1. Introduction

The course focuses on polynomial equations. The general problem, to understand solution sets of simultaneous polynomial equations in several variables (things like $x_1x_2 + 5x_3^2 - x_1x_3^3 = 2$), is enormously difficult. We shall only consider the one-variable case.

(1) The degree 1 (linear) case $ax + b = 0$ is trivial.

(2) The degree 2 (quadratic) case $ax^2 + bx + c = 0$ was solved in ancient times. The trick is to *complete the square*:

$$ax^2 + bx + c = a(x + \tfrac{b}{2a})^2 + (c - \tfrac{b^2}{4a})$$

giving the well-known formula $x = -(b/2a) \pm \sqrt{(b^2/4a^2) - (c/a)}$ for the roots.

(3) The degree 3 and 4 (cubic and quartic) cases were not solved until the 16th century, when several Italian mathematicians made the decisive breakthroughs.

(5) Abel (1824) showed that it is impossible to produce a formula for the solutions of a quintic equation of the kind that had been found for the lower degrees. Specifically, there is no such formula built up from operations of addition, subtraction, multiplication, division and nth root extraction. That is to say, quintic and higher degree polynomial equations are generally *not soluble by radicals*.

(6) Galois (1830) (who, coincidentally, was a radical in the political sense) found necessary and sufficient conditions for any given polynomial equation to be soluble by radicals.

Our primary objective in this course is to develop the modern algebraic machinery which is now used to express and prove Galois' results, as well as to facilitate analysis of many other algebraic problems. In particular, we shall demonstrate the insolubility of the quintic.

2. Solving the cubic without any theory

Our aim is to find a formula for the solutions of

$$x^3 - S_1x^2 + S_2x - S_3 = 0.$$

Let the solutions be t_1, t_2 and t_3. Then

$$x^3 - S_1x^2 + S_2x - S_3 = (x - t_1)(x - t_2)(x - t_3)$$

and equating coefficients gives

$$\begin{aligned} S_1 &= t_1 + t_2 + t_3 \\ S_2 &= t_1t_2 + t_1t_3 + t_2t_3 \\ S_3 &= t_1t_2t_3. \end{aligned} \tag{1}$$

Ideally, we would like to find expressions

$$\begin{aligned} t_1 &= f_1(S_1, S_2, S_3) \\ t_2 &= f_2(S_1, S_2, S_3) \\ t_3 &= f_3(S_1, S_2, S_3). \end{aligned}$$

However, if the f_i are to be functions in the normal sense, this seems to be impossible. The equations (1) are symmetrical in t_1, t_2 and t_3. For example, swapping t_1 and t_2 does not change S_1, S_2 and S_3, and so cannot change $f_1(S_1, S_2, S_3)$, which is meant to equal t_1. Since swapping t_1 and t_2 changes

t_1 to t_2, the only way out of this is for t_1 and t_2 to be equal, and it is easy to find examples of cubic equations for which the roots are not coincident. The most convenient way round this difficulty is to look for a deliberately ambiguous, or many-valued, expression $f(S_1, S_2, S_3)$, whose different values give the different roots t_i. To facilitate this we shall allow $\sqrt[n]{D}$ to denote any of the solutions of $x^n = D$, so that $\sqrt[n]{D}$ is an n-valued expression. With this convention the formula for the roots of the quadratic equation $x^2 - px + q = 0$ is the two-valued expression $x = \frac{1}{2}(p + \sqrt{p^2 - 4q})$.

Polynomial expressions in S_1, S_2 and S_3—that is, expressions built up from the S_i using only addition, multiplication and multiplication by constants—are also polynomials in t_1, t_2 and t_3, and since S_1, S_2 and S_3 are completely symmetrical, in the sense that all permutations of t_i leave them unchanged, it follows that every polynomial in the S_i will also be completely symmetrical in the t_i. Conversely, it turns out that every polynomial in the t_i which is symmetrical in these variables can be expressed as a polynomial in the S_i. For the time being there is no need for us to prove this, but merely be able to express a few given symmetrical polynomials in the t_i in terms of the S_i. For example,

$$\begin{aligned} t_1^2 + t_2^2 + t_3^2 &= (t_1 + t_2 + t_3)^2 - 2(t_1t_2 + t_1t_3 + t_2t_3) \\ &= S_1^2 - 2S_2 \end{aligned}$$

and similarly

$$\begin{aligned} t_1^2t_2 + t_2^2t_1 + t_1^2t_3 + t_3^2t_1 + t_2^2t_3 + t_3^2t_2 &= (t_1 + t_2 + t_3)(t_1t_2 + t_1t_3 + t_2t_3) - 3t_1t_2t_3 \\ &= S_1S_2 - 3S_3. \end{aligned}$$

The clue to solving the cubic is to investigate expressions in the t_i which are partially symmetrical, in some sense, but not completely symmetrical. In particular, we look at expressions which are unchanged by the cyclic permutation $t_1 \mapsto t_2 \mapsto t_3 \mapsto t_1$, which we shall denote by ρ, and its inverse. For example,

$$\alpha = t_1t_2^2 + t_2t_3^2 + t_3t_1^2$$

has this property, and interchanging t_2 and t_3 changes α to

$$\beta = t_1t_3^2 + t_3t_2^2 + t_2t_1^2,$$

which also has the property of invariance under the cyclic permutations. Now $\alpha + \beta$ and $\alpha\beta$ are completely symmetrical, and, to be specific, if we put $\alpha + \beta = P$ and $\alpha\beta = Q$ then

$$\begin{aligned} P &= S_1S_2 - 3S_3 \\ Q &= S_1^3S_3 + S_2^3 - 6S_1S_2S_3 + 9S_3^2. \end{aligned}$$

It follows that

$$(x - \alpha)(x - \beta) = x^2 - (\alpha + \beta)x - \alpha\beta = x^2 - Px + Q.$$

The significance of this is that α and β are the roots of a quadratic equation whose coefficients are polynomial expressions in S_1, S_2 and S_3. This demonstrates that it is possible to find formulas in terms of the S_i for expressions which are not completely symmetrical in the t_i.

Clearly, however, some further idea will we required to progress from partially symmetrical quantities to totally unsymmetrical ones. Furthermore, it is clear that cube roots will have to be involved in some way. Define $\omega = -\frac{1}{2} + \frac{\sqrt{-3}}{2}$, a complex cube root of 1. It is reasonable to expect that this number will have a role to play in any context where cube roots arise, since if C is one of the cube roots of a number D, the other cube roots are ωC and $\omega^2 C$. The trick is to consider

$$\theta = t_1 + \omega t_2 + \omega^2 t_3.$$

Observe that the cyclic permutation ρ takes θ to $t_2+\omega t_3+\omega^2 t_1 = \omega^2\theta$. So θ is not partially symmetrical in the sense considered above, but is close to being so: a cyclic permutation has the effect of multiplying θ by a cube root of 1. Applying ρ to $\theta^3 = \theta\theta\theta$ will give $(\omega^2\theta)(\omega^2\theta)(\omega^2\theta) = \omega^6\theta^3 = \theta^3$. Thus θ^3 is partially symmetrical, and it will be susceptible to the same kind of analysis as applied to α and β above. Indeed, interchanging t_2 and t_3 takes θ^3 to ψ^3, where

$$\psi = t_1 + \omega t_3 + \omega^2 t_2,$$

and $A = \theta^3 + \psi^3$ and $B = \theta^3\psi^3$ are both totally symmetrical. We find also that ρ takes ψ to $\omega\psi$, and, as we saw above for θ, it follows that ψ^3 is fixed by ρ. With a little calculation we find that in fact

$$\begin{aligned}\theta^3 &= S_1^3 - 3S_1S_2 + 9S_3 + 3\omega^2\alpha + 3\omega\beta\\ \psi^3 &= S_1^3 - 3S_1S_2 + 9S_3 + 3\omega^2\beta + 3\omega\alpha\end{aligned}$$

and since $\omega + \omega^2 = -1$ we obtain that

$$\begin{aligned}A = \theta^3 + \psi^3 &= 2S_1^3 - 6S_1S_2 + 18S_3 - 3(\alpha+\beta)\\ &= 2S_1^3 - 9S_1S_2 + 27S_3.\end{aligned}$$

The calculation of B is simplified by the fact that $\theta\psi$ is completely symmetrical; for example, interchanging t_2 and t_3 interchanges θ and ψ, while ρ multiplies θ by ω^{-1} and ψ by ω. It is readily checked that $\theta\psi = S_1^2 - 3S_2$, and hence $B = (\theta\psi)^3 = (S_1^2 - 3S_2)^3$. As θ^3 and ψ^3 are the roots of $x^2 - Ax + B = 0$, we have

$$\theta^3 = \tfrac{1}{2}(A + \sqrt{A^2 - 4B})$$

and thus

$$\theta = \sqrt[3]{\left(\tfrac{1}{2}(A + \sqrt{A^2 - 4B})\right)}.$$

Having found θ we can use $\theta\psi = S_1^2 - 3S_2$ to determine ψ, and now finding the the roots t_1, t_2 and t_3 themselves is simply a matter of solving a system of linear equations. Specifically,

$$\begin{pmatrix}1 & 1 & 1\\ 1 & \omega & \omega^2\\ 1 & \omega^2 & \omega\end{pmatrix}\begin{pmatrix}t_1\\ t_2\\ t_3\end{pmatrix} = \begin{pmatrix}S_1\\ \theta\\ \psi\end{pmatrix},$$

and inverting the coefficient matrix we find that

$$\begin{pmatrix}t_1\\ t_2\\ t_3\end{pmatrix} = \frac{1}{3}\begin{pmatrix}1 & 1 & 1\\ 1 & \omega^2 & \omega\\ 1 & \omega & \omega^2\end{pmatrix}\begin{pmatrix}S_1\\ \theta\\ \psi\end{pmatrix}.$$

It is somewhat complicated, but it is a formula for the roots of the original cubic in terms of the coefficients.

It is obvious that symmetry and partial symmetry played key roles in the above analysis. However, we need to develop more concepts to properly clarify matters. In particular, what properties of symmetry made the process work, and for what other equations will similar processes be successful? We can also hope that a deeper understanding will lessen the reliance on algebraic computation.

3. Ruler and compass problems

The same theory which is used to analyse polynomial equations can also be applied to study three classical geometrical problems, which were posed by ancient Greek mathematicians. Given only an unmarked ruler (for drawing straight lines) and compasses (for drawing circles) is it possible to perform the following geometrical constructions?

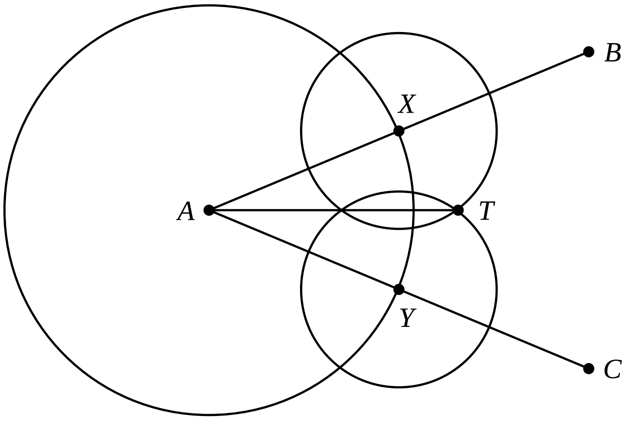
1. Trisect any given angle.
2. Construct a line segment whose length is $\sqrt[3]{2}$ times the length of a given line segment.
3. Construct a square with area equal to the area of a given circle.

The Greeks were unable to perform these constructions without resort to instruments for drawing other kinds of curves. However, they were unable to prove the impossibility of ruler-and-compass constructions. It turns out that these impossibility proofs are greatly simplified by the use of concepts of modern algebra. We shall be able to deal completely with the first two of the three, and show that the third follows from a famous theorem of Lindemann (1882), that the number π is not a root of any nontrivial polynomial equation with integer coefficients. We shall not prove Lindemann's theorem, as to do so would take us too far afield.

In order to prove that some things cannot be done with ruler and compasses, we need to figure out what *can* be done with those tools. Much of what follows may be familiar to you already.

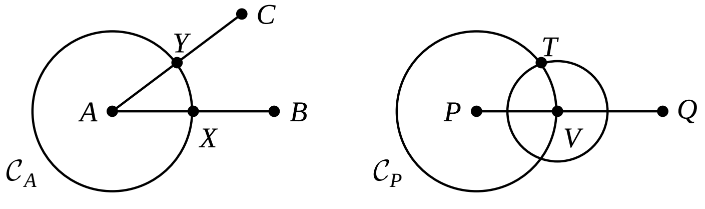
Bisecting an angle

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.



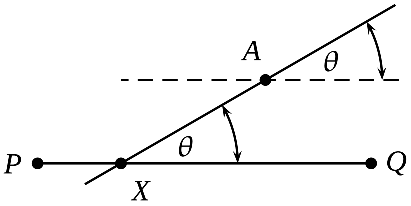
Copying an angle

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 QPT equals the angle BAC.



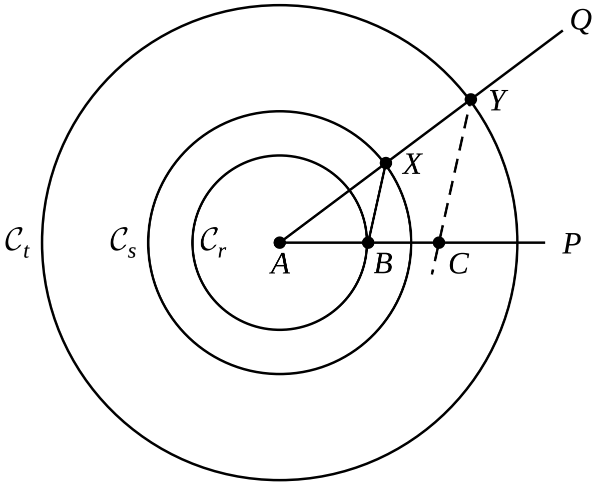
Drawing a parallel to a given line through a given point

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.



Multiplying a length by the ratio of two lengths

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.



It is easy to construct line segments of lengths equal to the sum and difference of the lengths of given line segments, and so the above construction also enables one to construct a/n and na, where a is a given length and n a given positive integer.

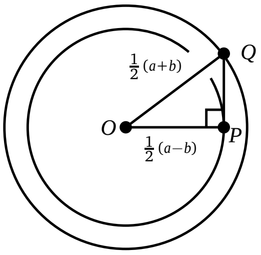
Trisecting π

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.

Since angles can be bisected, one can also construct angles of $\pi/6$, and hence also right-angles (since one can easily add two angles, by copying one alongside the other).

Squaring a rectangle

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. Let this perpendicular meet the large circle at Q. Then PQ has the required length.



We are now ready to analyse the general question of exactly what can be constructed with ruler and compasses. Start with two points—let us call them O and P—and consider an orthogonal coordinate system such that O is the origin and P the point $(1, 0)$. Define the notion of *constructibility* recursively as follows:

(1) O and P are constructible points;
(2) A straight line is constructible if it passes through two constructible points;
(3) A circle is constructible if its centre is a constructible point and its radius is the distance between two constructible points.
(4) Points of intersection of two constructible lines, or two constructible circles, or a constructible line and a constructible circle, are constructible points.

Thus, for example, the x-axis is constructible, since it is the line defined by the constructible points O and P. The circles $x^2 + y^2 = 1$ and $(x - 1)^2 + y^2 = 1$ are both constructible, since their centres are the constructible points O and P and they each have radius 1, which is the distance between the constructible points O and P. The points of intersection of these two circles give us two more constructible points, namely $(1/2, \sqrt{3}/2)$ and $(1/2, -\sqrt{3}/2)$. From here we can go on to construct more points, lines and circles in a multitude of different ways.

We now start to move away from geometry and towards algebra, by extending the concept of constructibility to numbers.

Definition (3.1): Let K be the set of all those real numbers which occur as either x-coordinates or y-coordinates of constructible points. Elements of K are called *constructible numbers*.

If a point Q is constructible then the points of intersection of the line through O perpendicular to OQ and the circle centred at O with radius OQ are also constructible. In particular, (a, b) is constructible if and only if (b, a) is constructible. Thus a number which occurs as an x-coordinate of a constructible point also occurs as a y-coordinate of a constructible point, and vice versa. Furthermore, if (a, y) and (x, b) are constructible points then so is (a, b), since it is the point of intersection of the line through (a, y) parallel to the y-axis and the line through (x, b) parallel to the x-axis (both of which are constructible lines). Thus a point (a, b) is constructible if and only if both a and b are constructible numbers.

We can now prove a theorem which gives a characterization of constructible numbers.

Theorem (3.2): *A number x is constructible if and only if there exists a sequence $x_0, x_1, \ldots, x_n$ such that*

(1) $x_0 = 0$ and $x_1 = 1$,
(2) $x_n = x$, and
(3) each term x_i of the sequence is a sum of earlier terms, or a product of earlier terms, or the negative, reciprocal or square root of an earlier term.

Proof. Let K' be the set of numbers that appear in sequences of the kind described in the theorem statement. We must show that $K' = K$; let us start by showing that $K' \subseteq K$.

Let $x \in K'$. Then there exists a sequence $x_0, x_1, \ldots, x_n = x$ of the kind described above. We use induction on n to show that $x \in K$. This is trivial if $n = 0$ or $n = 1$, since this would mean that

$x = 0$ or $x = 1$, and certainly $0, 1 \in K$. If $n > 1$ then x has one of the forms $a + b$, ab, $-a$, a^{-1} or $\sqrt{a}$, where $a, b \in \{x_0, x_1, \ldots, x_{n-1}\}$. Our inductive hypothesis yields that $a, b \in K$. So $(a, 0)$ and $(b, 0)$ are constructible points. It is now an easy matter to use the constructions described above to show that in each case the point $(x, 0)$ is also constructible, and hence $x \in K$.

It remains to show that $K \subseteq K'$. Observe that, by definition, K' is closed with respect to addition, subtraction, multiplication, division and extraction of square roots. We show, by induction, that

(a) every constructible line has an equation of the form $ax + by + c = 0$, with $a, b, c \in K'$;
(b) every constructible circle has an equation of the form $x^2 + y^2 + ax + by + c = 0$, with $a, b, c \in K'$;
(c) every constructible point has the form (a, b), with $a, b, \in K'$. (Let us express this more briefly by saying that all constructible points, lines and circles have coefficients in K'.)

The initial points, O and P, certainly have the form (a, b) with $a, b \in K'$, since $0, 1 \in K'$. This starts the induction. Now suppose that at some stage in a ruler and compass construction, all the points, lines and circles thus far constructed have coefficients in K'. If the next step is to draw a line, it will be the line L through two points (a_1, b_1) and (a_2, b_2) previously constructed. So $a_1, a_2, b_1, b_2 \in K'$. But L has equation

$$(b_2 - b_1)x + (a_1 - a_2)y = a_1b_2 - a_2b_1,$$

and, by the closure properties of K' mentioned above, we see that all the coefficients involved are in K'. Similarly, if the next step is to draw a circle, it will be the circle with centre a previously constructed point (a, b), and radius the distance between two previously constructed points (c, d) and (e, f). The equation of this circle is

$$(x - a)^2 + (y - b)^2 = (c - e)^2 + (d - f)^2,$$

and since $a, b, c, d, e, f \in K'$ it follows that all the coefficients in the equation are in K'. Finally, if the next step is to mark in a point of intersection, the coordinates of that point will be a solution of a pair of simultaneous equations, of the form

$$\begin{aligned} ax + by + c &= 0 \\ dx + ey + f &= 0 \end{aligned}$$

if it is the point of intersection of two lines, or

$$\begin{aligned} x^2 + y^2 + ax + by + c &= 0 \\ x^2 + y^2 + dx + ey + f &= 0 \end{aligned}$$

for the case of two circles, or

$$\begin{aligned} x^2 + y^2 + ax + by + c &= 0 \\ dx + ey + f &= 0 \end{aligned}$$

for the case of a circle and a line. In each case the inductive hypothesis tells us that the coefficients a, b, c, d, e and f are in K'. Since these equations can be solved using only the operations of addition, subtraction, multiplication, division and square root extraction, it follows that the newly constructed point has coordinates in K', as claimed. This completes the induction. But since by definition elements of K are coordinates of constructible points, it follows that every element of K is in K', as required. □

What this theorem shows us is, essentially, that with ruler and compasses one can add, subtract, multiply, divide and take square roots, and nothing else. So $\sqrt[3]{2}$ could only be constructed if it were possible to find a formula for $\sqrt[3]{2}$ involving only the integers 0 and 1 and the above operations. It seems rather unlikely that such a formula would exist; however, "seems unlikely" is not good enough for us. The history of mathematics has no shortage of examples of things that seemed unlikely once, but nevertheless turned out to be true. Before we can prove the nonexistence of such a formula, we need to develop rather a lot of algebraic machinery.

4. Commencing the theory

Definition (4.1): A *ring* is a set equipped with two operations, which we shall call addition (to be indicated by $+$) and multiplication (usually indicated by simple juxtaposition of the arguments), satisfying the following axioms.

(i) Addition is associative. That is, $(a + b) + c = a + (b + c)$ for all $a, b, c \in R$, where R is the set in question.

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

(iii) There is a zero element. That is, there exists an element $z \in R$ such that $a + z = a$ for all $a \in R$. Note immediately that z is uniquely determined, since if z_1 and z_2 both satisfy this property then $z_1 = z_2 + z_1 = z_1 + z_2 = z_2$. The zero element will usually be denoted by 0 (or 0_R, if we wish to specify the ring involved).

(iv) All elements have negatives. That is, for each $a \in R$ there exists $b \in R$ with $a + b = 0$. Note immediately that the negative of a given element a is uniquely determined, since if b_1 and b_2 both have the given property then

$$b_1 = b_1 + 0 = b_1 + (a + b_2) = (b_1 + a) + b_2 = b_2 + (a + b_1) = b_2 + 0 = b_2.$$

The negative of a will be written as $-a$, and $x + (-y)$ will often be abbreviated as $x - y$.

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

(vi) The distributive laws, $a(b + c) = ab + ac$ for all $a, b, c \in R$ and $(a + b)c = ac + bc$ for all $a, b, c \in R$, are both satisfied.

Although the definition as stated above says "A ring is a set . . . ", this way of stating things is a little lacking in rigour, since the operations involved are an important part of the definition. Change the operations and you change the ring, even if you do not change the set. A more formal treatment might define a ring to be a triple $(R, +, \cdot)$ such that R is a set and $+$ and $\cdot$ operations on R satisfying the axioms as listed. The set R would then be called the *underlying set* of the ring $(R, +, \cdot)$. But nobody maintains this level of formality for long, and in due course everyone uses the same name for the ring as for its underlying set. When it is necessary to say that that a specified pair of operations on a set R make R into a ring, it is usual to say that R is a ring under the operations in question.

All axioms in Definition (4.1) resemble standard familiar properties of ordinary addition and multiplication of numbers. So we can say at once that $\mathbb{R}$, the set of all real numbers, is a ring under the usual operations of addition and multiplication of real numbers. So also are $\mathbb{Z}$ (integers), $\mathbb{Q}$ (rational numbers) and $\mathbb{C}$ (complex numbers). (Recall that $\mathbb{Q}$ is the set of all real numbers of the form p/q, where p and q are integers. We shall describe later how one may formally construct $\mathbb{Q}$ from $\mathbb{Z}$, and prove that it is a ring. Giving formal mathematical constructions of the other number systems mentioned above and proving that they satisfy the ring axioms is a nontrivial task, but it is not part of this course. We just assume that everyone knows what numbers are, and knows also that they satisfy the properties which appear in the definition above.) We shall meet several less familiar examples of rings later on.

We have not actually said what we mean by an *operation* on a set. In the sense used in the above definition, an operation on a set S is a function of two variables from S to itself.† In other words, it is a rule which takes as input a pair of elements of S and outputs an element of S. For example, addition of integers is an operation on $\mathbb{Z}$; feeding the pair of integers $(5, 10)$ to this operation as input yields the integer $5 + 10 = 15$ as output. In other contexts a more general concept of "operation" might be appropriate; for example, the scalar multiplication operation of vector space theory takes as input a scalar and a vector and yields a vector as output, and sometimes people consider operations which take a triple of elements as input rather than a pair, and so on. But these will not arise in this course.

Since an operation is just a function, one could argue that it is silly to use the word "operation" at all, since we could use "function" instead. The point is, however, that all the functions which we choose to call "operations" satisfy properties rather like those that appear in the definition of a ring. So the use of the special word lets people know, in some sense, what kind of function we are talking about.

Exercise 1. Let R be a ring and $a \in R$. Show that $z = 0$ is the only solution of the equation $a + z = a$.

Exercise 2. Show that if R is a ring and $a \in R$ then $a0 = 0a = 0$. (Hint: consider $a(0 + 0)$, and make use of the negative $-(a0)$.)

More definitions An *identity element* in a ring R is an element $e \in R$ such that $ea = ae = a$ for all $a \in R$. It is easily shown that if a ring has an identity element it is unique. When R has an identity we shall usually denote it by 1. This does not say that the identity element of R is the same as the real number 1—it is probably a totally different object—but we find it convenient to use the same symbol for all identity elements as is used for the number 1. Furthermore, saying "Let R be a ring with 1", means exactly the same thing as saying "Let R be a ring with an identity element", and does not mean that the number 1 is in R. When it is necessary to distinguish notationally between the identity elements of different rings, we shall write 1_R for the identity of R, and 1_S for the identity of S, etc. (just as we do for the zero elements).

Note that if R has an identity element which happens to coincide with the zero element of R then $a = a1 = a0 = 0$ for all $a \in R$; so R is the trivial one-element ring. Thus in all interesting cases it is required that the identity be nonzero.

A *commutative ring* is a ring R satisfying $ab = ba$ for all $a, b \in R$. Note that the rings $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Z}$ and $\mathbb{C}$ all satisfy this extra property; hence they are examples of commutative rings. Matrices (see below) provide examples of noncommutative rings.

An *integral domain* is a commutative ring R which has a nonzero identity element, and has no zero divisors. That is, the following condition holds: for all $a, b \in R$, if $ab = 0$ then $a = 0$ or $b = 0$. (The terminology derives from the fact that the set of all integers is an integral domain with respect to the usual operations of addition and multiplication. We shall meet other examples of integral domains later.) An important property of integral domains is the *cancellation law*, which can be derived readily from the assumption that there are no zero divisors:

Cancellation Law: *Let R be an integral domain and $a, b, c \in R$. If $ab = ac$ and $a \neq 0$ then $b = c$.*

This is such a familiar looking property that one may be tempted to assume that it holds in arbitrary ring R. However, it is important to realize that it is not true in that generality: there are many examples of rings (which are not integral domains) in which the cancellation property fails.

† By a function of two variables from S to itself I mean a function whose domain is the set $S \times S = \{ (a, b) \mid a, b \in S \}$ and whose codomain is the set S.

Exercise 3. Prove that integral domains satisfy the cancellation law.

Exercise 4. Give an example of a ring in which the cancellation law fails.

(This last exercise may be too hard at this point, since we do not yet have enough examples of rings at our disposal. It will become easy enough after the subsections on matrix rings and direct products (see below).)

If R is a ring with an identity element and $a \in R$, an element $b \in R$ is called a (multiplicative) *inverse* of a if $ab = ba = 1$. It is not in general true that if $ab = 1$ then $ba = 1$, although it is true that if $ab = 1$ and $ca = 1$ then $b = c$. In particular, if a has an inverse it has only one. In such circumstances the inverse of a is denoted by a^{-1}.

A *field* is a commutative ring with a nonzero identity such that every nonzero element has an inverse. Note that the inverse is required to be an element of the ring in question. The real numbers and the complex numbers are the most important examples of fields. However, $\mathbb{Z}$ is not a field since (for example) the element $2 \in \mathbb{Z}$ does not have an inverse in Z.

Proposition (4.2): *Every field is an integral domain.*

Proof. Let R be a field, and suppose that $a, b \in R$ satisfy $ab = 0$. If $a \neq 0$ then a^{-1} exists (by definition of a field), and now

$$b = 1b = (a^{-1}a)b = a^{-1}(ab) = a^{-1}0 = 0.$$

So R has no zero divisors. Since it is also a commutative ring with 1, it is an integral domain. □

The associative and commutative laws for addition mean that in a sum with many terms there is no harm in permuting and regrouping the terms in any way we please. In particular, it is legitimate to use sigma notation: $\sum_{i=1}^{n} a_i$ for $(\cdots((a_1 + a_2) + a_3) + \cdots) + a_n$. Note that when double sums arise, the order of summation can be changed as usual. For example, it is true in any ring that $\sum_{i=1}^{n}\sum_{j=1}^{i} a_{ij} = \sum_{j=1}^{n}\sum_{i=j}^{n} a_{ij}$. Similarly, it is easily shown that the generalized distibutive laws, $x(\sum_{i=1}^{n} y_i) = \sum_{i=1}^{n} xy_i$ and $(\sum_{i=1}^{n} y_i)x = \sum_{i=1}^{n} y_i x$, are valid in any ring.

If $a_i = a$ for all i from 1 to n, then $\sum_{i=1}^{n} a_i$ is also written as na. In the case that n is a negative integer, na is defined to be equal to $(-n)(-a)$ (whenever a is an element of any ring). Since empty sums (like $\sum_{i=1}^{0} a_i$) are always defined to be zero, na is by definition the zero of R when the integer n is zero. That is, $0_{\mathbb{Z}}a = 0_R$ for all $a \in R$. The function $\mathbb{Z} \times R \to R$ given by $(n, a) \mapsto na$ can be termed *natural multiplication*: the elements na ($n \in \mathbb{Z}$) are called the *natural multiples* of the ring element a. This natural multiplication is certainly different from the multiplication operation in the ring R itself, as that is a function $R \times R \to R$. However, it is unnecessary to distinguish the two notationally as there is no possibility of ambiguity ever arising. It is a tiresome but trivial task to prove the following familiar properties:

$$\begin{aligned}
na + ma &= (n + m)a \\
na + nb &= n(a + b) \\
(nm)a &= n(ma) \\
n(ab) &= (na)b = a(nb)
\end{aligned}$$

(where n, m are arbitrary integers and a, b arbitrary ring elements). Note in particular that if R has an identity then $na = (n1)a = a(n1)$ for all $a \in R$.

The familiar exponent laws are multiplicative analogues of the laws we have just described for natural multiples. If $a \in R$ and n is a positive integer then a^n is defined to be the arbitrarily bracketed product $aa \cdots a$, where there are n factors. If R has an identity element then a^0 is defined to be the identity (in keeping with the universal rule that empty products are defined to be 1). If R has a 1 and $a \in R$ has an inverse then a^n is defined also for negative integers n, by the rule $a^n = (a^{-1})^{(-n)}$. Note that there is no ambiguity when $n = -1$. Now whenever $n, m \in \mathbb{Z}$ and $a \in R$ we have $a^{n+m} = a^n a^m$ and $(a^n)^m = a^{nm}$, provided that the powers concerned are defined. However, the property $(ab)^n = a^n b^n$ will not usually hold, unless $ab = ba$.

If R is a ring and X, Y are subsets of R then we define

$$X + Y = \{\, x + y \mid x \in X \text{ and } y \in Y \,\}.$$

Thus, the sum of two subsets of R is another subset of R. We also define the product of two subsets of R, but the definition of XY is not, as you might at first expect, the set of all products xy where $x \in X$ and $y \in Y$. Rather, it is defined to include also every element of R that is expressible as a sum of an arbitrary number of products of the form xy with $x \in X$ and $y \in Y$. That is,

$$XY = \{\, \sum_{i=1}^{n} x_i y_i \mid 0 \le n \in \mathbb{Z} \text{ and } x_i \in X,\ y_i \in Y \text{ for all } i \,\}.$$

Given these definitions it is easy to prove various desirable properties of addition and multiplication of subsets of an arbitrary ring. Notably,

$$\begin{gathered}(X + Y) + Z = X + (Y + Z)\\ (XY)Z = X(YZ)\\ X(Y + Z) \subseteq XY + XZ\\ (Y + Z)X \subseteq YX + ZX\end{gathered}$$

for all $X, Y, Z \subseteq R$. Equality will hold in these last two if the subsets Y and Z both contain the zero element of R.

5. Homomorphisms and subrings

Definition (5.1): Let R and S be sets which are both equipped with operations of addition and multiplication. A function $f\colon R \to S$ is called a *homomorphism* if $f(a + b) = fa + fb$ and $f(ab) = (fa)(fb)$ for all $a, b \in R$.†

In other words, a homomorphism is a function which preserves the operations in question. More precisely, for each operation in question on the set R there is a corresponding operation on S, and the function must intertwine them, in the following sense: one can apply the operation on two given elements of R and then apply the function to move across to S, or else move to S via the function first, then apply the corresponding operation on S, and the end result will be the same. This is conveniently illustrated by the following diagram, in which the operation on R is denoted by $*$ and that on S by $\odot$:

$$\begin{tikzcd}
R \times R & S \times S \\
(a,b) \arrow[r, mapsto] \arrow[d, mapsto] & (fa,fb) \arrow[d, mapsto] \\
a * b \arrow[r, mapsto] & f(a * b) = (fa) \odot (fb) \\
R & S
\end{tikzcd}$$

† We shall usually write the value of a function f at an element a as fa rather than $f(a)$. The primary use of parentheses is for grouping, indicating the order in which various operations are performed, and we do not wish to clutter up our formulas with unnecessary parentheses.

