = 0 = 0 + 0 = c(a)_i + c(b)_i = (c(a) + c(b))_i$$
$$c(ab)_i = 0 = \sum_{r+s=i} c(a)_r c(b)_s = (c(a)c(b))_i$$

(since in each term of the sum either $c(a)_r = 0$ or $c(b)_s = 0$).

It remains to check that S is indeed a subring of $R[X]$. Now clearly S is nonempty, since it contains the zero polynomial. If x and y are arbitrary elements of S then $x = c(a)$, $y = c(b)$ for some $a, b \in R$, and

$$x + y = c(a) + c(b) = c(a+b) \in S$$
$$xy = c(a)c(b) = c(ab) \in S$$
$$-x = -c(a) = c(-a) \in S$$

(the last line following from the fact that the negative of a polynomial is obtained by taking the negatives of all the coefficients). By Theorem 5.2, S is a subring. □

Comment ▷▷▷

6.7.1 Since the constant polynomials form a ring isomorphic to R no harm will come if we identify constant polynomials with the corresponding elements of R. In other words, we regard the ring S of 6.7 as being equal to R, so that R is a subring of $R[X]$. (This fits nicely with our notation which does not distinguish between constant polynomials and elements of R.) Now $R[X]$ can be thought of as a ring obtained by adjoining to R a new element X which satisfies no more equations than it is forced to satisfy to make $R[X]$ a commutative ring. ▷▷▷

§6d Polynomial functions

Any polynomial $p(X) = a_0 + a_1X + \cdots + a_nX^n$ in $R[X]$ determines a function $R \to R$ by the rule

$$c \longmapsto a_0 + a_1c + \cdots + a_nc^n$$

for all $c \in R$.

Polynomial functions are no doubt very familiar to the reader, but for us it is important to distinguish between polynomials, sometimes called *polynomial forms*, and polynomial functions. Note, for instance, that two distinct polynomials can give the same function. For example, if $p(X) = X^2$ and $q(X) = X$ in $\mathbb{Z}_2[X]$ then

$$p(\overline{0}) = (\,\overline{0}\,)^2 = \overline{0} = q(\overline{0}) \qquad \text{and} \qquad p(\overline{1}) = (\,\overline{1}\,)^2 = \overline{1} = q(\overline{1}),$$

so that $p(c) = q(c)$ for all $c \in \mathbb{Z}_2$. So the functions $c \mapsto p(c)$ and $c \mapsto q(c)$ are equal. However the polynomials p and q themselves are not equal since they have different coefficients.

§6e Evaluation homomorphisms

If c is any element of R there is a function

$$e_c : R[X] \longrightarrow R$$

defined by

$$p(X) \longmapsto p(c)$$

for all $p \in R[X]$. In other words, $e_c\big(p(X)\big) = p(c)$ for all polynomials p.

6.8 THEOREM *For each $c \in R$ the map $e_c : R[X] \to R$ is a homomorphism.*

Proof. Let $c \in R$ and let p, $q \in R[X]$. Then, in the notation we have been using for the i^{th} coefficient of a polynomial,

$$\begin{aligned} e_c(pq) &= (pq)_0 + (pq)_1 c + (pq)_2 c^2 + \cdots \\ &= p_0 q_0 + (p_0 q_1 + p_1 q_0)c + (p_0 q_2 + p_1 q_1 + p_2 q_0)c^2 + \cdots \\ &= (p_0 + p_1 c + p_2 c^2 + \cdots)(q_0 + q_1 c + q_2 c^2 + \cdots) \\ &= e_c(p) e_c(q) \end{aligned}$$

and similarly

$$\begin{aligned} e_c(p+q) &= (p+q)_0 + (p+q)_1 c + (p+q)_2 c^2 + \cdots \\ &= (p_0 + q_0) + (p_1 + q_1)c + (p_2 + q_2)c^2 + \cdots \\ &= (p_0 + p_1 c + p_2 c^2 + \cdots)(q_0 + q_1 c + q_2 c^2 + \cdots) \\ &= e_c(p) + e_c(q). \end{aligned}$$

□

Comments ▷▷▷

6.8.1 The function e_c is called an *evaluation homomorphism* since it maps $p(X) \in R[X]$ to $p(X)$ evaluated at c (that is, to $p(c)$).

6.8.2 To say that e_c preserves addition is to say that the result of adding two polynomials and then putting $X = c$ is the same as first putting $X = c$ in each and then adding. A similar statement applies for multiplication. The reason it works is because we have defined addition and multiplication of polynomials to make it work—when adding or multiplying polynomials the indeterminate X is treated as though it is an element of R.

6.8.3 If R is a subring of a ring S then $R[X]$ is a subring of $S[X]$. For instance, the set of all polynomials with rational coefficients is a subring of the set of all polynomials with real coefficients. Hence if c is any element of S the homomorphism $e_c : S[X] \to S$ may be restricted to $R[X]$ to yield a homomorphism from $R[X]$ to S. Thus, for instance, the map $\phi : \mathbb{Q}[X] \to \mathbb{R}$ given by $\phi(p(X)) = p(\sqrt[3]{2})$ is a homomorphism. ▷▷▷

§6f The division algorithm for polynomials over a field

Our chief application of polynomials in this course will be to study field extensions. Roughly speaking, if F is a field we wish to be able to make a larger field by adjoining extra elements to F, in much the way that the complex numbers are obtained from the real numbers by adjoining a square root of -1. We have already seen how $F[X]$ can be regarded as a ring obtained by adjoining the element X to F. However, $F[X]$ is not a field, and to obtain fields extending F we will have to deal with quotient rings of $F[X]$—rings obtained from $F[X]$ in the same way as $\mathbb{Z}_n$ is obtained from $\mathbb{Z}$. Once we have developed the theory of field extensions we will be able to prove things about the field **Con** of constructible numbers, which is an extension field of $\mathbb{Q}$ (rational numbers).

We start by investigating properties of divisibility and factorization for polynomials—properties analogous to those properties of $\mathbb{Z}$ which were used in our construction of $\mathbb{Z}_n$.

6.9 THEOREM *Let F be a field and $f(X)$, $g(X)$ elements of $F[X]$, with $g(X) \neq 0$. Then there exist unique $q(X)$ and $r(X)$ in $F[X]$ such that both the following hold:*

(i) $f(X) = q(X)g(X) + r(X)$.

(ii) Either $r(X) = 0$ *or* $\deg\big(r(X)\big) < \deg\big(g(X)\big)$.

Proof. We first prove the existence of such polynomials $q(X)$ and $r(X)$.

Let $S = \{\, f(X) - k(X)g(X) \mid k(X) \in F[X] \,\}$. If $0 \in S$ then there is a polynomial $k(X) \in F[X]$ with $f(X) = k(X)g(X)$, and we may take $q(X) = k(X)$ and $r(X) = 0$. Assume therefore that $0 \notin S$. The set of non-negative integers $K = \{\, \deg(p(X)) \mid p(X) \in S \,\}$ is nonempty, and therefore has a least element d. Let $r(X) \in S$ be such that $\deg(r(X)) = d$, and let $q(X)$ be such that $f(X) - q(X)g(X) = r(X)$ (possible since $r(X) \in S$).

It suffices to prove that $d < \deg(g(X))$; so suppose that this is not true. Let $\deg(g(X)) = m$ and let the leading coefficients of $r(X)$ and $g(X)$ be a and b respectively. Now $ab^{-1}X^{d-m}g(X)$ has degree $(d-m)+\deg(g) = d$ and leading coefficient (ab^{-1})(leading coefficient of g) $= a$; thus it has the same degree and leading coefficient as $r(X)$. It follows that the d^{th} and all higher coefficients of $s(X) = r(X) - ab^{-1}X^{d-m}g(X)$ are zero. Moreover,

$$s(X) = f(X) - q(X)g(X) - ab^{-1}X^{d-m}g(X) = f(X) - k(X)g(X) \in S$$

where $k(X) = q(X) + ab^{-1}X^{d-m}g(X)$. Thus $s(X)$ is a element of S of smaller degree than $r(X)$, contradicting the choice of $r(X)$. Thus the assumption that $d \geq m$ leads to a contradiction, and $d < m$, as required.

We have still to prove the uniqueness of q and r; so assume that q_1 and r_1 satisfy the same two properties; that is, $f = q_1 g + r_1$ and either $r_1 = 0$ or $\deg(r_1) < \deg(g)$. Then $q_1 g + r_1 = qg + r$, and so $r_1 - r = (q - q_1)g$. Now if $q - q_1 \neq 0$ then by Theorem 6.6 (i)

$$\deg(r_1 - r) = \deg(q - q_1) + \deg(g) \geq \deg(g)$$

which is impossible since the i^{th} coefficients of both r_1 and r are zero for $i > \deg(g)$. Hence $q_1 = q$, and this gives $r_1 = f - q_1 g = f - qg = r$ also.

□

Comment ▷▷▷

6.9.1 The polynomial $q(X)$ is called the *quotient* and $r(X)$ the *remainder*. They can be calculated by a process called the *division algorithm*, as follows. Firstly, if $\deg(f) < \deg(g)$ then the quotient is 0 and the remainder is equal to f. Otherwise we subtract from f a multiple of g with the same leading coefficient as f—that is, $ab^{-1}X^{n-m}g(X)$ where a, b are the leading coefficients of f, g and $n = \deg(f)$, $m = \deg(g)$—thereby reducing the degree, add the term $ab^{-1}X^{n-m}$ to the quotient, and repeat the process:

$$\begin{array}{rl} & b_m^{-1}a_n X^{n-m} + \cdots \\ b_m X^m + b_{m-1}X^{m-1} + \cdots + b_0 \,\Big|\, & a_n X^n + a_{n-1}X^{n-1} + \cdots + a_0 \\ \text{(subtract)} & \underline{a_n X^n + b_{m-1}b_m^{-1}a_n X^{n-1} + \cdots} \\ & a'_{n-1}X^{n-1} + \cdots \ \big(= \text{new } a(X)\big) \end{array}$$

—repeat.

▷▷▷

As a corollary of Theorem 6.9 we obtain the Remainder Theorem, which provides a quick method for calculating remainders on dividing by polynomials of degree one.

6.10 The Remainder Theorem *Let $c \in F$ and $f(X) \in F[X]$, where F is a field. Then the remainder in the division of $f(X)$ by $X - c$ is $f(c)$.*

Proof. By 6.9 we have $f(X) = (X - c)q(X) + r$, where either $r = 0$ or $\deg(r) < 1$. In either case r must be a constant polynomial; that is, an

element of F. Evaluating at c gives

$$f(c) = e_c(f(X)) = e_c(X - c)e_c(q(X)) + e_c(r) = 0q(c) + r = r.$$

□

NOTATION. If $a(X)$, $b(X) \in F[X]$ then '$a(X)\big|b(X)$' means 'there exists $q(X) \in F[X]$ with $b(X) = a(X)q(X)$'. If this holds we say that $a(X)$ is a *factor* of $b(X)$.

By Theorem 6.9, $a(X)$ is a factor of $b(X)$ if and only if the remainder on dividing $b(X)$ by $a(X)$ is zero.

As an immediate consequence of 6.10 we have the Factor Theorem, which provides an easy method for determining whether or not a given polynomial of degree 1 is a factor of $f(X)$.

6.11 THE FACTOR THEOREM *If $f(X) \in F[X]$ then $X - c$ is a factor of $f(X)$ if and only if $f(c) = 0$.*

Proof. Since $(X - c)\big|f(X)$ if and only if the remainder on dividing $f(X)$ by $X - c$ is zero, 6.10 yields that $(X - c)\big|f(X)$ if and only if $f(c) = 0$. □

§6g The Euclidean Algorithm

Throughout this section, F will be a field.

6.12 DEFINITION (i) Two polynomials $f(X)$ and $g(X)$ in $F[X]$ are said to be *associates* if $f(X) = cg(X)$ for some nonzero $c \in F$.

(ii) A polynomial $f(X) \in F[X]$ is said to be *monic* if it is nonzero and has leading coefficient 1.

Comment ▷▷▷

6.12.1 Obviously for any nonzero polynomial $f(X)$ there is a unique monic polynomial which is an associate of $f(X)$—namely, $a^{-1}f(X)$, where a is the leading coefficient of $f(X)$. ▷▷▷

6.13 PROPOSITION *Nonzero polynomials $f(X)$ and $g(X)$ in $F[X]$ are associates if and only if $f(X)|g(X)$ and $g(X)|f(X)$.*

Proof. If f and g are associates then for some $c \in F$ we have $f = cg$ and $g = c^{-1}f$, so that $g|f$ and $f|g$. Conversely, assume that $f|g$ and $g|f$. Then $f = q_1 g$ and $g = q_2 f$ for some q_1, $q_2 \in F[X]$, both of which are nonzero since f and g are. Thus by Theorem 6.6 (i) we have

$$\deg(f) = \deg(q_1) + \deg(g) \geq \deg(g) = \deg(q_2) + \deg(f).$$

Hence $\deg(q_2) = 0$, and therefore q_2 is a nonzero element of F, showing that f and g are associates. □

6.14 THEOREM *If $a(X)$ and $b(X)$ are polynomials in $F[X]$ which are not both zero then there exists a unique monic polynomial $d(X) \in F[X]$ such that both the following conditions are satisfied:*

(i) $d(X)|a(X)$ and $d(X)|b(X)$.

(ii) If $c(X)|a(X)$ and $c(X)|b(X)$ then $c(X)|d(X)$.

Moreover, there exist $m(X)$, $n(X) \in F[X]$ with

$$d(X) = m(X)a(X) + n(X)b(X).$$

Proof. Let $a(X)$ and $b(X)$ be elements of $F[X]$ which are not both zero. We first prove the existence of a $d(X)$ with the required properties.

Define S to be the set of all nonzero polynomials $p(X)$ in $F[X]$ such that $p(X) = m(X)a(X) + n(X)b(X)$ for some $m(X), n(X) \in F[X]$, and observe that $S \neq \emptyset$ since it must contain either $a(X)$ or $b(X)$. Hence the set of nonnegative integers

$$K = \{\, \deg(p(X)) \mid p(X) \in S \,\}$$

is nonempty, and must therefore contain a least element, k. Let $d(X)$ be an element of S which is monic and has degree k. (By 6.12.1 we can choose $d(X)$ to be monic, since associates of elements of S are also in S.)

Since $d(X) \in S$ the definition of S yields the existence $m(X)$ and $n(X)$ with $d(X) = m(X)a(X) + n(X)b(X)$, and from this it follows that if $c(X)|a(X)$ and $c(X)|b(X)$ then $c(X)\big|(m(X)a(X) + n(X)b(X)) = d(X)$. Thus we have established two of the properties of $d(X)$ and have only to prove that $d(X)|a(X)$ and $d(X)|b(X)$.

Suppose that $d(X) \not| a(X)$, and let $r(X)$ be the remainder on division of $a(X)$ by $d(X)$. Then $r(X) \neq 0$, and since $r(X) = a(X) - q(X)d(X)$ for

some $q(X)$, we obtain

$$\begin{aligned} r(X) &= a(X) - q(X)(m(X)a(X) + n(X)b(X)) \\ &= (1 - q(X)m(X))a(X) - q(X)b(X) \end{aligned}$$

so that $r(X) \in S$. But this contradicts the definition of k, since the degree of $r(X)$ is less than $\deg(d(X)) = k$. Thus $d(X)|a(X)$ and, by a similar argument, $d(X)|b(X)$ also.

It remains to prove uniqueness. So, let d_1 and d_2 be monic polynomials such that conditions (i) and (ii) are satisfied with d replaced by d_1 and also with d replaced by d_2. By (i) for d_1 and (ii) for d_2 it follows that $d_1|d_2$, and by (i) for d_2 and (ii) for d_1 it follows that $d_2|d_1$. By 6.13 we deduce that d_1 and d_2 are associates of each other, and hence, by 6.12.1, $d_1 = d_2$. □

Comment ▷▷▷

6.14.1 The polynomial $d(X)$ in 6.14 is called the *greatest common divisor* of $a(X)$ and $b(X)$. ▷▷▷

As in the case of integers, the greatest common divisor of two polynomials can be calculated by use of the Euclidean Algorithm (which is almost exactly the same for polynomials as integers):

Given $a, b \in F[X]$ with $a \neq 0$, $b \neq 0$ and $\deg(a) \geq \deg(b)$ (or $b = 0$, $a \neq 0$),

```
while b ≠ 0 do
        [a, b] := [b, a − b ∗ (adivb)]
enddo
α := leading coefficient of a
a := (1/α) a
end
```

At the end of the process, a is the gcd of the initial two polynomials.

Alternatively, let a_1, a_2 be the initial polynomials, and define a_3, a_4, ... by

$$\begin{aligned} a_1 &= q_3 a_2 + a_3 && \deg(a_3) < \deg(a_2) \\ a_2 &= q_4 a_3 + a_4 && \deg(a_4) < \deg(a_3) \\ &\vdots \\ a_{k-2} &= q_k a_{k-1} + a_k && \deg(a_k) < \deg(a_{k-1}) \\ a_{k-1} &= q_{k+1} a_k && (a_{k+1} = 0). \end{aligned}$$

The algorithm must terminate eventually since the remainder on dividing a_{i-1} by a_i is either zero or a polynomial of degree strictly less than that of a_i. Since the degree of a nonzero polynomial is always a nonnegative integer, and it is impossible to have an infinite decreasing sequence of nonnegative integers, it must eventually happen that we get a remainder of zero. (For instance, if $\deg(a_i) = 0$ then we will certainly find that $a_{i+1} = 0$; a polynomial of degree 0 is always a divisor of any other polynomial.)

As for integers, the set of common divisors of a_{i-1} and a_i remains unchanged throughout the algorithm, and hence

$$\gcd(a_1, a_2) = \gcd(a_2, a_3) = \cdots = \gcd(a_k, a_{k+1}).$$

But $\gcd(a_k, a_{k+1}) = \gcd(a_k, 0)$, which is the unique monic associate of a_k. (The greatest common divisor is always monic, by definition; so, for instance, $\gcd(2X + 3,\, 0\,) = X + \frac{3}{2}$.) Thus we conclude that the gcd of a_1 and a_2 is the unique monic associate of the last nonzero remainder obtained in the Euclidean Algorithm.

—***Example***—

#1 Find $m, n \in \mathbb{R}[X]$ such that

$$(*) \qquad m(X)\big(X^3 + X - 1\big) + n(X)\big(X^2 + 4\big) = 1.$$

$\gg\!\!\rightarrowtriangle$ By division we find

$$\begin{aligned} X^3 + X - 1 &= X(X^2 + 4) + (-3X - 1) \\ X^2 + 4 &= (-\tfrac{1}{3}X + \tfrac{1}{9})(-3X - 1) + \tfrac{37}{9}. \end{aligned}$$

(The Euclidean Algorithm terminates at the next step, since $\frac{37}{9} \mid -3X - 1$.)

So we have

$$\begin{aligned} \tfrac{37}{9} &= (X^2 + 4) + (\tfrac{1}{3}X - \tfrac{1}{9})(-3X - 1) \\ &= (X^2 + 4) + (\tfrac{1}{3}X - \tfrac{1}{9})[(X^3 + X - 1) - X(X^2 + 4)] \\ &= [1 - X(\tfrac{1}{3}X - \tfrac{1}{9})](X^2 + 4) + (\tfrac{1}{3}X - \tfrac{1}{9})(X^3 + X - 1) \\ &= (-\tfrac{1}{3}X^2 + \tfrac{1}{9}X + 1)(X^2 + 4) + (\tfrac{1}{3}X - \tfrac{1}{9})(X^3 + X - 1). \end{aligned}$$

Thus

$$1 = (-\tfrac{3}{37}X^2 + \tfrac{1}{37}X + \tfrac{9}{37})(X^2 + 4) + (\tfrac{3}{37}X - \tfrac{1}{37})(X^3 + X - 1).$$

So if we define

$$m_0(X) = \tfrac{3}{37}X - \tfrac{1}{37}$$
$$n_0(X) = -\tfrac{3}{37}X^2 + \tfrac{1}{37}X + \tfrac{9}{37}$$

then $m(X) = m_0(X)$, $n(X) = n_0(X)$ is a solution to $(*)$.

Notice that the solution is not unique. In fact, for any $p(X) \in \mathbb{R}[X]$ we can obtain another solution by putting

$$m(X) = m_0(X) + p(X)\big(X^2 + 4\big)$$
$$n(X) = n_0(X) - p(X)\big(X^3 + X - 1\big).$$

◁—≪

---

§6h Irreducible polynomials

Let F be a field and let $p \in F[X]$. Then for any nonzero $c \in F$ the equation $p(X) = c\big(c^{-1}p(X)\big)$ shows that c is a divisor of p. Similarly all associates of p are divisors of p. Polynomials which have only these trivial divisors are of considerable theoretical importance.

6.15 Definition A polynomial $p \in F[X]$ is said to be *irreducible* (or *prime*) if $\deg(p) \geq 1$ and the only divisors of p in $F[X]$ are polynomials of degree 0 and associates of p.

Comments ▷▷▷

6.15.1 If p is irreducible and $p(X) = d_1(X)d_2(X)$ then either d_1 is an associate of p, in which case $\deg(d_2) = 0$, or $\deg(d_1) = 0$, in which case d_2 is an associate of p.

6.15.2 If p is reducible (that is, not irreducible) and $\deg(p) \geq 1$ then p has a divisor d_1 satisfying

(i) $\deg(d_1) \geq 1$

(ii) d_1 is not an associate of p.

Since d_1 is a divisor of p we have $p(X) = d_1(X)d_2(X)$ for some d_2, and (ii) above implies that $\deg(d_2) \geq 1$. This combined with (i) above and the equation

$$\deg(p) \;=\; \deg(d_1) + \deg(d_2)$$

yields that

$$1 \leq \deg(d_i) \leq \deg(p) - 1$$

for $i = 1$ and $i = 2$.

6.15.3 If $\deg(p) = 1$ then it follows from 6.15.2 above that p is irreducible. For if p were reducible we could find d_1 and d_2 with $p(X) = d_1(X)d_2(X)$ and $1 \leq \deg(d_i) \leq \deg(p) - 1$ (for $i = 1, 2$). But this is impossible since $\deg(p) - 1 = 0$. (The point is that if neither d_1 nor d_2 is a constant polynomial then $\deg(p) = \deg(d_1) + \deg(d_2) \geq 1 + 1 = 2$.)

6.15.4 If $\deg(p)$ is 2 or 3 and p is reducible then p has a zero in F. For it follows from 6.15.2 that $p(X) = d_1(X)d_2(X)$ with

$$\deg(d_1) + \deg(d_2) = \deg(p) \ (= 2 \text{ or } 3)$$

and

$$\deg(d_i) \geq 1 \qquad \text{for } i = 1 \text{ and } i = 2.$$

Now if both $\deg(d_1) \geq 2$ and $\deg(d_2) \geq 2$ then $\deg(d_1) + \deg(d_2) \geq 4$, contradiction. So either d_1 or d_2 has degree 1. So p has a factor of the form $aX + b$ with $a, b \in F$ and $a \neq 0$. Thus, for some $d \in F[X]$,

$$\begin{aligned} p(X) &= (aX + b)d(X) \\ &= a\big(X - (-a^{-1}b)\big)d(X). \end{aligned}$$

By the Factor Theorem, $-a^{-1}b$ is a zero of $p(X)$. ▷▷▷

§6i Some examples

It will be convenient in the future to denote the elements of $\mathbb{Z}_n$ by '0', '1', '2', ... and so on, instead of '$\bar{0}$', '$\bar{1}$', '$\bar{2}$', ..., since it becomes rather clumsy to have bars over all the coefficients when dealing with polynomials over $\mathbb{Z}_n$. Of course we will have to be careful to remember things like $7{=}2$ in $\mathbb{Z}_5$, $-1 = +1$ in $\mathbb{Z}_2$, and so on.

#2 In $\mathbb{Z}_3[X]$ the polynomial $p(X) = X^2 - X - 1$ is irreducible. For by 6.15.4 above, if $X^2 - X - 1$ were reducible it would have a zero in $\mathbb{Z}_3$. But

$$\begin{aligned} p(0) &= -1 = 2 \neq 0 \\ p(1) &= -1 = 2 \neq 0 \\ p(2) &= 1 \neq 0, \end{aligned}$$

and since 0, 1, 2 are the only elements of $\mathbb{Z}_3$ we see that $p(X)$ has no zeros in $\mathbb{Z}_3$.

#3 In $\mathbb{R}[X]$ the polynomial $X^2 - X - 1$ is reducible. Indeed

$$X^2 - X - 1 = (X - \alpha)(X - \beta)$$

where $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$.

#4 In $\mathbb{R}[X]$ the polynomial $X^2 + 1$ is irreducible. For if it were reducible it would have a factor of degree 1, and hence a zero in $\mathbb{R}$. But $a^2 + 1 \neq 0$ for all $a \in \mathbb{R}$.

#5 In $\mathbb{C}[X]$ the polynomial $X^2 + 1$ is reducible. Indeed we have the factorization $X^2 + 1 = (X - \boldsymbol{i})(X + \boldsymbol{i})$.

#6 In $\mathbb{R}[X]$ we have $X^2 - 3 = (X - \sqrt{3})(X + \sqrt{3})$, and so $X^2 - 3$ is reducible. But in $\mathbb{Q}[X]$ the polynomial $X^2 - 3$ is irreducible, since it has no zeros in $\mathbb{Q}$. (There is no rational number α such that $\alpha^2 - 3 = 0$.)

#7 Polynomials of degree four or more may have no zeros and yet be reducible. For instance, $X^4 + X^2 + 1$ has no zeros in $\mathbb{R}$ but is nevertheless a reducible element of $\mathbb{R}[X]$. In fact $X^4 + X^2 + 1 = (X^2 + X + 1)(X^2 - X + 1)$.

#8 Irreducibles in $\mathbb{C}[X]$

The "Fundamental Theorem of Algebra" states that every polynomial p in $\mathbb{C}[X]$ of degree at least one has a zero in $\mathbb{C}$. By the Factor Theorem it follows that $p(X) = (X - c)q(X)$ for some $c \in \mathbb{C}$ and $q(X) \in \mathbb{C}[X]$. So if p is irreducible the degree of q must be zero, making $X - c$ an associate of p. It follows that the only irreducible polynomials in $\mathbb{C}[X]$ are the polynomials of degree one.

#9 Irreducibles in $\mathbb{R}[X]$

Suppose that $p \in \mathbb{R}[X]$, p is irreducible, and $\deg(p) > 1$. Since p has no factors of degree 1 in $\mathbb{R}[X]$ it has no zeros in $\mathbb{R}$. But by the Fundamental Theorem of Algebra $p(X)$ has a zero $a + b\boldsymbol{i}$ in $\mathbb{C}$. We must have $b \neq 0$ since this zero is not in $\mathbb{R}$. Observe that $a + b\boldsymbol{i}$ is also a zero of the polynomial $X^2 - 2aX + a^2 + b^2 \in \mathbb{R}[X]$. By Theorem 6.9

$$p(X) = q(X)\big(X^2 - 2aX + a^2 + b^2\big) + (r_0 + r_1 X)$$

for some $r_0,\ r_1 \in \mathbb{R}$. Substituting $X = a + b\boldsymbol{i}$ gives

$$0 = p(a+b\boldsymbol{i}) = q(a+b\boldsymbol{i})0 + \big(r_0 + r_1(a+b\boldsymbol{i})\big) = (r_0 + r_1 a) + (r_1 b)\boldsymbol{i}.$$

Equating real and imaginary parts gives $r_1 b = 0$ and $r_0 + r_1 a = 0$. Since $b \neq 0$ this gives $r_1 = 0$, and hence $r_0 = 0$. Thus $X^2 - 2aX + a^2 + b^2$ is a factor of $p(X)$, and since $p(X)$ is irreducible it must be an associate of $X^2 - 2aX + a^2 + b^2$. Hence all irreducibles in $\mathbb{R}[X]$ are of degree 1 or 2.

#10 In $\mathbb{Q}[X]$ there are irreducibles of all degrees. In fact Eisenstein's Criterion, to be proved in §6k below, shows that $X^n - 2$ is irreducible in $\mathbb{Q}[X]$ for all $n \in \mathbb{Z}$.

§6j Factorization of polynomials

The proofs of the following facts are very similar to the corresponding proofs for $\mathbb{Z}$, and are omitted.

6.16 Theorem *Let a, b, p be polynomials over the field F, and suppose that p is irreducible and $p|ab$. Then $p|a$ or $p|b$.*

6.17 Lemma *Suppose that p, q_1, q_2, $\dots q_s$ are monic irreducible polynomials in $F[X]$ and that for some nonzero $d \in F$ we have*

$$p(X)\big|dq_1(X)q_2(X)\dots q_s(X).$$

Then $p(X) = q_j(X)$ for some j.

6.18 Unique Factorization Theorem *(i) Suppose that $f(X)$ is a polynomial of degree greater than one with coefficients in the field F, and let c be the leading coefficient of f. Then there exist monic irreducible polynomials $p_1(X)$, $p_2(X)$, $\dots$, $p_r(X)$ in $F[X]$ such that*

$$f(X) = cp_1(X)p_2(X)\dots p_r(X).$$

(ii) If $cp_1(X)p_2(X)\dots p_r(X) = dq_1(X)q_2(X)\dots q_s(X)$ where c, d are nonzero elements of F and the p_i, q_j are monic irreducible polynomials, then $c = d$, $r = s$, and

$$p_1 = q_{i_1},\ p_2 = q_{i_2}, \dots,\ p_r = q_{i_r}$$

where i_1, i_2, $\dots i_r$ are the numbers 1, 2, $\dots r$ in some order.

—Examples—

#11 Since $\mathbb{Z}_{17}$ is a field (because 17 is prime) the Unique Factorization Theorem holds in $\mathbb{Z}_{17}[X]$. So, for instance, $X^2 - 6X + 5 = (X-1)(X-5)$, and this is the unique way of writing X^2-6X+5 as a product of irreducibles. On the other hand, $\mathbb{Z}_{16}$ is not a field, and in $\mathbb{Z}_{16}[X]$ we find that

$$\begin{aligned}(X-1)(X-5) &= X^2 - 6X + 5\\ &= X^2 + 10X + 21\\ &= (X+7)(X+3).\end{aligned}$$

Unique factorization does not hold in $\mathbb{Z}_{16}[X]$.

#12 In $\mathbb{Z}_2[X]$ there are eight polynomials of degree three. We list them all and express each as a product of irreducibles:

$$\begin{aligned}X^3 &= XXX\\ X^3+1 &= (X+1)(X^2+X+1)\\ X^3+X &= X(X+1)(X+1)\\ X^3+X+1 &\text{ is irreducible}\\ X^3+X^2 &= XX(X+1)\\ X^3+X^2+1 &\text{ is irreducible}\\ X^3+X^2+X &= X(X^2+X+1)\\ X^3+X^2+X+1 &= (X+1)(X+1)(X+1).\end{aligned}$$

---

§6k Irreducibility over the rationals

6.19 Proposition *Suppose that $f(X)$, $g(X) \in \mathbb{Z}[X]$ and $p \in \mathbb{Z}$ is a prime integer which divides all the coefficients of $f(X)g(X)$. Then either p divides all the coefficients of $f(X)$ or all the coefficients of $g(X)$.*

Proof. Let

$$\begin{aligned}f(X) &= a_0 + a_1X + \cdots + a_nX^n\\ g(X) &= b_0 + b_1X + \cdots + b_mX^m.\end{aligned}$$

and define

$$\phi(X) = \overline{a_0} + \overline{a_1}X + \cdots + \overline{a_n}X^n \in \mathbb{Z}_p[X]$$
$$\gamma(X) = \overline{b_0} + \overline{b_1}X + \cdots + \overline{b_m}X^m \in \mathbb{Z}_p[X]$$

where we have reverted to the bar notation for elements of $\mathbb{Z}_p$ to avoid confusion. Then

$$f(X)g(X) = a_0b_0 + (a_0b_1 + a_1b_0)X + (a_0b_2 + a_1b_1 + a_2b_0)X^2 + \cdots$$

and by a similar calculation

$$\begin{aligned}\phi(X)\gamma(X) &= \overline{a_0}\,\overline{b_0} + (\overline{a_0}\,\overline{b_1} + \overline{a_1}\,\overline{b_0})X + (\overline{a_0}\,\overline{b_2} + \overline{a_1}\,\overline{b_1} + \overline{a_2}\,\overline{b_0})X^2 + \cdots \\ &= \overline{a_0b_0} + \overline{(a_0b_1 + a_1b_0)}X + \overline{(a_0b_2 + a_1b_1 + a_2b_0)}X^2 + \cdots \\ &= \overline{0}\end{aligned}$$

since all the coefficients of $f(X)g(X)$ are divisible by p. But $\mathbb{Z}_p$ is an integral domain (by Theorem 4.10) and so $\mathbb{Z}_p[X]$ is also an integral domain (by Theorem 6.6), and therefore has no zero divisors. It follows that either $\phi(X) = \overline{0}$, in which case $\overline{a_0}$, $\overline{a_1}$, ... are all zero, and a_0, a_1, ... are all divisible by p, or $\gamma(X) = \overline{0}$, in which case all the coefficients of $g(X)$ are divisible by p. □

6.20 Gauss' Lemma *Suppose that $a(X) \in \mathbb{Q}[X]$ is reducible and has all its coefficients in $\mathbb{Z}$. Then $a(X)$ has a nontrivial factorization in $\mathbb{Z}[X]$.*

Proof. Let $a(X) = f(X)g(X)$, where $f(X)$, $g(X) \in \mathbb{Q}[X]$ and both f and g have degree less than the degree of a. Observe that there exist integers m and n such that all the coefficients of $mf(X)$ and $ng(X)$ lie in $\mathbb{Z}$—for instance, if $f(X) = (r_0/s_0) + (r_1/s_1)X + \cdots + (r_d/s_d)X^d$ with the r_i and s_i in $\mathbb{Z}$, then taking $m = s_0s_1 \ldots s_d$ would suffice. Hence if $k = mn$ the following property is satisfied:

(P) There exist polynomials f_1 and g_1 which have integral coefficients and are associates of f and g respectively, such that $ka(X) = f_1(X)g_1(X)$.