This is an example of what is known as a *commutative diagram*, meaning that the two possible paths from $R \times R$ to S agree.

Although we have formulated the definition of homomorphism for the case when R and S are equipped with two operations, the same word is also commonly used in the case of algebraic systems which have only one operation, or three (or even more); it is usually clear from the context what kind of homomorphism is meant. In this course the sets R and S will almost always be rings, and so the homomorphisms we encounter will normally be ring homomorphisms.

In the next section we shall describe various examples of rings. Only then, with more rings at our disposal, shall we be able to give interesting examples of homomorphisms. For the time being, we investigate some basic theoretical properties of homomorphisms.

The following is trivial (and clearly generalizes to any number of operations):

Theorem (5.2): *Let R, S and T be sets equipped with addition and multiplication, and let $f\colon R \to S$ and $g\colon S \to T$ be homomorphisms. Then $gf\colon R \to T$, defined by $(gf)a = g(fa)$ for all $a \in R$, is a homomorphism.*

Proof. For all $a, b \in R$,

$$\begin{aligned}(gf)(a+b) &= g(f(a+b))\\ &= g(fa+fb) \quad \text{(since } f \text{ preserves addition)}\\ &= g(fa)+g(fb) \quad \text{(since } g \text{ preserves addition)}\\ &= (gf)a+(gf)b,\end{aligned}$$

and so gf preserves addition. A similar argument shows that gf preserves multiplication. □

Suppose now that R is a ring and $f\colon R \to S$ a homomorphism, where S is a set equipped with addition and multiplication operations (but not necessarily a ring). Let $T = \operatorname{im} f = \{\, fa \mid a \in R \,\}$, a subset of S. It is not hard to show that T must be a ring: for each axiom, the fact that the axiom is satisfied in R combines with the fact that f preserves the operations to ensure that the axiom is satisfied in T. For example, if x, y and z are arbitrary elements of T, then since $T = \operatorname{im} f$ there exist a, b and $c \in R$ with $x = fa$, $y = fb$ and $z = fc$, and now

$$\begin{aligned}x(y+z) &= (fa)(fb+fc)\\ &= (fa)(f(b+c)) \quad \text{(since } f \text{ preserves addition)}\\ &= f(a(b+c)) \quad \text{(since } f \text{ preserves multiplication)}\\ &= f(ab+ac) \quad \text{(by the left distributive law in } R\text{)}\\ &= f(ab)+f(ac) \quad \text{(since } f \text{ preserves addition)}\\ &= (fa)(fb)+(fa)(fc) \quad \text{(since } f \text{ preserves multiplication)}\\ &= xy+xz,\end{aligned}$$

and it follows that the left distributive law is satisfied in T. Clearly similar arguments apply for the other distributive law and for associativity and commutativity of addition and associativity of multiplication. Similarly also, if 0 is the zero element of R then $f0 \in T$ has the properties required of a zero element for T. Finally, if $x \in T$ is arbitrary then we may choose an element $a \in R$ with $fa = x$, and if we now define $y = f(-a) \in T$ we see that

$$x+y = fa+f(-a) = f(a+(-a)) = f0 = f((-a)+a) = f(-a)+fa = y+x,$$

showing that y is a negative of x.

Proposition (5.3): *(i) Let R and S be sets equipped with addition and multiplication operations and let $f\colon R \to S$ be a homomorphism. If R is a ring then* $\operatorname{im} f$ *is a ring.*
(ii) If R and S are rings and $f\colon R \to S$ a homomorphism then $f(0_R) = 0_S$, and $f(-a) = -(fa)$ for all $a \in R$.

Proof. Part (i) was proved in the discussion above. Part (ii) was not quite proved above, since we showed only that $f(0_R)$ is a zero element for $\operatorname{im} f$ rather than for the whole of S. The situation for negatives is similar, highlighting a possible ambiguity of notation that we may be forced to attend to. But the proofs required here are easy. If we put $z = f(0_R)$ then

$$z + z = f(0_R) + f(0_R) = f(0_R + 0_R) = f(0_R) = z,$$

and so

$$z = 0_S + z = (-z + z) + z = -z + (z + z) = -z + z = 0_S,$$

proving the first assertion. And now if $a \in R$ is arbitrary then

$$f(-a) + fa = f(-a + a) = f(0_R) = 0_S,$$

so that by uniqueness of negatives it follows that $f(-a) = -(fa)$ (in S). □

We remark that if R is a ring with an identity element and $\phi\colon R \to S$ is a homomorphism, it is not necessarily true that $\phi(1_R)$ is an identity element for S. For this property to be guaranteed, the homomorphism ϕ has to be surjective.

Recall that a function $f\colon X \to Y$ is bijective if and only if there is a function $g\colon Y \to X$ (called the *inverse* of f and usually denoted by f^{-1}) such that for all $x \in X$ and $y \in Y$,

$$fx = y \text{ if and only if } gy = x.$$

The symmetry of these conditions shows that if f is bijective and $g = f^{-1}$ then g is bijective and $f = g^{-1}$. Now suppose that $f\colon X \to Y$ is bijective and that $*$ is an operation on X. Then an operation $\odot$ can be defined on Y by the rule

$$s \odot t = f(gs * gt) \quad \text{for all } s, t \in Y. \tag{2}$$

Since this equation can be rewritten as

$$g(s \odot t) = gs * gt \quad \text{for all } s, t \in Y, \tag{3}$$

the definition of $\odot$ guarantees that g intertwines $*$ and $\odot$. Conversely, if it is assumed that $\odot$ is an operation on Y satisfying Eq.(3), then Eq.(2) must hold too. Hence the operation $\odot$ is uniquely determined by the operation $*$ and the condition that g intertwines the two. Furthermore, if a and b are arbitrary elements of X and we let $s = fa$ and $t = fb$, so that $a = gs$ and $b = gt$, Eq.(2) becomes $fa \cdot fb = f(a * b)$. Thus f also intertwines the operations, and it follows also that $*$ is determined by $\odot$ in just the same way as $\odot$ is determined by $*$.

Definition (5.4): A bijective homomorphism is called an *isomorphism*. If there is an isomorphism from R to S then R and S are said to be *isomorphic*, and we write $R \cong S$.

By our discussion above, if f is an isomorphism from R to S then the operations on S are determined by f and the operations on R. Furthermore, f^{-1} is an isomorphism from S to R. In view of our earlier result (Proposition (5.3)) that the image of a homomorphism is always a ring, we can say that if R is a ring then S must also be a ring.

Proposition (5.5): *(i) If R is a ring and $f: R \to S$ is a bijective function then there are uniquely determined addition and multiplication operations on S that make S into a ring and f an isomorphism.*
(ii) If $f: R \to S$ is an isomorphism of rings then so is $f^{-1}: S \to R$.

Part (ii) of this proposition shows that $\cong$ is a symmetric relation on rings: if $R \cong S$ then $S \cong R$. It is trivially reflexive ($R \cong R$ for all rings R) since the identity function from R to itself is bijective and preserves addition and multiplication. Furthermore, if $f: R \to S$ and $g: S \to T$ are isomorphisms then $gf: R \to T$ is also, since the composite of two bijective functions is necessarily bijective, and (as shown above) the composite of two homomorphisms is a homomorphism. So $R \cong S$ and $S \cong T$ yield $R \cong T$, whence $\cong$ is transitive. Thus isomorphism of rings is an equivalence relation. (In case you have not encountered equivalence relations before, there is a discussion of them in the next section.)

In situations where we have a fixed isomorphism f from the ring R to the ring S, it is often convenient to regard f as identifying elements of R with elements of S, and actually think of R and S as being equal. As far as the ring structure is concerned there is no harm in doing this, since all properties of the addition and multiplication operations in R will be mirrored in S. That is to say, for every equation that holds in R there is a corresponding equation that holds in S, obtained by just applying f to everything. Conversely, applying f^{-1} transform equations in S to equations in R. In some (admittedly, ill-defined) sense, R and S are realizations of the same abstract ring: all that changes in passing from R to S are the names of the elements.

Definition (5.6): Let R be a ring. A *subring* of R is a subset S of R which is also a ring, with respect to addition and multiplication operations which are the restrictions of the addition and multiplication operations on R.

Thus, if S is a subring of R and $x, y \in S$, then $x + y$ must have the same value whether the addition used is S's addition or R's, and similarly xy must give the same value whether the multiplication is S's or R's. Since $z = 0_S$ is an element of R which is a solution of the equation $0_S + z = 0_S$, it follows from Exercise 1 above that $0_S = 0_R$. So there is no ambiguity in just writing 0 for the zero element. It is also clear by uniqueness of negatives that if $x \in S$ then the negative of x in S coincides with the negative of x in R; so the notation $-x$ will not be ambiguous.

We have already noted that $\mathbb{R}$ and $\mathbb{Z}$ are rings, and we also know that $\mathbb{Z} \subseteq \mathbb{R}$. So to check that $\mathbb{Z}$ is a subring of $\mathbb{R}$ it remains to note that addition and multiplication of integers is consistent with addition and multiplication of real numbers. For example, 2 times 3 is 6, whether you are thinking of the numbers concerned as integers or as real numbers.

Note that if S is a subring of R then the sum and product in R of two elements of S always yield elements of S. That is, a subset S of a ring R cannot be a subring unless it is closed under the operations of R, in the sense of the following definition:

Definition (5.7): If $*$ is an operation on a set R and S is a subset of R, we say that S is *closed under* $*$ if $x * y \in S$ for all $x, y \in S$.

In this situation (when the subset S is closed under the operation $*$), we can define an operation $\odot$ on S by the rule that $a \odot b = a * b$ for all $a, b \in S$. Normally one would not use different notations for these two operations: $\odot$ would simply be written as $*$. We say that S *inherits* the operation $*$ from R.

In accordance also with the definition above, if S is a subset of a ring R we shall say that S is *closed under taking negatives* if $-a \in S$ whenever $a \in S$. If S is a subring of R then uniqueness of negatives in R and in S combine to show that S necessarily satisfies this closure property also.

In the converse direction, we have the following theorem.

Theorem (5.8): *A subset S of a ring R is a subring of R if $S \neq \emptyset$ and S is closed under addition, multiplication and taking negatives.*

We leave the proof of Theorem (5.8) as an exercise.

Let $R = \mathrm{Mat}_3(\mathbb{R})$, the set of all 3×3 matrices over $\mathbb{R}$. We shall prove below (or, at least, give an indication of the proof) that R is a ring under the usual operations of matrix addition and multiplication. Let us use the criterion above to show that

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

is a subring of R.

Observe first that S is nonempty, since the zero matrix is an element of S. Now let A and B be arbitrary elements of S. Then

$$A = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} a' & b' & c' \\ 0 & d' & e' \\ 0 & 0 & f' \end{pmatrix}$$

for some real numbers $a, a', b, b', c, c', d, d', e, e', f$ and f'. We find that

$$A + B = \begin{pmatrix} a + a' & b + b' & c + c' \\ 0 & d + d' & e + e' \\ 0 & 0 & f + f' \end{pmatrix}$$

$$AB = \begin{pmatrix} aa' & ab' + bd' & ac' + be' + cf' \\ 0 & dd' & de' + ef' \\ 0 & 0 & ff' \end{pmatrix}$$

and

$$-A = \begin{pmatrix} -a & -b & -c \\ 0 & -d & -e \\ 0 & 0 & -f \end{pmatrix}$$

are all in the set S since the below diagonal entries are zero in each case. Hence the required closure properties hold, and so S is a subring.

Since a field is just a special kind of ring, it is permissible for fields to have subrings. Indeed, we have already noted that $\mathbb{Z}$ is a subring of $\mathbb{R}$. If a subring of a field happens to itself be a field (under the inherited operations) then it is called a *subfield*. Of course, $\mathbb{Z}$ is not a subfield of $\mathbb{R}$, because $\mathbb{Z}$ is not a field, but examples of subfields are not hard to find. For example, $\mathbb{Q}$ (rational numbers) is a subfield of $\mathbb{R}$, and $\mathbb{R}$ is a subfield of $\mathbb{C}$. Analogous to Theorem (5.8), we have the following criterion for a subset of a field to be a subfield.

Theorem (5.9): *A subset S of a field F is a subfield of F if and only if 0_F, $1_F \in S$ and the following closure properties are satisfied:*
(i) *$x + y$, xy and $-x$ are in S whenever $x, y \in S$;*
(ii) *$x^{-1} \in S$ whenever $x \in S$ and $x \neq 0$.*

The proof of this is also left as an exercise for the reader. The next result gives a glimpse of the way in which this theory can be applied to the geometrical problems we discussed in Section 3 above.

Theorem (5.10): *The set K of all constructible numbers is a subfield of $\mathbb{R}$*

Proof. Suppose that $a \in K$. By Theorem (3.2) there exists a sequence $0, 1, a_2, a_3, \ldots, a_n$ such that $a_n = a$ and each a_i in the sequence is the sum, product, negative, inverse or square root of a term or terms earlier in the sequence. If also $b \in K$ then there is a sequence $0, 1, b_2, \ldots, b_m = b$ with the same property. And now the sequence $0, 1, a_2, \ldots, a_n, b_2, \ldots, b_m, a + b$ also has this property, since the final term, $a + b$, is the sum of the earlier terms a_n and b_m, while each a_i or b_j can be obtained from earlier ones in the required way (since this is so for the original two sequences). Hence by Theorem (3.2) it follows that $a + b \in K$. Similar arguments show that ab, $-a \in K$, and $a^{-1} \in K$ (if $a \neq 0$). So K has all the required closure properties, and since Theorem (3.2) also guarantees that $0, 1 \in K$, it follows from Theorem (5.9) that K is a subfield of $\mathbb{R}$. □

The next theorem is another for which the reader should be able to provide a proof.

Theorem (5.11): *Let R be a ring and $S_1, S_2, \ldots, S_n$ subrings of R. Then $S_1 \cap S_2 \cap \cdots \cap S_n$ is a subring of R.*

The next result is little more than a rephrasing of Proposition (5.3).

Proposition (5.12): *If $\phi\colon R \to S$ is a ring homomorphism then $\operatorname{im}\phi$ is a subring of S.*

Any function $\phi\colon R \to S$ determines another function $\widehat{\phi}\colon R \to \operatorname{im}\phi$ by the rule that $\widehat{\phi}x = \phi x$ for all $x \in R$. If ϕ is surjective then ϕ and $\widehat{\phi}$ are exactly the same as each other. Otherwise, the only difference between ϕ and $\widehat{\phi}$ is that their codomains are different; in particular, the codomain of $\widehat{\phi}$ is a proper subset of the codomain of ϕ. In fact, most people would not bother to introduce a separate name for the function $\widehat{\phi}$, instead speaking of "ϕ regarded as a function from R to $\operatorname{im}\phi$", or some such phraseology. Strictly speaking, though, they are different functions. And, indeed, whereas ϕ need not be surjective, $\widehat{\phi}$ is necessarily surjective, since for every y in the codomain of $\widehat{\phi}$ there is an $x \in R$ with $y = \phi x$ (since the codomain of $\widehat{\phi}$ is $\operatorname{im}\phi$), and this gives $y = \widehat{\phi}x$. In the definition of $\widehat{\phi}$ all superfluous elements of the codomain have been excluded, and a surjective function results.

In the case that the original function ϕ is injective, the function $\widehat{\phi}$ is injective also, and consequently $\widehat{\phi}$ is bijective in this case. Furthermore, if R and S are equipped with addition and multiplication operations, and if ϕ is a homomorphism, then for all $x, y \in R$,

$$\widehat{\phi}(x + y) = \phi(x + y) = \phi x + \phi y = \widehat{\phi}x + \widehat{\phi}y$$
$$\widehat{\phi}(xy) = \phi(xy) = (\phi x)(\phi y) = (\widehat{\phi}x)(\widehat{\phi}y).$$

It follows readily that $\operatorname{im}\phi$ is closed under the addition and multiplication of S, so that it inherits addition and multiplication from S; moreover, $\widehat{\phi}$ is a homomorphism from R to $\operatorname{im}\phi$. In particular, if R and S are rings and ϕ is an injective homomorphism from R to S then R is isomorphic to the subring $\operatorname{im}\phi$ of S, the function $\widehat{\phi}$ being an isomorphism. Bearing in mind our earlier remarks that it is sometimes useful to regard two isomorphic rings as actually being equal, it is often convenient in this context to identify R with $\operatorname{im}\phi$, and say that R is a subring of S. A more conservative terminology would be to say that $\operatorname{im}\phi$ is a copy of R embedded in S. In this spirit, an injective homomorphism is often called an *embedding*.

To illustrate this, consider the fields $\mathbb{R}$ and $\mathbb{C}$. Some people define $\mathbb{C}$ to be the set of all ordered pairs of real numbers, with addition and multiplication defined by the rules

$$(x, y) + (u, v) = (x + u, y + v)$$
$$(x, y)(u, v) = (xu - yv, xv + yu)$$

for all real numbers x, y, u and v. It is probably more usual to use the notation $x + iy$, where i is a so-called imaginary number satisfying $i^2 = -1$, rather than (x, y). The $x + iy$ notation certainly makes it look as though $\mathbb{R}$ is a subset of $\mathbb{C}$, since surely the real number x is the same as the complex number $x + 0i$. But if the ordered pair notation is used then the question of whether $\mathbb{R}$ is a subset of $\mathbb{C}$ becomes more problematic, since it is not clear that we can say that x is the same as the ordered pair $(x, 0)$. The formal solution to this problem is to say simply that $x \mapsto x + i0$ (or $(x, 0)$) is an embedding of $\mathbb{R}$ in $\mathbb{C}$. For practical purposes this is good enough to allow us to say that $\mathbb{R}$ is a subring of $\mathbb{C}$.

Suppose that the ring R is embedded in the ring S, as above. If we really felt uneasy about saying that R is a subring of S, and wanted to find a ring that contains R, rather than an isomorphic copy of R, we could proceed as follows. First, find a set T which is disjoint from R and in one to one correspondence with the set $\{\, s \in S \mid s \notin \operatorname{im}\phi \,\}$. This one to one correspondence can then be extended to a one to one correspondence between $T \cup R$ and S, so that each $x \in R$ corresponds to $\phi r \in \operatorname{im}\phi$. In other words, we now have a set $S' = T \cup R$ and a bijective function $\psi\colon S' \to S$ such that the restriction of ψ to R is the originally given embedding $\phi\colon R \to S$. Now by Proposition (5.5) there is a unique way to define addition and multiplication on S' so that S' becomes a ring and the bijective function ψ a ring isomorphism. It is fairly easy to see that, when this is done, S' is a ring having R as a subring.

If $\phi\colon R \to S$ is a ring homomorphism we define the *kernel* of ϕ to be the set

$$\ker\phi = \{\, x \in R \mid \phi x = 0_S \,\}.$$

Kernels are extremely important in ring theory, and in algebra in general. They are subrings with an important extra closure property.

Theorem (5.13): *Let R and S be rings and $\phi\colon R \to S$ a homorphism. Then $\ker\phi$ is a subring of R; moreover, if $x \in \ker\phi$ and a is an arbitrary element of R then ax, $xa \in \ker\phi$.*

Proof. Note first that $\ker\phi \neq \emptyset$ since $0_R \in \ker\phi$. Furthermore, if x, $y \in \ker\phi$ then

$$\phi(x + y) = \phi x + \phi y = 0 + 0 = 0,$$
$$\phi(xy) = (\phi x)(\phi y) = 0\,0 = 0,$$

and

$$\phi(-x) = -(\phi x) = -0 = 0,$$

whence $\phi(x + y)$, $\phi(xy)$, $\phi(-x) \in \ker\phi$. So, by Theorem (5.8), $\ker\phi$ is a subring.

Now suppose that $x \in \ker\phi$ and $a \in R$. Then

$$\phi(ax) = (\phi a)(\phi x) = (\phi a)0 = 0,$$
$$\phi(xa) = (\phi x)(\phi a) = 0(\phi a) = 0,$$

whence ax, $xa \in \ker\phi$, as required. □

The following result provides a convenient way to check whether or not a homomorphism is one-to-one.

Proposition (5.14): *Let $\phi\colon R \to S$ be a ring homomorphism. Then ϕ is injective if and only if* $\ker\phi = \{0\}$.

Proof. We have already shown in Proposition (5.12) (ii) that $0_R \in \ker\phi$. Hence $\ker\phi = \{0_R\}$ if and only if $\ker\phi \subseteq \{0_R\}$.

Suppose first that ϕ is injective, and let $a \in \ker\phi$. Then

$$\phi a = 0_S = \phi 0_R$$

(by Proposition (5.12) (ii)), and since ϕ is injective it follows that $a = 0_R$. Thus 0_R is the only element of $\ker\phi$, and so $\ker\phi = \{0_R\}$, as required.

Conversely, suppose that $\ker\phi = \{0_R\}$, and suppose that $a, b \in R$ with $\phi a = \phi b$. Since ϕ is a homomorphism we have by Proposition (5.12) (ii) that

$$\phi(a - b) = \phi(a + (-b)) = \phi a + \phi(-b) = \phi a + (-\phi b) = 0_S,$$

and hence $a - b \in \ker\phi$. By hypothesis 0_R is the only element of $\ker\phi$; so $a - b = 0_R$, and thus $a = b$. We have shown that $\phi a = \phi b$ implies $a = b$; that is, ϕ is injective, as required. □

6. Constructing new rings from old

Direct products

If R_1 and R_2 are are rings then

$$R_1 \times R_2 = \{\, (r_1, r_2) \mid r_1 \in R_1, r_2 \in R_2 \,\}$$

becomes a ring if we define

$$\begin{aligned}(r_1, r_2) + (s_1, s_2) &= (r_1 + s_1, r_2 + s_2),\\ (r_1, r_2)(s_1, s_2) &= (r_1 s_1, r_2 s_2).\end{aligned}$$

More generally, let $(R_i)_{i\in I}$ be an indexed family of rings. This just means that for each $i \in I$ there is given a ring R_i. Here I can be any set. Define

$$P = \{\, (r_i)_{i\in I} \mid r_i \in R_i \text{ for all } i \in I \,\}.$$

(As an aid to intuition, it may help to think of the case when I is the set of all positive integers; then $(R_i)_{i\in I}$ is a sequence of rings R_1, R_2, R_3, ... , and elements of P are sequences $(r_1, r_2, r_3, \ldots)$ such that $r_i \in R_i$ for each i.) Define addition and multiplication on P by

$$\begin{aligned}(r_i)_{i\in I} + (s_i)_{i\in I} &= (r_i + s_i)_{i\in I}\\ (r_i)_{i\in I}(s_i)_{i\in I} &= (r_i s_i)_{i\in I}.\end{aligned}$$

Then P becomes a ring, called the *direct product* of the rings R_i. To prove this is a routine exercise in checking that the ring axioms are satisfied in P, given that they are satisfied in all the rings R_i. We do one case as an illustration, leaving the others as exercises.

Let x, y, $z \in P$. Then there exist elements r_i, s_i, $t_i \in R_i$ (for each $i \in I$) such that $x = (r_i)_{i\in I}$, $y = (s_i)_{i\in I}$ and $z = (t_i)_{i\in I}$. Now by the definitions of addition and multiplication in P we find that

$$\begin{aligned}x(y + z) &= (r_i)_{i\in I}((s_i)_{i\in I} + (t_i)_{i\in I})\\ &= (r_i)_{i\in I}((s_i + t_i)_{i\in I})\\ &= (r_i(s_i + t_i))_{i\in I},\end{aligned}$$

and similarly

$$\begin{aligned} xy + xz &= (r_i)_{i\in I}(s_i)_{i\in I} + (r_i)_{i\in I}(t_i)_{i\in I}) \\ &= (r_is_i)_{i\in I} + (r_it_i)_{i\in I} \\ &= (r_is_i + r_it_i)_{i\in I}. \end{aligned}$$

Since each of the rings R_i satisfies the distributive law we have that $r_i(s_i + t_i) = r_is_i + r_it_i$ for all $i \in I$, and hence $x(y + z) = xy + xz$.

Let S be the subset of P consisting of all $(r_i)_{i\in I}$ such that r_i is nonzero for at most finitely many values of i. It can be shown readily that S is closed under addition, multiplication and taking negatives; hence S is a subring of P.

Matrices

Let R be any ring and n a positive integer. Let $\mathrm{Mat}_n(R)$ be the set of all $n \times n$ matrices with entries from R, and define multiplication and addition as usual for matrices. Thus, denoting the (i, j) entry of a matrix A by A_{ij}, the definitions of addition and multiplication are as follows:

$$\begin{aligned} (X + Y)_{ij} &= X_{ij} + Y_{ij}, \\ (XY)_{ij} &= \sum_{k=1}^{n} X_{ik}Y_{kj}, \end{aligned}$$

for all $X, Y \in \mathrm{Mat}_n(R)$ and $i, j \in \{1, 2, \ldots, n\}$.

It can be shown that $\mathrm{Mat}_n(R)$ is a ring. Again, the proof is simply a matter of verifying that $\mathrm{Mat}_n(R)$ satisfies the ring axioms, given that R does. We shall do only the associative law for multiplication (which is the hardest).

Let $X, Y, Z \in \mathrm{Mat}_n(R)$. By the definitions of the matrix operations we find that

$$((XY)Z)_{ij} = \sum_{k=1}^{n}(XY)_{ik}Z_{kj} = \sum_{k=1}^{n}\Big(\sum_{l=1}^{n} X_{il}Y_{lk}\Big)X_{kj} = \sum_{k=1}^{n}\sum_{l=1}^{n}(X_{il}Y_{lk})X_{kj},$$

where the last step follows from a generalized distributive law. Similarly,

$$(X(YZ))_{ij} = \sum_{l=1}^{n} X_{il}(YZ)_{lj} = \sum_{l=1}^{n} X_{il}\Big(\sum_{k=1}^{n} Y_{lk}X_{kj}\Big) = \sum_{l=1}^{n}\sum_{k=1}^{n} X_{il}(Y_{lk}X_{kj}).$$

Now since R satisfies the associative law for multiplication we have that $(X_{il}Y_{lk})X_{kj} = X_{il}(Y_{lk}X_{kj})$ for all i, j, k and l, and hence $(XY)Z = X(YZ)$.

Exercise 5. Show that if R and S are any rings then $\mathrm{Mat}_2(R \times S) \cong \mathrm{Mat}_2(R) \times M_2(S)$ (where '$\times$' means 'direct product', and '$\cong$' means 'is isomorphic to').

Exercise 6. Show that $\mathrm{Mat}_2(\mathrm{Mat}_3(\mathbb{Z})) \cong \mathrm{Mat}_6(\mathbb{Z})$.

Formal power series

Let R be any ring and let $R[[x]] = \{\, (r_0, r_1, r_2, \ldots) \mid r_i \in R \,\}$, the set of all infinite sequences of elements of R. We have already seen one way of defining addition and multiplication on this set that produces a ring: the direct product $\prod_{i=0}^{\infty} R_i$, where $R_i = R$ for each i. However, alternative definitions of the operations are possible, and in particular we can define

$$\begin{aligned} (r_0, r_1, r_2, \ldots) + (s_0, s_1, s_2, \ldots) &= (r_0 + s_0, r_1 + s_1, r_2 + s_2 \ldots) \\ (r_0, r_1, r_2, \ldots)(s_0, s_1, s_2, \ldots) &= (r_0s_0, r_0s_1 + r_1s_0, r_0s_2 + r_1s_1 + r_2s_0, \ldots). \end{aligned}$$

Thus, addition is defined in the same way as it was for the direct product, but multiplication is given by a formula which may seem rather strange at first. To make it seem more natural, and to avoid confusion with the direct product, we shall use a different notation. When considered as an element of $R[[x]]$ the sequence $(r_0, r_1, r_2, \ldots)$ will be written as $r_0 + r_1x + r_2x^2 + \cdots$. The plus signs and the x's that appear here are to be regarded as meaningless symbols: $r_0 + r_1x + r_2x^2 + \cdots$ is just a bizarre notation for $(r_0, r_1, r_2, \ldots)$. The point of it is that with this notation the rules for addition and multiplication in $R[[x]]$ become

$$\begin{aligned}(r_0 + r_1x + r_2x^2 + \cdots) + (s_0 + s_1x + s_2x^2 + \cdots) &= (r_0 + s_0) + (r_1 + s_1)x + (r_2 + s_2)x^2 + \cdots \\ (r_0 + r_1x + r_2x^2 + \cdots)(s_0 + s_1x + s_2x^2 + \cdots) &= r_0s_0 + (r_0s_1 + r_1s_0)x + \\ &\qquad (r_0s_2 + r_1s_1 + r_2s_0)x^2 + \cdots ;\end{aligned}$$

the right hand side is obtained by expanding the left hand side as one would if the plus signs did stand for addition and the symbol x did stand for a ring element, and then collecting like terms. But infinite sums cannot be defined in a general ring; so one must be aware that if the symbol x is replaced by an element of R the result will probably be nonsense.

Checking that the ring axioms are satisfied is again routine. We call $R[[x]]$ the ring of *formal power series* over R in the *indeterminate* x. It is easy to show that the function $\psi\colon R \to R[[x]]$ defined by $\psi a = a + 0x + 0x^2 + \cdots$ is a ring homomorphism. It is also injective. The image of ψ is thus a subring of $R[[x]]$ isomorphic to R. Of course we would normally write the formal power series $a + 0x + 0x^2 + \cdots$ simply as a, which has the effect of making $\operatorname{im}\psi$ indistinguishable from R. This is no cause for concern, as it is usually desirable to identify R with the image of ψ, and thus regard R as a subring of $R[[x]]$.

Note that if the ring R is commutative then so is $R[[x]]$, as can be seen readily from the formula for the product of two elements of $R[[x]]$. If R has a 1 then so does $R[[x]]$: it is easily checked that the series $1 = 1 + 0x + 0x^2 + \cdots$ is an identity for the multiplication as defined above, given that 1 is an identity for multiplication in R. Furthermore, if R has a 1 then the symbol x can be regarded as an element of $R[[x]]$ by identifying it with the series $0 + 1x + 0x^2 + \cdots$. (It is somewhat annoying that R has to have an identity element before this identification can be made. Formal power series are thus more pleasant to contemplate if R has a 1 than if it does not).

Polynomials

Define $R[x]$ to be the subset of $R[[x]]$ consisting of those formal power series $\sum_{i=0}^{\infty} r_ix^i$ such that r_i is nonzero for at most finitely many values of i. Thus each element of $R[x]$ can be written in the form $\sum_{i=0}^{n} r_ix^i$, where n is a nonnegative integer. The set $R[x]$ is nonempty (it contains the zero series $\sum_{i=1}^{\infty} 0x^i$) and closed under addition, multiplication and taking negatives; hence it is subring of $R[[x]]$. It is called the ring of *polynomials* over R in the indeterminate x.

Polynomials are of central theoretical importance to the study of rings and fields, since, as we shall see later, they provide a method of constructing extension fields. (A field E is an *extension* of F if F is a subfield of E.) Closely tied in with this construction are a class of homomorphisms, known as *evaluation homomorphisms*, which we now describe.

Let T be a commutative ring and R a subring of T, and let α be an arbitrary element of T. Define a function $\operatorname{eval}_\alpha\colon R[x] \to T$ by the rule that

$$\operatorname{eval}_\alpha(r_0 + r_1x + r_2x^2 + \cdots + r_dx^d = r_0 + r_1\alpha + r_2\alpha^2 + \cdots + r_d\alpha^d$$

for all nonnegative integers d and $r_0, r_1, \ldots, r_d \in R$. In other words, if p is a polynomial then $\operatorname{eval}_\alpha(p)$ is what you get by replacing the indeterminate x by the element α. We shall frequently find it convenient to use the notation $p(x)$ (rather than just p) for a polynomial, and then write $p(\alpha)$ for $\mathrm{eval}_\alpha(p(x))$.

Theorem (6.1): *Let R be a subring of the commutative ring T, and let $\alpha \in T$. Then the function* $\mathrm{eval}_\alpha\colon R[x] \to T$ *is a homomorphism.*

The proof of this is an immediate consequence of the way addition and multiplication of polynomials is defined. Thus, suppose that $p(x) = r_0 + r_1x + \cdots + r_dx^d$ and $q(x) = s_0 + s_1x + \cdots s_ex^e$ are elements of $R[x]$. Then by the definition of multiplication in $R[x]$,

$$p(x)q(x) = r_0s_0 + (r_1s_0 + r_0s_1)x + (r_2s_0 + r_1s_1 + r_0s_2)x^2 + \cdots + r_ds_ex^{d+e},$$

and thus we see that

$$\begin{aligned}\mathrm{eval}_\alpha(p(x)q(x)) &= r_0s_0 + (r_1s_0 + r_0s_1)\alpha + (r_2s_0 + r_1s_1 + r_0s_2)\alpha^2 + \cdots + r_ds_e\alpha^{d+e}\\ &= (r_0 + r_1\alpha + \cdots + r_d\alpha^d)(s_0 + s_1\alpha + \cdots s_e\alpha^e)\\ &= \mathrm{eval}_\alpha(p(x))\mathrm{eval}_\alpha(q(x)).\end{aligned}$$

So eval_α preserves multiplication. The proof that it preserves addition is similar.

Note that the above theorem is false if the assumption that the ring T is commutative is dropped. This is because in $R[x]$ the product $(ax^i)(bx^j)$ is defined to be abx^{i+j}; however, if the indeterminate x is replaced by the ring element t the resulting equation $(at^i)(bt^j) = abt^{i+j}$ is false (probably) unless $tb = bt$. When dealing with noncommutative rings the best we can say is that an evaluation map $\mathrm{eval}_t\colon R[x] \to T$ is a homomorphism if t commutes with all elements of R.