Let K be the set of all positive integers k for which (P) is satisfied.

Since K is nonempty it has a least element, h. It suffices to prove that $h = 1$, for then (P) shows that $a(X)$ has a factorization of the required kind. So, suppose that $h > 1$. By Theorem 3.7 there exists a prime p which is a factor of h, and since the coefficients of $a(X)$ are integral it follows that the coefficients of $ha(X)$ are all divisible by p. Since $h \in K$ there exist f_1, $g_1 \in \mathbb{Z}[X]$

with $\deg(f_1) = \deg(f)$, $\deg(g_1) = \deg(g)$ and $ha(X) = f_1(X)g_1(X)$, and by Proposition 6.19 either $(1/p)f_1$ has integral coefficients or $(1/p)g_1$ does. It follows that (h/p) is in K, since

$$(h/p)a(X) = \big((1/p)f_1(X)\big)g_1(X) = f_1(X)\big((1/p)g_1(X)\big).$$

This contradicts the fact that h is the smallest element in K, proving that $h = 1$, as required. □

As a corollary of Theorem 6.20 we obtain an easy way of listing all rational numbers which can possibly be roots of a given integral polynomial.

6.21 Rational Roots Theorem *If $a_0 + a_1X + \cdots + a_dX^d \in \mathbb{Z}[X]$ then all zeros of $a(X)$ in $\mathbb{Q}$ have the form $\pm(m/n)$ where m and n are integers such that $m|a_0$ and $n|a_d$.*

Proof. If $a(X)$ has a zero in $\mathbb{Q}$ it has a linear factor $m - nX$ in $\mathbb{Q}[X]$. Thus there is a factorization

$$(m - nX)(b_0 + b_1X + \cdots + b_{d-1}X^{d-1}) = a_0 + a_1X + \cdots + a_dX^d$$

and by Theorem 6.20 we may assume that all the coefficients of both factors are integral. But since $a_0 = mb_0$ and $a_d = -nb_{d-1}$ we deduce that $m|a_0$ and $n|a_d$. This proves the claim, since the zero corresponding to the linear factor $m - nX$ is (m/n). (The $\pm$ appears in the theorem statement merely to emphasize that m and n may be negative.) □

—***Example***—

#13 By Theorem 6.20 the only rational numbers which can be zeros of $3 - 13X - 7X^2 + 2X^3$ are ±1, ±3, $\pm(1/2)$, $\pm(3/2)$. Trying them all one finds that in fact $-(3/2)$ is the only rational zero.

6.22 Eisenstein's Irreducibility Criterion *If p is a prime integer and $a(X) = a_0 + a_1X + \cdots + a_dX^d \in \mathbb{Z}[X]$ is such that $p|a_i$ for all $i \neq d$, $p \not| a_d$ and $p^2 \not| a_0$, then $a(X)$ is irreducible in $\mathbb{Q}[X]$.*

Proof. Suppose that there is a nontrivial factorization of $a(X)$ in $\mathbb{Q}[X]$:

$$(*) \quad a_0 + a_1X + \cdots + a_dX^d = (b_0 + b_1X + \cdots + b_rX^r)(c_0 + c_1X + \cdots + c_sX^s)$$

where $r \geq 1$, $s \geq 1$ and $r + s = d$. By 6.20 we may assume that b_i, $c_i \in \mathbb{Z}$.

On expanding $(*)$ we obtain the equations

$$\begin{aligned} a_0 &= b_0 c_0 \\ a_1 &= b_0 c_1 + b_1 c_0 \\ a_2 &= b_0 c_2 + b_1 c_1 + b_2 c_0 \\ &\vdots \end{aligned}$$

and in $\mathbb{Z}_p$ we therefore have similar equations

$$\begin{aligned} \overline{a_0} &= \overline{b_0}\,\overline{c_0} \\ \overline{a_1} &= \overline{b_0}\,\overline{c_1} + \overline{b_0}\,\overline{c_1} \\ \overline{a_2} &= \overline{b_0}\,\overline{c_2} + \overline{b_1}\,\overline{c_1} + \overline{b_2}\,\overline{c_0} \\ &\vdots \end{aligned}$$

which in $\mathbb{Z}_p$ yield the factorization

$$(**) \quad \overline{a_0} + \overline{a_1}X + \cdots + \overline{a_d}X^d = (\overline{b_0} + \overline{b_1}X + \cdots + \overline{b_r}X^r)(\overline{c_0} + \overline{c_1}X + \cdots + \overline{c_s}X^s).$$

But $\overline{a_0} = \overline{a_1} = \cdots = \overline{a_{d-1}} = \overline{0}$ and $\overline{a_d} \neq 0$; so the left hand side above is an associate of X^d. But $\mathbb{Z}_p$ is a field, and so the Unique Factorization Theorem 6.18 applies to $\mathbb{Z}_p[X]$. The only monic irreducible factor of X^d is X; so it follows that the factors in $(**)$ are associates of powers of X. Moreover, the right hand side of $(**)$ cannot have degree less than $d = r + s$; so $\overline{b_r} \neq 0$ and $\overline{c_s} \neq 0$. Thus

$$\begin{aligned} \overline{b_0} + \overline{b_1}X + \cdots + \overline{b_r}X^r &= \lambda X^r \\ \overline{c_0} + \overline{c_1}X + \cdots + \overline{c_s}X^s &= \mu X^s \end{aligned}$$

for some λ, $\mu \in \mathbb{Z}_p$. Since $r \geq 1$ and $s \geq 1$ it follows that $\overline{b_0} = \overline{0}$ and $\overline{c_0} = \overline{0}$. Thus for some integers h and k we have $b_0 = ph$ and $c_0 = pk$, giving

$$a_0 = b_0 c_0 = p^2 hk$$

contrary to the assumption that $p^2 \not| a_0$. □

For example, application of Eisenstein's Criterion with $p = 3$ shows that $X^4 - 9X^2 - 36X - 33$ is irreducible over $\mathbb{Q}$, but gives no information about $X^4 - 9X^2 - 36X - 36$.

—Examples—

#14 Prove that $\sqrt[3]{2}$ is irrational.

≫▷ By Eisenstein's Criterion with $p = 2$ we see that $X^3 - 2$ is irreducible in $\mathbb{Q}[X]$ and so, by the Factor Theorem, it has no zeros in $\mathbb{Q}$. Hence $\sqrt[3]{2} \notin \mathbb{Q}$.

Alternatively, by the Rational Roots Theorem the only possible rational roots of $X^3 - 2$ are ± 1 and ± 2. Since $\sqrt[3]{2}$ is a root, it is not rational. ◁≪

#15 Let a, b, c be rational numbers which are not all zero, and let

$$t = a + b\sqrt[3]{2} + c\left(\sqrt[3]{2}\right)^2.$$

Prove that $t \neq 0$.

≫▷ If c and b are both zero then a cannot be zero, and $t = a \neq 0$. If $c = 0$ and $b \neq 0$ then $t = 0$ would give $\sqrt[3]{2} = -(a/b)$, contradicting the irrationality of $\sqrt[3]{2}$. Thus we may assume that $c \neq 0$.

Let $p(X) = a + bX + cX^2$, and let $r(X)$ be the remainder obtained when $X^3 - 2$ is divided by $p(X)$. Thus

($) $$X^3 - 2 = p(X)q(X) + r(X)$$

for some $q(X) \in \mathbb{Q}[X]$, and by Theorem 6.9 $r(X) = c + dX$ for some c, $d \in \mathbb{Q}$. If $t = 0$ then substituting $X = \sqrt[3]{2}$ in ($) gives $r(\sqrt[3]{2}) = 0$, and this contradicts the irrationality of $\sqrt[3]{2}$ unless $c = d = 0$. Hence $X^3 - 2$ is divisible by $p(X)$. But this contradicts the fact that $X^3 - 2$ is irreducible in $\mathbb{Q}[X]$, proved in #14 above. Hence $t \neq 0$. ◁≪

Exercises

1. Find the greatest common divisor $d(X)$ of the polynomials

and
$$f(X) = X^3 - 6X^2 + X + 4$$
$$g(X) = X^5 - 6X + 1$$

in $\mathbb{Q}[X]$, as well as polynomials $s(X)$ and $t(X)$ such that

$$d(X) = s(X)f(X) + t(X)g(X)$$

2. Find the greatest common divisor $d(X)$ of

$$\text{and} \qquad \begin{aligned} f(X) &= X^5 + X^4 + 2X^3 - X^2 - X - 2 \\ g(X) &= 2X^4 + 4X^3 + 3X + 3 \end{aligned}$$

where $f(X)$, $g(X) \in \mathbb{Z}_5[X]$. Also find polynomials $s(X)$, $t(X) \in \mathbb{Z}_5[X]$ such that

$$d(X) = s(X)f(X) + t(X)g(X).$$

3. Write $X^4 + 2X^3 + X^2 + 2X + 2$ over $\mathbb{Z}_3$ as a product of irreducible polynomials.

4. Let $f(X) = a_0 + a_1X + \cdots + a_{n-1}X^{n-1} + X^n \in \mathbb{Q}[X]$ be a monic polynomial with integral coefficients. Show that if $\overline{a_0} + \overline{a_1}X + \cdots + X^n \in \mathbb{Z}_p[X]$ is irreducible (for some prime p), then $f(X)$ is irreducible over $\mathbb{Q}$.

5. List all the polynomials of degree 4 in $\mathbb{Z}_2[X]$, and express them as products of irreducibles.

6. Test the following polynomials for irreducibility over $\mathbb{Q}$:

(*i*) $5X^3 + X^2 + X - 4$

(*ii*) $3X^5 + 2X^3 - 6X + 6$

(*iii*) $X^4 + 7X^3 - 10X^2 + 2X + 5$ (Hint: Apply Exercise 4 with $p = 2$.)

(*iv*) $2X^4 + 7X^3 - 15X^2 + 3X + 6$

7. Give an example of a quadratic polynomial in $\mathbb{Z}_6[X]$ which has more than two roots.

8. Prove that $\sqrt[5]{6}$ is irrational.

9. Find all monic irreducible quadratic polynomials over $\mathbb{Z}_5$.

10. Let $R[X]$ be a commutative ring with 1. Prove that the polynomial ring $R[X]$ cannot be a field.

(Hint: Show that X does not have an inverse.)

11. Let $f(X)$ and $g(X)$ be monic polynomials of degree 5 in $\mathbb{Z}_7[X]$ with the property that $f(c) = g(c)$ for all $c \in \mathbb{Z}_7$. Show that $f(X) = g(X)$.

12. Let F be a field and $f(X)$, $g(X)$ and $h(X)$ distinct monic polynomials in $F[X]$ with $f(X)$ and $g(X)$ irreducible. Show that the gcd of $h(X)$ and $f(X)g(X)$ is 1 if $h(X)$ is irreducible, but need not be 1 otherwise.

13. Prove that $\psi\colon \mathbb{R}[X] \to \mathrm{Mat}(3, \mathbb{R})$ defined by

$$\psi(a_0 + a_1X + \cdots + a_nX^n) = \begin{pmatrix} a_0 & a_1 & a_2 \\ 0 & a_0 & a_1 \\ 0 & 0 & a_0 \end{pmatrix}$$

is a homomorphism.

7
More Ring Theory

In this chapter we return to the general setting and develop the remaining theory which will be needed in this course. Our principal objectives are the definition of quotient rings and the Fundamental Homomorphism Theorem.

§7a More on homomorphisms

It will be convenient for us to temporarily generalize slightly the concept of a homomorphism. Suppose that S and T are sets each equipped with operations called addition and multiplication, about which nothing is assumed. (Thus, in particular, S and T need not be rings.) We still refer to a map $\theta\colon S \to T$ satisfying $\theta(xy) = \theta(x)\theta(y)$ and $\theta(x+y) = \theta(x) + \theta(y)$ as a *homomorphism*.

7.1 Lemma *Let R be a ring and T any set equipped with addition and multiplication, and suppose that $\theta\colon R \to T$ is a homomorphism. Then*

$$S = \{\, \theta(x) \mid x \in R \,\}$$

is a subset of T which forms a ring under the operations of T, and

$$K = \{\, x \in R \mid \theta(x) = \theta(0) \,\}$$

is an ideal of R.

Proof. Let a, b, $c \in S$. Then there exist x, y, $z \in R$ such that $a = \theta(x)$, $b = \theta(y)$ and $c = \theta(z)$, and the associativity of addition in R together with the fact that θ preserves addition gives

$$\begin{aligned}(a+b)+c &= \big(\theta(x)+\theta(y)\big)+\theta(z) = \theta(x+y)+\theta(z) = \theta\big((x+y)+z\big)\\ &= \theta\big(x+(y+z)\big) = \theta(x)+\theta(y+z) = \theta(x)+\big(\theta(y)+\theta(z)\big) = a+(b+c).\end{aligned}$$

Similar arguments show that $a+b = b+a$, $(ab)c = a(bc)$, $a(b+c) = ab+ac$ and $(a+b)c = ac+bc$. Hence S satisfies Axioms (i), (iv), (v) and (vi) of Definition 2.2.

The element $\theta(0) \in S$ is a zero element for S, since if $a \in S$ is arbitrary then (for some $x \in R$)

$$\text{and} \qquad \begin{aligned} a &= \theta(x) = \theta(x+0) = \theta(x) + \theta(0) = a + \theta(0) \\ a &= \theta(x) = \theta(0+x) = \theta(0) + \theta(x) = \theta(0) + a. \end{aligned}$$

Furthermore, since

$$\text{and} \qquad \begin{aligned} a + \theta(-x) &= \theta(x) + \theta(-x) = \theta(x + (-x)) = \theta(0) \\ \theta(-x) + a &= \theta(-x) + \theta(x) = \theta((-x) + x) = \theta(0) \end{aligned}$$

we see that a has a negative. Thus Axioms (ii) and (iii) are satisfied.

To prove that K is an ideal we use Proposition 5.8. Since $0 \in K$ we have that $K \neq \emptyset$. Now if x, $y \in K$ and $r \in R$ then

$$\begin{gathered} \theta(x+y) = \theta(x) + \theta(y) = \theta(0) + \theta(0) = \theta(0+0) = \theta(0) \\ \theta(-x) = \theta(-x+0) = \theta(-x) + \theta(0) = \theta(-x) + \theta(x) = \theta(-x+x) = \theta(0) \\ \theta(xr) = \theta(x)\theta(r) = \theta(0)\theta(r) = \theta(0r) = \theta(0) \\ \theta(rx) = \theta(r)\theta(x) = \theta(r)\theta(0) = \theta(r0) = \theta(0) \end{gathered}$$

so that $x+y$, $-x$, rx and xr are all in K. So all the required closure properties hold. □

Comment ▷▷▷

7.1.1 Since the element $\theta(0)$ of S is the zero of S the definition of K is, effectively,

$$K = \{\, x \in R \mid \theta(x) = 0 \,\}.$$

The set K is usually called the 'kernel' of θ. (The set S is called the 'image' of θ—see §0b.) ▷▷▷

7.2 Definition If R and S are rings and $\theta\colon R \to S$ a homomorphism, then the *kernel* of θ is the subset of R

$$\ker\theta = \theta^{-1}(0_S) = \{\, x \in R \mid \theta(x) = 0_S \,\}$$

where 0_S is the zero of S.

From Lemma 7.1 we have immediately:

7.3 THEOREM *The kernel of a ring homomorphism $\theta\colon R \to S$ is an ideal of R, the image a subring of S.*

It is trivial that a homomorphism $\theta\colon R \to S$ is surjective if and only if $\operatorname{im}\theta = S$. The next result provides an analogous criterion for injectivity:

7.4 THEOREM *A ring homomorphism $\theta\colon R \to S$ is injective if and only if* $\ker\theta = \{0_R\}$.

Proof. Suppose that θ is injective. It follows from Theorem 5.5 (i) that $0_R \in \ker\theta$. Suppose that x is another element of $\ker\theta$. Then

$$\theta(x) = 0_S = \theta(0_R)$$

and injectivity of θ gives $x = 0_R$. Thus 0_R is the only element of $\ker\theta$, as required.

Conversely, suppose that $\ker\theta = \{0_R\}$, and let x and y be elements of R with $\theta(x) = \theta(y)$. Then by Theorem 5.5

$$\theta(x - y) = \theta(x) - \theta(y) = 0_S$$

and therefore $x - y \in \ker\theta$. Hence $x - y = 0_R$, and $x = y$. Thus θ is injective. □

7.5 THEOREM *If $\theta\colon R \to S$ and $\psi\colon S \to T$ are ring homomorphisms, then the composite map $\psi\theta\colon R \to T$ defined by*

$$(\psi\theta)(x) = \psi\big(\theta(x)\big) \qquad \textit{for all } x \in R$$

is also a homomorphism.