Example

Let $\phi\colon \mathbb{Z}[x] \to \mathbb{R}$ be evaluation at $\sqrt{2}$. That is, $\phi = \mathrm{eval}_{\sqrt{2}}$ is the function which takes an arbitrary integer polynomial $f(x)$ to $f(\sqrt{2})$. Thus $\phi(x^2 - 2x + 1) = 2 - 2\sqrt{2} + 1 = 3 - \sqrt{2}$. The function ϕ is a homomorphism. Recall that this simply means that you get the same answer whether you replace x by $\sqrt{2}$ before or after performing an addition or multiplication.

Each nonzero $a(x) \in R[x]$ can be uniquely written in the form $a_0 + a_1x + \cdots + a_dx^d$ with $a_d \neq 0$. The integer d is called the *degree* of the polynomial $a(x)$, and a_d is called the *leading coefficient*. Note that the zero polynomial does not have a leading coefficient. For a reason to be explained below, we say that the degree of the zero polynomial is $-\infty$. Polynomials such that the coefficient of x^i is zero for all $i \geq 1$ are called *scalar polynomials* or *constant polynomials*. Observe that a polynomial has degree zero if and only if it is a nonzero scalar polynomial.

Degrees and leading coefficients are of most help in the case that the coefficient ring R is an integral domain. This is because of the following proposition.

Proposition (6.2): *Let R be an integral domain and let a, b be nonzero polynomials in $R[x]$. Then ab is nonzero,* $\deg(ab) = \deg a + \deg b$, *and the leading coefficient of ab is the product of the leading coefficients of a and b.*

Proof. Let $a = a_0 + a_1x + \cdots + a_dx^d$ and $b = b_0 + b_1x + \cdots + b_ex^e$ where a_d and b_e are nonzero, so that d and e are the degrees of a and b respectively, and a_d and b_e the leading coefficients. Expanding and collecting like terms gives

$$ab = a_0b_0 + (a_1b_0 + a_0b_1)x + \cdots + a_db_ex^{d+e}$$

(the coefficient of x^n being $\sum_{i+j=n} a_ib_j = \sum_{i=\max(n-e,0)}^{\min(n,d)} a_ib_{n-i}$, which has just one term if $n = d + e$ and no terms for $n > d + e$.) Now since a_d and b_e are nonzero a_db_e must be nonzero, since R has no zero divisors. Thus ab is nonzero (since at least one of its coefficients is nonzero), $a_d b_e$ is the leading coefficient of ab, and the degree of ab is $d + e = \deg a + \deg b$, as required. □

Observe that if a is the zero polynomial then ab is zero also, for any polynomial b. So if $\deg(ab) = \deg a + \deg b$ is to remain valid we require that $\deg 0 = \deg 0 + \deg b$ for all b; hence we define $\deg 0 = -\infty$.

An important corollary of Proposition (6.2) is that if R is an integral domain then so is $R[x]$.

Theorem (6.3): *Let R be an integral domain. Then $R[x]$ is an integral domain.*

Proof. We must show that $R[x]$ is a commutative ring with 1 having no zero divisors, given that R itself has these properties. We have already noted in the course of our previous discussions that $R[x]$ is commutative and has a 1 when R is commutative and has a 1; so all that remains is the issue of zero divisors. But Proposition (6.2) showed that the product of nonzero elements of $R[x]$ is nonzero, as required. □

The following exercise is not hard. The result will be used at some point later on in the course.

Exercise 7. Suppose that R and S are rings, and $\theta\colon R \to S$ a homomorphism. Show that there is a homomorphism $\widetilde{\theta}\colon R[x] \to S[x]$ given by

$$\widetilde{\theta}(a_0 + a_1x + \cdots + a_dx^d) = \theta a_0 + (\theta a_1)x + \cdots + (\theta a_d)x^d$$

for all nonnegative integers d and $a_0, a_1, \ldots, a_d \in R$. Show also that if θ is injective then so is $\widetilde{\theta}$, and if θ is surjective then so is $\widetilde{\theta}$.

The field of fractions of an integral domain

We shall prove the following theorem.

Theorem (6.4): *Let R be an integral domain. Then there exists a field F such that*
(a) R is a subring of F, and
(b) every element of F is expressible in the form ab^{-1} with $a, b \in R$ and $b \neq 0$.

Before starting the proof, let us consider a particular case. If $R = \mathbb{Z}$ (the integers) then R is an integral domain; moreover, it is easily seen that if we put $F = \mathbb{Q}$ (the rational numbers) then properties (a) and (b) are satisfied. But if we are given only $\mathbb{Z}$, and not $\mathbb{Q}$, is there a way that we can, in some sense, construct $\mathbb{Q}$ from $\mathbb{Z}$? Rational numbers are usually written as quotients of pairs of integers, that is, in the form a/b where $a, b \in \mathbb{Z}$ and $b \neq 0$. So our first attempt at constructing $\mathbb{Q}$ might be to define $\mathbb{Q} = \{\, (a,b) \mid a, b \in \mathbb{Z} \text{ and } b \neq 0 \,\}$, and then define addition and multiplication on this set in such a way that the rules for adding and multiplying ordered pairs correspond to the familiar rules for adding and multiplying fractions. So, we would define

$$(a,b) + (c,d) = (ad + bc, bd) \quad \text{and} \quad (a,b)(c,d) = (ac, bd).$$

However, this does not work, and a moment's thought tells us why: the correspondence between pairs of integers (a,b) (where $b \neq 0$) and rational numbers is not one to one. The same rational number can be written in many ways as a quotient of a pair of integers: $3/4 = 6/8 = 9/12$, for instance. The technical mathematical device for dealing with this is to define a notion of equivalence on the set of pairs (a,b), in such a way that two pairs are equivalent if they correspond to the same rational number. Rather than attempting to identify a rational number with a pair of integers, we should identify a rational number with a class of equivalent pairs of integers.

Definition (6.5): Let $\sim$ be a relation on a set S, so that for every pair of elements $x,\ y \in S$ either $x \sim y$ (x is related to y) or else $x \not\sim y$ (x is not related to y). Then $\sim$ is called an *equivalence relation* if

(1) $x \sim x$ for all $x \in S$,
(2) for all $x,\ y \in S$, if $x \sim y$ then $y \sim x$, and
(3) for all $x,\ y,\ z \in S$, if $x \sim y$ and $y \sim z$ then $x \sim z$.

A relation $\sim$ is said to be *reflexive* if it satisfies (1), *symmetric* if it satisfies (2) and *transitive* if it satisfies (3). Hence, some books call equivalence relations RST-relations.

It is fairly easy to see that if an equivalence relation $\sim$ on a set S always partitions S into non-overlapping subsets, in such a way that elements $a,\ b \in S$ lie in the same subset if and only if $a \sim b$. These subsets are called *equivalence classes*. For each $a \in S$ we define

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

the equivalence class of the element x. Every element of S lies in some equivalence class (since, indeed, $x \in \overline{x}$), and so S is the union of all the equivalence classes. Furthermore, $\overline{x} = \overline{y}$ if and only if $x \sim y$, and if $x \not\sim y$ then $\overline{x}$ and $\overline{y}$ are disjoint.

The set of all equivalence classes is called the *quotient* of S by the equivalence relation $\sim$. That is, the quotient is the set

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

The notation $S/{\sim}$ is also commonly used for the quotient of S by $\sim$, but be aware that despite the use of the division sign / and the term “quotient”, this business has nothing much to do with ordinary division of numbers! The terminology seems to stem from the fact that an equivalence relation on S divides S up into equivalence classes.

There is an obvious function from S onto $\overline{S}$ given by $x \mapsto \overline{x}$, and called the *canonical surjection* from S to $\overline{S}$. Note that since $\overline{x} = \overline{y}$ whenever x and y are equivalent, this function from S to $\overline{S}$ is not likely to be one to one, although it is always onto. The set $\overline{S}$ is smaller than S. Intuitively, one should think of $\overline{S}$ as what you get from S if you identify equivalent elements.

Incidentally, we have just introduced a set whose elements are themselves sets. Admittedly, this is a rather abstract notion, and it may take a little getting used to. However, be warned that in general an algebraist would think nothing of having a set whose elements are sets whose elements are sets whose elements are sets ..., to any level you care to mention. So get used to the idea! Incidentally, the customary approach to the foundations of mathematics is to base everything on set theory; so, at least in the view of some people, every mathematical object is a set.

It is high time we proved Theorem (6.4). The proof will occupy rather a lot of space.

Proof of Theorem (6.4). We are given that R is an integral domain. Define

$$S = \{\, (a,b) \mid a,\ b \in R \text{ and } b \neq 0 \,\}.$$

Now let us define a relation on S as follows: if $(a,b),\ (c,d) \in S$ we say that (a,b) is *proportional* to (c,d) if and only if $ad = bc$. Write $(a,b) \sim (c,d)$ if (a,b) is proportional to (c,d).

It turns out that proportionality is an equivalence relation on S. Firstly,

(i) $x \sim x$ for all $x \in S$,

since x must have the form (a,b) for some $a,\ b \in R$ with $b \neq 0$, and the condition that $(a,b) \sim (a,b)$ (found by putting $c = a$ and $d = b$ in the definition above) is $ab = ba$, which is satisfied since R, being an integral domain, is a commutative ring. So proportionality is reflexive. Next,

(ii) if $x,\ y \in S$ and $x \sim y$ then $y \sim x$.

For, if $x = (a, b)$ and $y = (c, d)$ then $x \sim y$ gives $ad = bc$, whence $cb = da$, giving $(c, d) \sim (a, b)$, as required. Thus proportionality is symmetric. Finally,

(iii) if x, y, $z \in S$ and $x \sim y$ and $y \sim z$, then $x \sim z$.

To see this, let $x = (a, b)$, $y = (c, d)$ and $z = (e, f)$. Then $x \sim y$ and $y \sim z$ give $ad = bc$ and $cf = de$. We multiply the first of these equations by f and the second by b, and deduce (using the associative law) that

$$(ad)f = (bc)f = b(cf) = b(de).$$

Using the fact that R is commutative and using the associative law again we can rearrange this equation as $d(af) = d(be)$. Now by the cancellation law for integral domains, bearing in mind that $d \neq 0$ (since $y \in S$), we deduce that $af = be$, and hence $(a, b) \sim (e, f)$. Thus proportionality is transitive.

Having now proved that proportionality is an equivalence relation, let us define F to be the quotient of F by this equivalence relation. Recall that this means that F is the set of all equivalence classes; each element of F is an equivalence class. We now introduce some notation which will have the effect of making some things appear familiar when they actually are not: whenever $(a, b) \in S$ we define

$$a/b \overset{\text{def}}{=} \{\, (c, d) \in S \mid (a, b) \sim (c, d) \,\}.$$

In other words, we are using the notation a/b to stand for the equivalence class containing (a, b); this is the object that would be written as $\overline{(a, b)}$ if we were to use the conventions we used above in our discussion of arbitrary equivalence relations. The point of using this notation is that it highlights the fact that the algebraic properties of these equivalence classes that we are about to investigate mirror familiar properties of rational numbers. Note, first of all, the following criterion for equality of these pseudo-fractions:

Fact (6.6): *If p, p', q, $q' \in S$ with q, q' nonzero, then $p/q = p'/q'$ if and only if $pq' = qp'$.*

This follows from the fact, noted in our discussion of equivalence relations, that the equivalence class containing (p, q) equals that containing (p', q') if and only if $(p, q) \sim (p', q')$. That is (by the definitions), $p/q = p'/q'$ if and only if $pq' = qp'$, as claimed.

The following lemma will permit us to define operations of addition and multiplication on the set F.

Lemma (6.7): *Let p, p', q, q' r, r', s, s' be elements of R, with q, q', s, s' nonzero. If $p/q = p'/q'$ and $r/s = r'/s'$ then $(ps + qr)/qs = (p's' + q'r')/q's'$ and $pr/qs = p'r'/q's'$.*

Proof. Assuming $p/q = p'/q'$ and $r's = r'/s'$ gives $pq' = qp'$ and $rs' = sr'$. Hence we deduce, successively, that

$$\begin{aligned} pq'ss' &= qp'ss', \\ qq'rs' &= qq'sr', \\ pq'ss' + qq'rs' &= p'qss' + qq'sr', \\ (ps + qr)q's' &= (p's' + q'r')qs, \end{aligned}$$

and thus $(ps + qr, q's') = (p's' + q'r', qs)$, which in turn gives $(ps + qr)/qs = (p's' + q'r')/q's'$. The other part is similar (and easier). □

In view of the Lemma (6.7), we can make the following definitions: whenever p/q and r/s are elements of F, define

$$\begin{aligned} p/q + r/s &\overset{\text{def}}{=} (ps + qr)/qs \\ (p/q)(r/s) &\overset{\text{def}}{=} pr/qs \end{aligned} \tag{4}$$

(which of course are exactly the same as the familiar rules for addition and multiplication of rational numbers). It would not have been legitimate to make these definitions before proving the lemma. For, it is quite possible to have $p/q = p'/q'$ and $r/s = r'/s'$ without having $p = p'$, $q = q'$, $r = r'$ and $s = s'$, and the definition Eq.(4) simultaneously stipulates that $p/q + r/s \overset{\text{def}}{=} (ps + qr)/qs$ and that $p'/q' + r'/s' \overset{\text{def}}{=} (p's' + q'r')/q's'$. In other words, the same object is defined in two different ways. Lemma (6.7) shows that the two ostensibly different right hand sides are in fact the same, so that the definition is, after all, not ambiguous. Of course, the same comments apply for the definition of multiplication of these objects. To use the conventional terminology of mathematicians, the lemma shows that addition and multiplication of equivalence classes (given by Eq.(4)) is well-defined. (I do not like this terminology, since it seems to suggest that something can be defined without being well-defined. Indeed, some misguided authors would state Eq.(4) as a definition before proving the lemma, and then go on to say "These operations are well defined because of . . . ", and then state and prove the lemma. To be rigorous, though, the lemma needs to be done first.)

We still have not completed the proof of the Theorem (6.4). Recall that the first assertion of the theorem is that the given integral domain is a subring of some field F. So far we have produced a set F (the quotient of S by the proportionality relation) and defined operations of addition and multiplication on it. The next step is to show that these operations make F into a field. This has to be done by checking that all the field axioms are satisfied, and since there are quite a lot of axioms, this is a somewhat tedious process. However, it is not difficult. For example, to check that the associative law for addition in F is satisfied we must show that

$$(p/q + r/s) + t/u = p/q + (r/s + t/u)$$

whenever $p, q, r, s, t, u \in R$ and q, s and u are nonzero. Using the definition Eq.(4) and the associative and distributive laws in R gives

$$(p/q + r/s) + t/u = (ps + qr)/qs + t/u = ((ps + qr)u + (qs)t)/(qs)u = (psu + qru + qst)/qsu,$$

and similarly

$$p/q + (r/s + t/u) = p/q + (ru + st)/su = (p(su) + q(ru + st))/q(su) = (psu + qru + qst)/qsu,$$

which is the same thing. The other axioms work in very much the same way. One can see that it is bound to be OK, since the formal calculations involved are just the same as they would be if the objects a/b under consideration were ordinary fractions, and in that context we know that all the axioms are satisfied. So we omit the rest of these calculations, except for mentioning that the zero element of F is the equivalence class $0/1$, and the identity element of F is $1/1$ (where, of course, 0 and 1 stand for the zero and identity elements of the integral domain R).

To prove the theorem exactly we should really have constructed a field F which has R as a subring. The field F that we have in fact constructed does not literally have this property. But we can produce an injective homomorphism $\phi\colon R \to F$, which is good enough for our purposes since we may regard this homomorphism as an embedding of R in F (in accordance with our previous discussion of such matters).

Define $\phi\colon R \to F$ by $\phi a = a/1$ for all $a \in R$. Then by Eq.(4) we find that for all $a, b \in R$,

$$\begin{gathered}\phi a + \phi b = (a/1) + (b/1) = (a1 + 1b)/1 = (a + b)/1 = \phi(a + b),\\ (\phi a)(\phi b) = (a/1)(b/1) = ab/1 = \phi(ab),\end{gathered}$$

so that ϕ is a homomorphism. If $\phi a = \phi b$ then $a/1 = b/1$, which gives $a = b$ (by the criterion proved above); hence ϕ is injective, as required.

The one assertion of the theorem which still remains to be proved is that every element of F can be written in the form ab^{-1}, where $a, b \in R$ and $b \neq 0$. Note, however, that identification of R with the copy of R contained in F is implicit in this statement. What it really means is that every element of F is expressible in the form $(\phi a)(\phi b)^{-1}$, with $a, b \in R$ and $b \neq 0$. Since $(1/b)(b/1) = b/b = 1/1$ it follows that $(\phi b)^{-1} = (b/1)^{-1} = (1/b)$, and hence $(\phi a)(\phi b)^{-1} = (a/1)(1/b) = a/b$. Since, by the definition of F, every element of F has this form the assertion is correct, and the proof of Theorem (6.4) is complete (finally). □

The field F that appears in Theorem (6.4) is called the *field of fractions* of the integral domain R. Our next result, whose proof is left as an exercise, is that F is unique to within isomorphism.

Theorem (6.8): *Suppose that R is an integral domain and F_1, F_2 fields. Suppose that there exist embeddings $\eta_1\colon R \to F_1$ and $\eta_2\colon R \to F_2$ such that every element of F_1 can be expressed in the form $\eta_1 a(\eta_1 b)^{-1}$ for some $a, b \in R$, and every element of F_2 can be expressed in the form $\eta_2 a(\eta_2 b)^{-1}$ for some $a, b \in R$. Then there is an isomorphism $\phi\colon F_1 \to F_2$ such that $\phi(\eta_1 a) = \eta_2 a$ for all $a \in R$.*

Exercise 8. Let R be a commutative ring which is not the trivial ring $\{0\}$, and suppose that R has no zero divisors but does not have a 1. Examine all the steps of the proof of Theorem (6.4) and check that everything still works: there is still a field of fractions. (Note that the identity element is the equivalence class a/a, where a is any nonzero element of R.)

7. Ideals and quotient rings

Definition (7.1): A subset I of a ring R is called an *ideal* if I is a subring of R and, for all a and x, if $a \in R$ and $x \in I$ then $ax \in I$ and $xa \in I$.

Exercise 9. Show that in any ring R the sets $\{0\}$ and R are ideals of R.

Exercise 10. Let R be a ring with 1 and I an ideal of R. Prove that if there is an element $u \in I$ which has an inverse in R then $I = R$. Hence prove that if R is a field then the only ideals of R are $\{0\}$ and R.

Theorem (5.13) can be restated as follows:

Theorem (7.2): *If $\phi\colon R \to S$ is a ring homomorphism then the kernel of ϕ is an ideal of R.*

This result is the basic reason why kernels are important in ring theory.

Exercise 11. Let F and E be fields, and $\phi\colon F \to E$ a homomorphism which is not the zero function. Show that ϕ is injective. (Use Theorem (7.2) and Exercise 10.) Show also that $\phi(1_F) = 1_E$. (Hint: In a field, if $t^2 = 1$ then $t = 0$ or 1.)

We have already seen that a subset of a ring is a subring if and only if it is nonempty and closed under addition and multiplication. Combining this with the definition of ideal, we obtain the following criterion for a subset of a ring to be an ideal.

Theorem (7.3): *A subset I of a ring R is an ideal of R if and only if the following conditions all hold:*
(i) $I \neq \emptyset$,
(ii) $x + y \in I$ *for all* $x, y \in I$,
(iii) $-x \in I$ *for all* $x \in I$, *and*
(iv) $ax, xa \in I$ *for all* $x \in I$ *and* $a \in R$.

In other words, I is an ideal of the ring R if and only if I is a nonempty subset of R which is closed under addition ((ii) above), closed under taking negatives ((iii) above) and closed under multiplication on either side by elements of R ((iv) above). Note also that since $I \neq \emptyset$ there exists at least one element $a \in I$, and now by closure under taking negatives $-a$ must also be in I, whence $a + (-a) \in I$ by closure under addition. But $a - a = 0$, and so we have shown that an ideal of R must always contain the zero element of R. (We knew this anyway, since an ideal is a subring, and the zero element of a subring must coincide with the zero element of the ring.)

Examples

(i) The set of all even integers is an ideal in $\mathbb{Z}$, the ring of all integers. More generally, for every integer n the set $n\mathbb{Z} = \{\, nk \mid k \in \mathbb{Z} \,\}$ (integers which are multiples of n) is an ideal of $\mathbb{Z}$. This follows easily from Theorem (7.3). Firstly, $n\mathbb{Z} \neq \emptyset$ since 0 is a multiple of n. Now suppose that $x, y \in n\mathbb{Z}$ and $a \in \mathbb{Z}$. Then $x = nb$ and $y = nc$ for some integers b and c, and we see that $x+y = n(b+c)$, $-x = n(-b)$ and $ax = xa = n(ab)$ are all elements of $n\mathbb{Z}$. So $n\mathbb{Z}$ is nonempty, closed under addition and closed under multiplication by arbitrary elements of $\mathbb{Z}$, as required.

(ii) $(x^2+1)\mathbb{R}[x] = \{\, (x^2+1)p \mid p \in \mathbb{R}[x] \,\}$, the set of all polynomials over $\mathbb{R}$ which have x^2+1 as a factor, is an ideal in the ring $\mathbb{R}[x]$.

(iii) Let $I \subseteq \mathbb{Z}[x]$ consist of those integer polynomials whose constant term is even. That is, I consists of all polynomials $n_0 + n_1x + n_2x^2 + \cdots + n_dx^d$, where the coefficients n_i are integers, and n_0 is even. (The nonnegative integer d is allowed to vary.) Then I is an ideal in $\mathbb{Z}[x]$.

(iv) Let R be the set of all upper triangular 3×3 matrices over $\mathbb{R}$:

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

We have seen that R is a subring of the ring $\mathrm{Mat}_3(\mathbb{R})$. Let us show that

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

is an ideal in R.

Clearly $I \subseteq R$, and $I \neq \emptyset$ since the zero matrix is in I. Now let $X, Y \in I$ and $A \in R$. Then we have

$$A = \begin{pmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & g \end{pmatrix}, \quad X = \begin{pmatrix} 0 & h & i \\ 0 & 0 & j \\ 0 & 0 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & k & l \\ 0 & 0 & m \\ 0 & 0 & 0 \end{pmatrix}$$

for some real numbers a, b etc., and a little calculation yields that

$$X+Y = \begin{pmatrix} 0 & h+k & i+l \\ 0 & 0 & j+m \\ 0 & 0 & 0 \end{pmatrix} \qquad AX = \begin{pmatrix} 0 & ah & ai+bj \\ 0 & 0 & ej \\ 0 & 0 & 0 \end{pmatrix}$$
$$-X = \begin{pmatrix} 0 & -h & -i \\ 0 & 0 & -j \\ 0 & 0 & 0 \end{pmatrix} \qquad XA = \begin{pmatrix} 0 & he & hf+ig \\ 0 & 0 & jg \\ 0 & 0 & 0 \end{pmatrix}$$

all of which are elements of I. This establishes all of the required closure properties, and thus shows that I is an ideal in R.

(v) Let R be an arbitrary commutative ring, and a any element of R. Then $aR = \{\, ax \mid x \in R \,\}$ is an ideal in R. This can be proved by reasoning totally analogous to that used in (i) above, dealing with the case $R = \mathbb{Z}$. Note, however, that the assumption that the ring R is commutative is necessary, since in the noncommutative case aR may not be closed under multiplication on the left by elements of R. (If ax is an element of aR, and r is an element of R, there is no guarantee that the element $r(ax)$ can be written in the form ay, for any value of y.)

(vi) Let R be a commutative ring and let $a, b, c \in R$. The set $aR + bR + cR$ defined by

$$aR + bR + cR = \{\, ax + by + cy \mid x, y, z \in R \,\}$$

is an ideal in R. The proof of this is not hard; for example, the sum of two arbitrary elements $ax + by + cz$ and $ax' + by' + cz'$ of $aR + bR + cR$ is $a(x + x') + b(y + y') + c(z + z')$, which is also in $aR + bR + cR$, which is therefore closed under addition. We leave the other parts of the proof to the reader.

The ideal of the commutative ring R described in (vi) above is said to be *generated* by the three elements a, b and c. It is clear that ideals generated by any number of elements can be defined similarly. An ideal which is generated by a single element—that is, an ideal of the form aR where $a \in R$—is called a *principal ideal*. Remember though that this concept applies only to commutative rings.

Exercise 12. Suppose that R is a commutative ring that has no ideals other than $\{0\}$ and R, and assume that it is not the case that $xy = 0$ for all $x, y \in R$.

(i) Show that R has no zero divisors. (Hint: Suppose that R has at least one zero divisor. Show that if r is any zero divisor, then $I = \{\, s \in R \mid rs = 0 \,\}$ is a nonzero ideal of R, and deduce that $rs = 0$ for all $s \in R$. Next, show that if $J = \{\, r \in R \mid rs = 0 \text{ for all } s \in R \,\}$ then J is an ideal of R; furthermore, J is nonzero since it contains all the zero divisors, of which there is at least one. But $J = R$ contradicts the assumption that multiplication in R is nontrivial.)

(ii) Show that R has a 1. (Hint: Note that R has at least one nonzero element, and make a fixed choice of one such element, a. Show that aR is a nonzero ideal of R, and deduce that $aR = R$. Conclude that every element $t \in R$ can be expressed in the form $t = ax$ with $x \in R$, and in particular let $b \in R$ be such that $a = ab$. Show that $ax = (ax)b$ for all $x \in R$, and deduce that $t = tb$ for all $t \in R$.)

(iii) Show that every nonzero element of R has an inverse. (Hint: As in Part (ii), if $a \in R$ is nonzero then each element of R is expressible in the form ax with $x \in R$. In particular, the identity element can be expressed in this way.)

(iii) Show that R is a field.

Theorem (7.4): *Every ideal in the ring $\mathbb{Z}$ is principal.*

Proof. Let I be an ideal in $\mathbb{Z}$; we must show that $I = n\mathbb{Z}$ for some integer n. Definition (7.1) requires ideals to be nonempty, and so I has at least one element. If 0 is the only element of I then $I = \{0\} = 0\mathbb{Z}$, since 0 is the only multiple of 0, and so the required conclusion holds with $n = 0$. Suppose instead that I contains at least one nonzero element a. Since ideals must be closed under taking negatives, it follows that $-a \in I$ also. Since one of the integers a and $-a$ must be positive, it follows that I has at least one positive element.

Let n be the least positive integer in I. (Here we are making use of a basic property of the set of positive integers, known as the "Least Integer Principle", which asserts that every nonempty set of positive integers has a least element. The principle of mathematical induction is based upon the Least Integer Principle, and, conversely, one can easily use induction to show that a set of positive integers with no least element has to be empty.) Thus if $r \in I$ and $0 \leq r < n$ then $r = 0$ (since otherwise r would be a positive integer in I and less than n, which is supposed to be the least positive integer in I). Since I must be closed under multiplication by arbitrary elements of $\mathbb{Z}$ we have that $qn \in I$ for all integers q.

Now let $m \in I$ be arbitrary. By the well known division of property of integers we can divide m by the positive integer n and obtain a quotient q and a remainder r. In other words, q and r are integers which satisfy $m = qn + r$ and $0 \leq r < n$. By hypothesis $m \in I$, and we have already noted that $qn \in I$. By closure of I under taking negatives, it follows that $-qn \in I$, and hence by closure of I under addition we obtain that $m - qn \in I$. That is, $r \in I$. But as $0 \leq r < n$ it follows that $r = 0$, and hence $m = qn = nq$. So we have shown that an arbitrary element m of I has to be a multiple of n. As we have already noted that all multiples of n are in I, the conclusion is that I consists exactly of the multiples of n:

$$I = \{\, nq \mid q \in \mathbb{Z} \,\} = n\mathbb{Z}$$

as required. □

Note: We have shown in the above proof that if I is any nonzero ideal in $\mathbb{Z}$ then I is generated by the least positive integer it contains (in the sense that $I = n\mathbb{Z}$, where n is this least positive integer). Accordingly, whenever I is an ideal in Z, we define the *canonical generator* of I to be the least positive integer in I, or 0 if $I = \{0\}$.

Exercise 13. Let S be a subset of the set of positive integers and suppose that S has no least element. Use induction on k to prove that $k \notin S$ for all positive integers k, and deduce that $S = \emptyset$.

Let I be an ideal in the ring R. We define a relation on R called *congruence modulo I*, and denoted by $\equiv$ (mod I), as follows:

$$a \equiv b \text{ (mod } I) \text{ if and only if } a - b \in I.$$

We shall show that this is an equivalence relation on R.

Firstly, for all $a \in R$ we have that $a - a \in I$ (since the ideal I must contain the zero element of R), and so $a \equiv a$ (mod I). Hence congruence modulo I is a reflexive relation.

Secondly, suppose that $a \equiv b$ (mod I). Then $a - b \in I$, and by closure under taking negatives, $-(a - b) \in I$. But $-(a - b) = b - a$; so $b - a \in I$, and hence $b \equiv a$ (mod I). Thus congruence is symmetric.

Finally, suppose that $a \equiv b$ (mod I) and $b \equiv c$ (mod I). Then $a - b$, $b - c \in I$, and by closure under addition $a - c = (a - b) + (b - c) \in I$. So $a \equiv c$ (mod I), and we deduce that congruence is transitive.

Recall that the importance of equivalence relations comes from the fact that an equivalence relation always partitions the set on which it is defined into equivalence classes. The equivalence classes are disjoint from each other and cover the entire set in question. That is, every element of the set lies in exactly one equivalence class. For congruence modulo the ideal I of the ring R, the equivalence classes (or *congruence classes*) are also called the *cosets* of I in R. For example, $4\mathbb{Z}$ (the set of all multiples of 4) is an ideal in $\mathbb{Z}$. It is easy to see that congruence modulo $4\mathbb{Z}$ partitions $\mathbb{Z}$ into exactly four congruence classes:

$$\begin{aligned}
C_0 &= \{\ldots, -8, -4, 0, 4, 8, 12, \ldots\} \\
C_1 &= \{\ldots, -7, -3, 1, 5, 9, 13, \ldots\} \\
C_2 &= \{\ldots, -6, -2, 2, 6, 10, 14, \ldots\} \\
C_3 &= \{\ldots, -5, -1, 3, 7, 11, 15 \ldots\}.
\end{aligned}$$

Every integer lies in one or other of these four sets, and it is easily seen that any two numbers which lie in the same set C_i differ by a multiple of 4. The four congruence classes are the cosets of the ideal $4\mathbb{Z}$ in the ring $\mathbb{Z}$. Note that the coset C_0 is just the ideal $4\mathbb{Z}$ itself, and (more generally) the coset C_i consists of all the integers that are i greater than a multiple of 4.

Returning to the general situation, observe that if $a \in R$ is arbitrary then the coset of I containing a is

$$\begin{aligned}\{\, b \in R \mid b \equiv a \pmod{I} \,\} &= \{\, b \in R \mid b - a \in I \,\} \\ &= \{\, a + t \mid t \in I \,\}.\end{aligned}$$

It is natural to denote this coset by $a + I$. It is important to remember that the coset containing a is equal to the coset containing b if and only if $a \equiv b \pmod{I}$. That is, if I is an ideal in R and $a, b \in R$ then

$$a + I = b + I \text{ if and only if } a - b \in I. \tag{5}$$

In particular, $a + I = b + I$ does not imply that $a = b$ (unless I happens to be $\{0\}$), just as fractions p/q and p'/q' can be equal without p equalling p' and q equalling q'.

Our next objective is to define addition and multiplication on the set $R/I = \{\, a + I \mid a \in R \,\}$ (which is the quotient of R by the equivalence relation congruence modulo I). The set R/I can be thought of as the thing that R becomes if congruent elements are regarded as equal. So whole sets of congruent elements of R are coalesced to form single elements of R/I. We would like to define the addition and multiplication operations on R/I in a manner that is consistent with this viewpoint. That is, it should make no difference whether you coalesce elements before adding or multiplying them, or after. We therefore want addition and multiplication of cosets to satisfy

$$\begin{aligned}(a + I) + (b + I) &= (a + b) + I \\ \text{and} \qquad (a + I)(b + I) &= ab + I\end{aligned} \tag{6}$$

for all $a, b \in R$.[†] The possible obstruction to this is that since we can have $a + I = a' + I$ and $b + I = b' + I$ without having $a = a'$ and $b = b'$, it is conceivable that Eq.(6) might yield more than one definition of the same object. To prove that there are well defined operations on R/I satisfying Eq.(6) therefore requires proving the following lemma.

Lemma (7.5): *Suppose that I is an ideal in R, and let $a, a', b, b' \in R$. If $a + I = a' + I$ and $b + I = b' + I$ then $(a + b) + I = (a' + b') + I$ and $ab + I = a'b' + I$.*

Proof. If $a' + I = a + I$ and $b' + I = b + I$ then by Eq.(5) there exist $s, t \in I$ such that $a' = a + s$ and $b' = b + t$. This gives

$$\begin{aligned}a' + b' &= (a + s) + (b + t) \\ &= (a + b) + (s + t),\end{aligned}$$

and since $s + t \in I$ (by closure of I under addition) it follows that $a' + b' \equiv a + b \pmod{I}$, and thus $(a' + b') + I = (a + b) + I$. Similarly we find by the associative and distributive laws that

$$\begin{aligned}a'b' &= (a + s)(b + t) \\ &= ab + (at + sb + st).\end{aligned}$$

But now closure of I under left multiplication by elements of R gives $at \in I$ (since $t \in I$), closure of I under right multiplication by elements of R gives $sb \in I$ (since $s \in I$), and likewise $st \in I$, so

[†] Unfortunately, the definition of multiplication of cosets is not consistent with the definition we gave earlier for the product of two arbitrary subsets of a ring. Since the previous definition is unlikely to arise again in this course, it can be ignored.

that closure of I under addition gives $at + sb + st \in I$, hence showing that $a'b' \equiv ab \pmod{I}$. So we also have that $a'b' + I = ab + I$, as required. □

Recall that there is a surjective mapping $\phi\colon R \to R/I$ (the canonical surjection) given by $\phi a = a + I$ for all $a \in R$. We have defined addition and multiplication in R/I so that Eq.(6) holds; that is, so that $\phi(a + b) = \phi a + \phi b$ and $\phi(ab) = (\phi a)(\phi b)$ for all $a, b \in R$. It follows from Theorem (5.3) that $\operatorname{im}\phi$ is a ring and ϕ a ring homomorphism from R to $\operatorname{im}\phi$. Since ϕ is surjective, $\operatorname{im}\phi = R/I$. Thus we have shown that our definitions of addition and multiplication on R/I make R/I a ring. We call rings constructed in this way *quotient rings*. (Usually R/I is called just "R over I", but if you prefer to be more elaborate then it is the quotient of R by the ideal I.) The zero element of the ring R/I is $\phi 0$, which is the coset $0 + I = \{\, 0 + t \mid t \in I \,\} = I$. Note also that the kernel of the homomorphism ϕ is the set

$$\begin{aligned}\{\, a \in R \mid \phi a = 0_{R/I} \,\} &= \{\, a \in R \mid a + I = 0 + I \,\}\\ &= \{\, a \in R \mid a \in I \,\}\\ &= I\end{aligned}$$

The next theorem is a summary of what we have just done.

Theorem (7.6): *Let I be an ideal in the ring R. Then $R/I = \{\, a + I \mid a \in R \,\}$ is a ring under operations satisfying Eq.(6) above, and the mapping $\phi\colon R \to R/I$ given by $\phi a = a + I$ is a surjective homomorphism with kernel I.*

Example

If $R = \mathbb{Z}$, the ring of integers, and $I = 4\mathbb{Z}$, the principal ideal of R generated by 4, then as we have seen above there are exactly four cosets of I in R. In the notation we used previously, these cosets are C_0, C_1, C_2 and C_3, where C_i is the set of all numbers congruent to i modulo 4. In the notation we have introduced since, $C_i = i + 4\mathbb{Z}$. So $R/I = \mathbb{Z}/4\mathbb{Z} = \{\, 4\mathbb{Z}, 1 + 4\mathbb{Z}, 2 + 4\mathbb{Z}, 3 + 4\mathbb{Z} \,\}$. Addition and multiplication in this ring are given by the following tables (derived from Eq.(6)).

$+$	$4\mathbb{Z}$	$1 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$
$4\mathbb{Z}$	$4\mathbb{Z}$	$1 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$
$1 + 4\mathbb{Z}$	$1 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$	$4\mathbb{Z}$
$2 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$	$4\mathbb{Z}$	$1 + 4\mathbb{Z}$
$3 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$	$4\mathbb{Z}$	$1 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$

$\cdot$	$4\mathbb{Z}$	$1 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$
$4\mathbb{Z}$	$4\mathbb{Z}$	$4\mathbb{Z}$	$4\mathbb{Z}$	$4\mathbb{Z}$
$1 + 4\mathbb{Z}$	$4\mathbb{Z}$	$1 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$3 + 4\mathbb{Z}$
$2 + 4\mathbb{Z}$	$4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$4\mathbb{Z}$	$2 + 4\mathbb{Z}$
$3 + 4\mathbb{Z}$	$4\mathbb{Z}$	$3 + 4\mathbb{Z}$	$2 + 4\mathbb{Z}$	$1 + 4\mathbb{Z}$

For example, using Eq.(6) we find that $(2+4\mathbb{Z})(3+4\mathbb{Z}) = 6+4\mathbb{Z} = 2+4\mathbb{Z}$ (since $6 \equiv 2 \pmod{4\mathbb{Z}}$), and $(2 + 4\mathbb{Z}) + (3 + 4\mathbb{Z}) = 5 + 4\mathbb{Z} = 1 + 4\mathbb{Z}$ (since $5 \equiv 1 \pmod{4\mathbb{Z}}$). The traditional terminology in number theory is to say $6 \equiv 2 \pmod{4}$, rather than mod $4\mathbb{Z}$, and we will adopt this convention in future. The ring $\mathbb{Z}/4\mathbb{Z}$ is called the *ring of integers modulo 4*. It will soon become inconvenient to write the elements of $\mathbb{Z}/4\mathbb{Z}$ as $4\mathbb{Z}$, $1 + 4\mathbb{Z}$, $2 + 4\mathbb{Z}$ and $3 + 4\mathbb{Z}$: we will want something shorter. Sometimes we will revert to the notation that we first introduced in our discussion of equivalence classes, and write the elements as $\overline{0}$, $\overline{1}$, $\overline{2}$ and $\overline{3}$. But in due course even this will become too messy, and we will omit the bar, using the same notation for elements of $\mathbb{Z}/4\mathbb{Z}$ as for elements of $\mathbb{Z}$. When we do this, the reader will have to be aware from the context that the objects in question are integers modulo 4 rather than ordinary integers, and remember, for example, that $1 = 5 = 9 = -3$ in $\mathbb{Z}/4\mathbb{Z}$.

Clearly there was nothing special about the integer 4 in the above example. We could have started with any positive integer n, formed the ideal nZ, and then constructed the quotient ring

$\mathbb{Z}/n\mathbb{Z}$. This ring, which henceforth we will usually denote by $\mathbb{Z}_n$, is known as the ring of integers modulo n. It has exactly n elements, namely $\overline{0}, \overline{1}, \overline{2}, \dots, \overline{n-1}$, where $\overline{i}$ is the coset of $n\mathbb{Z}$, or congruence class, consisting of all integers that are congruent to i modulo n—that is, integers which differ from i by a multiple of n. There is a surjective homomorphism $\mathbb{Z} \to \mathbb{Z}_n$ given by $i \mapsto \overline{i}$, and $\overline{i} = \overline{j}$ if and only if $i \equiv j \pmod{n}$.

Exercise 14. Show that if the integer n is not prime then $\mathbb{Z}_n$ has at least one zero divisor.

Exercise 15. Let I be an ideal in the ring R. Show that if R is commutative then so is R/I, and show that if R has a 1 then so does R/I.

Example

Let F be a field and let $R = F[x]$, the ring of polynomials over F in the indeterminate x. Let $I = xF[x]$, the principal ideal generated by the polynomial x. That is, I consists of all polynomials which have x as a factor, and it can be seen that this is just the same as the set of polynomials with zero constant term. To say that two polynomials $a(x)$ and $b(x)$ are congruent modulo this ideal is to say that $a(x) - b(x)$ has zero constant term, or equivalently that $a(x)$ and $b(x)$ have the same constant term. For example, if F were the real field $\mathbb{R}$ and $a(x)$ the polynomial $5 - 2x - x^2$, then the polynomials in the coset $a(x) + I$ would be all the polynomials with 5 as their constant term. So $5 + 3x + x^2 + x^3$ and $5 + x^{71}$ would be examples of polynomials $b(x)$ with $a(x) + I = b(x) + I$. So too would be the constant polynomial 5. In general, if $t \in F$ is arbitrary then the set of polynomials

$$\{\, b(x) \in F[x] \mid b(x) = t + a_1x + a_2x^2 + \cdots + a_dx^d \text{ for some } a_i \in F \,\} \tag{7}$$

is a coset of I in R, and every coset has this form for some $t \in F$. We could perhaps write the coset in Eq.(7) as $t + \{*\}$, meaning the set of polynomials of the form t plus terms in higher powers of x. Now to add or multiply two cosets you just addor multiply representative elements of the coset, and it is clear that adding or multiplying two polynomials just involves adding or multiplying their constant terms. So to continue with the notation as above, for all $s, t \in F$

$$\begin{aligned}(s + \{*\}) + (t + \{*\}) &= s + t + \{*\}\\ (s + \{*\})(t + \{*\}) &= st + \{*\}\end{aligned}$$

and we see that $t + \{*\} \leftrightarrow t$ is a one to one correspondence between the elements of R/I (the cosets) and elements of F, and this correspondence preserves addition and multiplication. Hence $F[x]/xF[x] \cong F$.

8. The First Isomorphism Theorem

A consequence of Theorem (7.6) is that whenever I is an ideal in R one can find a homomorphism whose kernel is I. We had seen previously (Theorem (7.2)) that the kernel of a homomorphism is always an ideal. So ideals are the same as kernels. Furthermore, whenever we have an ideal we can construct a quotient ring. So in particular, if R and S are rings and $\phi\colon R \to S$ a homomorphism, then we can form the quotient ring $R/\ker\phi$. The theorem we are about to state, which is the most important theorem of introductory ring theory, examines the connection between this quotient ring and the homomorphism we started with.

The First Isomorphism Theorem *Let R and S be rings and $\phi\colon R \to T$ a homomorphism. Then the kernel of ϕ is an ideal of R, the image of ϕ is a subring of T, and there is an isomorphism $\psi\colon R/\ker\phi \to \operatorname{im}\phi$ such that $\psi(a + \ker\phi) = \phi a$ for all $a \in R$.*

Proof. Let $K = \ker\phi$ and $S = \mathrm{im}\,$■. We have already proved in Proposition (5.12) and Theorem (7.2) that S is a subring of T and K an ideal of R. Our main task now is to prove that there is a well defined function $\psi\colon R/K \to S$ satisfying $\psi(a + K) = \phi a$ for all $a \in R$. As always when discussing cosets it has to be remembered that $a + K = a' + K$ does not imply that $a = a'$, and so to show that ψ is well defined we must show that if $a + K = a' + K$ then $\phi a = \phi a'$. Once this is done we will know that the object ϕa depends only on the coset $a + K$ and not on the choice of representative element a in that coset.

If $a + K = a' + K$ then $a \equiv a' \pmod{K}$, and so $a - a' \in K = \ker\phi$. Hence $\phi(a - a') = 0$. But since ϕ is a homomorphism we have that $\phi(a - a') = \phi a - \phi a'$, and so we conclude that $\phi a = \phi a'$. The rule which determines the function $\psi\colon R/K \to S$ can now be spelt out as follows. Given $\alpha \in R/K$ we can choose an element $a \in R$ such that $\alpha = a + K$ (since the canonical map $R \to R/K$ is surjective); then $\phi a \in \mathrm{im}\,\phi = S$, and we define $\psi\alpha = \phi a$, noting that all choices of a lead to the same element of S.

Having established that ψ is well-defined, it follows easily from the fact that ϕ preserves addition and multiplication that ψ does also. Indeed, let $\alpha, \beta \in R/K$; then there exist $a, b \in R$ with $\alpha = a + K$ and $\beta = b + K$, and we have

$$\psi(\alpha + \beta) = \psi((a + K) + (b + K) = \psi((a + b) + K) = \phi(a + b) = \phi a + \phi b = \psi\alpha + \psi\beta,$$

and similarly

$$\psi(\alpha\beta) = \psi((a + K)(b + K) = \psi(ab + K) = \phi(ab) = (\phi a)(\phi b) = (\psi\alpha)(\psi\beta).$$

Thus ψ is a homomorphism, and all that remains is to prove that ψ is bijective.

It is clear that ψ is surjective, for if $s \in S$ is arbitrary then there exists $a \in R$ with $s = \phi a$ (since $S = \mathrm{im}\,\phi$), and this gives $s = \psi(a + K)$. Suppose now that $\alpha \in \ker\phi$. Choose $a \in R$ such that $\alpha = a + K$; then $\phi a = \psi\alpha = 0$, which shows that $a \in \ker\phi = K$, so that $\alpha = a + K = 0 + K$, the zero element of R/K. Hence we have shown that the zero of R/K is the only element of $\ker\psi$, whence by Proposition (5.14) it follows that ψ is injective. So ψ is a bijective homomorphism—that is, an isomorphism—as required. □

There are two main points to note about the First Isomorphism Theorem, at least as far as this course is concerned. First of all, it is applicable in every situation in which there is a homomorphism. So whenever you encounter a homomorphism you should immediately ask the following questions:

(a) What is the kernel of this homomorphism?
(b) What is the image of this homomorphism?
(c) What, in this case, is the isomorphism which the First Isomorphism Theorem gives us?

Secondly, remember that the First Isomorphism Theorem provides a method for proving that two things are isomorphic. So if you are asked to prove an isomorphism, you should ask yourself whether the First Isomorphism Theorem might be useful. In particular, if you are asked to prove that some ring S is isomorphic to some quotient ring R/K, then almost certainly the way to do it is to find a homomorphism from R to S whose kernel is K.

Examples

(i) We shall use the First Homomorphism Theorem to show that if $K = (x^2 + 1)\mathbb{R}[x]$ (the principal ideal of the ring $\mathbb{R}[x]$ generated by the element $x^2 + 1 \in \mathbb{R}[x]$) then the quotient ring $\mathbb{R}[x]/K$ is isomorphic to the field $\mathbb{C}$ (complex numbers).

Since $\mathbb{R}$ is a subfield of $\mathbb{C}$ and $i = \sqrt{-1}$ is an element of $\mathbb{C}$, there is an evaluation homomorphism $\theta\colon \mathbb{R}[x] \to \mathbb{C}$ given by $\theta(a(x)) = a(i)$ for all polynomials $a(x) \in \mathbb{R}[x]$. Everything we are seeking to prove, and more, will be given to us by determining exactly what the First Isomorphism Theorem says when applied to this homomorphism θ.

The first task is to determine the kernel of θ. To facilitate this we state a result about division of polynomials, concerning which we will have more to say later: if $a(x)$, $b(x) \in \mathbb{R}[x]$ with $b(x)$ nonzero, then there exist $q(x)$, $r(x) \in \mathbb{R}[x]$, with the degree of $r(x)$ less than the degree of $b(x)$, satisfying $a(x) = q(x)b(x) + r(x)$. consisting of replacing $a(x)$ by $a(x) - cx^i b(x)$, where i and the scalar c are chosen so that the leading terms cancel out (so that the degree is reduced), and repeating this until the degree is less than that of $b(x)$. We illustrate this with $a(x) = 2x^5 - x^4 - x^3 + 3x^2 + x + 7$ and $b(x) = x^2 + 1$.

$$
\begin{array}{r@{\,}l}
 & 2x^3 - x^2 - 3x + 4 \\ \cline{2-2}
x^2 + 1 \big) & 2x^5 - x^4 - x^3 + 3x^2 + x + 7 \\
 & 2x^5 \phantom{{}-x^4} + 2x^3 \\ \cline{2-2}
 & - x^4 - 3x^3 + 3x^2 + x + 7 \\
 & - x^4 \phantom{{}-3x^3} - x^2 \\ \cline{2-2}
 & - 3x^3 + 4x^2 + x + 7 \\
 & - 3x^3 \phantom{{}+4x^2} - 3x \\ \cline{2-2}
 & 4x^2 + 4x + 7 \\
 & 4x^2 \phantom{{}+4x} + 4 \\ \cline{2-2}
 & 4x + 3.
\end{array}
$$

The calculations show that $2x^5 - x^4 - x^3 + 3x^2 + x + 7 = (2x^3 - x^2 - 3x - 4)(x^2 + 1) + 4x + 3$. In the current context the point is that this allows easy evaluation at $x = i$: replace x by i in the above and the answer is $4i + 3$ (since $x^2 + 1$ evaluates to 0 at $x = i$). In general, if $a(x) \in \mathbb{R}[x]$ is arbitrary then division by $x^2 + 1$ gives a remainder of degree at most 1, so that $a(x) = q(x)\,(x^2 + 1) + cx + d$ for some c, $d \in \mathbb{R}$, and applying the evaluation map $\theta = \mathrm{eval}_i$ we obtain $a(i) = ci + d$. This is zero if and only if $c = d = 0$; hence the kernel of the evaluation homomorphism is precisely the set of all $a(x) \in \mathbb{R}[x]$ which yield a zero remainder on division by $x^2 + 1$. That is, $\ker\theta = (x^2 + 1)\mathbb{R}[x] = K$, the set of all polynomials in $\mathbb{R}[x]$ that have $x^2 + 1$ as a factor. The First Isomorphism Theorem tells us that this is an ideal of $\mathbb{R}[x]$—which we knew anyway, since it is the principal ideal generated by $x^2 + 1$—and that $\mathbb{R}[x]/K \cong \operatorname{im}\theta$.

It remains to check that $\operatorname{im}\theta$ is the whole of $\mathbb{C}$ and not some proper subset. So we must show that for every complex number α there is a polynomial $a(x) \in \mathbb{R}[x]$ with $a(i) = \alpha$. This is trivial, since $\alpha = ci + d$ for some c, $d \in \mathbb{R}$, and then the polynomial $a(x) = cx + d$ has the required property. We comment that these considerations also show that every coset of K in $\mathbb{R}[x]$ contains a representative of the form $cx + d$, and indeed that this representative is unique. The isomorphism $\mathbb{C} \to \mathbb{R}[x]/K$ inverse to the one obtained by the First Isomorphism Theorem is given by $ci + d \mapsto (cx + d) + K$.

(ii) Let $\mathbb{Z}_n = \{\overline{1}, \overline{2}, \ldots, \overline{n}\}$ be the ring of integers modulo n, and let $\phi\colon \mathbb{Z} \to \mathbb{Z}_n$ be defined by $\phi i = \overline{i}$ for all $i \in \mathbb{Z}$. Addition and multiplication in $\mathbb{Z}_n$ are defined so that $\overline{i + j} = \overline{i} + \overline{j}$ and $\overline{ij} = \overline{i}\,\overline{j}$ for all i, $j \in \mathbb{Z}$; so ϕ is a homomorphism. Since by definition $\overline{i} = \overline{j}$ if and only if $i \equiv j \pmod{n}$ we see that the kernel of ϕ is

$$\{\, i \in \mathbb{Z} \mid \overline{i} = \overline{0} \,\} = \{\, i \in \mathbb{Z} \mid i = nk \text{ for some } k \in \mathbb{Z} \,\} = n\mathbb{Z}.$$

It is clear that the image of ϕ is the whole of $\mathbb{Z}_n$. Of course, this homomorphism ϕ is simply the canonical surjection $\mathbb{Z} \to \mathbb{Z}_n$.

Applying the First Isomorphism Theorem we find that $Z/\ker\phi = \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}_n$, and that there is an isomorphism satisfying $i + n\mathbb{Z} \mapsto \bar{i}$. This is reassuring rather than interesting, since by definition $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$, and the function $i + n\mathbb{Z} \mapsto \bar{i}$ is simply the identity function on $\mathbb{Z}_n$.

(iii) Let R be the ring of upper triangular 3×3 matrices over $\mathbb{R}$, and let $I \subset R$ consist of those upper triangular matrices that have zeros on the diagonal as well as below. We shall show that I is an ideal of R and that $R/I \cong \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ (the direct product of three copies of $\mathbb{R}$). (In fact we showed earlier that I is an ideal of R, but we shall do so again by a more sophisticated means.)

Recall that $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ is the set of all ordered triples of real numbers, with addition and multiplication of triples being performed component by component. Define a function $\phi\colon R \to \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ by

$$\phi\begin{pmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & g \end{pmatrix} = (a, e, g).$$

A little calculation shows that

$$\begin{pmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & g \end{pmatrix}\begin{pmatrix} h & i & j \\ 0 & k & l \\ 0 & 0 & m \end{pmatrix} = \begin{pmatrix} ah & * & * \\ 0 & ek & * \\ 0 & 0 & gm \end{pmatrix}$$

where the $*$'s replace entries whose precise value is currently irrelevant to our purposes, and thus if $\phi A = (a, e, g)$ and $\phi B = (h, k, m)$ then

$$\phi(AB) = (ah, ek, gm) = (a, e, g)(h, k, m) = (\phi A)(\phi B).$$

Similarly, adding A and B involves adding their diagonal entries, and so $\phi(A + B) = \phi A + \phi B$. Thus ϕ is a homomorphism, and since $\phi A = (0, 0, 0)$ if and only if A has 0's on the diagonal, we see that $\ker\phi = I$. The image of ϕ is the whole of $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$, since the arbitrary element $(a, b, c) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ equals ϕD, where D is the diagonal matrix with diagonal entries a, b and c. So the first isomorphism theorem gives $R/I \cong \mathbb{R} \times \mathbb{R} \times \mathbb{R}$, as claimed.

Note that all the matrices in the coset $A + I$ have the same diagonal entries as the matrix A:

$$\begin{pmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & g \end{pmatrix} + I = \left\{ \begin{pmatrix} a & x & y \\ 0 & e & z \\ 0 & 0 & g \end{pmatrix} \,\middle|\, x,\, y,\, z \in \mathbb{R} \right\} = \left\{ \begin{pmatrix} a & * & * \\ 0 & e & * \\ 0 & 0 & g \end{pmatrix} \right\}$$

to use a fairly natural notation. The isomorphism which the theorem yields is given by

$$\left\{ \begin{pmatrix} a & * & * \\ 0 & e & * \\ 0 & 0 & g \end{pmatrix} \right\} \mapsto (a, e, g).$$

(iv) Let R be the ring of 5×5 lower triangular matrices over the ring $\mathbb{Z}[x]$. (One can prove that R is a subring of $\mathrm{Mat}_5(\mathbb{Z}[x])$ in much the same way as we proved that the upper triangular matrices form a subring of $\mathrm{Mat}_3(\mathbb{R})$.) The subset I of R consisting of those matrices whose diagonal entries have even constant term is an ideal of R, and $R/I \cong (\mathbb{Z}_2)^5$ (the direct product of five copies of the ring of integers modulo 2). This is proved by considerations analogous to those used in (iii) above: it is shown that the map $R \to (\mathbb{Z}_2)^5$ given by

$$\begin{pmatrix} a_1(x) & * & * & * & * \\ 0 & a_2(x) & * & * & * \\ 0 & 0 & a_3(x) & * & * \\ 0 & 0 & 0 & a_4(x) & * \\ 0 & 0 & 0 & 0 & a_5(x) \end{pmatrix} \mapsto (\overline{a_1(0)}, \overline{a_2(0)}, \overline{a_3(0)}, \overline{a_4(0)}, \overline{a_5(0)})$$

is a surjective homomorphism whose kernel is I. (Note that if $a(x) = n_0 + n_1x + \cdots n_dx^d \in \mathbb{Z}[x]$ then $a(0)$ is the constant term n_0, and $\overline{a(0)} = \overline{0}$, the zero of $\mathbb{Z}_2$, if and only if n_0 is even.)

(v) The set $\mathbb{Q}[\sqrt{2}] \overset{\text{def}}{=} \{\, a + b\sqrt{2} \mid a,\, b \in \mathbb{Q} \,\}$ is a subring of $\mathbb{R}$ isomorphic to the quotient ring $\mathbb{Q}[x]/(x^2-2)\mathbb{Q}[x]$. This is another example of the use of evaluation homomorphisms. Noting that $\mathbb{Q}$ is a subring of $\mathbb{R}$ (which is a commutative ring), we define $\theta\colon \mathbb{Q}[x] \to \mathbb{R}$ to be evaluation at $\sqrt{2}$. That is, $\theta(a(x)) = a(\sqrt{2})$ for all $a(x) \in \mathbb{R}[x]$. The theory we have discussed tells us that this is a homomorphism. By division of polynomials it can be seen that for every $a(x) \in \mathbb{Q}[x]$ there exist $q(x) \in \mathbb{Q}[x]$ and $c,\, d \in \mathbb{Q}$ with $a(x) = q(x)\,(x^2-2) + cx + d$, and we find that $a(\sqrt{2}) = c\sqrt{2} + d$. In particular this shows that $\theta(a(x)) \in \mathbb{Q}[\sqrt{}]$, and so $\operatorname{im}\theta \subseteq \mathbb{Q}[\sqrt{2}]$. On the other hand, an arbitrary element of $\mathbb{Q}[\sqrt{2}]$ has the form $c\sqrt{2} + d = \theta(cx + d)$ for some $c,\, d \in \mathbb{Q}$, and so the image of θ is the whole of $\mathbb{Q}[\sqrt{2}]$.

If $a(x) \in \ker\theta$ and $a(x) = q(x)\,(x^2-2) + cx + d$ as above, then $c\sqrt{2} + d = a(\sqrt{2}) = 0$. Since c and d are rational numbers this forces $c = d = 0$. (This is a fact that we will prove later, equivalent to the fact that $\sqrt{2}$ is irrational. For the time being, though, let us just assume it.) It follows that $a(x) \in K = (x^2-2)\mathbb{Q}[x]$. Conversely, if $a(x) \in K$ then $a(x) = q(x)\,(x^2-2)$ for some $q(x)$, and hence $\theta(a(x)) = 0$ (since $\theta(x^2-2) = (\sqrt{2})^2 - 2 = 0$). So $\ker\theta = K$.

Having determined the kernel and image of θ, we can apply the First Isomorphism Theorem. The conclusions are that $\mathbb{Q}[\sqrt{2}]$ (the image of θ) is a subring of $\mathbb{R}$, that $K = (x^2-2)\mathbb{Q}[x]$ (the kernel of θ) is an ideal of $\mathbb{Q}[x]$, and that there is an isomorphism $\mathbb{Q}[x]/K \to \mathbb{Q}[\sqrt{2}]$ such that $(cx + d) + K \mapsto c\sqrt{2} + d$ for all $c,\, d \in \mathbb{Q}$.

(vi) Another reassuring but not particularly interesting example is provided by the identity function from a ring R to itself: the function id defined by $\mathrm{id}a = a$ for all a. It is trivially a homomorphism—indeed, an isomorphism—from R to itself, and so the First Isomorphism Theorem applies. The kernel of id is the subset of R consisting of the zero element alone, and the image of id is the ring R itself. The conclusions of the theorem are that the kernel, $\{0_R\}$, is an ideal of R, the image (namely, R) is a subring of R, and $R/\{0_R\} \cong R$, with an isomorphism satisfying $a + \{0_R\} \mapsto a$. Observe that the coset $a + \{0_R\}$ equals $\{\, a + t \mid t \in \{0_R\} \,\} = \{a\}$, the subset of R consisting of the single element a. Since addition and multiplication of cosets are performed by adding or multiplying representative elements, and in this case there is always only one choice for the representative, we see that $\{a\} + \{b\} = \{a + b\}$ and $\{a\}\{b\} = \{ab\}$. So our isomorphism $R/\{0_R\}$ is the obvious bijective correspondence between singleton subsets and elements given by $\{a\} \mapsto a$, the set of singleton subsets having been made into a ring by the equally obvious definitions of addition and multiplication of singletons.

(vii) Recall that for any ring R there is a natural multiplication function $\mathbb{Z} \times R \to R$, for which we use the notation $(n, a) \mapsto na$. Suppose that R has a 1, and define a mapping $\mu\colon \mathbb{Z} \to R$ by $\mu n = n1$ for all $n \in \mathbb{Z}$. By properties of natural multiplication that we have already discussed, we know that for all $n,\, m \in \mathbb{Z}$,

$$\mu(n + m) = (n + m)1 = n1 + m1 = \mu n + \mu m$$
$$\mu(nm) = (nm)1 = n(m1) = (n1)(m1) = (\mu n)(\mu m),$$

whence μ is a homomorphism. By Theorem (7.4) (and the note following it) the kernel of μ has the form $m\mathbb{Z}$ for some nonnegative integer m. We call m the *characteristic* of the ring R. (Thus, the characteristic of R is the canonical generator of the kernel of μ.) The image of μ is the subset P of R consisting of all the natural multiples of 1. By the First Isomorphism Theorem P is a subring of R isomorphic to $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$, and the isomorphism which the theorem guarantees satisfies $\overline{n} \mapsto n1$ for all $n \in \mathbb{Z}$.

In the case where the characteristic is 0 then the kernel of μ is $\{0\}$, and μ is an embedding of $\mathbb{Z}$ in R. When the characteristic is nonzero it is the ring of integers modulo n that is embedded in R.

Observe that the characteristic of a ring R that has a 1 is the least positive integer m such that $m1 = 0_R$, or zero if there is no such positive integer. Note that if $k \in \mathbb{Z}$ satisfies $k1 = 0_R$ then $ka = (k1)a = 0_R a = 0_R$ for all $a \in R$. If the ring R does not have a 1 we define the characteristic of R to be the least positive integer m such that $ma = 0_R$ for all $a \in R$, or zero if there is no such positive integer.

The ring of integers modulo n provides the most obvious example of a ring of nonzero characteristic. Its characteristic is clearly n. Overfamiliarity with this example sometimes causes students to say that the characteristic of a ring is the number of elements in it (since $\mathbb{Z}_n$ also has n elements). Note, however, that the ring $\mathbb{Z}_n[x]$ of polynomials over $\mathbb{Z}_n$ in the indeterminate x also has characteristic n, yet has infinitely many elements. The reader can check that the ring $Z_3[x]/(x^2+1)\mathbb{Z}_3[x]$ (the quotient of $\mathbb{Z}_3[x]$ by the principal ideal generated by $x^2 + 1$) is a ring of characteristic 3 having nine elements. (Show, by considerations similar to those used in Example (i) above, that every element of $Z_3[x]/(x^2 + 1)\mathbb{Z}_3[x]$ is uniquely expressible in the form $(cx + d) + (x^2 + 1)\mathbb{Z}_3[x]$ with $c, d \in \mathbb{Z}_3$).

9. Principal Ideal Domains

Recall that a principal ideal, in a commutative ring R, is a single-generator ideal: an ideal of the form aR for some $a \in R$.

Definition (9.1): An integral domain in which every ideal is principal is called a *principal ideal domain*.

We saw in Theorem (7.4) that $\mathbb{Z}$ is a principal ideal domain. The proof made use of the division property: whenever $a, n \in \mathbb{Z}$ with $n \neq 0$ there exist unique $q, r \in \mathbb{Z}$ with $a = qn + r$ and $0 \leq r < |n|$. We did not prove this, but let us do so now. Define $S = \{\, a - kn \mid k \in \mathbb{Z} \,\}$, and $S' = \{\, s \in S \mid s \geq 0 \,\}$. It is clear that S' is nonempty; for example, if we put $k = -|a|n$ then $a - kn = a + |a|n^2 \geq -|a| + |a|n^2 = (n^2 - 1)|a| \geq 0$, since $n^2 \geq 1$ and $|a| \geq 0$. By the Least Integer Principle a nonempty set of nonnegative integers must have a least element, and so S' must have a least element. Choose $q \in \mathbb{Z}$ so that $r = a - qn$ is this least element. Then $a = qn + r$, and $r \geq 0$ since $r \in S'$. Furthermore, since $r - n = a - (q + 1)n$ is not in S', it must be negative. So $r < n$. As for uniqueness, observe that if $a = qn + r = q'n + r'$ with r and r' both in $\{0, 1, \ldots, n - 1\}$, then $|(q - q')n| = |r' - r| < n$; this forces $(q - q')n = 0$, giving $q = q'$ and hence $r = r'$.

If F is any field then $F[x]$, the ring of all polynomials over F, has a similar division property, which can be proved similarly. Since F is an integral domain (by Proposition (4.2)) we know that $F[x]$ is an integral domain (by Theorem (6.3)), and the division property will enable us to prove that $F[x]$ is a principal ideal domain (by imitating the proof of Theorem (7.4)).

Lemma (9.2): *Let F be a field and $f, a \in F[x]$ polynomials, with a nonzero. Then there exist unique polynomials $q, r \in F[x]$ such that $f = qa + r$ and $\deg r < \deg a$.*

Proof. Recall that we have defined the degree of the zero polynomial to be $-\infty$. For the purposes of this proof, the term "number" includes $-\infty$, and we adopt the natural conventions that $-\infty < n$ and $-\infty + n = -\infty = -\infty - \infty$ whenever n is a nonnegative integer. A trivial extension of the Least Integer Principle tells us that any nonempty subset of the set $\{-\infty\} \cup \{\, n \in \mathbb{Z} \mid n \geq 0 \,\}$ must have a least element.

Define $S = \{\, \deg(f - ga) \mid g \in F[x] \,\}$. That is, S is the set of all numbers that occur as degrees of polynomials of the form $f - ga$, as g varies over all elements of $F[x]$. (Note that f, g and a are all polynomials, despite the fact that we have chosen not to write them as $f(x)$, $g(x)$ and $a(x)$, as we could have done. We are under no compulsion to use the $f(x)$ notation for polynomials if a single letter will suffice, and in this proof the extra x's would only clutter the place up.) The set S is nonempty, containing (for example) the number $\deg f = \deg(f - 0a)$, and so it must have a least element, m (say). Choose $q \in F[x]$ so that $r = f - qa$ has degree m.