The proof of this is left to the exercises.

—***Example***—

#1 Define $f\colon \mathbb{Z}[X] \to \mathbb{Z}$ by

$$f(a_0 + a_1X + \cdots + a_nX^n) = a_0.$$

Observe that f coincides with the evaluation map e_0 (see §6e): the result of putting $X = 0$ in $a_0 + a_1X + \cdots + a_nX^n$ is a_0. So f is a homomorphism. Now let $g\colon \mathbb{Z} \to \mathbb{Z}_3$ be the natural homomorphism (given by $a \mapsto \bar{a}$ for all $a \in \mathbb{Z}$ (see §5b#7)). The composite map $gf\colon \mathbb{Z}[X] \to \mathbb{Z}_3$ is given by

$$a_0 + a_1X + \cdots + a_nX^n \longmapsto \overline{a_0}$$

and by 7.5 it is a homomorphism. By 7.3 the kernel of this homomorphism is an ideal of $\mathbb{Z}[X]$. It can be seen that this ideal consists of all polynomials over $\mathbb{Z}$ with constant term divisible by three.

---

§7b More on ideals

If R is any ring then the subsets $\{0\}$ and R are ideals. For some rings these are the only ideals; for instance, fields have no ideals other than these trivial ones.

Let I be an ideal of R, and suppose that $a \in I$. Then I contains every element of the form ra for $r \in R$. In particular, if R has an identity and a has an inverse then I contains $t = (ta^{-1})a$ for any $t \in R$. This observation gives us the following theorem:

7.6 Theorem *(i) An ideal which contains an element with an inverse must be the whole ring.*
(ii) If F is a field then the only ideals in F are $\{0\}$ and F.

Proof. The first part is immediate from the preceding remarks, and the second part follows from the first since all nonzero elements of fields have inverses. □

Notation. If R is a commutative ring the set $\{\, ar \mid r \in R \,\}$ will be denoted by 'aR' or 'Ra'.

7.7 Theorem *If R is a commutative ring and $a \in R$ then aR is an ideal of R.*

Proof. Since $a0 \in aR$ it is immediate that $aR \neq \emptyset$. Now let x, $y \in aR$ and $r \in R$. Then $x = as$, $y = at$ for some s, $t \in R$, and hence

$$x + y = as + at = a(s + t) \in aR$$
$$-x = -(as) = a(-s) \in aR$$
$$xr = (as)r = a(sr) \in aR.$$

Since R is commutative we deduce also that $rx = xr \in aR$. By Proposition 5.8 it follows that aR is an ideal. □

Comment ▷▷▷

7.7.1 The above proof uses commutativity of R, and it is impossible to avoid this. If R is not commutative then aR is not necessarily an ideal.

▷▷▷

7.8 DEFINITION Let R be a commutative ring with 1 and let $a \in R$. Then aR is called the *principal ideal generated by* a.

—Examples—

#2 Let $\phi\colon \mathbb{Z} \to \mathbb{Z}_n$ be the homomorphism given by $\phi(a) = \bar{a}$ (see §5b#7). The kernel of ϕ must be an ideal of $\mathbb{Z}$ (by 7.3), and in fact

$$\begin{aligned}\ker\phi &= \{\, a \in \mathbb{Z} \mid \bar{a} = \bar{0} \,\} \\ &= \{\, a \in \mathbb{Z} \mid a \equiv 0 \pmod{n} \,\} \\ &= \{\, a \in \mathbb{Z} \mid a = nk \quad \text{for some } k \in \mathbb{Z} \,\} \\ &= n\mathbb{Z}.\end{aligned}$$

That is, $\ker\phi$ is the principal ideal generated by n.

#3 Define $\sigma\colon \mathbb{R}[X] \to \mathbb{R}$ by $\sigma\big(f(X)\big) = f(3)$ for all $f \in \mathbb{R}[X]$. (That is, in the notation of §6e, $\sigma = e_3$.) Then

$$\begin{aligned}\ker\sigma &= \{\, f(X) \mid f(3) = 0 \,\} \\ &= \{\, f(X) \mid X - 3 \text{ is a factor of } f(X) \,\} \quad \text{(by Theorem 6.11)} \\ &= \{\, (X-3)g(X) \mid g \in \mathbb{R}[X] \,\} \\ &= (X-3)\mathbb{R}[X].\end{aligned}$$

That is, $\ker\sigma$ is the principal ideal generated by $X - 3$.

#4 There are ideals which are not principal. We saw in #1 that the subset I of $\mathbb{Z}[X]$ consisting of those polynomials which have constant term divisible by three is an ideal of $\mathbb{Z}[X]$. However, I is not a principal ideal: there is no $p \in \mathbb{Z}[X]$ such that $I = p(X)\mathbb{Z}[X]$. For suppose that such a p exists. Then since $3 \in I$ we have $3 = p(X)q(X)$ for some $q \in \mathbb{Z}[X]$. By Theorem 6.6

$$\deg(p) + \deg(q) = \deg(3) = 0$$

and so $\deg(p) = \deg(q) = 0$. So p and q are constant polynomials; that is, p and q are elements of $\mathbb{Z}$ (regarding $\mathbb{Z}$ as a subring of $\mathbb{Z}[X]$—see 6.7.1).

Since 3 is the product of p and q and since 3 is prime we see that the only possibilities for p are ± 1 and ± 3. But if $p = \pm 1$ then $p\mathbb{Z}[X] = \mathbb{Z}[X] \neq I$, and if $p = \pm 3$ then $p\mathbb{Z}[X] = 3\mathbb{Z}[X]$ consists of those polynomials for which **every** coefficient is divisible by three, not just the 0^{th}. So our assumption that $p\mathbb{Z}[X] = I$ is contradicted, showing that no such p can exist.

#5 There is a homomorphism $\rho\colon \mathbb{R}[X] \to \mathbb{C}$ given by

$$\rho\big(p(X)\big) = p(\boldsymbol{i}) \qquad \text{for all } p \in \mathbb{R}[X].$$

By Theorem 6.9, for any $p \in \mathbb{R}[X]$ there exists $q \in \mathbb{R}[X]$ and $a, b \in \mathbb{R}$ with

$$p(X) = q(X)\big(X^2 + 1\big) + bX + a,$$

and putting $X = \boldsymbol{i}$ we get $p(\boldsymbol{i}) = b\boldsymbol{i} + a$. So if $p(\boldsymbol{i}) = 0$ we must have both $a = 0$ and $b = 0$, in which case $X^2 + 1$ is a factor of $p(X)$. Thus the kernel of ρ is the set of all p which have $X^2 + 1$ as a factor:

$$\ker \rho = (X^2 + 1)\mathbb{R}[X].$$

---

§7c Congruence modulo an ideal

If n is a positive integer then, as we have seen, the set $n\mathbb{Z}$ of all integers divisible by n is an ideal of $\mathbb{Z}$, and on $\mathbb{Z}$ there is an equivalence relation 'congruence modulo n' given by

$$a \equiv b \pmod{n} \qquad \text{if and only if} \qquad a - b \in n\mathbb{Z}.$$

The same works for any ideal in any ring.

7.9 THEOREM *Let I be an ideal in a ring R, and define a relation on R by*

$$a \equiv b \pmod{I} \qquad \textit{if and only if} \quad a - b \in I.$$

The relation so obtained is an equivalence relation.

Proof. For all $x \in R$ we have $x - x = 0 \in I$, since I is a subring and must therefore contain the zero of R (by 5.2.1). Hence $x \equiv x \pmod{I}$, and the Reflexive Law is satisfied.

Suppose that x, $y \in I$ and $x \equiv y$. Then $x - y \in I$, and so

$$y - x = -(x - y) \in I$$

by Proposition 5.8. Thus $y \equiv x$, and the Symmetric Law is satisfied.

Finally, suppose that x, y, $z \in R$ and $x \equiv y$ and $y \equiv z$. Then $x - y \in I$ and $y - z \in I$, and so

$$x - z = (x - y) + (y - z) \in I$$

by Proposition 5.8. Thus $x \equiv z$, and the Transitive Law holds. □

The relation defined in 7.9 is called *congruence modulo I*. By the results of §4a we see that it partitions R into equivalence classes. These equivalence classes are called the *cosets* of I. Reformulating this slightly gives the following:

7.10 Definition If I is an ideal in the ring R and $a \in R$ then the *coset* of I containing a is the set $I + a = \{\, b \mid b \in R \text{ and } a - b \in I \,\}$.

Comment ▷▷▷

7.10.1 The notation '$I+a$' derives from the fact that the coset containing a is alternatively described as the set $\{\, x + a \mid x \in I \,\}$. ▷▷▷

Occasionally we will use the same notation as we used in §4a for equivalence classes, and write '$\bar{a}$' for '$I + a$'. The advantage of the bar notation is that it is shorter, the disadvantage that it suppresses any mention of I, so that the reader has to remember which ideal is being used.

§7d Quotient rings

If I is an ideal in the ring R define

$$R/I = \{\, I + a \mid a \in R \,\}.$$

In other words, R/I is the set of all equivalence classes of R under the relation of congruence modulo I. That is, in accordance with Definition 4.3, R/I is the quotient of R by this equivalence relation. We wish to define operations of addition and multiplication on R/I to make R/I into a ring. We do this in exactly the same way as we did it for $\mathbb{Z}_n$.

7.11 THEOREM *Let I be an ideal in the ring R. Then there exist well-defined operations of addition and multiplication on R/I such that*

$$(I + a) + (I + b) = I + (a + b)$$
$$(I + a)(I + b) = I + ab$$

for all $a, b \in R$.

Proof. Since every element of R/I has the form $I + a$ for some $a \in R$, the given equations define the sum and product of every pair of elements of R/I. The problem is that since elements of R/I may be expressible in this form in several ways the equations may be inconsistent. We must prove, therefore, that if $I + a' = I + a$ and $I + b' = I + b$ then $I + a'b' = I + ab$ and $I + (a' + b') = I + (a + b)$.

Assume that $I + a' = I + a$ and $I + b' = I + b$. Then $a \equiv a'$ and $b \equiv b' \pmod{I}$, and so $a' = a + x$, $b' = b + y$ for some $x, y \in I$. Now

$$a' + b' = (a + x) + (b + y) = (a + b) + (x + y) \equiv a + b \pmod{I}$$
$$a'b' = (a + x)(b + y) = ab + (xb + ay + xy) \equiv ab \pmod{I}$$

since by 5.8 the fact that $x, y \in I$ gives that $x + y$, xb, ay, xy and hence $xb + ay + xy$ are all in I. Thus $I + (a' + b') = I + (a + b)$ and $I + a'b' = I + ab$, as required. □

Comment ▷▷▷

7.11.1 By these definitions, to add or multiply two cosets one picks elements in the cosets and adds or multiplies the elements. The theorem shows that the coset containing the result is independent of the elements chosen.

◁◁◁

7.12 THEOREM *Let I be an ideal in the ring R. Then R/I is a ring under the operations of addition and multiplication defined in 7.11. Furthermore, the map $\nu: R \to R/I$ defined by $\nu(a) = I + a$ (for all $a \in R$) is a surjective homomorphism.*

Proof. The definitions of addition and multiplication yield immediately that

$$\nu(a)\nu(b) = (I + a)(I + b) = I + ab = \nu(ab)$$
$$\nu(a) + \nu(b) = (I + a) + (I + b) = I + (a + b) = \nu(a + b)$$

and therefore ν is a homomorphism. It is trivial that ν is surjective, since every element of R/I has the form $I + a$ with $a \in R$. Since by Lemma 7.1 the image of ν is a ring we deduce also that R/I is a ring. □

7.13 Definition The map $\nu: R \to R/I$ defined in Theorem 7.12 is called the *natural homomorphism* from R to the quotient ring R/I.

—Examples—

#6 If $R = \mathbb{Z}$ and $I = n\mathbb{Z}$ (where $n \in \mathbb{Z}^+$) then the cosets $n\mathbb{Z} + a \quad (a \in \mathbb{Z})$ are exactly the congruence classes $\bar{a}$ $(a \in \mathbb{Z})$ as defined in §4b, and the quotient ring $\mathbb{Z}/n\mathbb{Z}$ is exactly the same as the ring $\mathbb{Z}_n$.

#7 Let $R = \mathbb{R}[X]$ and $I = (X^2 + 1)\mathbb{R}[X]$. As we have seen (#5 above), for each $p \in \mathbb{R}[X]$ there exists $q \in \mathbb{R}[X]$ and $a, b \in \mathbb{R}$ with

$$p(X) = (X^2 + 1)q(X) + bX + a.$$

This gives

$$p(X) \equiv bX + a \pmod{I}$$

and shows that every equivalence class contains a polynomial of the form $bX + a$. Thus

$$R/I = \{\, I + (bX + a) \mid a,\, b \in \mathbb{R} \,\}.$$

Observe that

$$\begin{aligned}(I + bX + a)(I + dX + c) &= I + (bX + a)(dX + c) \\ &= I + \big(bdX^2 + (ad + bc)X + ac\big) \\ &= I + bd(X^2 + 1) + (ad + bc)X + (ac - bd).\end{aligned}$$

But since

$$bd(X^2 + 1) + (ad + bc)X + (ac - bd)$$

is congruent to

$$(ad + bc)X + (ac - bd)$$

modulo the ideal $I = (X^2 + 1)\mathbb{R}[X]$, this gives

$$(\spadesuit) \qquad \big(I + (a + bX)\big)\big(I + (c + dX)\big) = I + \big((ac - bd) + (ad + bc)X\big).$$

We also have

$$(\heartsuit) \qquad \big(I + (a + bX)\big) + \big(I + (c + dX)\big) = I + \big((a + c) + (b + d)X\big).$$

Comparing ($\spadesuit$) and ($\heartsuit$) with the rules for multiplication and addition of complex numbers,

$$\begin{aligned}(a + b\boldsymbol{i})(c + d\boldsymbol{i}) &= (ac - bd) + (ad + bc)\boldsymbol{i}\\ (a + b\boldsymbol{i}) + (c + d\boldsymbol{i}) &= (a + c) + (b + d)\boldsymbol{i}\end{aligned}$$

we see readily that the ring $R/I = \mathbb{R}[X]/(X^2 + 1)\mathbb{R}[X]$ is isomorphic to $\mathbb{C}$.

From an intuitive point of view the construction of a quotient ring R/I amounts to regarding all elements of I as being equal to zero. As a consequence of this we must regard two elements as equal if they differ by an element of I; that is, if they are in the same coset. This process is sometimes called "factoring out I".

§7e The Fundamental Homomorphism Theorem

7.14 The Fundamental Homomorphism Theorem *Let R and S be rings and let $\theta\colon R \to S$ be a homomorphism. Then*

$$R/\ker\theta \cong \operatorname{im}\theta.$$

Indeed there is an isomorphism

$$\psi\colon R/\ker\theta \longrightarrow \operatorname{im}\theta$$

satisfying

$$\psi(\ker\theta + a) = \theta(a)$$

for all $a \in R$.

Proof. Since every element of $R/\ker\theta$ is expressible in the form $\ker\theta + a$, possibly in more than one way, the formula $\psi(\ker\theta + a) = \theta(a)$ will define a function from R to $R/\ker\theta$ provided that it is consistent with itself.

Thus we must show that if $\ker\theta + a' = \ker\theta + a$ then $\theta(a') = \theta(a)$. But if $\ker\theta + a' = \ker\theta + a$ then $a' - a \in \ker\theta$, and hence, by 5.5,

$$\theta(a') - \theta(a) = \theta(a' - a) = 0$$

giving the result. So ψ is well defined.

If α, $\beta \in R/\ker\theta$ then there exist a, $b \in R$ such that $\alpha = \ker\theta + a$ and $\beta = \ker\theta + b$, and we obtain

$$\begin{aligned}
\psi(\alpha)\psi(\beta) &= \psi(\ker\theta + a)\psi(\ker\theta + b) \\
&= \theta(a)\theta(b) \\
&= \theta(ab) \qquad \text{(since } \theta \text{ preserves multiplication)} \\
&= \psi(\ker\theta + ab) \\
&= \psi\big((\ker\theta + a)(\ker\theta + b)\big) \\
&= \psi(\alpha\beta).
\end{aligned}$$

A similar argument based on the fact that θ preserves addition shows that $\psi(\alpha)+\psi(\beta) = \psi(\alpha+\beta)$. Thus ψ is a homomorphism, and it remains to show that ψ is bijective.

Let $x \in \operatorname{im}\theta$. Then $x = \theta(a)$ for some $a \in R$, and from this it follows that $\psi(\ker\theta+a) = \theta(a) = x$. Hence ψ is a surjective mapping from $R/\ker\theta$ to $\operatorname{im}\theta$. Now suppose that $\psi(\alpha) = \psi(\beta)$ for some α and β in $R/\ker\theta$. Choosing a and b in R with $\alpha = \ker\theta + a$ and $\beta = \ker\theta + b$ we find that

$$\theta(a - b) = \theta(a) - \theta(b) = \psi(\alpha) - \psi(\beta) = 0$$

and therefore $a - b \in \ker\theta$. That is, $a \equiv b \pmod{\ker\theta}$, and

$$\alpha = \ker\theta + a = \ker\theta + b = \beta.$$

Therefore ψ is injective. □

Comment ▷▷▷

7.14.1 If $\theta\colon R \to S$ is a homomorphism with $\ker\theta = I$ then θ maps two elements of R to the same element of $\operatorname{im}\theta \subseteq S$ if and only if the two given elements of R differ by an element of I. Since factoring out I amounts to regarding two elements of R as equal if and only if they differ by an element of I, this means that each element of $\operatorname{im}\theta$ corresponds to just one element of R/I. So the homomorphism $R \to \operatorname{im}\theta$ becomes an isomorphism $R/I \to \operatorname{im}\theta$.

◁◁◁

—Examples—

#8 For any ring R the identity map $\iota\colon R \to R$ (given by $\iota(x) = x$ for all $x \in R$) is a homomorphism. Clearly $\ker \iota = \{0\}$ and $\operatorname{im} \iota = R$, and so the Fundamental Homomorphism Theorem says that $R/\{0\} \cong R$. The isomorphism guaranteed by 7.14 is $\{0\} + a \mapsto \iota(a) = a$.

#9 For any rings R and S the zero map $R \to S$, defined by $x \mapsto 0_S$ (for all $x \in R$), is a homomorphism. Its kernel is the whole of R and its image is the zero subring of S. By 7.14,

$$R/R \;\cong\; \{0\}.$$

(Note that R/R has just one element, since $R + x = R$ for all $x \in R$.)

#10 Let $\rho\colon \mathbb{R}[X] \to \mathbb{C}$ be the homomorphism considered in #5 above, namely

$$\rho\big(p(X)\big) = p(\boldsymbol{i}) \qquad \text{for all} \quad p \in \mathbb{R}[X].$$

Clearly ρ is surjective (since every element of $\mathbb{C}$ is of the form $a + b\boldsymbol{i}$ for some a, $b \in \mathbb{R}$, and $a + b\boldsymbol{i} = \rho\big(a + bX\big)\big)$. So $\operatorname{im} \rho = \mathbb{C}$. As we saw in #5, $\ker \rho = (X^2 + 1)\mathbb{R}[X]$. Hence 7.14 gives

$$\mathbb{R}[X] \Big/ (X^2 + 1)\mathbb{R}[X] \cong \mathbb{C}.$$

Furthermore there is a isomorphism

$$\psi\colon \mathbb{R}[X] \Big/ (X^2 + 1)\mathbb{R}[X] \longrightarrow \mathbb{C}$$

satisfying

$$\psi\big(I + p(X)\big) = \rho\big(p(X)\big) = p(\boldsymbol{i})$$

for all $p \in \mathbb{R}[X]$ (where we have written 'I' for '$(X^2+1)\mathbb{R}[X]$'). In particular this says

$$\psi\big(I + (a + bX)\big) = a + b\boldsymbol{i},$$

in agreement with #7 above.

#11 Let R and S be rings. Recall that the direct sum $R \dot{+} S$ of R and S is the set $\{\,(r, s) \mid r \in R,\ s \in S\,\}$ under componentwise addition and multiplication (see §2c#5).

There is a homomorphism $\eta\colon R \to R \dot{+} S$ given by $\eta(r) = (r, 0)$ for all $r \in R$. Since $\ker \eta = \{0\}$ it follows that $\operatorname{im} \theta \cong R/\{0\} \cong R$ (by #8 above).

The image of θ is the set $R' = \{\,(r,0) \mid r \in R\,\}$; we have thus shown that R' is a subring of $R \dot{+} S$ isomorphic to R. Similarly the set $S' = \{\,(0,s) \mid s \in S\,\}$ is a subring of $R \dot{+} S$ isomorphic to S.

Now define $\pi\colon R \dot{+} S \to R$ by $\pi(r,s) = r$. It is easily seen that π is a homomorphism and that $\operatorname{im}\pi = R$ and $\ker\pi = \{\,(0,s) \mid s \in S\,\} = S'$. So in fact S' is an ideal of $R \dot{+} S$, and $R \dot{+} S/S' \cong R$. Similarly, R' is an ideal and $R \dot{+} S/R' \cong S$.

The isomorphism $\psi\colon R \dot{+} S/S' \longrightarrow R$ given by 7.14 can be described explicitly in the following way. If $(r,s) \in R \dot{+} S$ then the coset $S' + (r,s)$ is

$$\begin{aligned}
&\{\,(x,y) \mid x \in R,\ y \in S \text{ and } (x,y) - (r,s) \in S'\,\}\\
&= \{\,(x,y) \mid x \in R,\ y \in S \text{ and } (x-r, y-s) \in S'\,\}\\
&= \{\,(x,y) \mid x \in R,\ y \in S \text{ and } x - r = 0\,\}\\
&= \{\,(r,y) \mid y \in S\,\}.
\end{aligned}$$

By 7.14 we have $\psi\big(S' + (r,s)\big) = \pi(r,s)$; that is,

$$\psi\Big(\{\,(r,y) \mid y \in S\,\}\Big) = r.$$

In other words, what we have shown is this:

> For each $r \in R$ there is a coset of S' consisting of all ordered pairs (x,y) in $R \dot{+} S$ such that the first component, x, is equal to r. This gives a one-to-one correspondence between elements of R and cosets of S'; that is, between elements of R and elements of $R \dot{+} S/S'$. This correspondence is an isomorphism.

Exercises

1. (*i*) Prove that $\mathbb{Q}[\sqrt[3]{2}]$ as defined in Exercise 11 of Chapter Five is a subfield of $\mathbb{R}$.

(Hint: By Exercise 11 of Chapter Five, $\mathbb{Q}[\sqrt[3]{2}]$ is an integral domain. To prove that all nonzero elements of $\mathbb{Q}[\sqrt[3]{2}]$ have

inverses in $\mathbb{Q}[\sqrt[3]{2}]$ use §6k#15 of Chapter 6 and prove the formula

$$\left(a + b\sqrt[3]{2} + c\left(\sqrt[3]{2}\right)^2\right)^{-1} = (e/d) + (f/d)\sqrt[3]{2} + (g/d)\left(\sqrt[3]{2}\right)^2$$

where

$$\begin{aligned} d &= a^3 + 2b^3 + 4c^3 - 6abc \\ e &= a^2 - 2bc \\ f &= 2c^2 - ab \\ g &= b^2 - ac \end{aligned}$$

for inverses of elements of $\mathbb{Q}[\sqrt[3]{2}]$.)

(*ii*) Prove that $\mathbb{Q}[\sqrt[3]{2}]$ is isomorphic to the quotient ring $\mathbb{Q}[X]/K$, where K is the set of all $p(X) \in \mathbb{Q}[X]$ such that $p(\sqrt[3]{2}) = 0$.

(Hint: Use the Fundamental Homomorphism Theorem and the evaluation homomorphism $p(X) \mapsto p(\sqrt[3]{2})$.)

2. Calculate the kernel and image of the homomorphism ψ in Exercise 13 of Chapter Six.

3. Prove Theorem 7.5.

4. Let R be a commutative ring with 1. For $x, y \in R$ let '$x|y$' mean 'there exists $z \in R$ with $y = xz$'. Prove that if $a, b \in R$ then $aR = bR$ if and only if $a|b$ and $b|a$.

5. Let θ be the homomorphism defined in Exercise 13 of Chapter Five:

$$\theta \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & f & g \end{pmatrix} = \begin{pmatrix} d & e \\ f & g \end{pmatrix}$$

where the domain R of θ is a subring of $\mathrm{Mat}(3, \mathbb{Z})$ and the codomain is $\mathrm{Mat}(2, \mathbb{Z})$. Prove that the kernel of θ is equal to the set I of all matrices of the form $\begin{pmatrix} a & b & c \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ with $a, b, c \in \mathbb{Z}$. Deduce that $R/I \cong \mathrm{Mat}(2, \mathbb{Z})$, and give an explicit isomorphism.

(Hint: R/I is the set of equivalence classes of the relation $\sim$ defined in Exercise 13 of Chapter Five.)

6. In each case show that I is an ideal of the given ring R, and find a homomorphism which has kernel equal to I:

(*i*) $R = \mathbb{Z}$, $I = 5\mathbb{Z}$

(*ii*) $R = \left\{ \begin{pmatrix} a & 0 \\ b & c \end{pmatrix} \middle| \, a, b, c \in \mathbb{R} \right\}$, $I = \left\{ \begin{pmatrix} a & 0 \\ b & 0 \end{pmatrix} \middle| \, a, b \in \mathbb{R} \right\}$

(*iii*) $R = \mathbb{Z}[X]$
$I = \{ a_0 + a_1X + \cdots + a_nX^n \mid a_i \in \mathbb{Z}, \ \sum_{i=0}^{n} a_i = 0 \}$

(*iv*) $R = \mathbb{Z}[X]$
$I = \{ a_0 + a_1X + \cdots + a_nX^n \mid a_i \in \mathbb{Z}, \ \sum_{i=0}^{n} a_i \equiv 0 \pmod 2 \}$

(*v*) $R = \mathbb{Z}[X]$
$I = \{ a_0 + a_1X + \cdots + a_nX^n \mid a_i \in \mathbb{Z}, \ \sum_{i=0}^{n} (-1)^i a_i \equiv 0 \pmod 3 \}$.

7. Let I be an ideal and S a subring in the ring R.

(*i*) Show that $S + I = \{ s + x \mid s \in S, \ x \in I \}$ is a subring of R.

(*ii*) Show that I is an ideal in $S + I$.

(*iii*) Show that $\theta \colon S \to (S + I)/I$ given by $\theta(s) = I + s$ is a homomorphism.

(*iv*) Deduce that $S \cap I$ is an ideal in S and $S/(S \cap I)$ is isomorphic to $(S + I)/I$. (Hint: Use 7.14 and 7.3.)

(*v*) Prove directly that $S \cap I$ is an ideal in S.

8. Prove that if $A = n\mathbb{Z}$ and $B = m\mathbb{Z}$ then $A + B = d\mathbb{Z}$ and $A \cap B = l\mathbb{Z}$, where $d = \gcd(n, m)$ and $l = \operatorname{lcm}(n, m)$ and $A + B$ is as defined in the previous exercise.

8
Field Extensions

In §7d#7 we saw that the polynomial ring $\mathbb{R}[X]$ has a quotient ring isomorphic to the field $\mathbb{C}$ of complex numbers. This is an example of a phenomenon we wish to study in more detail, as a method of constructing fields containing a given field F as a subfield. The fields to be constructed will be quotient rings of $F[X]$. Our first step in this program is to study ideals in $F[X]$.

§8a Ideals in polynomial rings

From now on we will only be concerned with polynomials over fields.

8.1 Theorem *Let F be a field and let I be an ideal of $F[X]$. Then there exists a polynomial $f(X)$ such that $I = f(X)F[X]$.*

Proof. By 5.2.1 we know that $0 \in I$. If 0 is the only element of I then the assertion of the theorem holds with $f(X) = 0$. Thus we may assume that I contains nonzero elements.

Of all nonzero elements of I choose $f(X)$ to be one of minimal degree,† and let $J = f(X)F[X]$. If $p(X) \in J$ then $p(X) = q(X)f(X)$ for some q, and, since $f(X) \in I$, 5.8 (iv) yields that $p(X) \in I$. Thus $J \subseteq I$. Conversely, let $p(X) \in I$, and let $r(X)$ be the remainder on dividing $p(X)$ by $f(X)$. Then for some polynomial q we have $r(X) = p(X) - q(X)f(X)$, and since $p(X)$ and $f(X)$ are both in I we deduce (by 5.8) that $r(X) \in I$. By the choice of $f(X)$ we know therefore that the degree of $r(X)$ cannot be less than the degree of $f(X)$; hence by Theorem 6.9 it follows that $r(X) = 0$. Thus $p(X) = f(X)q(X) \in f(X)F[X] = J$, and we conclude that $I \subseteq J$. Hence $I = J$, as required. □

† Note the use of the Least Integer Principle in this step

Comment ▷▷▷

8.1.1 This says that all ideals of $F[X]$ are principal. Furthermore, in the above proof we have in fact shown that a nonzero ideal in $F[X]$ is generated by any nonzero element of minimal degree contained in it. ▷▷▷

As a corollary of 8.1 we obtain the following proposition:

8.2 PROPOSITION *Let I be an ideal in $F[X]$ with $I \neq F[X]$, and suppose that I contains an irreducible polynomial $p(X)$. Then $I = p(X)F[X]$.*

Proof. By Theorem 8.1 there exists $f(X) \in F[X]$ with $I = f(X)F[X]$. Since $p(X) \in I$ it follows that $f(X)|p(X)$. Since $p(X)$ is irreducible $f(X)$ must be either an associate of $p(X)$ or of degree zero. But if $\deg(f) = 0$ then by 7.6 (i) we obtain $I = F[X]$, contrary to hypothesis. So f and p are associates, and therefore $f(X)F[X] = p(X)F[X]$ (by Exercise 4 of Chapter Seven). □

§8b Quotient rings of polynomial rings

Continuing with the notation of 8.1, let $I = f(X)F[X]$. We wish to investigate the ring $Q = F[X]/I$. For simplicity we will use the bar notation for cosets: $\overline{g(X)} = I + g(X)$ for all $g \in F[X]$.

8.3 THEOREM *Suppose that $f(X) = c_0 + c_1X + \cdots + c_nX^n$, where $n \geq 1$, $c_i \in F$ for each i, and $c_n \neq 0$. Then we have the following:*

(i) Each element of $Q = F[X]/I$ is uniquely expressible in the form

$$\overline{a_0 + a_1X + \cdots + a_{n-1}X^{n-1}}$$

with $a_0, a_1, \ldots, a_{n-1} \in F$.

(ii) The set $\overline{F} = \{\, \overline{a} \mid a \in F \,\}$ is a subring of $F[X]/I$ isomorphic to F.

(iii) The element $\overline{X}$ of Q satisfies the equation

$$\overline{c_0} + \overline{c_1}\overline{X} + \cdots + \overline{c_n}\overline{X}^n = \bar{0}.$$

Proof. (i) An arbitrary element of Q is a coset of I, and hence equal to $\overline{g(X)}$ for some polynomial $g \in F[X]$. By Theorem 6.9

$$(*) \qquad g(X) = q(X)f(X) + (a_0 + a_1X + \cdots + a_{n-1}X^{n-1})$$

for uniquely determined $a_0, a_1, \ldots, a_{n-1} \in F$. Since I is the set of all polynomials of the form $q(X)f(X)$ it follows that equation $(*)$ is equivalent to $g(X) \equiv a_0 + a_1X + \cdots + a_{n-1}X^{n-1} \pmod{I}$, and hence to

$$(**) \qquad \overline{g(X)} = \overline{a_0 + a_1X + \cdots + a_{n-1}X^{n-1}}.$$

So $(**)$ holds for unique a_i, as required.

(ii) Define a mapping $\theta\colon F \to F[X]/I$ by $\theta(a) = \overline{a}$. Then θ is a homomorphism, since it is the restriction to the subring F of $F[X]$ of the natural homomorphism $F[X] \to Q$. (See 7.12 and 5.5.2.) If $a \in F$ is in $\ker\theta$ then $\overline{a} = \overline{0}$, and since $a \in F$ it follows from (i) that $a = 0$. (Alternatively, $\overline{a} = \overline{0}$ means that $a \in I$, and hence a is divisible by $f(X)$. Since a is a constant and $\deg(f) \geq 1$ we must have $a = 0$.) Thus $\ker\theta = \{0\}$, and since $\operatorname{im}\theta = \{\,\theta(a) \mid a \in F\,\} = \overline{F}$ Theorem 7.14 gives

$$\overline{F} \cong F/\ker\theta \cong F$$

in view of §7e#8.

(iii) By the definition of addition and multiplication in a quotient ring,

$$\overline{r(X)} + \overline{s(X)} = \overline{r(X) + s(X)}$$
$$\overline{r(X)}\,\overline{s(X)} = \overline{r(X)s(X)}$$

for all $r(X), s(X) \in F[X]$. Hence

$$\overline{c_0} + \overline{c_1}\overline{X} + \cdots + \overline{c_n}\overline{X}^n = \overline{c_0 + c_1X + \cdots + c_nX^n} = \overline{f(X)}$$

which is equal to $\overline{0}$ since $f(X) \in I$. □

Comment ▷▷▷

8.3.1 Part (ii) of 8.3 permits us to regard F as a subring of Q, in the same way as we have identified F with the set of constant polynomials in $F[X]$. That is, we identify $\bar{a}$ with a for each $a \in F$. Thus Q is a ring which contains F as a subring and also contains an element $\overline{X}$ satisfying $c_0 + c_1\overline{X} + \cdots + c_n\overline{X}^n = 0$. That is, $\overline{X} \in Q$ is a zero of the polynomial $f(Y) = c_0 + c_1Y + \cdots + c_nY^n \in Q[Y]$. Furthermore, by (i) of 8.3, every element of Q is of the form $a_0 + a_1\overline{X} + \cdots + a_{n-1}\overline{X}^{n-1}$ with coefficients $a_i \in F$. Hence $Q = F[X]/f(X)F[X]$ can be regarded as a ring obtained from F by adjoining to F a new element $\overline{X}$ which is to be a zero of f.