Suppose that $m \geq d = \deg a$, and remember that $d \geq 0$ since a is nonzero (by hypothesis). Write $a = a_0 + a_1x + \cdots + a_dx^d$ and $r = r_0 + r_1x + \cdots + r_mx^m$. Then a_d is the leading coefficient of a, and is therefore nonzero. Since F is a field, a_d^{-1} exists, and we see that

$$r_ma_d^{-1}x^{m-d}a = r_ma_0a_d^{-1}x^{m-d} + r_ma_1a_d^{-1}x^{m-d+1} + \cdots + r_ma_da_d^{-1}x^{m-d+d}$$

has the same degree m and leading coefficient r_m as has r. (Note that if m were less than d the above expression would not be a legitimate polynomial, since it would involve negative powers of x). It follows that $f - (q + r_ma_d^{-1}x^{m-d})a = r - r_ma_d^{-1}x^{m-d}a$ has degree less than m, since the terms r_mx^m cancel. But this contradicts the minimality of m, and so we conclude that $m < d$; that is, $\deg r < \deg a$.

It remains only to prove that q and r are unique. Accordingly, suppose also that $f = q_1a + r_1$, where $\deg r_1 < \deg a$. Then it follows that $\deg(r - r_1) < \deg a$ (since both r and r_1 have degree less than a), and furthermore $r - r_1 = (q_1 - q)a$ (since $q_1a + r_1 = f = qa + r$). Hence $\deg(q - q_1) + \deg a = \deg(r - r_1) \leq \deg a$, and so $\deg(q - q_1) < 0$. But the only polynomial with negative degree is the zero poynomial; so we conclude that $q - q_1 = 0$, which gives $q = q_1$ and hence $r = r_1$ also. □

We leave to the reader the proof of the following easy consequence of this division property.

Theorem (9.3): *(i) Let $f(x) \in F[x]$ and $t \in F$, where F is a field. Then $f(x) = (x - t)q(x) + r$ for some $q(x) \in F[x]$ and $r \in F$; moreover, the remainder r is equal to $f(t)$.*
(ii) With $f(x)$ and t as in Part (i), $x - t$ is a factor of $f(x)$ if and only if $f(t) = 0$.

Part (i) of Theorem (9.3) is sometimes called the "Remainder Theorem", and Part (ii) the "Factor Theorem". However, "Factors-of-degree-one Theorem" would be better, since it only tells one about factors of degree 1. It is of course quite possible for a polynomial to have a nontrivial factorization in $F[x]$ without having any factors of degree 1. For example, over the real field $\mathbb{R}$ the polynomial $x^4 + 5x^2 + 4$ factorizes as $(x^2 + 1)(x^2 + 4)$, but has no factors of degree 1, and hence no roots in R. This is a point worth emphasizing: a polynomial which has a root has a factor of degree 1 and therefore is not irreducible, but polynomials with no roots are not necessarily irreducible since they may have nontrivial factors of degree greater than 1. In general, it is not easy to prove that a polynomial is irreducible, and to do so involves more than merely proving that it has no roots.

Theorem (9.4): *If F is a field then $F[x]$ is a principal ideal domain. Moreover, if I is a nonzero ideal in $F[x]$ then any nonzero element of I of minimal degree will generate I.*

Proof. As explained above, $F[x]$ is an integral domain; so we only have to prove that all the ideals of $F[x]$ are principal. So let I be an ideal in $F[x]$, and suppose first that $I \neq \{0\}$. Since $0 \in I$, this assumption yields that the set $I' = \{\, p \in I \mid p \neq 0 \,\}$ is nonempty, and so by the Least Integer Principle we may choose an element $p \in I'$ of minimal degree.

Since $p \in I$ it follows that $pq \in I$ for all $q \in F[x]$, and hence $pF[x] \subseteq I$. But if $f \in I$ is arbitrary then since $p \neq 0$ we may use Lemma (9.2) to deduce that $f = pq + r$ for some q, $r \in F[x]$, with $\deg r < \deg p$. Since f, $pq \in I$ the closure properties of ideals yield that $r = f - pq \in I$. But $r \notin I'$, since $\deg r < \deg p$, and by definition p is an element of I' of minimal degree. This forces $r = 0$, the only element of I that is not in I'. Hence $f = pq \in pF[x]$, and since f was an arbitrary element of I it follows that $I \subseteq pF[x]$. Hence $I = pF[x]$, a principal ideal. As the only other possibility is $I = \{0\} = 0F[x]$, which is also principal, it follows that all ideals of $F[x]$ are principal, as required. □

It is a familiar fact that every integer greater than 1 can be factorized uniquely as a product of primes. The idea of factorizing polynomials is also familiar. If F is a field and $p \in F[x]$ a polynomial of degree greater than 0, we say that p is irreducible if it has no factors other than nonzero scalar polynomials and nonzero scalar multiples of itself. It turns out that every nonzero polynomial can be factorized as a product of irreducible polynomials, and the irreducible polynomials that arise are unique to within scalar factors. These properties of factorization of integers and polynomials are two examples of a general fact about principal ideal domains, which can be expressed succinctly as follows: every principal ideal domain is a unique factorization domain. Before embarking on the proof of this, we need to make a few definitions. Although some of these definitions will be stated for an arbitrary ring with 1, some for an arbitrary commutative ring, some for a commutative ring with 1, and some for an integral domain, it is always the case that we are primarily interested in integral domains. You may, if you wish, assume throughout that R is an integral domain.

Definition (9.5): Let R be a ring with 1. An element $u \in R$ is called a *unit* if it has an inverse. That is, u is a unit if there exists $v \in R$ with $uv = vu = 1$.

Recall that if u has an inverse then that inverse is unique, and so we may denote it by u^{-1}. It is clear that if $v = u^{-1}$ then also $v = u^{-1}$; so the inverse of a unit is a unit. Furthermore, if u and t are both units then ut is also a unit, its inverse being $t^{-1}u^{-1}$. Note that in the ring $\mathbb{Z}$ the only units are 1 and -1, in a field F all nonzero elements are units, and in the ring $F[x]$ (where F is a field) the units are precisely the polynomials of degree 0.

Exercise 16. Show that if R is a commutative ring with 1 then $a \in R$ is a unit if and only if $aR = R$.

Definition (9.6): Let R be a commutative ring with 1, and let $a, b \in R$. We say that b is an *associate* of a if $b = ua$ for some unit u.

Note that if $b = ua$ where u is a unit, then $a = u^{-1}b$. Since u^{-1} is also a unit we conclude that a is an associate of b.

Exercise 17. Let R be a commutative ring with 1, and define a relation $\approx$ on R by $a \approx b$ if and only if a is an associate of b. Prove that $\approx$ is an equivalence relation.

Definition (9.7): If R is a commutative ring and $a, b \in R$, we say that *a divides b*, or that b is *divisible* by a, or that a is a *factor* or *divisor* of b, if $b = ca$ for some $c \in R$. Alternatively, we could say that b is a *multiple* of a.

We use the notation $a \mid b$ to mean that a divides b. Thus, for example, in the ring $\mathbb{Z}$ we have $4 \mid 20$.

There is a mild inconsistency in our terminology at this point. Observe that $a \mid 0$ is true for all $a \in R$, since $0 = 0a$. So according to Definition (9.7) we can say that a is a divisor of 0. However, most algebraists would interpret the statement “a is a divisor of zero” as meaning that a is a zero divisor in the sense that $a \neq 0$ and $ab = 0$ for some $b \neq 0$. In view of this, we will avoid ever saying “a is a divisor of 0”, although we will still, if the need arises, write “$a \mid 0$” to mean “$0 = ax$ for some x” (which will always be true!), and we will continue to use the term “zero divisors” for nonzero elements whose product is 0.

Exercise 18. Prove that the “divides” relation is transitive: if $a \mid b$ and $b \mid c$ then $a \mid c$.

Exercise 19. Prove that if $a \mid b$ and $a = 0$ then $b = 0$.

Exercise 20. Let R be a commutative ring with 1, and $a, b \in R$. Prove that $a \mid b$ if and only if $bR \subseteq aR$.

Exercise 21. Let R be an integral domain, and $a, b \in R$. Prove that a and b are associates if and only if $a \mid b$ and $b \mid a$.

Exercise 22. Prove that if R is an integral domain and $a, b \in R$, then $aR = bR$ if and only if a and b are associates.

Definition (9.8): Let R be an integral domain. A nonzero element of R is said to be *irreducible* if it is not a unit and its only divisors are units and associates of itself.

Thus if $a \in R$ is irreducible, and $a = bc$ for some $b, c \in R$, then either b or c must be a unit, and the other an associate of a.

Exercise 23. Show that an associate of an irreducible element must also be irreducible.

Definition (9.9): We say that an integral domain R is a *unique factorization domain* if the following two conditions are satisfied:

(i) for every non-unit $a \in R$ there exists a positive integer n and elements $p_1, p_2, \ldots, p_n \in R$ which are irreducible and satisfy $a = p_1 p_2 \cdots p_n$;

(ii) if n, m are nonnegative integers and $p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m$ for some irreducible elements $p_1, p_2, \ldots, p_n$ and $q_1, q_2, \ldots, q_m$ of R, then $m = n$, and the q_i can be renumbered so that q_i is an associate of p_i for all i from 1 to n.

Part (ii) of this definition says that factorization into irreducibles is as unique as it could possibly be, given that multiplication is commutative in an integral domain, and given that there may be units in R. A factorization can be trivially altered by rearranging the order of the factors, or by multiplying one factor by a unit and another factor by the inverse of that unit. So in $\mathbb{Z}$, for example, 30 can be factorized as $2 \times 3 \times 5$, or as $(-5) \times 2 \times (-3)$, and in $\mathbb{Q}[x]$ the element $x^4 + x^3 - x^2 + x - 2$ can be factorised as $(x^2 + 1)(x + 2)(x - 1)$, or as $(\frac{1}{2}x - \frac{1}{2})(\frac{1}{3}x + \frac{2}{3})(6x^2 + 6)$. Part (ii) of the definition requires that if R is a unique factorization domain then trivial modifications of this kind produce the only alternative factorizations that are ever possible.

Let R be an integral domain, and suppose that a is a nonzero element of R which is not a unit. If it is not irreducible then it has a divisor which is neither a unit nor an associate of itself, and if that divisor is not irreducible then it in turn has a proper divisor, and so on. Is it guaranteed that this process must halt, after a finite number of steps, with an irreducible element? Or is it possible that there might be an infinite sequence of elements, $a_1, a_2, a_3, \ldots$, with the property that $a_{i+1} \mid a_i$, and a_{i+1} is not a unit or an associate of a_i, for all $i \geq 1$? We shall show that this situation cannot arise in a principal ideal domain. Observe that by Exercise 20 above the condition $a_{i+1} \mid a_i$ is equivalent to $a_i R \subseteq a_{i+1} R$, while combining Exercise 20 and Exercise 21 we see that $a_i R = a_{i+1} R$ if and only if a_i and $a_{i+1} R$ are associates. So the question becomes whether it is possible to have an infinite strictly increasing chain of principal ideals $a_1 R \subsetneq a_2 R \subsetneq a_3 R \subsetneq \cdots$ in the domain R.

Definition (9.10): A ring R is said to satisfy the *ascending chain condition* on ideals if whenever $I_1, I_2, I_3, \ldots$ is an infinite sequence of ideals of R there exists a positive integer n such that $I_m = I_n$ for all $m \geq n$.

Thus the ascending chain condition says that ascending chains cannot go on getting strictly bigger indefinitely; at some point the sequence effectively terminates.

Lemma (9.11): *Let $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$ be an ascending chain of ideals in the ring R. Then $I = \bigcup_{k=1}^{\infty} I_k$ is an ideal in R.*

Proof. We use the criterion from Theorem (7.3) for a subset of a ring to be an ideal. Suppose that $a, b \in I$. Since I is the union of the I_k's, this means that $a \in I_h$ and $b \in I_k$ for some h and k. If we put $m = \max\{h, k\}$ then since the chain of ideals ascends it follows that $I_h \subseteq I_m$ and $I_k \subseteq I_m$, and therefore $a, b \in I_m$. But since I_m is an ideal it is closed under addition; so $a + b \in I_m$, and hence $a + b \in I$ (since $I_m \subseteq I$). It follows that I is closed under addition.

The rest of the proof of Lemma (9.11) is left to the reader. □

The ideal I in Lemma (9.11) can be termed the *limit* of the increasing sequence $(I_j)_{j=1}^{\infty}$; it clearly is the smallest set containing all the I_j.

We are now able to prove that principal ideal domains satisfy the ascending chain condition on ideals. In fact, since no extra effort is required, we prove a slightly stronger result.

Theorem (9.12): *Let R be a commutative ring with 1, and suppose that R has the property that every ideal of R is finitely generated. Then R satisfies the ascending chain condition on ideals.*

Proof. Recall that an in ideal of R is finitely generated if it has the form $a_1R + a_2R + \cdots + a_kR$ for some positive integer k; principal ideals satisfy this with $k = 1$. If $I = a_1R + a_2R + \cdots + a_kR$ then $I = \{\, a_1r_1 + a_2r_2 + \cdots + a_kr_k \mid r_i \in R \,\}$. We call the a_i *generators* of the ideal I; note that the generators are certainly elements of I, since $a_i = \sum_j a_jr_j$ holds if we put $r_i = 1$ and $r_j = 0$ for all $j \neq i$.

Let $I_1 \subseteq I_2 \subseteq I_3 \cdots$ be an increasing chain of ideals in R. Put $I = \bigcup_{j=1}^{\infty} I_j$, and note by Lemma (9.11) that I is an ideal. Since all ideals of R are finitely generated we can find $a_1, a_2, \ldots, a_k \in R$ that generate I. Since $a_i \in I = \bigcup_{j=1}^{\infty} I_j$ for each i, we can find a positive integer $j(i)$ such that $a_i \in I_{j(i)}$, and if we put $n = \max_{1 \le i \le k}\{j(i)\}$ then the fact that the sequence $(I_j)_{j=1}^{\infty}$ is increasing ensures that $I_{j(i)} \subseteq I_n$, and hence $a_i \in I_n$, for all $i \in \{1, 2, \ldots, k\}$. Now closure of the ideal I_n under multiplication by elements of R yields that $a_ir_i \in I_n$ for any choice of elements $r_i \in R$, and hence closure of I_n under addition yields that $a_1r_1 + a_2r_2 + \cdots + a_kr_k \in I_n$ for all r_i. But since I is precisely the set of all elements of the form $a_1r_1 + a_2r_2 + \cdots + a_kr_k$, we have proved that $I \subseteq I_n$. Now because I is the limit of the increasing sequence $(I_j)_{j=1}^{\infty}$ we deduce that

$$I \subseteq I_n \subseteq I_m \subseteq \bigcup_{j=1}^{\infty} I_j = I$$

whenever $m \geq n$. Thus $I_m = I_n$ for all $m \geq n$, as required. □

It is a consequence of Theorem (9.12) that every nonzero non-unit in a principal ideal domain has an irreducible divisor. (We should note that it is part of Definition (9.8) that an irreducible element is never a unit. Since a unit of R is a divisor of every element of R, the proposition we are about to prove would be trivial if there were such a thing as an irreducible unit!)

Proposition (9.13): *Let R be a principal ideal domain and $a \in R$ a nonzero element which is not a unit. Then there exists an irreducible element $b \in R$ such that $b \mid a$.*

Proof. Suppose to the contrary that a does not have any irreducible factor. We define $a_1 = a$, and, proceeding inductively, prove that the following condition holds for all positive integers i:

(C) if $i > 1$ then there exists a nonzero element $a_i \in R$ which not a unit, and is a divisor but not an associate of a_{i-1}.

Observe that (C) holds by default in the case $i = 1$.

Suppose that $k > 1$, and that (C) holds for all $i \in \{1, 2, \ldots, k-1\}$. Since $a_{k-1} \mid a_{k-2} \mid \cdots \mid a_1$, it follows (by repeated use of Exercise 18) that $a_{k-1} \mid a_1 = a$, and since a has no irreducible factors, we deduce that a_{k-1} is not irreducible. Accordingly, since a_{k-1} is nonzero and not a unit, there exists an element $a_k \in R$ such that a_k is is a divisor of a_{k-1} but not a unit or an associate of a_{k-1}. Obviously $a_k \neq 0$, since $a_{k-1} \neq 0$ (see Exercise 19). Thus we have shown that (C) also holds for $i = k$, and by induction therefore it holds for all positive integers i.

We have thus obtained an infinite sequence a_1, a_2, a_3, ... such that each a_j is a divisor but not an associate of its predecessor. If we now consider the principal ideals they generate then we have, in view of Exercise 20, that

$$a_1R \subseteq a_2R \subseteq a_3R \subseteq \cdots ,$$

and hence by Theorem (9.12) there exists an n such that $a_{n+1}R = a_nR$. But by Exercise 20 this implies that $a_n \mid a_{n+1}$ as well as $a_{n+1} \mid a_n$, so that by Exercise 21 it follows that a_{n+1} is an associate of a_n, contrary to the construction. This contradiction shows that our original assumption must be false, and therefore a has an irreducible factor. □

A similar kind of argument can now be used to show that every nonzero non-unit in a principal ideal domain has a factorization as a product of irreducible elements.

Proposition (9.14): *Let R be a principal ideal domain and $a \in R$ a nonzero element which is not a unit. Then there exists a positive integer k and irreducible elements p_1, p_2, ..., $p_k \in R$ such that $a = p_1p_2 \cdots p_k$.*

Proof. Suppose to the contrary that a cannot be expressed as a product of irreducibles. Note in particular that a is not itself irreducible, or else taking $k = 1$ and $p_1 = a$ would give $a = p_1p_2 \cdots p_k$ with all p_i irreducible. We now define $a_1 = a$, and, proceeding inductively, prove that the following condition holds for all positive integers i:

(C′) there exist nonzero elements p_i, $a_{i+1} \in R$ such that p_i is irreducible and not an associate of a_i, and $a_i = p_ia_{i+1}$.

Remember that p_i, being irreducible, cannot be a unit.

Suppose first that $i = 1$. Since $a_1 = a$ is nonzero and not a unit, Lemma (9.13) guarantees the existence of an irreducible element $p_1 \in R$ such that $a_1 = p_1a_2$ for some $a_2 \in R$, and in view of Exercise 23 and the fact that a_1 is not irreducible, p_1 cannot be an associate of a_1. Hence condition (C′) holds in this case.

Suppose that $k > 1$, and that (C′) holds for all $i \in \{1, 2, \ldots, k-1\}$. Observe that we have

$$a = a_1 = p_1a_2 = p_1(p_2a_3) = p_1p_2(p_3a_4) = \cdots = p_1p_2 \cdots p_{k-1}a_k.$$

If a_k were irreducible this would contradict our assumption that a cannot be expressed as a product of irreducibles. Furthermore, a_k is not a unit, for if it were then the fact that $a_{k-1} = p_{k-1}a_k$ would imply that p_{k-1} is an associate of a_{k-1}, contrary to (C′) for $i = k-1$. Hence Lemma (9.13) yields that there exists $p_k \in R$ such that $a_k = p_ka_{k+1}$ for some $a_{k+1} \in R$. By Exercise 23 and the fact that a_k is not irreducible, p_k is not an associate of a_k. Thus we have shown that (C′) also holds for $i = k$, and by induction therefore it holds for all positive integers i.

We have thus obtained an infinite sequence a_1, a_2, a_3, ... such that each a_j is a divisor of its predecessor. If we now consider the principal ideals they generate then we have, in view of Exercise 20, that

$$a_1R \subseteq a_2R \subseteq a_3R \subseteq \cdots ,$$

and hence by Theorem (9.12) there exists an n such that $a_{n+1}R = a_nR$. But by Exercise 20 this implies that $a_{n+1} = q_na_n$ for some $q_n \in R$, and combined with $a_n = p_na_{n+1}$ this gives $a_n = p_nq_na_n$. Since $a_n \neq 0$ we can use Exercise 3 to conclude that $p_nq_n = 1$. But this shows that p_n is a unit, contrary to the fact that it is an irreducible element. This contradicts the fact that p_n is irreducible, since an irreducible element cannot be a unit. This contradiction shows that our original assumption must be false, and therefore a has an expression as a product of irreducible factors. □

Before we can complete the proof that a principal ideal domain has to be a unique factorization domain, we need to discuss primes and greatest common divisors. It is a familiar property of $\mathbb{Z}$ that if an integer p is a prime, and if p is a divisor of the product ab of integers a and b, then p must be a divisor of a or a divisor of b. We use this property as the definition of primality in other commutative rings.

Definition (9.15): Let R be a commutative ring, and $p \in R$. We say that p is *prime* if it is nonzero and not a unit, and the following condition holds: for all $a, b \in R$, if $p \mid ab$ then either $p \mid a$ or $p \mid b$.

The elements of $\mathbb{Z}$ which satisfy this are exactly the prime numbers in the usual sense, together with their negatives. This is exactly the set of irreducible elements of $\mathbb{Z}$. It is not the case that the concepts of “prime” and “irreducible” coincide for all integral domains, although it is easily seen that a prime element of an integral domain has to be irreducible. In the case of principal ideal domains the converse does hold, as we shall shortly prove: irreducible elements have to be prime. This is rather convenient, since it means that principal ideal domains are very similar in behaviour and spirit to the familiar integral domain $\mathbb{Z}$, and the integral domains that are most important in this course are all principal ideal domains.

Exercise 24. Give an example of integers n, a and b such that $n \mid ab$ but $n \nmid a$ and $n \nmid b$.

Exercise 25. Prove that a prime element of an arbitrary integral domain R is necessarily irreducible.

Definition (9.16): Let R be a principal ideal domain and a, b nonzero elements of R. An element $d \in R$ is called a *greatest common divisor* of a and b if

(i) $d \mid a$ and $d \mid b$, and

(ii) for all $e \in R$, if $e \mid a$ and $e \mid b$ then $e \mid d$.

Condition (i) says that d is a common divisor; Condition (ii) says that every other common divisor is smaller than d, in the sense of being a divisor of d. It is an important property of principal ideal domains that greatest common divisors always exist.

Theorem (9.17): *Let R be a principal ideal domain, and $a, b \in R$ nonzero elements. Then*

(i) there is an element $d \in R$ which is a greatest common divisor of a and b, and every associate of d is also a greatest common divisor of a and b.

(ii) if d and d' are both greatest common divisors of a and b then d and d' are associates of each other;

(iii) an element $d \in R$ is a greatest common divisor of a and b if and only if d is a divisor of a and b such that $d = ar + bs$ for some $r, s \in R$;

(iv) an element $d \in R$ is a greatest common divisor of a and b if and only if $aR + bR = dR$.

Proof. We prove (ii) first. If d is a greatest common divisor of a and b then by Definition (9.16) we have

(A) $d \mid a$ and $d \mid b$;

(B) if $e \mid a$ and $e \mid b$ then $e \mid d$.

If d' is also a greatest common divisor of a and b then similarly

(A′) $d' \mid a$ and $d' \mid b$;

(B′) if $e \mid a$ and $e \mid b$ then $e \mid d'$.

Now (A) and (B′) together imply that $d \mid d'$, while (A′) and (B) together imply that $d' \mid d$. By Exercise 21 it follows that d and d' are associates, as claimed.

We saw in an example in Section 7 that $aR + bR = \{\, ax + by \mid x,\ y \in R \,\}$ is always an ideal (given that R is a commutative ring). Since we are assuming that R is a principal ideal domain, $aR + bR$ must be a principal ideal. That is, there exists some $d \in R$ such that $dR = aR + bR$. By Exercise 22, if d' is any associate of d then $d'R = aR + bR$ also. If Part (iv) of the theorem is true then it will follow that d and allits associates are greatest common divisors of a and b. So Part (i) of the theorem will follow as a consequence of Part (iv). We will prove Parts (iii) and (iv) together.

Suppose that d is a divisor of a and of b, and that $d = ax + by$ for some $x,\ y \in R$. Then certainly Condition (i) of Definition (9.16) is satisfied. Furthermore, if $e \mid a$ and $e \mid b$ then we can write $a = er$ and $b = es$ for some $r,\ s \in R$, and we obtain

$$d = ax + by = (er)x + (es)y = e(rx) + e(sy) = e(rx + sy),$$

which shows that $e \mid d$. Hence Condition (ii) of Definition (9.16) is also satisfied, and we conclude that d is a greatest common divisor of a and b. This has proved one of the implications required for Part (iii).

Suppose next that d is a greatest common divisor of a and b. By Condition (i) of Definition (9.16) there exist $r,\ s \in R$ with $a = dr$ and $b = ds$. If now t is an arbitrary element of $aR + bR$ then we have $t = ax + by$ for some $a,\ b \in R$, and hence

$$t = (dr)x + (ds)y = d(rx) + d(sy) = d(rx + sy) \in dR,$$

and so it follows that $aR + bR \subseteq dR$. On the other hand, since R is a principal ideal domain and $aR + bR$ is an ideal, we know that $aR + bR = eR$ for some $e \in R$, and since

$$a = a1 + b0 \in \{\, ax + by \mid x,\ y \in R \,\} = aR + bR = eR,$$

it follows that $a = ez$ for some $z \in R$. Hence $e \mid a$. Similarly,

$$b = a0 + b1 \in \{\, ax + by \mid x,\ y \in R \,\} = aR + bR = eR,$$

and so $e \mid b$ also. Now by Condition (ii) of Definition (9.16) it follows that $e \mid d$, and so Exercise 20 yields that $dR \subseteq eR = aR + bR$. We had proved the reverse inclusion above, and so we have shown that if d is a greatest common divisor of a and b then $dR = aR + bR$. This is one of the implications required for Part (iv).

To complete the proof of Parts (iii) and (iv) it remains to show that if $dR = aR + bR$ then d is a divisor of a and b such that $d = ax + by$ for some $x,\ y \in R$. But if $dR = aR + bR$ then

$$d = d1 \in dR = aR + bR = \{\, ax + by \mid x,\ y \in R \,\},$$

so that $d = ax + by$ for some $x,\ y \in R$, and furthermore

$$a = a1 + b0 \in \{\, ax + by \mid x,\ y \in R \,\} = aR + bR = dR$$

and

$$b = a0 + b1 \in \{\, ax + by \mid x,\ y \in R \,\} = aR + bR = eR,$$

show that $d \mid a$ and $d \mid b$, as required. □

Theorem (9.18): *Let R be a principal ideal domain, and let $r \in R$ be an irreducible element. Then r is prime. That is, for all $a, b \in R$, if r is a factor of ab then it is a factor of a or of b.*

Proof. Let $a, b \in R$ be such that $r \mid ab$. The $ab = rt$ for some $t \in R$. We must show that either $r \mid a$ or $r \mid b$, which is the same as showing that if $r \nmid a$ then $r \mid b$. So assume that $r \nmid a$.

Let d be a greatest common divisor of r and a. Then $d \mid r$, and since r is irreducible it follows that d is either an associate of r or a unit. If d is an associate of r then we can write $d = ru$ for some unit u. But since d is a common divisor of r and a we also know that $a = ds$ for some $s \in R$, and thus $a = (ru)s = r(us)$, contradicting our assumption that $r \nmid a$. So we are forced into the alternative scenario, in which d is a unit.

Let $d' = d^{-1}$. By Theorem (9.17) we have that $dR = rR + aR$, so that

$$1 = dd' \in dR = rR + aR = \{\, rx + ay \mid x, y \in R \,\},$$

and it follows that $1 = rx + ay$ for some $x, y \in R$. Multiplying through by b and using $ab = rt$ gives

$$b = (rx + ay)b = rxb + (ab)y = rxb + (rt)y = r(xb + ty),$$

which shows that $r \mid b$, as required. □

An easy induction based on Theorem (9.18) yields the following corollary, whose proof we leave to the reader.

Corollary (9.19): *Let R be a principal ideal domain, and let $p \in R$ be an irreducible element. If $q_1, q_2, \ldots, q_l \in R$ are such that p is a divisor of the product $q_1 q_2 \cdots q_l$, then $p \mid q_j$ for some j.*

We are now able to complete the proof of the main theorem of this section.

Theorem (9.20): *Let R be a principal ideal domain. Then R is also a unique factorization domain.*

Proof. We have already shown in Theorem (9.14) that every nonzero element of R which is not a unit can be expressed as a product of irreducible elements, and so all that remains is to prove the uniqueness property. We will do this somewhat informally, since it would become tedious otherwise.

Suppose that $p_1, p_2, \ldots, p_k$ and $q_1, q_2, \ldots, q_l$ are irreducible elements of R having the property that $p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_l$. We see that

$$p_1 \mid p_1(p_2 \cdots p_k) = q_1 q_2 \cdots q_l,$$

and since p_1 is irreducible we have by Corollary (9.19) that $p_1 \mid q_j$ for some j. As q_j is also irreducible, its only divisors are units and associates of itself, and therefore p_1 is an associate of q_j (since p_1, being irreducible, is not a unit).

Renumbering the q_i if necessary, we may assume that $j = 1$. We now have that $q_1 = p_1 u$ for some unit u, and therefore

$$p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_l = (p_1 u) q_2 \cdots q_l.$$

By Exercise 3 we deduce that $p_2 p_3 \cdots p_k = q_2' q_3 \cdots q_l$, where $q_2' = q_2 u$ is an associate of q_2 (which is still irreducible, by Exercise 23). We now repeat the argument. At each stage we find that one of the q's is an associate of one of the p's, and after possibly renumbering the factors and/or replacing them by associates, this pair can cancelled away, reducing the number of factors on both sides. Eventually there will be no factors left one one side. If there were then any factors left on the other side we would have a product of irreducible elements equalling 1, which is contrary to the fact that an irreducible element cannot have an inverse. So when the last factor disappears from one side, the last factor simultaneously disappears from the other side. So the p's and q's match up exactly, as required. □

Exercise 26. Prove that the real numbers $\sqrt{2}$ and $\sqrt[3]{2}$ are not rational. (Hint: Suppose that $\sqrt{2} = p/q$ with $p, q \in \mathbb{Z}$. Then $p^2 = 2q^2$; now if we express p and q as products of primes we find that the multiplicity with which 2 occurs in the prime factorization of p^2 is even (being twice its multiplicity in p), whereas its multiplicity as a factor of $2q^2$ is odd (twice its multiplicity in q plus one). This contradicts the uniqueness of prime factorizations in $\mathbb{Z}$.)

Calculating greatest common divisors: the Euclidean algorithm

Since every associate of a greatest common divisor of two elements is also a greatest common divisor of those two element, it is rare for greatest common divisors to be unique. Indeed, if the domain in question has any units other than the identity element, then greatest common divisors cannot be unique. However, it would certainly be nicer if we could talk of *the* greatest common divisor, rather than *a* greatest common divisor.

In $\mathbb{Z}$ there are exactly two units, 1 and -1. So for every pair of nonzero integers there are two gcd's, one positive and the other negative. In this case it is customary to call the positive one *the* gcd. In the case of $F[x]$, where F is a field, the units are the nonzero scalar polynomials, and it can be seen that every nonzero polynomial has a unique associate whose leading coefficient is 1 (obtained by multiplying through by the unit which is the inverse of the leading coefficient). A polynomial is said to be *monic* if its leading coefficient is 1, and it is customary to define the gcd of two nonzero polynomials to be the monic polynomial that satisfies Definition (9.16). For other principal ideal domains there may be no particularly natural way to choose a preferred gcd. Nevertheless, we will still find it convenient to say things like "let $d = \gcd(a,b)$" to mean "let d be one of the gcd's of a and b", remembering that d is only unique up to multiplication by a unit.

Some integral domains, notably $\mathbb{Z}$ and $F[x]$ for any field F, have the property that if a and b are arbitrary elements, with b nonzero, then one can find elements q (the *quotient*) and r (the *remainder*) such that $a = qb + r$, and r is in some sense smaller than b. More precisely, it should be possible to assign a nonnegative integer $\deg t$ to each nonzero element t of the domain, and the remainder t should either be zero or else satisfy $\deg r < \deg b$. Observe that $\mathbb{Z}$ has this property if we define $\deg n = |n|$ for all integers n, while for $F[x]$ we can let $\deg p$ be the degree, in the usual sense, of the polynomial p. Integral domains with this property are called *Euclidean domains*.† The same argument used to prove Theorems (7.4) and (9.4) can be applied to show that a Euclidean domain is necessarily a principal ideal domain.

In any Euclidean domain there is a procedure, known as the *Euclidean algorithm*, which can be used to calculate gcd's. Assume that R is a Euclidean domain, and for each pair of elements (a,b) that are not both zero define

$$D(a,b) \overset{\text{def}}{=} \{\, e \in R \mid e \mid a \text{ and } e \mid b \,\},$$

the set of all common divisors of a and b. Observe that since $e \mid 0$ is true for all $e \in R$, if one of a or b is zero then $D(a,b)$ is just the set of divisors of the other. Rephrasing Definition (9.16), we see that $d = \gcd(a,b)$ if and only if $d \in D(a,b)$ and $e \mid d$ for all $e \in D(a,b)$. Note that, obviously, $D(b,a) = D(a,b)$.

† If R is a Euclidean domain then $\deg 0$ may or may not be defined. If it is not otherwise defined, we may put $\deg 0 = -\infty$, as we did for polynomials.

The key to the Euclidean algorithm is the following lemma.

Lemma (9.21): *Let a, b, $m \in R$ with $b \neq 0$. Then*
(i) $D(a,b) = D(b, a+mb)$, *and*
(ii) $\gcd(a,b) = \gcd(b, a+mb)$.

Proof. Let $e \in D(a,b)$. Then $a = re$ and $b = se$ for some r, $s \in R$, giving $a + mb = (r + ms)e$. Thus $e \mid b$ and $e \mid (a + mb)$; so $e \in D(b, a+mb)$. Hence $D(a,b) \subseteq D(b, a+mb)$. Since this argument works for all a and m we can replace a by $a + mb$ and m by $-m$, to conclude that $D(a+mb, b) \subseteq D(b, (a+mb) - mb) = D(b,a)$. Since we now have both inclusions, it follows that $D(a,b) = D(b, a+mb)$.