These remarks should be compared with the remarks in 6.7.1. The polynomial ring $F[X]$ is a ring obtained from F by adjoining an element X which satisfies no nontrivial equations. Now Q is obtained by adjoining to F an element $\overline{X}$ satisfying the equation $c_0+c_1\overline{X}+\cdots+c_n\overline{X}^n=0$. Furthermore we know by 7.12 that there is a homomorphism

$$F[X] \to Q = F[X]/I$$

such that $X \mapsto \overline{X}$, and, in general,

$$a_0 + a_1X + \cdots + a_rX^r \mapsto a_0 + a_1\overline{X} + \cdots + a_r\overline{X}^r$$

for any polynomial $a_0 + a_1X + \cdots + a_rX^r \in F[X]$. This is reasonable, since factoring I out of $F[X]$ amounts to regarding elements of I, in particular the element $f(X) = c_0 + c_1X + \cdots + c_nX^n$, as being equal to 0. ▷▷▷

To complete our discussion of quotient rings of $F[X]$ it remains to say what happens when $p(X)$ is a constant polynomial. There are two cases, both trivial.

8.4 Theorem *Let $p(X) = a \in F$, and let $I = p(X)F[X]$.*
(i) If $a = 0$ then $I = \{0\}$ and $F[X]/I \cong F[X]$.
(ii) If $a \neq 0$ then $I = F[X]$ and $F[X]/I \cong \{0\}$.

Proof. Part (i) is immediate from §7e#8, and, in view of 7.6 (i), Part (ii) is immediate from §7e#9. □

—Examples—

#1 Let $R = \mathbb{Q}[X]/(X^2-3)\mathbb{Q}[X]$. We use the bar notation again: if $g(X) \in \mathbb{Q}[X]$ then $\overline{g(X)}$ denotes the element $(X^2-3)\mathbb{Q}[X] + g(X)$ of the ring R. By the discussion in 8.3.1 we know that R can be thought of as the result of adjoining to $\mathbb{Q}$ an element $\overline{X}$ satisfying $\overline{X}^2 - 3 = 0$. Every element of R will have the form $a + b\overline{X}$ for some a, $b \in \mathbb{Q}$, and the following rules hold for addition and multiplication in R:

$$\begin{aligned}(a + b\overline{X}) + (c + d\overline{X}) &= (a + c) + (b + d)\overline{X}\\ (a + b\overline{X})(c + d\overline{X}) &= ac + (ad + bc)\overline{X} + bd\overline{X}^2\\ &= (ac + 3bd) + (ad + bc)\overline{X}\end{aligned}$$

since $\overline{X}^2 = 3$.

There is another way to adjoin to $\mathbb{Q}$ an element whose square is three; namely, consider the set $\mathbb{Q}[\sqrt{3}]$ of all real numbers of the form $a + b\sqrt{3}$ with a and b in $\mathbb{Q}$. We saw in §5a#6 that $\mathbb{Q}[\sqrt{3}]$ is a subfield of $\mathbb{R}$. The above considerations suggest that this subfield of $\mathbb{R}$ ought to be isomorphic to $\mathbb{Q}[X]/(X^2 - 3)\mathbb{Q}[X]$. It is easy to prove this by using the Fundamental Homomorphism Theorem.

Define $\theta\colon \mathbb{Q}[X] \to \mathbb{R}$ by $\theta\big(g(X)\big) = g(\sqrt{3})$ for all $g \in \mathbb{Q}[X]$. Then θ is a homomorphism. Since $(\sqrt{3})^i$ is in $\mathbb{Q}$ if i is even and of the form $q\sqrt{3}$ with $q \in \mathbb{Q}$ if i is odd, we have

$$\begin{aligned}\operatorname{im}\theta &= \{\, a_0 + a_1\sqrt{3} + a_2(\sqrt{3})^2 + \cdots + a_n(\sqrt{3})^n \mid 0 \le n \in \mathbb{Z},\ a_i \in \mathbb{Q} \,\} \\ &= \{\, a + b\sqrt{3} \mid a,\ b \in \mathbb{Q} \,\} \\ &= \mathbb{Q}[\sqrt{3}].\end{aligned}$$

Moreover, since (by 6.9) any element of $\mathbb{Q}[X]$ is expressible in the form $(X^2 - 3)q(X) + (a + bX)$ with $q(X) \in \mathbb{Q}[X]$ and $a,\ b \in \mathbb{Q}$, it follows that

$$\begin{aligned}\ker\theta &= \{\, g \in \mathbb{Q}[X] \mid g(\sqrt{3}) = 0 \,\} \\ &= \{\, (X^2 - 3)q(X) + a + bX \mid q \in \mathbb{Q}[X],\ \ a,\ b \in \mathbb{Q},\ a + b\sqrt{3} = 0 \,\} \\ &= \{\, (X^2 - 3)q(X) \mid q \in \mathbb{Q}[X] \,\} \\ &= (X^2 - 3)\mathbb{Q}[X].\end{aligned}$$

By 7.14,

$$\mathbb{Q}[X]\Big/(X^2 - 3)\mathbb{Q}[X] \cong \mathbb{Q}[\sqrt{3}]$$

and there is an isomorphism satisfying

$$\overline{a + bX} = (X^2 - 3)\mathbb{Q}[X] + (a + bX) \longmapsto a + b\sqrt{3}$$

for all $a,\ b \in Q$.

#2 Prove that $\mathbb{Q}[\sqrt[3]{2}] = \{\, a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \mid a,\ b,\ c \in \mathbb{Q} \,\}$ is isomorphic to $\mathbb{Q}[X]/(X^3 - 2)\mathbb{Q}[X]$.

$\gg\!\!\rightarrow$ Define $\phi\colon \mathbb{Q}[X] \to \mathbb{R}$ by $\phi(p(X)) = p(\sqrt[3]{2})$. By Theorem 6.8 we know that ϕ is a homomorphism. Now any rational linear combination of powers of $\sqrt[3]{2}$ lies in $\mathbb{Q}[\sqrt[3]{2}]$, since $(\sqrt[3]{2})^i$ is in $\mathbb{Q}$ if $i \equiv 0 \pmod 3$, of the form $q\sqrt[3]{2}$

with $q \in \mathbb{Q}$ if $i \equiv 1 \pmod 3$ and of the form $q(\sqrt[3]{2})^2$ with $q \in \mathbb{Q}$ if $i \equiv 2 \pmod 3$. Thus

$$\begin{aligned}\operatorname{im}\phi &= \{\, a_0 + a_1\sqrt[3]{2} + \cdots + a_n(\sqrt[3]{2})^n \mid 0 \le n \in \mathbb{Z}\ a_i \in \mathbb{Q}\,\} \\ &= \{\, a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \mid a,\ b,\ c \in \mathbb{Q}\,\} \\ &= \mathbb{Q}[\sqrt[3]{2}].\end{aligned}$$

By Theorem 7.14 it follows that $\mathbb{Q}[\sqrt[3]{2}] \cong \mathbb{Q}[X]/\ker\phi$, and it remains for us to prove that $\ker\phi = (X^3 - 2)\mathbb{Q}[X]$. Now certainly $X^3 - 2 \in \ker\phi$, since $\phi(X^3-2) = (\sqrt[3]{2})^3 - 2 = 0$. But since $X^3 - 2$ is irreducible in $\mathbb{Q}[X]$ (by Eisenstein's Criterion—see §6k#14), it follows from 8.2 that $\ker\theta = (X^3-2)\mathbb{Q}[X]$, as required. ⊲⋘

#3 Let $Q = \mathbb{R}[X]/X^3\mathbb{R}[X]$. Then by 8.3 the set

$$\mathbb{R}' = \{\, X^3\mathbb{R}[X] + t \mid t \in \mathbb{R}\,\}$$

is a subring of Q isomorphic to $\mathbb{R}$, and $\alpha = X^3\mathbb{R}[X] + X$ is an element of Q satisfying $\alpha^3 = 0$. Every element of Q is uniquely expressible in the form $t_0 + t_1\alpha + t_2\alpha^2$ with $t_0,\ t_1,\ t_2 \in \mathbb{R}'$. The rule for addition in Q is obvious:

$$(s_0 + s_1\alpha + s_2\alpha^2) + (t_0 + t_1\alpha + t_2\alpha^2) = (s_0 + t_0) + (s_1 + t_1)\alpha + (s_2 + t_2)\alpha^2$$

for all $s_i,\ t_i \in \mathbb{R}$ $(i = 0,\ 1,\ 2)$. To multiply two elements of Q, simply expand the product and use $\alpha^3 = 0$:

$$(s_0+s_1\alpha+s_2\alpha^2)(t_0+t_1\alpha+t_2\alpha^2) = s_0t_0+(s_0t_1+s_1t_0)\alpha+(s_0t_2+s_1t_1+s_2t_0)\alpha^2.$$

#4 Let $T = \mathbb{R}[X]/X\,\mathbb{R}[X]$. Using the bar notation again, we have

$$\overline{X} = \overline{0} = \text{zero element of } T,$$

since X is in the ideal $X\,\mathbb{R}[X]$. (In factoring out $X\,\mathbb{R}[X]$ all elements of $X\,\mathbb{R}[X]$ are regarded as being zero, and this includes the element X.) So in accordance with 8.3 the ring T can be thought of as obtained from $\mathbb{R}$ by adjoining to $\mathbb{R}$ an element $\overline{X}$ satisfying $\overline{X} = 0$.

Adjoining 0 to $\mathbb{R}$ doesn't do a lot—0 is already an element of $\mathbb{R}$. So we should have that $T \cong \mathbb{R}$. Again this can be proved using 7.14. There is a

homomorphism $\phi\colon \mathbb{R}[X] \to \mathbb{R}$ given by $\phi\big(p(X)\big) = p(0)$ for all $p \in \mathbb{R}[X]$; that is,

$$\phi(a_0 + a_1X + \cdots + a_nX^n) = a_0.$$

Clearly $\operatorname{im}\phi = \mathbb{R}$ and $\ker\phi = X\,\mathbb{R}[X]$; so 7.14gives $\mathbb{R}[X]/X\,\mathbb{R}[X] \cong \mathbb{R}$. This can also be seen directly by observing that

$$a_0 + a_1X + \cdots + a_nX^n \equiv a_0 \pmod{X\,\mathbb{R}[X]},$$

and hence that every element of $\mathbb{R}[X]/X\,\mathbb{R}[X]$ is equal to $\bar{a}$ for some $a \in \mathbb{R}$.

§8c Fields as quotient rings of polynomial rings

8.5 Definition If F is a subfield of a field E then E is called an *extension* of F. More generally, if E has a subfield isomorphic to F we say that E is an extension of F.

We have seen that $\mathbb{R}[X]/(X^2+1)\mathbb{R}[X]$ is a field (isomorphic to $\mathbb{C}$) containing $\mathbb{R}$ as a subfield, and that $\mathbb{Q}[X]/(X^2-3)\mathbb{Q}[X]$ is a field (isomorphic to $\mathbb{Q}[\sqrt{3}]$) containing $\mathbb{Q}$ as a subfield. However, $\mathbb{R}[X]/X^3\mathbb{R}[X]$ is not a field, since it contains zero divisors. (The element $\alpha = X^3\mathbb{R}[X] + X$ is nonzero but satisfies $\alpha^3 = 0$, as we saw in #3.) We are led to wonder under what circumstances a quotient ring of $F[X]$ is a field.

Analogy with quotient rings of $\mathbb{Z}$ provides a clue to the answer. We have seen (Theorems 4.10 and 4.11) that $\mathbb{Z}/n\mathbb{Z}$ is a field if and only if n is prime; that is, the quotient is a field if and only if the ideal is generated by a prime. Exactly the same is true for quotients of $F[X]$.

8.6 Theorem *Let F be a field and $p(X) \in F[X]$. Then $F[X]/p(X)F[X]$ is a field if and only if p is irreducible.*

Proof. Let $I = F[X]/p(X)F[X]$, and suppose first that $p(X)$ is not irreducible. Then either $p(X)$ is a constant or else $\deg(p) > 1$ and $p(X)$ has a factorization $p(X) = s(X)t(X)$ for some s and t of degree less than $\deg(p)$.

If $p(X)$ is a constant then by 8.4 $F[X]/I$ is isomorphic either to $F[X]$, which is not a field since polynomials of degree greater than 1 do not have inverses in $F[X]$, or to the trivial ring $\{0\}$, which is not a field since it does not

have a nonzero identity element. We are left with the case $p(X) = s(X)t(X)$ with $1 \leq \deg(s) < \deg(p)$ and $1 \leq \deg(t) < \deg(p)$. Now since $p(X) \in I$ we find that

$$(I + s(X))(I + t(X)) = I + s(X)t(X) = I + p(X) = I = I + 0$$

the zero element of $F[X]/I$. Moreover, $p(X)$ cannot be a factor of $s(X)$ or $t(X)$ (since $1 \leq \deg(s) < \deg(p)$ and $1 \leq \deg(t) < \deg(p)$), and therefore $I + s(X) \neq I$ and $I + t(X) \neq I$. So the ring $F[X]/I$ has zero divisors, hence is not an integral domain, hence is not a field.

Suppose, on the other hand, that $p(X)$ is irreducible. We must prove that $Q = F[X]/I$ is a field. Since it is certainly a ring, it suffices to prove that it is commutative and has a nonzero identity, and that all nonzero elements have inverses.

Let $\alpha, \beta \in Q$. Then $\alpha = I + f(X)$, $\beta = I + g(X)$ for some $f(X)$ and $g(X)$ in $F[X]$, and

$$\begin{aligned}\alpha\beta = (I + f(X)(I + g(X)) &= I + f(X)g(X) \\ &= I + g(X)f(X) = (I + g(X))(I + f(X)) = \beta\alpha\end{aligned}$$

by commutativity of $F[X]$. Furthermore,

$$\begin{aligned}\alpha(I + 1) = (I + f(X))(I + 1) &= I + f(X)1 \\ &= I + 1f(X) = (I + 1)(I + f(X)) = (I + 1)\alpha\end{aligned}$$

where 1 is the identity of F. Thus Q is commutative and has an identity. The identity is nonzero since $I + 1 = I$ would imply that $1 \in I$ and hence that $p(X)|1$, which is impossible since $\deg(p) \neq 0$.

Let α be a nonzero element of Q, and let $f(X)$ be an element of $F[X]$ such that $\alpha = I + f(X)$. Then $f(X) \not\equiv 0 \pmod{I}$, since

$$I + f(X) \neq I = \text{zero element of } Q,$$

and so $p(X) \not| f(X)$. Thus the gcd of $p(X)$ and $f(X)$ cannot be an associate of $p(X)$, and since $p(X)$ is irreducible the only other divisors it has are polynomials of degree 0. So the gcd of $p(X)$ and $f(X)$ must be 1. Now by 6.14 there exist $m(X)$ and $n(X)$ with $m(X)p(X) + n(X)f(X) = 1$, and this gives

$$\begin{aligned}\alpha(I + n(X)) &= (I + f(X))(I + n(X)) \\ &= I + f(X)n(X) \\ &= I + (1 - m(X)p(X)) \\ &= I + 1\end{aligned}$$

since $m(X)p(X) \in I$. Thus α has an inverse, as required. □

—Examples—

#5 Let $F = \mathbb{R}$ and $p(X) = X^2 - 3X + 2 = (X-1)(X-2)$. Then $p(X)$ is not irreducible, and so $Q = \mathbb{R}[X]/p(X)\mathbb{R}[X]$ is not a field. Indeed,

$$\overline{X-1} = p(X)\mathbb{R}[X] + (X-1)$$

and

$$\overline{X-2} = p(X)\mathbb{R}[X] + (X-2)$$

are nonzero elements of Q with product zero:

$$\begin{aligned}\big(\overline{X-1}\big)\big(\overline{X-2}\big) &= \overline{X^2 - 3X + 2} \\ &= \overline{0},\end{aligned}$$

since $X^2 - 3X + 2 \equiv 0 \pmod{p(X)\mathbb{R}[X]}$.

#6 Let $F = \mathbb{Q}$ and $p(X) = X^2 - 3$. We have seen that $X^2 - 3$ is irreducible in $\mathbb{Q}[X]$ (§6i#6), and so $\mathbb{Q}[X]/(X^2-3)\mathbb{Q}[X]$ is a field. We had noted this already in #1 above.

#7 The polynomial $X^2 + 1$ is irreducible in $\mathbb{R}[X]$, hence the quotient ring $\mathbb{R}[X]/(X^2+1)\mathbb{R}[X]$ is a field (as we had seen in §7d#7).