As remarked above, $d = \gcd(a,b)$ is an element of $D(a,b)$ that is divisible by all elements of $D(a,b)$. So, by the first part, d is an element of $D(b, a+mb)$ divisible by all elements of $D(b, a+mb)$; hence $d = \gcd(b, a+md)$. □

If $a \neq 0$ the $D(a,0)$ is the set of all divisors of a, and so it makes sense to define $\gcd(a,0) = a$ (or any associate of a). The Euclidean algorithm is a trivial recursive procedure which starts with an arbitrary pair of elements of R which are not both zero, and replaces them with another pair which have the same gcd, while reducing the value of the deg function. The algorithm terminates when one of the pair of elements it yields is zero; the gcd is then the nonzero element of the pair.

Since R is a Euclidean domain, if $b \in R$ is nonzero then for any $a \in R$ there exists $q \in R$ such that $r = a - qb$, the remainder on division of a by b, satisfies $\deg r < \deg b$. Furthermore, Lemma (9.21) shows that $\gcd(a,b) = \gcd(b,r)$. Writing $\mathrm{Rem}(a,b)$ for the remainder on division of a by b, we can state the Euclidean algorithm as follows.

Euclidean Algorithm.

```
while $b \neq 0$ do
        $[a,b] := [b, \mathrm{Rem}(a,b)]$
enddo
return $a$
```

Examples

(i) We calculate $\gcd(132, 102)$ by the Euclidean algorithm:

$$132 = 1 \times 102 + 30 \tag{8}$$
$$102 = 3 \times 30 + 12 \tag{9}$$
$$30 = 2 \times 12 + 6 \tag{10}$$
$$12 = 2 \times 6.$$

This shows that $\gcd(132,102) = \gcd(102,30) = \gcd(30,12) = \gcd(12,6) = \gcd(6,0) = 6$. Note also that using Eq's (8), (9) and (10) in reverse order enables us to express 6 in the form $132r + 102s$ for some integers r and s:

$$\begin{aligned} 6 &= 30 - 2 \times 12 \\ &= 30 - 2 \times (102 - 3 \times 30) = 7 \times 30 - 2 \times 102 \\ &= 7 \times (132 - 102) - 2 \times 102 = 7 \times 132 - 9 \times 102. \end{aligned}$$

In other words, $r = 7$ and $s = -9$ is a solution. Note that it is not the only solution. In fact, $r = 7 + 17n$ and $s = -9 - 22n$ is a solution for each integer n. Check this!

(ii) Let $F = \mathbb{Z}_3$, the ring of integers modulo 3. It happens to be true that $\mathbb{Z}_p$ is a field whenever p is a prime number; so, in particular, F is a field. In this example we shall represent elements of F by integers, remembering that integers which are congruent modulo 3 represent the same element of F. (Thus $2 = -1 = 8$ in F, $4 = 1$, and so on.) Let us use the Euclidean algorithm to find the gcd of the elements $x^6 + 2x^5 + x^4 + 2x^3 + 1$ and $x^4 + 2x^3 + 2x$ of $F[x]$:

$$\begin{aligned}
x^6 + 2x^5 + x^4 + 2x^3 + 1 &= (x^2 + 1)(x^4 + 2x^3 + 2x) + (x^3 + x + 1) && (11)\\
x^4 + 2x^3 + 2x &= (x + 2)(x^3 + x + 1) + (2x^2 + 2x + 1) && (12)\\
x^3 + x + 1 &= (2x + 1)(2x^2 + 2x + 1).
\end{aligned}$$

The last nonzero remainder is the gcd. However, it is conventional to multiply through by a scalar factor chosen so that the leading coefficient becomes 1; if we do this, the gcd is $2(2x^2 + 2x + 1) = x^2 + x + 2$.

Each one of the steps above involves a polynomial division. Here are the calculations for the first of these:

$$\begin{array}{r|llllllll}
\multicolumn{1}{r}{} & x^2 & & +1 & & & & \\
\cline{2-8}
x^4 + 2x^3 + 2x & x^6 & +2x^5 & +x^4 & +2x^3 & & & +1 \\
 & x^6 & +2x^5 & & +2x^3 & & & \\
\cline{2-5}
 & & & x^4 & & & & +1 \\
 & & & x^4 & +2x^3 & & +2x & \\
\cline{4-8}
 & & & & x^3 & & +x & +1.
\end{array}$$

Doing arithmetic modulo 3 takes a little getting used to! However, after a small amount of practice things like $1-2 = 2$ and $2\times 2 = 1$ become quite routine. Please check the calculations above.

Again, the gcd can be written as $(x^6 + 2x^5 + x^4 + 2x^3 + 1)r(x) + (x^4 + 2x^3 + 2x)s(x)$ for some polynomials $r(x)$, $s(x) \in F[x]$, and suitable $r(x)$ and $s(x)$ can be found by utilizing Eq.(12) and Eq.(11):

$$\begin{aligned}
2x^2{+}2x + 1 &= (x^4 + 2x^3 + 2x) - (x + 2)(x^3 + x + 1)\\
&= (x^4 + 2x^3 + 2x) - (x + 2)((x^6 + 2x^5 + x^4 + 2x^3 + 1) - (x^2 + 1)(x^4 + 2x^3 + 2x))\\
&= (x^3 + 2x^2 + x)(x^4 + 2x^3 + 2x) - (x + 2)(x^6 + 2x^5 + x^4 + 2x^3 + 1),
\end{aligned}$$

whence $x^2 + x + 1 = (x + 2)(x^6 + 2x^5 + x^4 + 2x^3 + 1) + (2x^3 + x^2 + 2x)(x^4 + 2x^3 + 2x)$.

Exercise 27. Let R be a principal ideal domain and a, $b \in R$ nonzero elements. An element $m \in R$ is called a *least common multiple* of a and b if

(i) $a \mid m$ and $b \mid m$, and

(ii) for all $c \in R$, if $a \mid c$ and $b \mid c$ then $m \mid c$.

Prove that any two lcm's of a and b have to be associates of each other. Prove also that if $d = \gcd(a, b)$ then ab/d is a lcm of a and b. (Note: whenever x, y are elements of an integral domain such that $x \mid y$ there is an element z such that $y = xz$, and, provided x is nonzero, z is unique. Under these circumstances, we define y/x to be z.)

Exercise 28. Let a, b be nonzero elements of the principal ideal domain R, and let $d = \gcd(a, b)$. Show that if r, $s \in R$ satisfy $d = ra + sb$, then then most general solution of $d = xa + yb$, with x, $y \in R$, is given by

$$\begin{aligned} x &= r + (b/d)t \\ y &= s - (a/d)t \end{aligned}$$

where $t \in R$ is arbitrary. (You must show that this is a solution for all choices of t, and that every solution has this form for some value of t.)

In the following sequence of exercises we investigate properties of the *ring of Gaussian integers*, which is the subset $\mathbb{G}$ of the complex field $\mathbb{C}$ defined by

$$\mathbb{G} = \{\, n + mi \mid n, m \in \mathbb{Z} \,\}$$

(where $i = \sqrt{-1}$).

Exercise 29. Prove that $\mathbb{G}$ is a subring of $\mathbb{C}$. Show also that $\mathbb{G}$ is an integral domain (so that $\mathbb{G}$ is, in fact, a subdomain of $\mathbb{C}$).

Exercise 30. Let $a \in \mathbb{G}$ be nonzero. Using the Argand diagram identification of complex numbers with points in the Euclidean plane, check that 0, a, ia and $(1 + i)a$ are the vertices of a square, the diagonal of which has length $\sqrt{2}|a|$. Check furthermore that the square lattice generated by a and ia—that is, the points of the plane that can obtained by adding an integer multiple of a and an integer multiple of ia—consists precisely of the multiples of a in $\mathbb{G}$ (in the sense of Definition (9.7)). Conclude that for every $b \in \mathbb{G}$ there is a multiple of a whose distance from b is at most $|a|/\sqrt{2}$.

Exercise 31. For each $a \in \mathbb{G}$ define the *norm* of a to be $N(a) = |a|^2$. That is, if $a = n + mi$ then $N(a) = n^2 + m^2$. Use Exercise 30 to show that $\mathbb{G}$ is a Euclidean domain. (The function deg appearing in the definition of a Euclidean domain can be defined by $\deg a = N(a)$ for all $a \in \mathbb{G}$.)

Note that it is a consequence of Exercise 31 that $\mathbb{G}$ is a principal ideal domain.

Exercise 32. Show that if a, $b \in \mathbb{G}$ then $N(ab) = N(a)N(b)$. Deduce that if a is a unit of $\mathbb{G}$ then $N(a) = 1$. Show that the only elements $a \in \mathbb{G}$ with $N(a) = 1$ are ± 1 and $\pm i$, and deduce that these are the only units of $\mathbb{G}$.

Observe that $\mathbb{Z}$ is a subdomain of $\mathbb{G}$, and note that it is possible for an integer to be irreducible in $\mathbb{Z}$ yet reducible in $\mathbb{G}$. (For example, 17 has no nontrivial factoization in $\mathbb{Z}$, but in $\mathbb{G}$ we have $17 = (4 + i)(4 - i)$. In what follows, when we talk of a "prime integer" we mean it in the usual sense: an element of $\mathbb{Z}$ with no nontrivial factorization in $\mathbb{Z}$. Our aim (which we will not be able to achieve until after the development of some more theory) is to identify the irreducible elements of $\mathbb{G}$; in particular, we wish to find out whether prime integers remain irreducible when considered as elements of $\mathbb{G}$. Note that (fortunately) if a, $b \in \mathbb{Z}$ then "$a \mid b$" has the same meaning whether interpreted as referring to divisibility in $\mathbb{Z}$ or divisibility in $\mathbb{G}$. This is because the rational number a/b is in $\mathbb{G}$ if and only if it is in $\mathbb{Z}$.

Exercise 33. Use Exercise 32 to show that if $a \in \mathbb{G}$ is such that $N(a)$ is a prime integer, then a is an irreducible element of G.

Exercise 34. Let $a = n + mi \in \mathbb{G}$ be irreducible. Show that there exists a prime integer p such that $p \mid N(a)$ and $N(a) \mid p^2$, and deduce that either $N(a) = p$ or $N(a) = p^2$. (Hint: Let $n^2 + m^2 = p_1 p_2 \cdots p_k$ be the factorization of $N(a) \in \mathbb{Z}$ as a product of prime integers. Show that $a \mid N(a)$, and deduce that $a \mid p_j$ for some j. Use Exercise 32 to deduce that $N(a) \mid N(p_j) = p_j^2$.)

Exercise 35. Show that if $a \in G$ is irreducible and $N(a) = p^2$ for some prime integer p, then a is an associate of p in $\mathbb{G}$ (so that p is irreducible as an element of $\mathbb{G}$). Deduce that p cannot be expressed in the form $n^2 + m^2$ with n, $m \in \mathbb{Z}$ (Hint: For the first part, observe that $a \mid N(a) = p^2$;

so $a \mid p$. Writing $p = ab$ we see that $N(b) = 1$, and hence b is a unit. For the second part, note that $p = n^2 + m^2$ would lead to the factorization $p = (n + mi)(n - mi)$.)

Exercises 34 and 35 have shown that if $a \in G$ is irreducible then either $a = n + mi$ where $n, m \in \mathbb{Z}$ are such that $n^2 + m^2$ is a prime integer p, or else $a = \pm p$ or $\pm ip$ for some prime integer p. Conversely, if p is a prime integer then we may choose an irreducible element $a \in \mathbb{G}$ such that $a \mid p$, and then since $N(a) \mid p^2$ we find that either $N(a) = p$, so that p is the sum of two squares, or else $N(a) = p^2$, so that p is irreducible in $\mathbb{G}$ and not the sum of two squares.

To complete our study of irreducible elements of $\mathbb{G}$ we need to determine which prime integers can be expressed as the sum of two squares. We defer this for a while.

10. Constructing fields as quotient rings

We remarked above that the ring of integers modulo p is a field whenever $p \in \mathbb{Z}$ is prime. More generally, if R is any principal ideal domain and $a \in R$ an irreducible element, then the quotient ring R/aR is a field. We are primarily interested in the case $R = \mathbb{Z}$ and the case $R = F[x]$, where F is a field. But once again we find that there is not much extra work involved in proving the theorems in significantly greater generality, and, accordingly, we shall do so. Readers should keep those two special cases uppermost in their minds, and not worry too much about the extra generality.

Definition (10.1): Let R be a ring and I an ideal in R. We say that I is a *prime ideal* if the following condition holds: whenever $a, b \in R$ satisfy $ab \in I$, either $a \in I$ or $b \in I$.

For example (and this is the motivation for the definition), if R is a principal ideal domain and $p \in R$ nonzero, then the ideal pR is prime if and only if the element p is prime.

Exercise 36. Show that if R is a commutative ring then the zero ideal is prime if and only if R has no zero divisors.

Definition (10.2): Let R be an arbitrary ring. An ideal I of R is said to be *maximal* if $I \neq R$ and for all ideals J of R satisfying $I \subseteq J$, either $J = I$ or $J = R$.

Exercise 37. Let R be a principal ideal domain and $a \in R$. Use Exercises 16 and 20 to show that aR is maximal if and only if a is irreducible. Conclude that in a principal ideal domain a nonzero ideal is prime if and only if it is maximal.

Although we had quite a lot to say about the First Isomorphism Theorem, we have not as yet mentioned the results known as the Second and Third Isomorphism Theorems. These are in fact reasonably easy consequences of the First Isomorphism Theorem, and we will leave their proofs as exercises for the reader. The Second Isomorphism Theorem asserts that if S is a subring of R and I an ideal in R then the set $S + I$ is a subring of R, and I is an ideal of this subring; furthermore, $S \cap I$ is an ideal of S, and $S + I/I$ is isomorphic to $S/S \cap I$. (It is proved by applying the First Isomorphism Theorem to the homomorphism $S \to R/I$ given by $s \mapsto s + I$.) The Third Isomorphism Theorem says that if I is an ideal of R then $S \mapsto S/I$ provides a one to one correspondence between subrings of R containing I and subrings of R/I; furthermore, if S is an ideal of R containing I then S/I is an ideal of R/I, and $(R/I)/(S/I) \cong R/S$. (This last part is proved by applying the First Isomorphism Theorem to a homomorphism $R/I \to R/S$ satisfying $x + I \mapsto x + S$ for all $x \in R$.)

The following result is in fact a special instance of the Third Isomorphism Theorem, but since it is directly relevant to our purposes we include a proof.

Proposition (10.3): Let I be a maximal ideal of the ring R. Then the quotient ring R/I has no ideals other than the zero ideal and the ring R/I itself.

Proof. Let $\mathcal{J} \subseteq R/I$ be a nonzero ideal of R/I. Remembering that elements of R/I are cosets (of the form $r + I$, where $r \in R$), it follows that $\mathcal{J}$ is a set of cosets. Define $J = \{\, r \in R \mid r + I \in \mathcal{J} \,\}$. Observe that $I \subseteq J$, for if $x \in I$ then $x + I = 0 + I$ is the zero element of R/I, which has to be in $\mathcal{J}$ since ideals always contain the zero element, and so it follows that $x \in J$. Furthermore, if J were equal to I then all elements of $\mathcal{J}$ would have the form $x + I$ for $x \in I$; that is, $\mathcal{J}$ would have no elements other than zero, contrary to our assumption.

We prove that J is an ideal of R. Firstly, it is nonempty since $I \subseteq J$ and I is nonempty. Now let $x, y \in J$ and $r \in R$ be arbitrary. Then $x + I$, $y + I \in calJ$ and $r + I \in R/I$, and since $\mathcal{J}$ is an ideal it follows from the closure properties listed in Theorem (7.3) that $(x + I) + (y + I)$, $-(x + I)$, $(r + I)(x + I)$ and $(x + I)(r + I)$ must all be elements of $\mathcal{J}$. That is, by the definitions of addition and multiplication in R/I, the elements $(x + y) + I, (-x) + I, rx + I$ and $xr + I$ are all in $\mathcal{J}$. So $x + y$, $-x$, rx and xr are all in J. This establishes that J has all the required closure properties, and so J is an ideal of R (by Theorem (7.3)).

So J is an ideal of R containing I and not equal to I. Since I was assumed to be maximal, it follows that $J = R$, and hence $\mathcal{J} = R/I$. Since our original assumption was merely that $\mathcal{J}$ is a nonzero ideal in R/I, the conclusion must be that R/I has no ideals other than R/I itself and zero, as required. □

Suppose in particular that R is a commutative ring and I a maximal ideal of R. Then Proposition (10.3) shows combined with Exercise 15 shows that R/I is a commutative ring with no ideals other than itself and zero. Exercise 12 then shows that R/I is a field.

Combining some of the things we have noted above yields the following theorem, which plays an important role in this course.

Theorem (10.4): Let R be a principal ideal domain and $p \in R$ an irreducible element. Then R/pR is a field.

Proof. Since p is irreducible it is prime (Theorem (9.18)), and so the ideal pR is prime (exercise!) and therefore maximal (Exercise 37). So R/pR is a field (by the discussion above). □

In particular, Theorem (10.4) applies to the ring $\mathbb{Z}$, and tells us that $\mathbb{Z}_p$ is a field if p is prime. We have already seen in Exercise 14 that if n is not prime then $\mathbb{Z}_n$ has zero divisors—for example, $2 \times 3 = 0$ in $\mathbb{Z}_6$—and so is not even an integral domain, let alone a field.

Because our proof of Theorem (10.4) made rather heavy use of previous results, we shall give another, more direct, proof. Consider $\mathbb{Z}_n$ first. It is certainly a commutative ring (since $\mathbb{Z}$ is commutative) and has a 1 (since $\mathbb{Z}$ has a 1). To show that $\mathbb{Z}_n$ is a field if n is prime, it remains to show that, in this case, every nonzero element has an inverse. Every element of $\mathbb{Z}_n$ has the form $\overline{r}$ for some $r \in \mathbb{Z}$, and $\overline{r} = \overline{0}$ (the zero of $\mathbb{Z}_n$) if and only if $n \mid r$. So assume that $n \nmid r$, and let $d = \gcd(n, r)$. Then $d \mid n$, and since n is prime either $d = n$ or $d = 1$. But $d \mid r$ and $n \nmid r$; so $d \neq n$. Hence $d = 1$. By one of the basic properties of gcd's (Theorem (9.17) (iii)), there exist integers s and t such that $ns + rt = \gcd(n, r) = 1$. But this gives that $rt \equiv 1 \pmod{n}$, and so $\overline{r}\,\overline{t} = \overline{1}$. That is, $\overline{t}$ is an inverse for $\overline{r}$.

The proof we just gave for $\mathbb{Z}$ applies unchanged for any principal ideal domain R. If $n \in R$ is irreducible then R/nR is certainly a commutative ring with 1 (since R is), and so to show that it is a field it suffices to show that all nonzero elements of R/nR have inverses. Every such element has the form $r + nR$ for some $r \in R$ such that $n \nmid r$. The gcd of n and r cannot be an associate of n since $n \nmid r$; so since this gcd must be a divisor of n, which is irreducible, it follows that $\gcd(n, r) = 1$ (or any unit). Hence there exist $s, t \in R$ with $ns + rt = 1$, and this gives $(r + nR)(t + nR) = rt + nR = 1 + nR$ (since $rt \equiv 1 \pmod{nR}$). So $t + nR$ is an inverse for $r + nR$.

Theorem (10.4) is also applicable when $R = F[x]$ for some field F, and it is this application which is of most theoretical importance in this course. The theorem tells us that if $p(x) \in F[x]$ is irreducible, then the the quotient ring $F[x]/p(x)F[x]$ is a field. In order to be able to apply this result we need to have some understanding of irreducible polynomials, and so a large part of this section will be devoted to an investigation of these. First of all, let us look at an example in which the field F is the two-element field $\mathbb{Z}_2$.

Recall that the elements of $\mathbb{Z}_2$ are the modulo 2 congruence classes of $\mathbb{Z}$. There are just two of these: the set of all even integers and the set of all odd integers. For brevity we denote these two sets by 0 and 1. Addition and multiplication of these elements is defined by the following rules:

$$\begin{array}{ll} 0+0=0, & 0\times 0=0, \\ 0+1=1, & 0\times 1=1, \\ 1+0=1, & 1\times 0=1, \\ 1+1=0, & 1\times 1=0. \end{array} \tag{13}$$

(These become natural if you interpret "$1 + 1 = 0$" as "odd + odd = even", and so on.) Note that every field has to have at least a 0 and a 1. The field $\mathbb{Z}_2$ has no other elements.

We are going to be dealing with the polynomial ring $\mathbb{Z}_2[x]$. Let us list all of the elements of $\mathbb{Z}_2[x]$ that have degree at most 3.

$$\begin{array}{lrl} \text{degree} & -\infty: & 0 \\ \text{degree} & 0: & 1 \\ \text{degree} & 1: & x,\ x+1 \\ \text{degree} & 2: & x^2,\ x^2+1,\ x^2+x,\ x^2+x+1 \\ \text{degree} & 3: & x^3,\ x^3+1,\ x^3+x,\ x^3+x+1, \\ & & x^3+x^2,\ x^3+x^2+1,\ x^3+x^2+x,\ x^3+x^2+x+1. \end{array}$$

(Remember that the coefficients can only be 0 or 1. It is easily seen that there are 16 polynomials of degree 4 in $\mathbb{Z}_2[x]$, 32 of degree 5, and so on.)

Remember that the degree of the product of two polynomials is the sum of the degrees of the factors (Proposition (6.2)). Consequently if a polynomial $a(x)$ can be factorized, the degrees of the factors must be less than or equal to the degree of $a(x)$. Since polynomials of degree zero are units (and in $\mathbb{Z}_2[x]$ there is only one of these, namely 1), a factorization which is not of the form (unit) × (associate of $a(x)$) must have both factors of degree less than $\deg a(x)$. This is clearly impossible idf $\deg a(x) = 1$, since 1 cannot be expressed as a sum of quantities that are nonpositive. So polynomials of degree 1 are always irreducible (given that the coefficient ring F is a field). A polynomial of degree 2 is irreducible if it cannot be expressed as a product of two factors of degree 1. For $\mathbb{Z}_2[x]$ the possible products of two factors of degree 1 are $xx = x^2$, $x(x+1) = x^2 + x$ and $(x+1)^2 = x^2+2x+1 = x^2+1$ (since $2 = 0$ in $\mathbb{Z}_2$). The one remaining polynomial of degree 2, $x^2 + x + 1$, must therefore be irreducible. The polynomials of degree 3 that factorize nontrivially must either have three irreducible factors of degree 1, or one irreducible factor of degree 1 and one of degree 2. The possibilities are $xxx = x^3$, $xx(x+1) = x^3 + x^2$, $x(x+1)^2 = x^3 + x$, $(x+1)^3 = x^3 + x^2 + x + 1$, $x(x^2 + x + 1) = x^3 + x^2 + x$ and $(x+1)(x^2+x+1) = x^3 + 1$. The remaining two, $x^3 + x + 1$ and $x^3 + x^2 + 1$, must be irreducible.

It follows from the above that if $R = \mathbb{Z}_2[x]$ and K is the principal ideal $(x^3 + x + 1)\mathbb{Z}_2[x]$ of R, then R/K is a field. The elements of this field are the cosets of K in R. Every coset has the form $f(x) + K$ for some polynomial $f(x) \in \mathbb{Z}_2[x]$, and $f(x) + K = g(x) + K$ if and only if $f(x) \equiv g(x) \pmod{K}$. By the division property for polynomials over a field, for each $f(x) \in \mathbb{Z}_2[x]$

we can find $q(x)$, $r(x) \in \mathbb{Z}_2[x]$ such that $f(x) = (x^3 + x + 1)q(x) + r(x)$, and $\deg r(x) < 3$. Note that $f(x) - r(x) \in K$ (since $(x^3 + x + 1)q(x) \in (x^3 + x + 1)\mathbb{Z}_2[x]$). So every coset contains a representative of degree less than 3. Furthermore, if $r_1(x) + K = r_2(x) + K$ with $r_1(x)$ and $r_2(x)$ both of degree less than 3, then $r_1(x) - r_2(x)$ is a polynomial of degree less than 3 which is divisible by $x^3 + x + 1$. This forces $r_1(x) - r_2(x) = 0$ (since $\deg(x^3 + x + 1)a(x) = 3 + \deg a(x) \geq 3$ unless $a(x) = 0$). So in fact each coset of K in R contains a unique representative of degree less than 3.

We have seen that there are exactly 8 elements of degree less than 3 in R; so there are precisely 8 cosets:

$$\begin{array}{llll} K, & 1 + K, & x + K, & x + 1 + K, \\ x^2 + K, & x^2 + 1 + K, & x^2 + x + K, & x^2 + x + 1 + K. \end{array}$$

Every element of R must lie in one of these cosets. For example, we find that $x^4 \in x^2 + x + K$, since $x^4 - (x^2 + x) = (x^3 + x + 1)x \in K$. (Note that $-$ equals $+$ in rings of characteristic 2, since $2 = 0$.) So R/K is a field that has exactly 8 elements.

The zero and identity elements of R/K are $0 + K = K$ and $1 + K$ respectively. Let us follow our usual notational conventions for zero and identity elements, and denote $0 + K$ and $1 + K$ simply by 0 and 1. This will not cause any inconsistencies, since the formulas in Eq.(13) above remain true in R/K. Let us also denote $x + K$ by α. We can easily express all 8 elements in terms of α:

$$\begin{array}{llll} 0, & 1, & \alpha, & \alpha + 1, \\ \alpha^2, & \alpha^2 + 1, & \alpha^2 + \alpha, & \alpha^2 + \alpha + 1. \end{array}$$

Furthermore, $\alpha^3 + \alpha + 1 = (x^3 + x + 1) + K = 0 + K$ since $x^3 + x + 1 \in K$, and so the element α of R/I satisfies $\alpha^3 + \alpha + 1 = 0$. So we can think of α as a root of the polynomial $x^3 + x + 1$. In some sense, the field R/I is constructed from the field $\mathbb{Z}_2$ by "adjoining" to $\mathbb{Z}_2 = \{0, 1\}$ the new element α, which is assumed to be a root of $x^3 + x + 1$. Of course adjoining α forces one to admit further things, like α^2 and $\alpha + 1$, but the relation $\alpha^3 + \alpha + 1 = 0$ limits the number of different elements obtained to the 8 listed above.

The construction we have been through is totally analogous to the construction which creates the complex field from the real field. The polynomial $x^2 + 1 \in \mathbb{R}[x]$ is irreducible. We adjoin to $\mathbb{R}$ a new element i, which is assumed to be a root of $x^2 + 1$, and this automatically means that we must also admit elements of the form $x + yi$ for all $x, y \in \mathbb{R}$. Since $i^2 = -1$ no further elements are required, and we have obtained a new ring, $\mathbb{C}$, which contains the original ring $\mathbb{R}$ as a subring, as well as containing a root of the polynomial $x^2 + 1$.

Before continuing with the general theory, let us look at another example over $\mathbb{Z}_2$, doing things a little less formally this time. Let us adjoin to $\mathbb{Z}_2$ a new element β, which we want to be a root of the polynomial $x^3 + x^2 + 1$. What extra elements must we have? Besides 0, 1 and β, we will need $\beta + 1$, β^2, $\beta^2 + 1$, $\beta^2 + \beta$ and $\beta^2 + \beta + 1$. Using the fact that $1 + 1 = 0$ here, it is easily checked that this set of 8 elements is closed under addition. It is also closed under multiplication, since the equation $\beta^3 + \beta^2 + 1 = 0$ permits β^3 and higher powers of β to be expressed in terms of 1, β and β^2. (For example, $\beta^3 = \beta^2 + 1$, and hence $\beta^4 = \beta^3 + \beta = \beta^2 + \beta + 1$.) It is not totally clear from this informal approach that the 8 element system so constructed is actually a ring. The formal quotient ring approach that we used before provides the best method of proving that you do get a ring.

Although in all the examples above we started with an irreducible polynomial, we did not really have make this assumption. Indeed, if one starts with a polynomial which is not irreducible and goes through the same construction, then everything works in just the same manner, and a new ring is constructed which has the original field as a subring. This new ring will not be a field, however. Irreducibility of the original polynomial is required if one wants to construct a field, but not required otherwise.

Theorem (10.5): *Let F be a field and $p(x) \in F[x]$ an element of degree $d \geq 1$. Let $K = p(x)F[x]$, the principal ideal generated by $p(x)$, and let $E = F[x]/K$. Then*

(i) every element of E is uniquely expressible in form $(a_0 + a_1x + \cdots + a_{d-1}x^{d-1}) + K$ for some elements $a_0, a_1, \ldots, a_{d-1} \in F$,

(ii) the mapping $\eta\colon a \mapsto a + K$ is an embedding of F into E, and

(iii) if we write $\alpha = x + K$, and use the embedding η to identify F with a subring of E, then each element of E is uniquely expressible in the form $a_0 + a_1\alpha + \cdots + a_{d-1}\alpha^{d-1}$ with the $a_i \in F$, and α is a root in E of the polynomial $p(x) \in F[x]$.

Proof. By the uniqueness part of Lemma (9.2), for each $f(x) \in F[x]$ there is a unique $r(x) \in F[x]$ such that $p(x) \bigm| (f(x) - r(x))$ and $\deg r(x) < \deg p(x)$. Hence $f(x) + K$ is uniquely expressible in the form $r(x) + K$ with $\deg r(x) < d$, which proves Part (i).

By the definitions of addition and multiplication in $F[x]/K$, we have, for all r, $s \in F$,

$$(\eta r)(\eta s) = (r + K)(s + K) = rs + K = \eta(rs)$$
$$\eta r + \eta s = (r + K) + (s + K) = (r + s) + K = \eta(r + s),$$

showing that η is a homomorphism. So to check that η is an embedding it will suffice to show that $\ker\eta = \{0\}$ (by Proposition (5.14)). Recall that the zero element of $F[x]/K$ is the coset containing 0, namely K itself. So if $r \in \ker\eta$ then $r + K = K$, and so $r \in K$. Note that here r, which commenced life as an element of F, is being regarded as a scalar polynomial. To say that $r \in K = p(x)F[x]$ is to say that $p(x) \bigm| r$ in $F[x]$. But since the degree of $p(x)$ is at least 1, while $\deg r \leq 0$, this forces $r = 0$. So 0 is the only element of $\ker\eta$, and hence η is an embedding. So Part (ii) is proved.

Using the embedding η to identify F with a subring of E amounts to writing r for the coset $r + K$ whenever $r \in F$. Of course, when one does this one needs to be away that one is doing it, and that r as an element of $F[x]/K$ is not exactly the same thing as r as an element of F or as an element of $F[x]$ (although exceedingly similar). Anyway, with this notation, and writing $x + K = \alpha$, we see that

$$\begin{aligned} a_0 + a_1\alpha + \cdots + a_{d-1}\alpha^{d-1} &= (a_0 + K) + (a_1 + K)(x + K) + \cdots + (a_{d-1} + K)(x + K)^{d-1} \\ &= (a_0 + a_1x + \cdots + a_{d-1}x^{d-1}) + K, \end{aligned}$$

and by Part (i) each element of $F[x]/K$ is uniquely expressible in this form. Furthermore, if $p(x) = p_0 + p_1x + \cdots + p_dx^d$ then the same considerations show that

$$\begin{aligned} p(\alpha) &= (p_0 + K) + (p_1 + K)(x + K) + \cdots + (p_d + K)(x + K)^d \\ &= (p_0 + p_1x + \cdots + p_dx^d) + K \\ &= p(x) + K \\ &= 0 \end{aligned}$$

since $p(x) \in K$ (and therefore $p(x) + K = K$, which is the zero of $F[x]/K$). □

Restating Theorem (10.5) a little gives the following corollary.

Corollary (10.6): *Let F be a field and $p(x) \in F[x]$ with $\deg p(x) \geq 1$. Then there exists a ring E which contains F as a subring and contains an element α such that $p(\alpha) = 0$. Moreover, every element of E is uniquely expressible in the form $f(\alpha)$ with $f(x) \in F[x]$ of degree less than d. If $p(x)$ is irreducible in $F[x]$ then E is a field.*

Note that when investigating questions of irreducibility or factorization of polynomials, it is necessary to specify the particular polynomial ring in which the factors are to be sought. For example, over the real field $\mathbb{R}$ the polynomial $x^2 + 1$ is irreducible, but over the complex field $\mathbb{C}$ it factorizes as $(x + i)(x - i)$. Likewise $x^2 - 2$ is irreducible as an element of $\mathbb{Q}[x]$, but as an element of $\mathbb{R}[x]$ it factorizes as $(x - \sqrt{2})(x + \sqrt{2})$. For a third example, let F be the field with 8 elements that we discussed previously, formed by adjoining to $\mathbb{Z}_2$ an element α satisfying $\alpha^3 + \alpha + 1 = 0$. As an element of $\mathbb{Z}_2[x]$ the polynomial $x^3 + x + 1$ is irreducible, but it has a root in F, and hence has a nontrivial factorization in $F[x]$. Indeed, it turns out that $x^3+x+1 = (x+\alpha)(x+\alpha^2)(x+\alpha^2+\alpha)$. The reader should check this by expanding the right hand side and using $\alpha^3 + \alpha + 1 = 0$ to simplify the expression. (And do not forget that $1 + 1 = 0$, since we are working over a field of characteristic 2.)