#8 We saw in §6j#12 that $X^3 + X + 1$ is an irreducible polynomial in $\mathbb{Z}_2[X]$. Hence if $I = p(X)\mathbb{Z}_2[X]$ we have that $K = \mathbb{Z}_2[X]/I$ is a field containing $\mathbb{Z}_2$ as a subfield. By Theorem 8.3 each element of K is uniquely expressible in the form $a_0 + a_1\overline{X} + a_2\overline{X}^2$ with $a_0, a_1, a_2 \in \mathbb{Z}_2$ (where $\overline{X} = I + X$). Since there are exactly two choices (0 or 1) for each of a_0, a_1 and a_2, there are exactly eight possible expressions $a_0 + a_1\overline{X} + a_2\overline{X}^2$. So K is a field with eight elements.

---

§8d Field extensions and vector spaces

Let F be a field and $p(X)$ a polynomial over F, of degree $n \geq 1$. Let $Q = F[X]/I$, where $I = p(X)F[X]$. By Theorem 8.3 each element of Q is uniquely expressible in the form $a_0 + a_1\overline{X} + \cdots + a_{n-1}\overline{X}^{n-1}$ where the coefficients $a_0, a_1, \ldots, a_{n-1}$ are elements of F and $\overline{X} = I + X$. We deduce the following lemma:

8.7 LEMMA *The elements* $1, \overline{X}, \ldots, \overline{X}^{n-1}$ *form a basis for* Q *considered as a vector space over* F.

Proof. The proof that Q is a vector space over F is a straightforward checking of Property $(*)$ and Axioms (i)–(viii) listed in §0c, and is omitted. As observed above, every element of Q can be expressed as a linear combination of the given elements; that is, they span Q. Suppose that $a_0 + a_1\overline{X} + \cdots + a_{n-1}\overline{X}^{n-1} = 0$ with the $a_i \in F$. Then

$$a_0 + a_1\overline{X} + \cdots + a_{n-1}\overline{X}^{n-1} = 0 + 0\overline{X} + \cdots + 0\overline{X}^{n-1}$$

and by the uniqueness part of 8.3 (i) it follows that all the a_i are equal to 0. This proves linear independence. □

As a consequence of the lemma we have the following:

8.8 THEOREM *Let* $Q = F[X]/p(X)F[X]$ *where* F *is a field and* p *a polynomial over* F *of degree* $n \geq 1$. *Then* Q *is a vector space over* F, *and the dimension of this vector space is* n.

In the situation described above, if p is irreducible then Q is an extension field of F. In general, if E is any extension of a field F then $(*)$ and (i)–(viii) of §0c are satisfied, and so E may be regarded as a vector space over F. The dimension of this vector space is called the *degree* of the extension.

8.9 DEFINITION If E is an extension field of F the *degree* of E over F, denoted by '$[E : F]$', is the dimension of E as a vector space over F.

Comments ▷▷▷

8.9.1 There is no guarantee that $[E : F]$ is finite.

8.9.2 If $F[X]/p(X)F[X]$ is a field then its degree over F equals $\deg(p)$. ▷▷▷

§8e Extensions of extensions

8.10 THEOREM *Suppose that* F, E, K *are fields with* $F \subseteq E \subseteq K$, *and suppose that* $[K : E] = m$ *and* $[E : F] = n$. *Then* $[K : F] = mn$.

Proof. Let $x_1, x_2, \ldots, x_m$ be a basis for K over E and let $y_1, y_2, \ldots, y_n$ be a basis for E over F. We show that

8.10.1 $$x_1y_1, x_1y_2, \ldots, x_1y_n,\ x_2y_1, \ldots, x_2y_n, \ldots\ \ \ldots, x_my_1, \ldots, x_my_n$$

is a basis for K over F.

We prove first that the elements 8.10.1 span K over F. Let $t \in K$. Then since $\{x_1, \ldots, x_m\}$ spans K over E there exist $s_1, s_2, \ldots, s_m \in E$ with

$$t = s_1x_1 + s_2x_2 + \cdots + s_mx_m.$$

Now each s_i is an F-linear combination of $y_1, y_2, \ldots, y_n$, since these elements span E over F. So we have

$$\begin{aligned} s_1 &= u_{11}y_1 + u_{12}y_2 + \cdots + u_{1n}y_n \\ s_2 &= u_{21}y_1 + u_{22}y_2 + \cdots + u_{2n}y_n \\ &\vdots \\ s_m &= u_{m1}y_1 + u_{m2}y_2 + \cdots + u_{mn}y_n \end{aligned}$$

with the coefficients u_{ij} in F for all i and j. Now substituting gives

$$t = u_{11}y_1x_1 + u_{12}y_2x_1 + \cdots + u_{1n}y_nx_1 + \cdots\cdots + u_{mn}y_nx_m,$$

an F-linear combination of the elements 8.10.1.

Now we must show that the elements 8.10.1 are linearly independent over F. Suppose that u_{ij} $(1 \leq i \leq m,\ 1 \leq j \leq n)$ are elements of F such that $\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}y_jx_i = 0$. Then

$$\Big(\sum_{j=1}^{n} u_{1j}y_j\Big)x_1 + \Big(\sum_{j=1}^{n} u_{2j}y_j\Big)x_2 + \cdots + \Big(\sum_{j=1}^{n} u_{mj}y_j\Big)x_m = 0,$$

and since each coefficient $\sum_{j=1}^{n} u_{ij}y_j$ is an element of E and $x_1, x_2, \ldots, x_m$ are linearly independent over E we deduce that $\sum_{j=1}^{n} u_{ij}y_j = 0$ for each i. Now the linear independence over F of $y_1, y_2, \ldots, y_n$ gives $u_{ij} = 0$ for all i and j, as required. □

§8f Algebraic and transcendental elements

Let E be an extension field of F and let $a \in E$. Any subring of E containing F and containing a clearly must contain a^2, a^3, a^4, ..., and hence must contain everything of the form $b_0 + b_1 a + \cdots + b_n a^n$ with $0 \leq n \in \mathbb{Z}$ and $b_0, b_1, \ldots, b_n \in F$. That is, it must contain everything of the form $f(a)$ for $f(X) \in F[X]$.

8.11 DEFINITION Let $F[a]$ be the subset of E defined by

$$F[a] = \{ f(a) \mid f(X) \in F[X] \}.$$

Any subfield S of E which contains F and also a certainly contains $F[a]$ (since S is also a subring). So if $u, v \in F[a]$ and $v \neq 0$ then it follows that $uv^{-1} \in S$.

8.12 DEFINITION Let $F(a) = \{ uv^{-1} \mid u, v \in F[a] \text{ and } v \neq 0 \}$.

8.13 THEOREM *Let E be an extension field of F and let $a \in E$.*

(i) $F[a]$ is a subring of E containing F and a, and any subring of E containing F and a contains $F[a]$.

(ii) $F(a)$ is a subfield of E containing $F[a]$. Any subfield of E containing F and a contains $F(a)$.

Proof. (i) We proved in the discussion above that every subring containing F and a contains $F[a]$. That $F[a]$ is a subring follows from 7.3, since $F[a]$ is the image of the evaluation homomorphism $f(X) \mapsto f(a)$ from $F[X]$ to E.

(ii) It suffices to prove that $F(a)$ is a subfield, since the other assertion was proved above. We use Theorem 5.3.

Since F contains the zero and identity of E, so too does $F(a)$. Now if $x, y \in F(a)$ then $x = uv^{-1}$ and $y = st^{-1}$ for some $u, v, s, t \in F[a]$, and by the closure properties of the subring $F[a]$ we have that $ut + vs$, us and $-u$ are all in $F[a]$. Hence

$$x + y = uv^{-1} + st^{-1} = (ut + vs)(vt)^{-1} \in F(a)$$
$$xy = \left(uv^{-1}\right)\left(st^{-1}\right) = (us)(vt)^{-1} \in F(a)$$
$$-x = -\left(uv^{-1}\right) = (-u)v^{-1} \in F(a)$$

while if $u \neq 0$ then $vu^{-1} \in F(a)$. So all the requirements of Theorem 5.3 are satisfied, and $F(a)$ is a subfield. □

Theorem 8.13 justifies the following terminology:

$F[a]$ is the *subring of E generated by F and a*,

$F(a)$ is the *subfield of E generated by F and a*.

8.14 Definition Let E be an extension field of F, and let $a \in E$. If there exists a nonzero polynomial $f(X) \in F[X]$ with $f(a) = 0$ then a is said to be *algebraic* over F. Otherwise a is said to be *transcendental* over F.

8.15 Theorem *Let E be an extension field of F and let $a \in E$.*

(i) If a is transcendental over F then $F[a] \cong F[X]$ (where X is an indeterminate), and $F(a) \neq F[a]$.

(ii) If a is algebraic over F then

(a) there exists a unique monic irreducible polynomial $p(X) \in F[X]$ for which $p(a) = 0$,

(b) $F[X]/I \cong F[a]$, where $I = p(X)F[X]$ is the principal ideal of $F[X]$ generated by p,

(c) $F(a) = F[a]$.

Proof. Let ϕ be the evaluation homomorphism $F[X] \to E$ given by the rule $\phi\big(f(X)\big) = f(a)$ for all $f(X) \in F[X]$. Then by 7.14,

$$\begin{aligned} F[X]/\ker\phi &\cong \operatorname{im}\phi \\ &= \{\, \phi\big(f(X)\big) \mid f(X) \in F[X] \,\} \\ &= \{\, f(a) \mid f(X) \in F[X] \,\} \\ &= F[a]. \end{aligned}$$

(i) Suppose first that a is transcendental over F. Then there are no nonzero polynomials $f(X) \in F[X]$ with $f(a) = 0$, and so

$$\ker\phi = \{\, f(X) \mid \phi\big(f(X)\big) = 0 \,\} = \{0\}.$$

Thus $F[a] \cong F[X]/\{0\} \cong F[X]$, and we have proved the first assertion in (i). Since $F[X]$ is not a field we deduce (by §5b#11 in Chapter 5) that $F[a]$ is not a field, and therefore $F(a) \neq F[a]$.

(ii) Suppose now that a is algebraic over F. By Definition 8.14 there exist nonzero elements of $F[X]$ of which a is a zero, and therefore $\ker\phi \neq \{0\}$. Let $p(X)$ be a nonzero polynomial of minimal degree in $\ker\phi$. Since associates of

elements of $\ker\phi$ will also be in $\ker\phi$, we may choose $p(X)$ to be monic (by 6.12.1). We have $\ker\phi = p(X)F[X]$ (by 8.1.1), and also $p(a) = \phi(p(X)) = 0$ (since $p(X) \in \ker\phi$). Clearly $\deg(p) > 1$, since $p(X)$ is nonzero and $p(a) = 0$. If $s(X)$ and $t(X)$ are polynomials of smaller degree than $p(X)$ such that $p(X) = s(X)t(X)$ then since $s(a)t(a) = p(a) = 0$ and the field E can have no zero divisors it follows that either $s(a) = 0$ or $t(a) = 0$. But this contradicts the choice of $p(X)$ as a polynomial of minimal degree of which a is a zero. So $p(X)$ has no such factorization, and is therefore irreducible.

To complete the proof of (a) it remains to show that $p(X)$ is the unique monic irreducible element of $F[X]$ of which a is a zero. So, assume that $q(X) \in F[X]$ is irreducible, monic and satisfies $q(a) = 0$. Then $\phi(q(X)) = 0$, and consequently

$$q(X) \in \ker\phi = p(X)F[X].$$

Thus $p(X)|q(X)$, and since $q(X)$ is irreducible and $\deg(p) > 1$ it follows that $p(X)$ and $q(X)$ are associates. Because they are both monic this implies that $q(X) = p(X)$.

Since $F[X]/\ker\phi \cong F[a]$ and $\ker\phi = p(X)F[X]$, part (b) has been proved. We know by Theorem 8.6 and §5b#11 that $F[a]$ is a field; so the second assertion in 8.13 (ii) yields that $F[a]$ contains $F(a)$. But the first assertion in 8.13 (ii) gives the reverse inclusion, and therefore $F(a) = F[a]$, proving (c). □

Comment ▷▷▷

8.15.1 The polynomial $p(X)$ in 8.15 (ii) is called the *minimal polynomial* of the algebraic element a. Note that the minimal polynomial is always irreducible. Note also that if p is the minimal polynomial of a then $F[a]$ is an extension of F of degree equal to $\deg(p)$. ▷▷▷

—***Example***—

#9 If F is a subfield of $\mathbb{R}$ and $0 < a \in \mathbb{R}$ with $a \notin F$ then

$$F(a) = F[a] = \{\, x + y\sqrt{a} \mid x, y \in F \,\}$$

is a subfield of $\mathbb{R}$, and is an extension of F of degree 2.

---

§8g Ruler and compass constructions revisited

We finally have the machinery at hand to deal with the classical geometrical problems described in Chapter 1. Reformulating Theorems 1.1 and 1.2 gives the following characterization of constructible numbers:

8.16 Theorem *A real number t is constructible if and only if there is a finite sequence of subfields of $\mathbb{R}$*

$$\mathbb{Q} = F_1 \subset F_2 \subset \cdots \subset F_n$$

such that $t \in F_n$ and for each $i = 1, 2, \ldots, n-1$ there is an $a_i \in F_i$ such that $F_{i+1} = F_i\big(\sqrt{a_i}\big)$.

Proof. Suppose firstly that we are given such a chain of subfields F_i of $\mathbb{R}$: we will prove that all the elements of the subfields are constructible numbers. Since 1 is constructible it follows from 1.1 that all elements of $\mathbb{Q}$ are constructible; that is, $F_1 \subseteq \mathbf{Con}$. But if $F_i \subseteq \mathbf{Con}$ and $a, b \in F_i$ then $a, b, a_i \in \mathbf{Con}$, and by 1.1 it follows that $a + b\sqrt{a_i} \in \mathbf{Con}$. So $F_{i+1} \subseteq \mathbf{Con}$, and by induction all fields in the chain are contained in **Con**.

Conversely, let t be a constructible number, and let α be a constructible point one of whose coordinates is t. Let $\alpha_0, \alpha_1, \ldots, \alpha_n = \alpha$ be the points obtained in a ruler and compass construction, listed in the order obtained. (Thus $\alpha_0 = (0,0)$ and $\alpha_1 = (1,0)$.) For each i let F_i be the set of all real numbers obtainable from the coordinates of $\alpha_0, \alpha_1, \ldots, \alpha_i$ by finite sequences of operations of addition, subtraction, multiplication and division. Clearly each F_i is closed under these operations, and we see from 5.3 that F_i is a subfield of $\mathbb{R}$. Moreover, by 1.2 there exists $a_i \in F_i$ such that $\alpha_{i+1} = (s+t\sqrt{a_i}, u+v\sqrt{a_i})$ for some $s, t, u, v \in F_i$. If $t = v = 0$ we can replace a_i by 0, and in this case we have $F_{i+1} = F_i = F_i(\sqrt{a_i})$. If t or v is nonzero then all elements of $F_i(\sqrt{a_i})$ can be obtained from the coordinates of α and elements of F_i by finite sequences of field operations, and we see that the field F_{i+1} is equal to $F_i(\sqrt{a_i})$ in this case too. Hence we have a sequence of fields of the required kind with $t \in F_n$. □

Now we can dispose of the Delian Problem:

8.17 Theorem *The number $\sqrt[3]{2}$ is not constructible: a cube cannot be duplicated by ruler and compass.*

Proof. Suppose that $\sqrt[3]{2}$ is a constructible number. By 8.16 there is a sequence of fields $\mathbb{Q} = F_1 \subset F_2 \subset \cdots \subset F_n$, each a quadratic extension

of the preceding, with $\sqrt[3]{2} \in F_n$. Let $F_i = F_{i-1}\big(\sqrt{a_{i-1}}\big)$, and assume that $F_i \neq F_{i-1}$. (If $F_i = F_{i-1}$ simply delete F_i from the sequence.) Then $\sqrt{a_{i-1}}$ is not in F_{i-1}, and so $X^2 - a_{i-1}$ is an irreducible polynomial in $F_{i-1}[X]$. Hence

$$[F_i : F_{i-1}] = \deg(X^2 - a_{i-1}) = 2.$$

By Theorem 8.10 it follows that $[F_n : \mathbb{Q}] = 2^n$. But since $\sqrt[3]{2} \in F_n$ it follows that $\mathbb{Q}\big(\sqrt[3]{2}\big) \subseteq F_n$. Furthermore, as we have seen, $X^3 - 2$ is an irreducible element of $\mathbb{Q}[X]$, and therefore

$$[\mathbb{Q}\big(\sqrt[3]{2}\big) : \mathbb{Q}] = \deg(X^3 - 2) = 3.$$

By Theorem 8.10 we have

$$\begin{aligned} 2^n &= [F_n : \mathbb{Q}] \\ &= \big[F_n : \mathbb{Q}\big(\sqrt[3]{2}\big)\big]\big[\mathbb{Q}\big(\sqrt[3]{2}\big) : \mathbb{Q}\big] \\ &= 3\big[F_n : \mathbb{Q}\big(\sqrt[3]{2}\big)\big], \end{aligned}$$

—a contradiction, since the degree $\big[F_n : \mathbb{Q}\big(\sqrt[3]{2}\big)\big]$ must be an integer, but 2^n is not divisible by three. □

A similar proof applies for $\kappa = \cos(\frac{\pi}{9})$. Since $\cos 3\theta = 4\cos^3\theta - 3\cos\theta$ we have that $4\kappa^3 - 3\kappa = \frac{1}{2}$; that is, κ is a zero of $8X^3 - 6X - 1$. This polynomial is irreducible over $\mathbb{Q}[X]$, and so $[\mathbb{Q}(\kappa) : \mathbb{Q}] = 3$. Thus, by the same argument as above, κ cannot lie in an extension field of $\mathbb{Q}$ of degree a power of two. Thus we have proved

8.18 Theorem *The number* $\cos(\frac{\pi}{9})$ *is not constructible; an angle of sixty degrees cannot be trisected.*

I am forced now to confess that the third classical problem is beyond the scope of this course. The proof that $\sqrt{\pi}$ is not constructible depends on showing that π is transcendental over $\mathbb{Q}$; that is, π is not a zero of any polynomial equation over $\mathbb{Q}$. From this it follows that $\sqrt{\pi}$ is also transcendental, and hence not constructible (since it follows readily from 8.16 that every constructible number is algebraic over $\mathbb{Q}$). To prove that π is transcendental would require a lengthy digression into Number Theory. The interested reader is referred to Hardy and Wright "An Introduction to the Theory of Numbers" (4^{th} ed.) §11.14, p.173.