The Euclidean algorithm provides a means of calculating inverses of elements of quotient rings $F[x]/p(x)F[x]$, as illustrated in the following example. Let $p(x) = x^3 + x^2 + 1 \in \mathbb{Z}_2[x]$, and for brevity write K for the ideal generated by $p(x)$. Consider the element $(x^2+x+1)+K \in \mathbb{Z}_2[x]/K$. Its inverse will have the form $f(x) + K$, where $f(x) \in \mathbb{Z}_2[x]$ is such that $(x^2 + x + 1)f(x) + K = 1 + K$. So we need to find an element $f(x)$ such that $(x^2 + x + 1)f(x) - 1 \in K = p(x)F[x]$; that is, $(x^2 + x + 1)f(x) - 1 = (x^3 + x^2 + 1)g(x)$ for some $g(x) \in \mathbb{Z}_2[x]$. Applying the Euclidean algorithm to $x^3 + x^2 + 1$ and $x^2 + x + 1$ gives the following steps:

$$\begin{aligned} x^3 + x^2 + 1 &= (x^2 + x + 1)x + x + 1 && (14)\\ x^2 + x + 1 &= (x + 1)x + 1 && (15)\\ x + 1 &= 1(x + 1) + 0. \end{aligned}$$

The last nonzero remainder is 1, which shows that $\gcd(x^3 + x^2 + 1, x^2 + 1) = 1$. Of course we knew that the gcd would be 1, since $x^3 + x^2 + 1$ is irreducible, and hence has no divisors of degree 1 or 2. What is more relevant here is that combining Eq's (14) and (15) gives

$$\begin{aligned} 1 &= (x^2 + x + 1) - x(x + 1)\\ &= (x^2 + x + 1) - x((x^3 + x^2 + 1) - x(x^2 + x + 1))\\ &= (x^2 + 1)(x^2 + x + 1) - x(x^3 + x^2 + 1), \end{aligned}$$

so that $f(x) = x^2 + 1$ and $g(x) = x$ is a solution to $(x^2 + x + 1)f(x) - 1 = (x^3 + x^2 + 1)g(x)$. So $x^2 + 1 + K$ is the inverse of $x^2 + x + 1 + K$ in $\mathbb{Z}_2[x]/K$.

Another way to see that $F[x]/p(x)F[x]$ is a field if $p(x)$ is irreducible—analogous to an argument we gave for $\mathbb{Z}_p$—is as follows. It is clear that $F[x]/p(x)F[x]$ is a commutative ring with 1, since $F[x]$ is. But if $a(x) + p(x)F[x]$ is a nonzero element of $F[x]/p(x)F[x]$ then $p(x) \nmid a(x)$, and since $p(x)$ is irreducible this forces $\gcd(p(x), a(x))$ to be 1. Then an inverse can be found for $a(x)$ as in the above example.

For another example, consider the ring $\mathbb{Q}[x]$ (polynomials over the rational field $\mathbb{Q}$), and let K be the ideal generated by $x^3 - 2$. According to Theorem (10.5), every element of $\mathbb{Q}[x]/K$ is uniquely expressible in the form $(a + bx + cx^2) + K$ with a, b, $c \in \mathbb{Q}$. Furthermore, using the viewpoint of Corollary (10.6), the ring $\mathbb{Q}[x]/K$ can be thought of as the result of adjoining to $\mathbb{Q}$ an element α which is a root of the polynomial $x^3 - 2$.

Note that we already knew of a field containing $\mathbb{Q}$ as a subfield and containing also a root of $x^3 - 2$: the real field $\mathbb{R}$ has these properties. Moreover, there is an evaluation homomorphism $\theta = \mathrm{eval}_{\sqrt[3]{2}}\colon \mathbb{Q}[x] \to \mathbb{R}$ given by $\theta(f(x)) = f(\sqrt[3]{2})$ for all $f(x) \in \mathbb{Q}[x]$. Now for an arbitrary $f(x) \in \mathbb{Q}[x]$, dividing by $x^3 - 2$ gives

$$f(x) = (x^3 - 2)q(x) + a_0 + a_1x + a_2x^2$$

for some $q(x) \in \mathbb{Q}[x]$ and $a_0, a_1, a_2 \in \mathbb{Q}$. Thus

$$\theta(f(x)) = f(\sqrt[3]{2}) = a_0 + a_1\sqrt[3]{2} + a_2(\sqrt[3]{2})^2,$$

and it follows that the image of θ is the set of all real numbers of the form $a_0 + a_1\sqrt[3]{2} + a_2(\sqrt[3]{2})^2$ with $a_0, a_1, a_2 \in \mathbb{Q}$. We define

$$\mathbb{Q}[\sqrt[3]{2}] = \operatorname{im}\theta = \{\, a_0 + a_1\sqrt[3]{2} + a_2(\sqrt[3]{2})^2 \mid a_0, a_1, a_2 \in \mathbb{Q}\,\}.$$

Let us show that $x^3 - 2$ is irreducible in $\mathbb{Q}[x]$. Suppose, for a contradiction, that it is not. Then it must have a factorization in $\mathbb{Q}[x]$ as a product of a polynomial of degree 1 and a polynomial of degree 2:

$$x^3 - 2 = (ax - b)(cx^2 + dx + e) \tag{16}$$

for some $a, b, c, d, e \in \mathbb{Q}$. By first year calculus we know that $x^3 - 2$ has one real root and two non-real complex roots (the real root being $\sqrt[3]{2}$), yet Eq.(16) shows that the rational number b/a is a root of $x^3 - 2$. So $\sqrt[3]{2} = b/a$ is rational, contradicting Exercise 26.

Using the irreducibility in $\mathbb{Q}[x]$ of $x^3 - 2$ we can show that the kernel of θ is $K = (x^3 - 2)\mathbb{Q}[x]$. The kernel of θ is certainly an ideal in $\mathbb{Q}[x]$, and so it must have the form $p(x)\mathbb{Q}[x]$ for some $p(x) \in \mathbb{Q}[x]$ (since $\mathbb{Q}[x]$ is a principal ideal domain). But $x^3 - 2$ is certainly in the kernel of θ (since evaluating $x^3 - 2$ at $x = \sqrt[3]{2}$ gives 0). So $p(x) \mid x^3 - 2$, and since $p(x)$ cannot be a unit (since $p(\sqrt[3]{2}) = 0$) the irreducibility of $x^3 - 2$ forces $p(x)$ to be an associate of $x^3 - 2$, so that $(x^3 - 2)\mathbb{Q}[x] = p(x)\mathbb{Q}[x] = \ker\theta$, as claimed.

By the First Isomorphism Theorem we deduce that there is an isomorphism $\mathbb{Q}[x]/K \to \mathbb{Q}[\sqrt[3]{2}]$ satisfying $f(x) + K \mapsto f(\sqrt[3]{2})$ for all $f(x) \in \mathbb{Q}[x]$, and in particular

$$(a_0 + a_1x + a_2x^2) + K \mapsto a_0 + a_1\sqrt[3]{2} + a_2(\sqrt[3]{2})^2$$

for all $a_0, a_1, a_2 \in \mathbb{Q}$. This is very much in agreement with our previous analysis of $\mathbb{Q}[x]/K$, according to which $(a_0 + a_1x + a_2x^2) + K$ could be thought of as $a_0 + a_1\alpha + a_2\alpha_2$ where α is a root of $x^3 - 2$.

Note also that the irreducibility of $x^3 - 2$ over $\mathbb{Q}[x]$ guarantees that $\mathbb{Q}[x]/K$ is a field, and therefore that $\mathbb{Q}[\sqrt[3]{2}]$ is a field.

The ideas that arose in the above example can be employed in a wider context. To facilitate the discussion of this we need a few definitions.

Definition (10.7): A field E is called an *extension* of a field F if F is a subfield of E.

More generally we could call E an extension of F whenever we have an embedding of F into E, although we shall only do so if we intend to use the embedding to identify F with a subfield of E. Note that by Exercise 11 we know that a nonzero homomorphism from one field to another is necessarily an embedding.

Definition (10.8): Let E be an extension field of F and let $t \in E$. Define

$$F[t] = \{\, a_0 + a_1t + \cdots + a_dt^d \mid 0 \le d \in \mathbb{Z} \text{ and } a_0, a_1, \ldots, a_d \in F\,\}.$$

We call $F[t]$ the *subring of E generated by F and t*.

Observe that $F[t] = \{\, f(t) \mid f(x) \in F[x]\,\}$, and thus is the image of the evaluation homomorphism $\mathrm{eval}_t\colon F[x] \to E$.

Definition (10.9): Let E be an extension field of F and let $t \in E$. We say that t is *algebraic over F* if there is a nonzero polynomial $f(x) \in F[x]$ such that $f(t) = 0$. If t is not algebraic over F it is said to be *transcendental over F*.

It is clear that t is algebraic over F if and only if the kernel of the evaluation homomorphism $\mathrm{eval}_t \colon F[x] \to E$ is nonzero.

Definition (10.10): Let E be an extension field of F and let $t \in E$ be algebraic over F. The *minimal polynomial* of t over F is the polynomial $p(x) \in F[x]$ which is monic, satisfies $p(t) = 0$, and has minimal degree among all such polynomials.

Recall that to say that $p(x)$ is monic is to say that its leading coefficient is 1 (which of course implies that $p(x)$ is nonzero). Note that, by the Least Integer Principle, if t is algebraic over F there must exist a minimal degree nonzero polynomial $p(x)$ with $p(t) = 0$; moreover, such a polynomial will have a monic associate, which will then satisfy the requirements of Definition (10.10). Moreover, the minimal polynomial must be unique, since if $p_1(x)$ and $p_2(x)$ both satisfy the requirements then $f(x) = p_1(x) - p_2(x)$ has smaller degree (since the leading terms cancel) and also satisfies $f(t) = 0$, and this forces $f(x) = 0$ since $p_1(x)$ and $p_2(x)$ have minimal degree among the nonzero polynomials of which t is a root.

The next result is a crucial, yet almost trivial, property of minimal polynomials.

Theorem (10.11): *Let E, F be fields such that E is an extension of F, and let $t \in E$ be algebraic over F. Then the minimal polynomial of t over F is irreducible.*

Proof. Suppose to the contrary, and let $p(x)$ be the minimal polynomial. Then there is a nontrivial factorization $p(x) = r(x)s(x)$, where $r(x), s(x) \in F[x]$ have degree strictly less than $\deg p(x)$. Evaluating at t gives $0 = p(t) = r(t)s(t)$, and hence either $r(t) = 0$ or $s(t) = 0$, since a field has no zero divisors. But this contradicts the minimality of $p(x)$. □

The minimal polynomial of an element t can alternatively be characterized as the monic generator of the ideal $\ker \mathrm{eval}_t$.

Proposition (10.12): *Let E be an extension field of F and let $t \in E$ be algebraic over F. Then the set $I = \{\, f(x) \in F[x] \mid f(t) = 0 \,\}$ is an ideal of $F[x]$, and if $p(x)$ is the minimal polynomial of t over F then $I = p(x)F[x]$. Furthermore, every generator of I is an associate of $p(x)$.*

Proof. Since I is the kernel of the evaluation homomorphism eval_t, it is an ideal. However, we saw in Theorem (9.4) that the generators of a nonzero ideal in $F[x]$ are precisely the nonzero polynomials of minimal degree that it contains, and since $p(x)$ has minimal degree among the nonzero elements of I the result follows. □

We should also note the following simple results, whose proofs we leave as exercises.

Proposition (10.13): *Let R be a principal ideal domain and I an ideal in R. If I contains an irreducible element p then either $I = R$ or $I = pR$.*

Proposition (10.14): *Let E be an extension field of F and let $t \in E$ be algebraic over F. If $p(x) \in F[x]$ is irreducible and satisfies $p(t) = 0$ then $p(x)$ is an associate of the minimal polynomial of t (and hence if it is monic it is the minimal polynomial of t).*

Since in the course of the above discussion we have identified the kernel and the image of the evaluation homomorphism $\mathrm{eval}_t \colon F[x] \to E$, we would be negligent not to note what information the First Isomorphism Theorem yields.

Theorem (10.15): *Let E be an extension field of F and let $t \in E$.*

(i) *If t is transcendental over F then* $\mathrm{eval}_t \colon F[x] \to E$ *is an embedding which maps $F[x]$ isomorphically onto $F[t]$. In this case $F[t]$ is not a field.*

(ii) *If t is algebraic over F then $F[t]$ is a subfield of E isomorphic to the quotient ring $F[x]/I$, where $I = \{\, f(x) \in F[x] \mid f(t) = 0 \,\}$. Furthermore, there is an isomorphism $F[t] \to F[x]/I$ satisfying $f(t) \mapsto f(x) + I$ for all $f(x) \in F[x]$.*

Proof. The set I defined in Part (ii) is the kernel of eval_t, and hence it is an ideal. If t is transcendental then by definition $I = \{0\}$, and so eval_t is an embedding (by Proposition (5.14)). So in this case $f(x) \mapsto f(t)$ is an isomorphism $F[x] \to F[t]$, and since $F[x]$ is not a field—nonzero elements of $F[x]$ of degree greater than 1, such as x, do not have inverses in $F[x]$—it follows that $F[t]$ is not a field either.

Turning to the case of an algebraic element t, the kernel I of eval_t is generated by the minimal polynomial $p(x)$ of t, which is irreducible by Theorem (10.11). So by Theorem (10.4) $F[x]/I = F[x]/p(x)F[x]$ is a field. The First Isomorphism Theorem tells us that there is an isomorphism $F[x]/I \to F[t]$ satisfying $f(x) + I \mapsto f(t)$ for all $f(x) \in F[x]$ (and the inverse of this is given by $f(t) \mapsto f(x) + I$). Since $F[x]/I$ is a field it follows that $F[t]$ is a field. □

Exercise 38. Let E be an extension field of F and let $t \in E$. Define the *subfield of E generated by F and t* to be

$$F(t) = \{\, f(t)g(t)^{-1} \mid f(x),\, g(x) \in F[x] \text{ and } g(t) \neq 0 \,\}.$$

Show that $F(t)$ is a subfield of E, and show further that $F(t) = F[t]$ if t is algebraic over F.

Exercise 39. With E, F and $t$as in Exercise 38, assume that t is transcendental over F. Use Theorem (6.8) to show that $F(t)$ is isomorphic to the field of fractions of $F[x]$.

As a comment on the above exercise we mention that the field of fractions of $F[x]$ is called the *field of rational functions* over F in the indeterminate x, and denoted by $F(x)$. Its elements are objects of the form $f(x)/g(x)$ where $f(x)$, $g(x) \in F[x]$ and $g(x) \neq 0$. Note that, despite the name, elements of $F(x)$ are not functions. They are formal expressions

$$\frac{a_0 + a_1x + \cdots + a_nx^n}{b_0 + b_1x + \cdots + b_mx^m}$$

where the coefficients a_i, b_j are elements of F (and the b_j are not all zero), two such expressions $f_1(x)/g_1(x)$ and $f_2(x)/g_2(x)$ being identified if $f_1(x)g_2(x) = f_2(x)g_1(x)$ in $F[x]$. Addition and multiplication of these formal expressions is governed by the rules described in our discussion of fields of fractions. Nor is it even the case that a rational function $f(x)/g(x)$ determines a (genuine) function $F \to F$ by $t \mapsto f(t)/g(t)$ for all $t \in F$, unless it happens to be the case that $g(x)$ has no roots in F.

While on these matters we should also point out that polynomials themselves are also not functions. In fact we defined polynomials to be formal power series that have only finitely many nonzero coefficients. Effectively, this means that a polynomial is a formal expression

$$a_0 + a_1x + \cdots a_nx^n,$$

addition and multiplication being given by the rules we have previously stated. It is true that each polynomial $f(x) \in R[x]$ defines a *polynomial function* $R \to R$ by the rule $t \mapsto f(t)$ for all $t \in R$ (although the same is not true for general power series since infinite sums in R are not defined). Note, however, that it is quite possible for two different polynomial $f(x)$, $g(x) \in R[x]$ to give rise to the same polynomial function $R \to R$. The polynomials $f(x)$ and $g(x)$ are equal if and only if they have the same coefficients, but the polynomial functions they determine are equal if and only if $f(t) = g(t)$ for all $t \in R$, and this is a weaker condition.

Exercise 40. Give an example of two distinct polynomials over the field $\mathbb{Z}_2$ which determine the same polynomial function $\mathbb{Z}_2 \to \mathbb{Z}_2$.

Suppose that R is a ring with 1 and F a subring of R which contains the 1. Suppose also that F is a field. (These assumptions apply, in particular, whenever R is an extension field of F.) Then R may be regarded as a vector space over F: addition of "vectors" (elements of R) is given by the addition operation of R, and scalar multiplication is given by the multiplication operation of R (since both the "vectors"—elements of R—and the "scalars"—elements of F—can be regarded as elements of R). It is routine to check that the vector space axioms are satisfied, following easily from the ring axioms for R. It is a familiar fact, for example, that the complex field $\mathbb{C}$ can be regarded as a two-dimensional vector space over $\mathbb{R}$.

Let F be a field and $p(x) \in F[x]$. By Corollary (10.6) we know that the quotient ring $R = F[x]/p(x)F[x]$ can be regarded as the result of adjoining to F an element α which is a root of $p(x)$. Moreover, if $\deg p(x) = d$ then each element of R is uniquely expressible in the form $a_0 + a_1\alpha + a_2\alpha^2 + \cdots + a_{d-1}\alpha^{d-1}$, with $a_0, a_1, \ldots, a_{d-1}$ in F. In the terminology of vector space theory, regarding the elements of F as the scalars, this says that each element of R is uniquely expressible as a linear combination of the elements $1, \alpha, \alpha^2, \ldots, \alpha^{d-1}$. In other words, these elements form a basis of R considered as a vector space over F. In particular, R is finite dimensional as a vector space over F, and the dimension is d.

Proposition (10.16): *Let F be a field and $p(x) \in F[x]$ a nonzero polynomial. As well as being a ring, $F[x]/p(x)F[x]$ is a vector space over F of dimension $d = \deg p(x)$.*

Definition (10.17): Let F be a field and E an extension of F. The dimension of E considered as a vector space over F is called the *degree* of the extension, denoted by $[E : F]$.

If E is an extension of F and $t \in E$ is algebraic over F, then Theorem (10.15) tells us that $F[t]$ is a field isomorphic to $F[x]/p(x)F[x]$, where $p(x)$ is the minimal polynomial of t over F, and
Proposition (10.18): then tells us that the degree $[F[t] : F]$ equals $d = \deg px(x)$.

Proposition (10.19): *Let E be an extension of F and $t \in E$ algebraic over F. Then $F[t]$ is an extension field of F, and the degree $[F[t] : F]$ of this extension equals the degree of the minimal polynomial of t over F.*

Note that the degree of a field extension could well be infinite: it is possible that there is no finite set $t_1, t_2, \ldots, t_n$ of elements of E with the property that every element of E can be obtained as an F-linear combination of the t_i. This is the case, for example, for $\mathbb{R}$ considered as an extension of $\mathbb{Q}$. In the important special case that E does have a finite F-basis, so that the degree $[E : F]$ is finite, we say that E is a *finite extension* of F. It is an important fact that finite extensions are necessarily *algebraic extensions* (meaning that every element of the extension field E is algebraic over the subfield F).

Theorem (10.20): Let E be an extension field of F and suppose that $[E : F]$ is finite. Then every element of E is algebraic over F.

Proof. Let $d = [E : F] < \infty$. Since d is the dimension of E as a vector space over F it follows from a standard result in vector space theory that any sequence of $d + 1$ or more elements of E has to be linearly dependent over F. So for every $t \in E$ the elements $1, t, t^2, \ldots, t^d$ are linearly dependent over F. This means that there are $a_0, a_1, a_2, \ldots, a_d \in F$, which are not all zero, satisfying $a_0 + a_1t + a_2t^2 + \cdots + a_dt^d = 0$. This means that t is a root of the nonzero polynomial $a_0 + a_1x + a_2x^2 + \cdots + a_dx^d \in F[x]$. So by the definition, t is algebraic over F. □

A fact which is crucial for our theoretical applications is the following multiplicative property for degrees of extensions of extensions.

Theorem (10.21): *Let F, E and L be fields, with E an extension of F and L an extension of E. Then L is an extension of F, and $[L : F] = [L : E][E : F]$.*

Proof. Let $n = [E : F]$ and $m = [L : E]$. If n is finite we may choose $e_1, e_2, \ldots, e_n \in E$ which form a basis for E as a vector space over F. To say that n is infinite means that it is possible to choose an infinite sequence $e_1, e_2, e_3, \ldots$ of elements of E that are linearly independent over F. Similarly, we may choose $l_1, l_2, \ldots, l_m \in L$ which form a basis of L as a vector space over E, or an infinite sequence of linearly independent elements if m is infinite. We show first that the collection of all elements e_il_j (as i ranges from 1 to n and j ranges from 1 to m) is linearly independent.

Suppose that some F-linear combination of these elements is zero. That is, suppose that $\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_{ij}e_il_j = 0$ for some coefficients $\lambda_{ij} \in F$. (Note that in the case that either m or n is infinite, there are ostensibly infinitely many terms here, but since genuinely infinite sums are not allowed in this context it is assumed that only finitely many of the coefficients λ_{ij} are nonzero. We could avoid using these ostensibly infinite sums by instead considering a finite, though arbitrarily large, subset B_E of $\{e_1, e_2, \ldots\}$, and a finite, though arbitrarily large, subset B_L of $\{l_1, l_2, \ldots\}$. Our proof will show that the collection of all products el, for $e \in B_E$ and $l \in B_L$, is linearly independent over F, thus establishing the existence of arbitrarily large subsets of L which are linearly independent over F.) Returning to the proof, we have

$$(\sum_{i=1}^{n}\lambda_{i1}e_i)l_1 + (\sum_{i=1}^{n}\lambda_{i2}e_i)l_2 + \cdots + (\sum_{i=1}^{n}\lambda_{im}e_i)l_m = 0,$$

and since the coefficients $\sum_{i=1}^{n}\lambda_{ij}e_i$ are elements of E (since the λ_{ij} are in $F \subseteq E$ and the e_i are in E), the linear independence over E of $l_1, l_2, \ldots, l_m$ shows that all the coefficients are zero. That is,

$$\lambda_{1j}e_1 + \lambda_{2j}e_2 + \cdots + \lambda_{nj}e_n = 0$$

for all j. But here the coefficients λ_{ij} are in F, and since $e_1, e_2, \ldots, e_n$ are linearly independent over F, it follows that all the coefficients λ_{ij} are 0. So having started from the assumption that $\sum_{i,j}\lambda_{ij}e_il_j = 0$ we have proved that all the coefficients λ_{ij} have to be 0, which proves the linear independence over F of the e_il_j.

It remains to prove that the e_il_j span L as a vector space over F. Let $a \in L$ be arbitrary. Since the l_j span L over E there exist $b_1, b_2, \ldots, b_m \in E$ such that

$$a = b_1l_1 + b_2l_2 + \cdots + b_ml_m. \tag{17}$$

But since $e_1, e_2, \ldots, e_n$ span E over F, each b_j can be expressed as an F-linear combination of the e_i; that is,

$$b_j = \lambda_{1j}e_1 + \lambda_{2j}e_2 + \cdots + \lambda_{nj}e_n, \tag{18}$$

and substituting Eq.(18) into Eq.(17) gives

$$a = (\sum_{i=1}^{n}\lambda_{i1}e_i)l_1 + (\sum_{i=1}^{n}\lambda_{i2}e_i)l_2 + \cdots + (\sum_{i=1}^{n}\lambda_{im}e_i)l_m = 0,$$

showing that the arbitrarily chosen element a of L can be expressed as an F-linear combination of the e_il_j.

Since the mn elements e_il_j are both linearly independent and span L over F, they constitute an F-basis for L. So $[L : F] = mn = [L : E][E : F]$, as required. □

Suppose that K is an extension field of F, and $s, t \in K$ are algebraic over F. Put $E = F[s]$, and note that $[E : F]$ is finite (by Proposition (10.19)). Since t is algebraic over F there is a nonzero polynomial $f(x) \in F[x]$ such that $f(t) = 0$. Since F is a subfield of E, polynomials over F can also be regarded as polynomials over E, and in particular we can regard $f(x)$ as an element of $E[x]$. So t is a root of a is a nonzero polynomial in $E[x]$, which shows that t is algebraic over E. If we now put $L = E[t]$ then we have that $[L : E]$ is finite, and hence (by Theorem (10.21)) $[L : F] = [L : E][E : F]$ is finite too. Observe that $s \in L$ (since $s \in E \subseteq L$) and $t \in L$, and by the various closure properties of L (see Theorem (5.9)) we conclude that $s + t$, $s - t$, st and (if $t \neq 0$) st^{-1} are all elements of L. So by Theorem (10.20) these elements are all algebraic over F. Thus we have proved the following theorem.

Theorem (10.22): *Let s and t be elements of an extension field of F, and suppose that they are algebraic over F. Then $s + t$, $s - t$, st and (if $t \neq 0$) st^{-1} are algebraic over F.*

Example

Since $\sqrt[3]{2}$ and $\sqrt{5}$ are both algebraic over $\mathbb{Q}$, it follows that $\sqrt[3]{2} + \sqrt{5}$ is also algebraic over $\mathbb{Q}$. Indeed, $\sqrt[3]{2}$ is generates an extension E of $\mathbb{Q}$ of degree 3, and $\sqrt{5}$ generates an extension of E of degree 2; so it follows that $\sqrt[3]{2}$ and $\sqrt{5}$ are both elements of an extension of $\mathbb{Q}$ of degree 6. So there must be a polynomial in $\mathbb{Q}[x]$ of degree less than or equal to 6 that has $\sqrt[3]{2} + \sqrt{5}$ as a root. There are various ways to go about finding such a polynomial. The method suggested by the proofs given above is as follows. Calculate the powers $(\sqrt[3]{2} + \sqrt{5})^i$ for all i from 0 to 6, using the formulas $(\sqrt[3]{2})^3 = 2$ and $(\sqrt{5})^2 = 5$ to express them all in terms of $(\sqrt[3]{2})^i(\sqrt{5})^j$ for $i \in \{0, 1, 2\}$ and $j \in \{0, 1\}$. Then use linear algebra to find a linear relation between them. Writing $t = \sqrt[3]{2}$ and $u = \sqrt{5}$, a little calculation yields the following matrix equation.

$$\begin{pmatrix} (t+u)^0 \\ (t+u)^1 \\ (t+u)^2 \\ (t+u)^3 \\ (t+u)^4 \\ (t+u)^5 \\ (t+u)^6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 5 & 0 & 1 & 0 & 2 & 0 \\ 2 & 15 & 0 & 5 & 0 & 3 \\ 25 & 2 & 30 & 8 & 20 & 0 \\ 100 & 125 & 2 & 25 & 10 & 50 \\ 129 & 150 & 375 & 200 & 150 & 12 \end{pmatrix} \begin{pmatrix} 1 \\ t \\ t^2 \\ u \\ ut \\ ut^2 \end{pmatrix}$$

The aim is now to find a nonzero row vector v with rational entries satisfying the equation $vM = 0$, where M is the 7×6 matrix on the right hand side of the above equation. To do this, apply column operations to M to reduce it to column echelon form (or, if you prefer to use row operations, transpose M), and then solve the echelon system. This is just standard first year linear algebra, although somewhat tedious! The result is that

$$v = (-121, -60, 75, -4, -15, 0, 1)$$

is a solution. Hence $t + u$ is a root of the polynomial $-121 - 60x + 75x^2 - 4x^3 - 15x^4 + x^6$.

An easier method to get the answer is to make an educated guess. The polynomial $x^3 - 2$ has three roots in the complex field, namely t, ωt and $\overline{\omega} t$, where $\omega = -\frac{1}{2} + \frac{\sqrt{3}}{2}i$, a complex cube root of 1. However, from the point of view of the rational numbers, these three roots are indistinguishable from each other, in that they have the same minimal polynomial over $\mathbb{Q}$, and the extension fields they generate ($\mathbb{Q}[t]$, $\mathbb{Q}[\omega t]$ and $\mathbb{Q}[\overline{\omega} t]$) are isomorphic to each other. In the same way the two roots of $x^2 - 5$ (namely, $\pm\sqrt{5}$) are algebraically equivalent as far as $\mathbb{Q}$ is concerned. So it is reasonable to expect that an element of $\mathbb{Q}[x]$ that has $t + u$ as a root will also have $\omega t + u$, $\overline{\omega}t + u$, $t - u$, $\omega t - u$ and $\overline{\omega}t - u$ as roots. Multiplying it out, we find that

$$\begin{aligned}
(x - u - t)(x - u - \omega t)(x{-}u - \overline{\omega}t)(x + u - t)(x + u - \omega t)(x + u - \overline{\omega}t)& \\
&= ((x - u)^3 - 2)((x + u)^3 - 2) \\
&= (x^3 - 3x^2u + 15x - 5u - 2)(x^3 + 3x^2u + 15x + 5u - 2) \\
&= (x^3 + 15x - 2)^2 - (3x^2u + 5u)^2 \\
&= (x^6 + 30x^4 - 4x^3 + 225x^2 - 60x + 4) - (45x^4 + 150x^2 + 125) \\
&= x^6 - 15x^4 - 4x^3 + 75x^2 - 60x - 121
\end{aligned}$$

in agreement with the previously found answer.

Eisenstein's Irreducibility Criterion

This subsection is devoted to proving a useful criterion for irreducibility of elements of $\mathbb{Q}[x]$. Assume that F is a field and R a subring of F such that
(i) $1 \in R$,
(ii) every element of F is expressible in the form ab^{-1} with $a, b \in R$, and
(iii) R is a principal ideal domain.
Although all our proofs will be phrased in terms of F and R as above, in fact we will only be interested in the case that $R = \mathbb{Z}$ and $F = \mathbb{Q}$. (It is easy to check that items (i), (ii) and (iii) are then satisfied.) Indeed, for the rest of this subsection the reader is advised to assume that $R = \mathbb{Z}$ and $F = \mathbb{Q}$.

Temporarily, let us adopt the following notation: if $f(x), g(x) \in F[x]$ then "$f(x) \sim g(x)$" means "there exists a nonzero $\lambda \in F$ with $f(x) = \lambda g(x)$". In other words, $f(x) \sim g(x)$ if and only if $f(x)$ and $g(x)$ are associates in $F[x]$.

Note that the fact that R is a subring of F means that $R[x]$ is a subring of $F[x]$. (This is obvious, but can also viewed as an application of Exercise 7.)

Lemma (10.23): *For every $f(x) \in F[x]$ there exists $g(x) \in R[x]$ with $g(x) \sim f(x)$.*

Proof. We may write $f(x) = (b_0c_0^{-1}) + (b_1c_1^{-1})x + \cdots + (b_nc_n^{-1})x^n$ where the b_i and c_i are elements of R for all i, and the c_i are all nonzero. Then all the coefficients of $(c_0c_1\cdots c_n)f(x)$ lie in R. □

Lemma (10.24): *Let $f(x), g(x) \in R[x]$, and let p be an irreducible element of R. If all the coefficients of $f(x)g(x)$ are divisible by p then either all the coefficients of $f(x)$ are divisible by p or all the coefficients of $g(x)$ are divisible by p.*

Proof. Let $a \mapsto \overline{a}$ be the canonical surjective homomorphism from R to R/pR, and also let $a(x) \mapsto \overline{a(x)}$ be the homomorphism $R[x] \to (R/pR)[x]$ to which the homomorphism $R \to R/pR$ gives rise in the manner described in Exercise 7. That is, if

$$a(x) = a_0 + a_1x + \cdots + a_nx^n$$

then by definition

$$\overline{a(x)} = \overline{a_0} + \overline{a_1}x + \cdots + \overline{a_n}x^n.$$

Now $\overline{a_i} = \overline{0}$ if and only if $p \mid a_i$, and so $\overline{a(x)}$ is the zero polynomial in $(R/pR)[x]$ if and only if $p \mid a_i$ for all i.

Suppose that all the coefficients of $f(x)g(x)$ are divisible by p. Then

$$\overline{f(x)g(x)} = \overline{0}$$

and therefore

$$(\overline{f(x)})(\overline{g(x)}) = \overline{0}.$$

But R/pR is a field (Theorem (10.4)); so $(R/pR)[x]$ is an integral domain (Theorem (6.3)), and therefore

$$\overline{f(x)} = 0 \quad \text{or} \quad \overline{g(x)} = 0$$

since $(R/pR)[x]$ can have no zero divisors. But this says that either all the coefficients of $f(x)$ are divisible by p or else all the coefficients of $g(x)$ are divisible by p, as required. □

Our next lemma is a key step in the proof of Eisenstein's Criterion, and is interesting in its own right. It says that if a polynomial with integer coefficients can be factorized nontrivially in $\mathbb{Q}[x]$, then factors can be found which have integer coefficients. You never really need to use fractions to factorize an integer polynomial. Thus, for example, although $x^2 + 5x + 6 = (\frac{1}{2}x + 1)(2x + 6)$, fractions can be avoided by multiplying the factors by suitable scalars: $x^2 + 5x + 6 = (x + 2)(x + 3)$.