§8h Finite fields

Although we have answered the questions we set out to answer, it would be a shame to leave the subject without a few words on fields which have only finitely many elements. If $p \in \mathbb{Z}^+$ is prime then $\mathbb{Z}_p$ is a field with exactly p elements (by Theorems 4.10 and 4.11). We have also seen (#8) how a field can be constructed which has exactly eight elements. It is natural to wonder for which positive integers n a field can be found with exactly n elements.

8.19 Theorem *Let F be any field. Then the characteristic of F is either zero or a prime number.*

Proof. Suppose to the contrary that F has characteristic m and that m is composite (that is, not prime). Then $m = rs$ for some $r, s \in \{2, 3, \ldots, m-1\}$. By 5.11 the elements $r1$ and $s1$ of F are nonzero (since r and s are less than the characteristic of F), but $(r1)(s1) = m1 = 0$. This contradicts the fact that there are no zero divisors in a field. □

Now suppose that F is a field with exactly n elements. Then the subset S of F defined by

$$S = \{\, n1 \mid n \in \mathbb{Z} \,\}$$

has only a finite number of elements, and so it cannot be isomorphic to $\mathbb{Z}$. So by Theorem 5.12 it follows that $S \cong \mathbb{Z}_p$, where p is the characteristic of F, and by 8.19 we know that p must be prime. Therefore we have proved the following:

8.20 Proposition *A finite field F must be an extension of $\mathbb{Z}_p$ for some prime p.*

The field F can be regarded as a vector space over the subfield $S \cong \mathbb{Z}_p$, and since F has only finitely many elements this must certainly be a finite dimensional vector space. So the degree $[F : \mathbb{Z}_p]$ of the extension is finite. From this we deduce the next theorem.

8.21 Theorem *Suppose that F is an extension of $\mathbb{Z}_p$ (where p is prime), and suppose that $[F : \mathbb{Z}_p] = k$. Then F has exactly p^k elements.*

Proof. Let $a_1, a_2, \ldots, a_k$ be a basis for F over $\mathbb{Z}_p$. Then each element of F is uniquely expressible in the form $\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_k a_k$ with the λ_i in $\mathbb{Z}_p$, and since there are p choices for each λ_i there are p^k such expressions altogether. □

It is also possible to show that for each prime power p^k there is a field with p^k elements, and that any two fields with p^k elements are isomorphic. To construct such a field one simply has to find an irreducible polynomial $f \in \mathbb{Z}_p[X]$ with $\deg(f) = k$, for then 8.6 and 8.9.2 show that $F = \mathbb{Z}_p[X]/I$ (where $I = f(X)\mathbb{Z}_p[X]$) is a field satisfying $[F : \mathbb{Z}_p] = k$.

8.22 THEOREM *For each prime power q there is (up to isomorphism) a unique field with q elements.*

8.23 DEFINITION The field referred to in 8.22 is called the *Galois field* with q elements. It is commonly denoted by '$GF(q)$'.

We omit the proof of Theorem 8.22, but give several examples to illustrate Galois fields.

—***Examples***—

#10 $GF(4)$

The only polynomials of degree 1 in $\mathbb{Z}_2[X]$ are X and $X+1$. Therefore the only reducible polynomials of degree 2 are X^2, $X(X+1) = X^2 + X$ and $(X+1)^2 = X^2 + 1$. (Of course $(X+1)^2$ is equal to $X^2 + 2X + 1$, but $2 = 0$ in $\mathbb{Z}_2$.) Hence $p(X) = X^2 + X + 1$, the remaining polynomial of degree 2, must be irreducible. Let $I = p(X)\mathbb{Z}_2[X]$ and $Q = \mathbb{Z}_2[X]/I$. Then Q is an extension field of $\mathbb{Z}_2$ of degree equal to $\deg(p) = 2$, and if $\alpha = I + X$ then $1, \alpha$ is a basis for Q over $\mathbb{Z}_2$ (by 8.7). Thus Q has exactly 4 elements:

$$0 = 0 \cdot 1 + 0\alpha \qquad \alpha = 0 \cdot 1 + 1\alpha$$
$$1 = 1 \cdot 1 + 0\alpha \qquad \alpha + 1 = 1 \cdot 1 + 1\alpha.$$

The rules for addition and multiplication of these elements are completely determined by the equation $\alpha^2 + \alpha + 1 = 0$ together with the fact that the characteristic of Q is 2. Thus, for example,

$$\alpha(\alpha + 1) = \alpha^2 + \alpha = (\alpha^2 + \alpha + 1) + 1 = 1.$$

Similar calculations yield the complete addition and multiplication tables for $GF(4)$. (We have omitted 0 from the multiplication table since the rule for multiplying by 0 is trivial: $0x = 0$ for all x.)

$+$	0	1	α	$\alpha+1$
0	0	1	α	$\alpha+1$
1	1	0	$\alpha+1$	α
α	α	$\alpha+1$	0	1
$\alpha+1$	$\alpha+1$	α	1	0

$\cdot$	1	α	$\alpha+1$
1	1	α	$\alpha+1$
α	α	$\alpha+1$	1
$\alpha+1$	$\alpha+1$	1	α

#11 $GF(16)$

Suppose that $f(X) \in \mathbb{Z}_2[X]$ is reducible and has degree 4. Then one of the following cases must occur:

(a) $f(X)$ has four irreducible factors of degree 1. There are five possibilities for $f(X)$:

$$\begin{array}{ll} X^4 & X(X+1)^3 \\ X^3(X+1) & (X+1)^4 \\ X^2(X+1)^2 & \end{array}$$

(b) $f(X)$ has two irreducible factors of degree 1 and one of degree 2. There are three possibilities:

$$\begin{array}{ll} X^2(X^2+X+1) & (X+1)^2(X^2+X+1) \\ X(X+1)(X^2+X+1) & \end{array}$$

(c) $f(X)$ has one irreducible factor of degree 1 and one of degree 3. There are four possibilities:

$$\begin{array}{ll} X(X^3+X+1) & X(X^3+X^2+1) \\ (X+1)(X^3+X+1) & (X+1)(X^3+X^2+1) \end{array}$$

(d) $f(X)$ has two irreducible factors of degree 2. The only possibility is

$$f(X) = (X^2+X+1)^2.$$

So there are thirteen reducible polynomials of degree 4. Since there are sixteen polynomials of degree 4 altogether it follows that there are three irreducible ones, and they can be found by writing down all sixteen elements of $\mathbb{Z}_2[X]$ of degree 4 and crossing out the thirteen reducible ones above. The irreducibles are

$$X^4+X+1, \quad X^4+X^3+1, \quad X^4+X^3+X^2+X+1,$$

and we can use any of these in the construction of $GF(16)$. For instance, let $I = (X^4+X+1)\mathbb{Z}_2[X]$ and $t = I+X \in \mathbb{Z}_2[X]/I$. Then the sixteen elements of $\mathbb{Z}_2[X]/I$ are:

$$\begin{array}{ll} 0,\ 1,\ t^2+t,\ t^2+t+1 & \text{(forming a subfield with 4 elements)} \\ t,\ t^2,\ t+1,\ t^2+1 & \text{(which are zeros of } X^4+X+1) \\ t^3,\ t^3+t^2,\ t^3+t^2+t+1,\ t^3+t & \text{(zeros of } X^4+X^3+X^2+X+1) \\ t^3+1,\ t^3+t^2+1,\ t^3+t^2+t,\ t^3+t+1 & \text{(zeros of } X^4+X^3+1). \end{array}$$

#12 $GF(9)$

As in our previous examples we can explicitly determine the reducible polynomials of degree 2 in $\mathbb{Z}_3[X]$, and deduce that the remaining ones are irreducible. We find that there are three monic irreducibles, namely $X^2 - X - 1$, $X^2 + X - 1$ and $X^2 + 1$. Now $GF(9)$ can be constructed by adjoining to $GF(3) = \mathbb{Z}_3$ a zero of one of these polynomials (we can choose whichever we like). Thus, for instance, if s is a zero of $X^2 + 1$ then $GF(9)$ consists of

$$\begin{aligned} 0,\ 1,\ -1 &\quad \text{(lying in the subfield } \mathbb{Z}_3\text{)}, \\ s,\ -s &\quad \text{(zeros of } X^2 + 1\text{)}, \\ s+1,\ -s+1 &\quad \text{(zeros of } X^2 + X - 1\text{)}, \\ s-1,\ -s-1 &\quad \text{(zeros of } X^2 - X - 1\text{)}. \end{aligned}$$

The addition table is easy to write down provided that you remember that $1 + 1 = -1$ and $s + s = -s$ (since $3 = 0$ in a ring of characteristic three). The multiplication table is also straightforward, using $s^2 + 1 = 0$.

#13 $GF(27)$

There are eight monic irreducible polynomials of degree 3 over $GF(3)$, and $GF(27)$ contains three zeros for each of them (making 24 elements) along with the three elements of the subfield $GF(3)$. (Note that $GF(9)$ is not a subfield of $GF(27)$—in general, $GF(q_1)$ is a subfield of $GF(q_2)$ if and only if q_2 is a power of q_1.)

Exercises

1. Suppose that R is a ring of characteristic three with identity element 1.

(i) Is the set $\{0, 1, -1\}$ a subring of R?

(ii) Suppose that $t \in R$ and $t^2 + 1 = 0$. Show that there are at most nine elements of R given by polynomial expressions in t, and write down nine such expressions giving these elements. Prove that these elements form a subring S of R. Is S a field?

2. Prove that the ring $\mathbb{Q}[X]/(X^2 - 2)\mathbb{Q}[X]$ is isomorphic to the field of all real numbers of the form $a + b\sqrt{2}$ with a, $b \in \mathbb{Q}$.

3. Let $F = \mathbb{Q}[\sqrt{1+\sqrt{2}}]$.

(i) Show that the equation of the line joining the points $(1+\sqrt{2}, 1)$ and $(\sqrt{2}, \sqrt{1+\sqrt{2}})$ has coefficients in F.

(ii) Calculate the the coordinates of the point of intersection of the lines

$$\sqrt{2}x - \sqrt{1+\sqrt{2}}y = 1$$
$$(1+\sqrt{2})x + y = 1 + \sqrt{1+\sqrt{2}}$$

and show that they lie in F.

4. Find the smallest subfield F of $\mathbb{R}$ which contains the coordinates of the point of intersection of the circle with centre $(0,0)$ and radius $\sqrt{3}$ and the line joining $(1/2, 0)$ and $(4\sqrt{2}, \sqrt{2})$. Calculate also the degree $[F : \mathbb{Q}]$.

5. Let $J = (X^3 + X + 1)\mathbb{Z}_2[X]$. Express the following elements of $\mathbb{Z}_2[X]/J$ in the form $J + (aX^2 + bX + c)$ with a, b, $c \in \mathbb{Z}_2$:

(i) $J + X^5$. (ii) $J + (X^4 + X + 1)$.

6. Let I be the principal ideal in $\mathbb{Q}[X]$ generated by $X^4 - 2X^2 + 1$. For each of the following elements of $\mathbb{Q}[X]/I$ determine whether an inverse exists, and, if one does, find it.

(i) $I + (X^2 + X + 1)$. (ii) $I + (X^2 + X - 2)$.

7. For the given field F and polynomial $s(X) \in F[X]$, list all the elements of the ring $Q = F[X]/s(X)F[X]$ and construct a multiplication table for Q.

(i) $F = \mathbb{Z}_2$, $s(X) = X^2 + 1$

(ii) $F = \mathbb{Z}_3$, $s(X) = X^2 + 1$

(iii) $F = \mathbb{Z}_2$, $s(X) = X^3 + X + 1$

(iv) $F = \mathbb{Z}_3$, $s(X) = X^2 - X + 1$

8. Prove that $\sqrt{5} \notin \mathbb{Q}[\sqrt{2}]$ by attempting to solve $\sqrt{5} = x + y\sqrt{2}$ for rational numbers x and y.

9. (*i*) Prove that $\sqrt{5} \notin \mathbb{Q}[\sqrt[3]{2}]$ by attempting to solve

$$\sqrt{5} = x + y\sqrt[3]{2} + z(\sqrt[3]{2})^2$$

for x, y, $z \in \mathbb{Q}$.
(Hint: Use the fact that 1, $\sqrt[3]{2}$ and $(\sqrt[3]{2})^2$ are linearly independent over $\mathbb{Q}$.)

(*ii*) Prove that $\sqrt{5} \notin \mathbb{Q}[\sqrt[3]{2}]$ by considering degrees of field extensions.
(Hint: If $\sqrt{5} \in \mathbb{Q}[\sqrt[3]{2}]$ then $\mathbb{Q} \subseteq \mathbb{Q}[\sqrt{5}] \subseteq \mathbb{Q}[\sqrt[3]{2}]$.)

(*iii*) Use considerations of degree to prove that $\sqrt[5]{17} \notin \mathbb{Q}[\sqrt[7]{43}]$. (Note that a computational proof of this fact would be rather messy!)

10. Let $E = \mathbb{Q}[\sqrt{2}]$, $K = E[\sqrt{5}]$.

(*i*) Calculate the degree $[K : \mathbb{Q}]$.

(*ii*) Observe that $t = \sqrt{2} + \sqrt{5} \in K$. Can the numbers 1, t, t^2, t^3, t^4 be linearly independent over $\mathbb{Q}$?

(*iii*) Prove that $\sqrt{2} + \sqrt{5}$ is algebraic over $\mathbb{Q}$ and find a polynomial $f(X) \in \mathbb{Q}[X]$ such that $f(\sqrt{2} + \sqrt{5}) = 0$.

11. Let F be a finite field and S a subfield of F. Prove that the number of elements of F is a power of the number of elements of S.
(Hint: Imitate the proof of Theorem 8.21.)

12. Let F be a field of characteristic p.

(*i*) Prove that $(a + b)^p = a^p + b^p$ for all a, $b \in F$.

(*ii*) Prove that the function $\phi\colon F \to F$ defined by $\phi(a) = a^p$ (for all $a \in F$) is a ring homomorphism.

(*iii*) Prove that $x = 1$ is the only solution in F of the equation $x^p = 1$.

13. Let F be a field of characteristic p such that the polynomial $X^{p^3} - X$ has p^3 distinct roots in F. Prove that these roots form a subfield of F with p^3 elements. (Hint: Use Theorem 5.3.)

14. Let F be a field and $f(X) \in F[X]$ be irreducible polynomial. Let $Q = \mathbb{Z}_p[X]/I$ where $I = f(X)F[X]$ and let $\alpha = I + X \in Q$. Thus Q is an extension field of F and α a zero of $f(Y)$ in $Q[Y]$.

(*i*) Prove that $f(Y) = (Y - \alpha)g(Y)$ for some $g(Y) \in Q[Y]$.

(*ii*) Suppose that $\deg(g) \geq 1$ and let $h(Y)$ be an irreducible factor of $g(Y)$ in $Q[Y]$. Prove that there exists an extension field E of Q in which h has a zero.

(*iii*) Prove the field E in the previous part is an extension of F in which the polynomial f has at least two zeros.

(*iv*) Explain why there must exist an extension of F in which f has $\deg(f)$ zeros. Show that this is also true if f is not irreducible.
(Hint: Consider the irreducible factors of f separately.)

15. Use the previous two exercises to prove the existence of a field with p^3 elements, and generalize the argument to prove the existence of a field with p^k elements for any k.

Index of notation

Index of examples

Index of examples

Index of examples

Chapter 6

Chapter 7

Index of examples