Lemma (10.25): *If $f(x) \in R[x]$ and $f(x) = g(x)h(x)$ for some $g(x), h(x) \in F[x]$ then there exist $g_1(x), h_1(x) \in R[x]$, with $g_1(x) \sim g(x)$ and $h_1(x) \sim h(x)$, satisfying $f(x) = g_1(x)h_1(x)$.*

Proof. Choose $a, b \in R$ such that $ag(x), bh(x) \in R[x]$. Then $(ab)f(x) = (ag(x))(b(h(x))$, which shows that there exists at least one $c \in R$ such that $cf(x)$ factorizes as $g_1(x)h_1(x)$ with $g_1(x), h_1(x) \in R[x]$, and $g_1(x) \sim g(x)$ and $h_1(x) \sim h(x)$. Now since R is a unique factorization domain, every $c \in R$ which has this property can be expressed as a product of irreducible elements: $c = p_1p_2 \cdots p_n$. Amongst all possible choices for c, choose one for which n is as small as possible. That is, n is the least nonnegative integer such that there exist irreducible elements $p_1, p_2, \ldots, p_n \in R$ with $(p_1p_2 \cdots p_n)f(x) = g_1(x)h_1(x)$ for some $g_1(x), h_1(x) \in R[x]$ such that $g_1(x) \sim g(x)$ and $h_1(x) \sim h(x)$.

Suppose that $n > 0$. Then all the coefficients of $(p_1p_2 \cdots p_n)f(x)$ are divisible by p_n, since all the coefficients of $f(x)$ itself are in R. By Lemma (10.24), either p_n divides all the coefficents of $g_1(x)$ or else p_n divides all the coefficients of $h_1(x)$. In the former case we put $g_2(x) = p_n^{-1}g_1(x)$ and $h_2(x) = h_1(x)$, in the latter case we put $g_2(x) = g_1(x)$ and $h_2(x) = p_n^{-1}h_1(x)$. Then $g_2(x), h_2(x) \in R[x]$,

$$g_2(x)h_2(x) = p_n^{-1}(g_1(x)h_1(x)) = (p_1p_2 \cdots p_{n-1})f(x),$$

and since also $g_2(x) \sim g_1(x) \sim g(x)$ and $h_2(x) \sim h_1(x) \sim h(x)$ this contradicts the minimality of n. □

Example

We illustrate a useful application of Lemma (10.25): it enables one to find all the rational roots of an arbitrary integer polynomial.

Let $f(x) = 12x^4 - 4x^3 - 37x^2 + 14x + 15 \in \mathbb{Z}[x]$. We shall find all $\alpha \in \mathbb{Q}$ such that $f(\alpha) = 0$. By Theorem (9.3), if $f(\alpha) = 0$ then $f(x) = (x - \alpha)h(x)$ for some $h(x) \in \mathbb{Q}[x]$. By Lemma (10.25) this yields an integral factorization

$$f(x) = (qx - p)h_1(x) \quad \text{with } qx - p,\ h_1(x) \in \mathbb{Z}[x]$$

such that $qx - p$ is a scalar multiple of $x - \alpha$. That is, $\alpha = p/q$. Now on multiplying it out and equating coefficients,

$$12x^4 - 4x^3 - 37x^2 + 14x + 15 = (qx - p)(ax^3 + bx^2 + cx + d)$$

gives $12 = qa$ and $15 = -pq$. So $q \mid 12$ and $p \mid 15$. So p is one of ± 1, ± 3, ± 5, ± 15 and q is one of ± 1, ± 2, ± 3, ± 4, ± 6, ± 12. It is now easy to try the possibilities one at a time until the rational roots are found. (After finding one root the amount of calculation can be reduced by taking out the corresponding degree 1 factor of $f(x)$ and repeating the process with the lower degree polynomial that is left.) It turns out that in this example the values of p/q that give roots are 1, $3/2$, $-5/3$ and $-1/2$.

Since the same process obviously works for any integer polynomial, we can see that the following theorem holds.

Theorem (10.26): *If $f(x) = a_0 + a_1x + \cdots + a_dx^d \in \mathbb{Z}[x]$ then every root of $f(x)$ in $\mathbb{Q}$ has the form $\pm(p/q)$ for some $p, q \in \mathbb{Z}$ such that $p \mid a_0$ and $q \mid a_d$.*

Theorem (10.26) is known as the Rational Roots Theorem.

Numbers (real or complex) which are roots of monic polynomials in $\mathbb{Z}[x]$ are called *algebraic integers*. The Rational Roots Theorem has the following consequence, which, although we will not make use of it in this course, is nevertheless an important mathematical fact: an algebraic integer that is rational must be an element of $\mathbb{Z}$.

We now come to Eisenstein's Criterion.

Theorem (10.27): *Let $f(x) = a_0 + a_1x + \cdots + a_dx^d \in R[x]$, where $d \geq 1$, and suppose that there exists a prime $p \in R$ such that*
(i) $p \mid a_i$ for $i = 0, 1, \ldots, d-1$,
(ii) $p \nmid a_d$, and
(iii) $p^2 \nmid a_0$.
Then $f(x)$ is irreducible as an element of $F[x]$.

Before proving it we give an example to illustrate its application. Consider the polynomial $f(x) = 3x^5 + 4x^3 + 2x^2 - 16x - 10 \in \mathbb{Z}[x]$. Observe that the coefficients other than the leading coefficient are -10, -16, 2, 4 and 0, which are all divisible by the prime 2. The leading coefficient, 3,is not divisible by 2, and the constant coefficient, -10, is not divisible by 2^2. So Eisenstein's Criterion applies, and tells us that $f(x)$ is irreducible over $\mathbb{Q}$. (In consequence, if $K = f(x)\mathbb{Q}[x]$ then $F = \mathbb{Q}[x]/K$ is a field. Moreover, F can be alternatively obtained by adjoining a root of $f(x)$ to $\mathbb{Q}$.)

Proof of (10.27). Suppose that $f(x)$ is not irreducible. Then it follows that $f(x)$ has a factorization $f(x) = g(x)h(x)$ such that $g(x), h(x) \in F[x]$ have degrees less than d and greater than 0. Lemma (10.25) tells us that $g(x)$and $h(x)$ can be chosen to be in $R[x]$. Applying the homomorphism $R[x] \to (R/pR)[x]$ given by $a(x) = \sum_i a_ix^i \mapsto \overline{a(x)} = \sum_i \overline{a_i}x^i$, where $a \mapsto \overline{a}$ is the canonical surjective homomorphism $R \to R/pR$, we obtain

$$\overline{f(x)} = (\overline{g(x)})(\overline{h(x)}). \tag{19}$$

But all the coefficients of $f(x)$ except the leading coefficient are congruent to 0 modulo p; so $\overline{f(x)} = \overline{a_d}x^d$. Furthermore, R/pR is a field, whence $(R/pR)[x]$ is a unique factorization domain, and so the only irreducible factors of $\overline{a_d}x^d$ are scalar multiples of x. Now Eq.(19) and uniqueness of factorization in $(R/pR)[x]$ combine to show that the only irreducible factors of $\overline{g(x)}$ and $\overline{h(x)}$ are scalar multiples of x. Hence

$$\overline{g(x)} = \bar{b}x^n, \qquad \overline{h(x)} = \bar{c}x^m$$

for some nonnegative integers n, m and some c, $d \in R$ such that $\bar{c}$, $\bar{d}$ are nonzero. Moreover, $n \leq \deg g(x) < d$ and $m \leq \deg h(x) < d$, and since also

$$d = \deg \overline{f(x)} = \deg \overline{g(x)} + \deg \overline{h(x)} = n + m \leq \deg g(x) + \deg h(x) = \deg f(x) = d,$$

it follows that $n = \deg g(x)$ and $m = \deg h(x)$. Hence

$$\begin{aligned} g(x) &= b_0 + b_1x + \cdots + b_nx^n \\ h(x) &= c_0 + c_1x + \cdots + c_mx^m \end{aligned}$$

for some b_i, $c_j \in R$ such that $b_0, b_1, \ldots, b_{n-1}$ and $c_0, c_1, \ldots, c_{m-1}$ are congruent to 0 modulo p, and $b_n \equiv b \not\equiv 0$ and $c_m \equiv c \not\equiv 0 \pmod{p}$. Since n, $m > 0$, in particular $p \mid b_0$ and $p \mid c_0$. Now

$$a_0 + a_1x + \cdots + a_dx^d = f(x) = g(x)h(x) = (b_0 + b_1x + \cdots + b_nx^n)(c_0 + c_1x + \cdots + c_mx^m)$$

gives, on equating coefficients, that $a_0 = b_0c_0$. Since $p \mid b_0$ and $p \mid c_0$ this gives $p^2 \mid a_0$, contrary to hypothesis (iii) of the theorem. Hence the original assumption that $f(x)$ is reducible must be false. □

Exercise 41. Use the Rational Roots Theorem to prove that $\sqrt[5]{2}$ is irrational. Then use Eisenstein's Criterion to show that the minimal polynomial of $\sqrt[5]{2}$ over $\mathbb{Q}$ is $x^5 - 2$.

Example

Let p be a prime integer. Recall that the binomial coefficients $\binom{p}{i}$ are integers and satisfy the formula

$$i(i-1)(i-2)\cdots 3 \cdot 2 \cdot 1 \binom{p}{i} = p(p-1)(p-2)\cdots(p-i+1).$$

If $1 \leq i \leq p-1$ then the prime p appears as a factor on the right hand side, and so by Corollary (9.19) it must be a divisor of one of the factors on the left hand side. Clearly it is not a divisor of any of $1, 2, 3, \ldots, i$ (since $i < p$), and so it follows that $p \mid \binom{p}{i}$. We shall use this together with Eisenstein's Criterion and a change of variable argument to show that the polynomial $f(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$ is irreducible over $\mathbb{Q}$.

Suppose to the contrary that $f(x)$ is reducible. Then $f(x)$ has a factorization $f(x) = r(x)s(x)$ such that $r(x)$, $s(x) \in \mathbb{Q}[x]$ both have degree less than the degree of $f(x)$. Define new polynomials $f_1(x)$, $r_1(x)$ and $s_1(x)$ by replacing x by $x+1$ in $f(x)$, $r(x)$ and $s(x)$. That is, put $f_1(x) = f(x+1)$, $r_1(x) = r(x+1)$ and $s_1(x) = s(x+1)$. It is not hard to see that the degrees of $f_1(x)$, $r_1(x)$ and $s_1(x)$ are the same as the degrees of $f(x)$, $r(x)$ and $s(x)$ respectively, and since

$$f_1(x) = f(x+1) = r(x+1)s(x+1) = r_1(x)s_1(x)$$

it follows that $f_1(x)$ is not irreducible. But we shall show that this contradicts Eisenstein's Criterion.

Observe that $f(x) = (x^p - 1)/(x-1)$. It follows from the binomial theorem that

$$\begin{aligned} f_1(x) = f(x+1) &= \frac{(x+1)^p - 1}{(x+1) - 1} = \frac{\left(\sum_{i=0}^{p} \binom{p}{i} x^i\right) - 1}{x} \\ &= x^{-1} \sum_{i=1}^{p} \binom{p}{i} x^i = \binom{p}{1} + \binom{p}{2}x + \binom{p}{3}x^2 + \cdots + \binom{p}{p-1}x^{p-2} + x^{p-1} \end{aligned}$$

(since $\binom{p}{0} = \binom{p}{p} = 1$). The leading coefficient of $f_1(x)$ is 1, which is not divisible by p, the constant coefficient is p, which is divisible by p but not by p^2, and all the other coefficients are divisible by p since they are of the form $\binom{p}{i}$ with $2 \leq i \leq p-1$. Eisenstein's Criterion thus tells us $f_1(x)$ is irreducible over $\mathbb{Q}$, and this contradiction shows that our original assumption is false. That is, $f(x)$ is irreducible.

11. Equation solving and constructible numbers revisited

We saw in Theorem (3.2) that a real number t is constructible if and only if there exists a sequence of real numbers $t_0, t_1, \ldots, t_n$ such that $t_0 = 0$, $t_1 = 1$ and $t_n = t$, and for each i from 2 to n the number t_i is a sum, product, negative, reciprocal or square root of an earlier term or terms in the sequence. Given such a sequence, define a sequence of subfields of $\mathbb{R}$ recursively as follows: put $F_0 = \mathbb{Q}$, and for $i > 0$ put $F_i = F_{i-1}[t_i]$. Observe that if $t_i \in F_{i-1}$ then $F_i = F_{i-1}$, and this happens in particular if t_i is a sum, product, negative or reciprocal of anearlier term or terms, since fields are closed under these operations. Deleting repeated terms we thus obtain a sequence of subfields of $\mathbb{R}$

$$\mathbb{Q} = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_m \tag{20}$$

such that $t \in F_m$, and for each j from 0 to $m-1$,

$$F_{j+1} = F_j[\sqrt{a_j}]$$

for some positive number $a_j \in F_j$ whose square root is not in F_j. Note that the condition $\sqrt{a_j} \notin F_j$ shows that $\sqrt{a_j}$ is not a root of any polynomial in $F_j[x]$ of degree 1, and in consequence $x^2 - a_j$ is the minimal polynomial of $\sqrt{a_j}$ over F_j. So by Proposition (10.19) we conclude that $[F_{j+1} : F_j] = 2$.

These considerations allow us to reformulate our criterion for a number to be constructible.

Theorem (11.1): *A number $t \in R$ is constructible if and only if there exists a chain of subfields of $\mathbb{R}$ as in Eq.(20) such that $t \in F_m$ and $[F_{j+1} : F_j] = 2$ for each j from 0 to $m-1$.*

Proof. The discussion above showed that if t is constructible then there is a chain of fields of the required kind. Suppose, conversely, that $\mathbb{Q} = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_m$ is a chain of subfields of $\mathbb{R}$ with $[F_{j+1} : F_j] = 2$ for all j from 0 to $m-1$. We use induction on m to show that every element of F_m is constructible.

If $m = 0$ the result follows from the fact that all rational numbers are constructible (which holds since starting from 0 and 1 you can get to any rational number by a finite sequence of additions, multiplications, subtractions and divisions). Suppose now that $m > 0$ and that the result holds for shorter sequences of fields. In particular, then, the inductive hypothesis tell us that all elements of the field F_{m-1} are constructible, and we must show that all elements of $F_m = F_{m-1}[\sqrt{a_{m-1}}]$ are constructible. Now by Corollary (10.6) we know that each element of F_m is expressible in the form $a + b\sqrt{a_{m-1}}$ with a, b and a_{m-1} in F_{m-1}. Since a, b and a_{m-1} are in F_{m-1} they are constructible, and since $a + b\sqrt{a_{m-1}}$ is obtained from a, b and a_{m-1} numbers by use of operations of addition, multiplication and square root extraction, it too is constructible, as required. □

Corollary (11.2): *Suppose that $t \in \mathbb{R}$ is constructible. Then t is algebraic over $\mathbb{Q}$ and $[\mathbb{Q}[t] : \mathbb{Q}]$ is a power of 2.*

Proof. By Theorem (11.1) we can find a chain of fields as in Eq.(20) such that each degree $[F_{j+1} : F_j]$ is 2 and $t \in F_m$. By Theorem (10.21) we see that $[F_m : \mathbb{Q}] = 2^m$. But since $\mathbb{Q} \subseteq F_m$ and $t \in F_m$ we have that $\mathbb{Q}[t]$ is a subfield of F_m, and so Theorem (10.21) also yields

$$2^m = [F_m : \mathbb{Q}] = [F_m : \mathbb{Q}[t]][\mathbb{Q}[t] : \mathbb{Q}].$$

Thus $[\mathbb{Q}[t] : \mathbb{Q}]$ is a divisor of 2^m, and by unique factorization of integers it follows that it is a power of 2, as required. □

We now have at our disposal a proof of the impossibility of one of the three ruler and compasses constructions we mentioned in Section 3: the doubling of the cube. If this could be done by ruler and compasses then $\sqrt[3]{2}$ would have to be a constructible number, but since its minimal polynomial over $\mathbb{Q}$ is $x^3 - 2$ (since this is irreducible over $\mathbb{Q}$ by Eisenstein's Criterion) it follows that $[\mathbb{Q}[\sqrt[3]{2}] : \mathbb{Q}] = 3$, which is not a power of 2, contradicting Corollary (11.2).

Assuming Lindemann's Theorem that π is not a root of any nontrivial polynomial with integer coefficients, it follows that the same is true of $\sqrt{\pi}$, and in particular it follows that $\sqrt{\pi}$ is not algebraic over $\mathbb{Q}$. This shows that $\sqrt{\pi}$ is not a constructible number, and so also the problem of squaring the circle cannot be solved by a ruler and compasses construction.

The problem of trisecting an arbitrary given angle is similar to the problem of doubling the cube, in that to solve it one would have to solve a cubic equation. If we are given an angle POQ we can set up a coordinate system with O as the origin and P as $(1, 0)$, and then if R is the point on OQ such that PR and OQ are perpendicular, we see that OR has length equal to the cosine of $\angle POQ$. Constructing an angle equal to one third of $\angle POQ$ is easily seen to be equivalent to constructing a line segment whose length is $\cos\left(\frac{1}{3}\angle POQ\right)$. Since $\cos\frac{1}{3}\theta$ is a root of the polynomial $4x^3 - 3x - \cos\theta$, solution of a cubic is involved. This cubic could well be irreducible, which would mean that $\cos\frac{1}{3}\theta$ would generate a degree 3 extension of $\mathbb{Q}[\cos\theta]$, unachievable by ruler and compasses since 3 is not a power of 2.

Exercise 42. Show that if $f(x) \in F[x]$ has no roots in F and has degree 2 or 3 then it is irreducible.

Exercise 43. Use the trigonometric formula $\cos 3\theta = 4(\cos\theta)^3 - 3\cos\theta$ to show that the cosine of $\pi/9$ (or 20°) generates a degree 3 extension of $\mathbb{Q}$, and deduce that an angle of $\pi/3$ cannnot be trisected by ruler and compasses.

Regular polygons

Everyone knows that equilateral triangles, squares and regular hexagons can be constructed with ruler and compasses. Using the fact that it is possible to bisect angles, it is fairly easy to see that if a regular n-sided polygon can be constructed then so can a regular $2n$-sided polygon. So regular octagons, 16-gons, 32-gons, ... and 12-gons, 24-gons and so on can be constructed with ruler and compasses. What other regular n-gons can be constructed?

It is fairly well known that regular pentagons can constructed, and there are simple constructions. We leave it to the reader's ingenuity to devise one if (s)he does not know one already. In view of the theory we have been through, the algebraic explanation of why this can be done is that $\cos(2\pi/5)$ is a constructible number. This can be seen as follows. Since the roots of $x^5 - 1$ are the complex fifth roots of 1, $e^{2k\pi i/5} = \cos(2k\pi/5) + i\sin(2k\pi/5)$, we have that

$$x^5 - 1 = (x - 1)(x - e^{2\pi i/5})(x - e^{4\pi i/5})(x - e^{6\pi i/5})(x - e^{8\pi i/5}).$$

Note that $e^{8\pi i/5}$ is the inverse and complex conjugate of $e^{2\pi i/5}$, and $e^{4\pi i/5}$ and $e^{6\pi i/5}$ are similarly related. So pairing the second and fifth factors above and also the third and fourth we find that

$$x^5 - 1 = (x - 1)(x^2 - 2\cos(2\pi/5) + 1)(x^2 - 2\cos(4\pi/5) + 1),$$

whence

$$x^4 + x^3 + x^2 + x + 1 = (x^2 - ax + 1)(x^2 - bx + 1),$$

where $a = 2\cos(2\pi/5)$and $b = 2\cos(4\pi/5)$. Expanding the right hand side gives $a + b = -1$ and $ab = -1$, so that a and b are the roots of $x^2 + x - 1$. So a and b can be constructed by finding an appropriate square root ($\sqrt{5}$, in fact).

It is a remarkable fact, discovered by Gauss, that for odd values of n a regular n-gon can be constructed if and only if n is a Fermat prime. This means that n must be prime and $n-1$ a power of 2. Now observe that if r is odd then

$$x^r + 1 = (x+1)(x^{r-1} - x^{r-2} + x^{r-3} - \cdots + x^2 - x + 1)$$

and evaluating at $x = 2^k$ we see that $2^{rk}+1$ is divisible by 2^k+1 whenever r is odd. So 2^m+1 cannot be prime if m has any odd factor r. So Fermat primes must have the form 2^m+1 with m a power of 2. Examples are $2^0+1=2$, $2^1+1=3$, $2^2+1=5$, $2^4+1=17$, $2^8+1=257$ and $2^{16}=1=65537$. The reader is invited to check that $2^{32}+1 \equiv 0 \pmod{641}$, and is therefore not prime. (To do this, it is convenient to make use of the readily checked facts that $5 \cdot 2^8 \equiv -1 \pmod{641}$ and $5^4 \equiv -2^4 \pmod{641}$.) In fact, no more Fermat primes are known.

The following two exercises guide the reader through a proof that $\cos(2\pi/17)$ is a constructible number.

Exercise 44. Let $\theta = 2\pi/17$ and let $\omega = e^{i\theta} = \cos\theta + i\sin\theta$, a complex 17th root of 1. Prove that

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

Exercise 45. Define complex numbers as follows:

$$\begin{aligned}
\alpha_1 &= \frac{-1+\sqrt{17}}{2}, \quad \alpha_2 = \frac{-1-\sqrt{17}}{2} \\
\beta_1 &= \frac{1}{2}(\alpha_1 + \sqrt{\alpha_1^2 + 4}) \\
\beta_2 &= \frac{1}{2}(\alpha_1 - \sqrt{\alpha_1^2 + 4}) \\
\beta_3 &= \frac{1}{2}(\alpha_2 + \sqrt{\alpha_2^2 + 4}) \\
\beta_4 &= \frac{1}{2}(\alpha_2 - \sqrt{\alpha_2^2 + 4})
\end{aligned}$$

$$\begin{aligned}
\gamma_1 &= \frac{1}{2}(\beta_1 + \sqrt{\beta_1^2 - 4\beta_3}) & \gamma_5 &= \frac{1}{2}(\beta_3 + \sqrt{\beta_3^2 - 4\beta_1}) \\
\gamma_2 &= \frac{1}{2}(\beta_1 - \sqrt{\beta_1^2 - 4\beta_3}) & \gamma_6 &= \frac{1}{2}(\beta_3 - \sqrt{\beta_3^2 - 4\beta_1}) \\
\gamma_3 &= \frac{1}{2}(\beta_2 + \sqrt{\beta_2^2 - 4\beta_4}) & \gamma_7 &= \frac{1}{2}(\beta_4 + \sqrt{\beta_4^2 - 4\beta_2}) \\
\gamma_4 &= \frac{1}{2}(\beta_2 - \sqrt{\beta_2^2 - 4\beta_4}) & \gamma_8 &= \frac{1}{2}(\beta_4 - \sqrt{\beta_4^2 - 4\beta_2}).
\end{aligned}$$

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

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

and similarly

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

(ii) Check that

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

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 $\gamma_1, \gamma_2, \ldots, \gamma_8$ are equal to $2\cos\theta, 2\cos 2\theta, \ldots, 2\cos 8\theta$ (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 compasses.

Let us now turn our attention to solution of arbitrary polynomial equations. Our first task is to decide what it means for a polynomial equation to be soluble by radicals. We employ the following definition.

Definition (11.3): Let $f(x) \in F[x]$, where F is a field. We say that the equation $f(x) = 0$ is *soluble by radicals* if there is a chain of field extensions $F = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_m$ such that

(i) for each i from 0 to $m-1$ the field F_{i+1} is obtained by adjoining to F_i a k_i-th root of some element of F_i (where k_i is some positive integer), and

(ii) there exist $\alpha_1, \alpha_2, \ldots, \alpha_d \in F_m$ such that $f(x) = (x-\alpha_1)(x-\alpha_2)\cdots(x-\alpha_d)$.

It is probably not immediately apparent how this relates to our intuitive notion of what it means for an equation to be soluble. Normally, saying that something is soluble suggests the existence of an algorithmic procedure that will lead one to the solution. In the case of a polynomial equation, the algorithm should consist of performing various arithmetical operations, taking the coefficients of the polynomial as input and yielding its roots as output. If the equation is soluble by radicals, then the only arithmetical operations allowed—apart from addition, subtraction, multiplication and division—are k-th root extractions (for positive integers k).† Since addition, subtraction, multiplication and division of field elements yield other elements of the same field it is only the root extractions that necessitate field extensions. Forming such an extension for each root extraction involved in the solution process leads to a chain of extensions of the kind described in Definition (11.3). So it is at least in accord with intuition that if an equation is soluble by radicals then there should exist a chain of extensions as described in (11.3).

The other direction is perhaps more contentious. Is it right to say that an equation is soluble by radicals simply because some chain of field extensions exists, without there necessarily being any clear-cut procedure for finding the fields, or for finding the elements of the final field in the chain which are the actual solutions of the equation? Perhaps not. But then, we are aiming to show that certain equations are not soluble by radicals, and for this it will certainly be adequate to show that no chain of field extensions exists having the properties required in Definition (11.3). This being said, it nevertheless turns out to be the case—fortunately—that whenever a polynomial equation is soluble by radicals in the sense of Definition (11.3), procedures can be found for constructing the requisite field extensions, and formulas for the roots can be written down.

† The word “radical” is derived from the Latin *radix -icis*, meaning “root”; in the present context “radical” means either “root” (of an equation $x^n = a$) or the root sign “$\sqrt{\ }$”.

12. Symmetry

As we suggested in Section 2, symmetry plays an important role in the theoretical investigation of solutions of polynomial equations. By a *symmetry* of an object (of any kind) we mean a transformation of the object which preserves its essential features. Of course the identity (do-nothing) transformation is always a symmetry; furthermore, it should always be possible to undo a symmetry. That is, for a transformation to be a symmetry there should exist an inverse transformation (which will also be a symmetry). Finally, the composite effect of performing one symmetry and then following it with another will also be a transformation that preserves the essential features of the object, so that the composite of two symmetries is another. Thus the set of all symmetries of an object is a group, in the sense of the following definition.

Definition (12.1): A *group* is a set G equipped with an operation $(x, y) \mapsto xy$ such that
(i) $x(yz) = (xy)z$ for all x, y, $z \in G$,
(ii) there is an element $i \in G$ such that $ix = xi = x$ for all $x \in G$, and
(iii) with i as in (ii), for every $x \in G$ there exists a $y \in G$ such that $xy = yx = i$.

The use of the term "group" in this sense originated with Galois, who introduced it as an aid in explaining his ideas concerning the solution of polynomial equations. For the present course we do not need to go very deeply into group theory.

A group G is said to be *Abelian* or *commutative* if $xy = yx$ for all x, $y \in G$. The element i appearing in Axiom (ii) is called the *identity* element of G. It is easily proved that the identity element is unique. In most group theory books the identity element of a group is denoted by 1. We shall probably adopt this convention after a while, but for the time being we shall denote the identity of G by i_G. If x, $y \in G$ are related as in Axiom (iii) then they are said to be *inverses* of each other. It is also easily shown that the inverse of an element is unique, and the inverse of x is denoted by x^{-1}, as one would expect. If the group operation is written as addition—which will only ever be done if the group is Abelian—then the identity element is denoted by 0 and called the *zero* element, and the inverse of x is denoted by $-x$ and called the *negative* of x. Observe that the first four ring axioms say that a ring is an Abelian group under addition; the group derived from a ring in this way is called the *additive group* of the ring in question. Observe also that if F is a field then the nonzero elements of F form a group; this is called the multiplicative group of F.

Definition (12.2): A subset H of a group G is called a *subgroup* of G if H is a group under an operation compatible with the group operation of G. We write $H \leq G$ to indicate that H is a subgroup of G.

As explained in our discussion of subrings, if a set S has an operation $*$ defined on it then a subset T of S has a compatible operation if and only if T satisfies the closure condition: $x * y \in T$ for all x, $y \in T$. It is straightforward to derive the following criterion, analogous to Theorem (5.8) for rings, for a subset of a group to be a subgroup.

Theorem (12.3): *A subset H of a group G is a subgroup of G if (and only if) H is nonempty and for all x, $y \in G$*
(i) if x, $y \in H$ then $xy \in H$, and
(ii) if $x \in H$ then $x^{-1} \in H$.

Exercise 46. Show that if $H_1, H_2, \ldots, H_n$ are subgroups of the group G then the intersection $H_1 \cap H_2 \cap \cdots \cap H_n$ is also a subgroup.

A group G is said to be *cyclic* if there exists a $g \in G$ such that every element of G is a power of g. (Powers of g include negative powers—powers of g^{-1}—as well as $g^0 = i_G$.) An element g which has the property that every element of G is a power of g is called a *generator* of the cyclic group G. (Note that if the group operation is written as addition then g^n becomes ng, and the condition for g to be a generator becomes that every element of G is a natural multiple of g.) It turns out that the generator g of a cyclic group is not unique if the group has more than two elements. The additive group of the ring $\mathbb{Z}$ is an example of a cyclic group with infinitely many elements, and it is easily seen that 1 and -1 are the only two generators. The additive group of the ring $\mathbb{Z}_n$ is a cyclic group with exactly n elements. Again 1 and -1 are generators, but there are likely to be others too. For example, in $\mathbb{Z}_7$ the first seven multiples of 4 are (in order) 4, 1, 5, 2, 6, 3 and 0, and since this list includes every element of $\mathbb{Z}_7$ it follows that 4 is a generator. Similar calculations show that in fact all nonzero elements of $\mathbb{Z}_7$ are generators.

It is not hard to see that each element of an arbitrary group G is a generator of a cyclic subgroup of G.

Proposition (12.4): *Let G be a group and $x \in G$. Then the set $\{\, x^n \mid n \in \mathbb{Z} \,\}$ is a subgroup of G.*

The proof of Proposition (12.4) is a trivial application of Theorem (12.3), and is omitted. We shall often use the notation $\langle x \rangle$ to denote the cyclic subgroup generated by the element x.

A *homomorphism* from a group G to a group H is a function $\phi\colon G \to H$ having the property that $\phi(xy) = (\phi x)(\phi y)$ for all $x, y \in G$. A homomorphism which is bijective is called an *isomorphism*, and two groups are said to be isomorphic if there is an isomorphism from one to the other. It is easily shown that the identity function from a group to itself is an isomorphim, the inverse of an isomophism is always an isomorphism, and the composite of two isomorphisms is always an isomorphism. Hence the relation "is isomorphic to" is an equivalence relation.

The following proposition shows that, to within isomorphism, the additive groups $\mathbb{Z}$ and $\mathbb{Z}_n$ are the only cyclic groups.

Proposition (12.5): *Let x be a generator of a cyclic group C. If there is no positive integer n such that $x^n = i_G$ then the mapping $\phi\colon \mathbb{Z} \to G$ given by $\phi m = x^m$ is an isomorphism from the additive group of $\mathbb{Z}$ to G; otherwise, if n is the least positive integer such that $x^n = i_G$ then there is a well-defined mapping $\psi\colon \mathbb{Z}_n \to G$, which is an isomorphism from the additive group of $\mathbb{Z}_n$ to G, satisfying $\psi\overline{m} = x^m$ for all $m \in \mathbb{Z}$.*

Proof. Define $\phi\colon \mathbb{Z} \to G$ by $\phi m = x^m$, and note that ϕ is surjective since x is a generator of G. For all $r, s \in \mathbb{Z}$

$$\phi(r+s) = x^{r+s} = x^r x^s = (\phi r)(\phi s),$$

which shows that ϕ is a homomorphism from $\mathbb{Z}$ (under addition) to G. (Note that our discussion of exponent laws in Section 4 applies to groups just as well as to rings.) Now define $K = \{\, r \in \mathbb{Z} \mid \phi r = i_G \,\}$ (the kernel of ϕ). Observe that $0 \in K$, and if $r, s \in K$ then

$$\phi(r+s) = (\phi r)(\phi s) = (i_G)(i_G) = i_G,$$

and also

$$\phi(-r) = i_G\phi(-r) = (\phi r)(\phi(-r)) = \phi(r-r) = \phi 0 = i_G,$$

so that $-r, r+s \in K$. It follows from these properties that if $r \in K$ and $m \in \mathbb{Z}$ then $mr = rm \in K$ (because multiplication in the ring $\mathbb{Z}$ can be expressed in terms of addition and subtraction: $mr = \underbrace{r + r + \cdots + r}_{m\text{terms}}$ if $m \geq 0$, and $mr = (-m)(-r)$ if $m < 0$). Hence K is an ideal in $\mathbb{Z}$ (by Theorem (7.3)), and by Theorem (7.4) we deduce that $K = n\mathbb{Z}$ for some nonnegative integer n. Recall also from our discussion of ideals in $\mathbb{Z}$ that n is the least positive integer in K if there is a positive integer in K, and $n = 0$ otherwise.

If $x^r = x^s$ then $x^{r-s} = x^r x^{-s} = x^s x^{-s} = x^0 = i_G$, so that $r - s \in K$. Hence if $K = \{0\}$ (corresponding to $n = 0$) then ϕ is injective, whence it is a bijective and an isomorphism $\mathbb{Z} \to G$. If $n \neq 0$ then we have that $\phi r = \phi s$ if and only $r \cong s \pmod{n\mathbb{Z}}$; in other words, $\phi r = \phi s$ if and only if $\bar{r} = \bar{s} \in \mathbb{Z}_n$. So there is a well-defined and injective mapping $\psi\colon \mathbb{Z}_n \to G$ such that $\psi\bar{r} = \phi r$ for all $r \in \mathbb{Z}$. (It is well-defined since $\bar{r} = \bar{s}$ implies $\phi r = \phi s$, and injective since $\phi r = \phi s$ implies $\bar{r} = \bar{s}$.) Moreover, ψ is surjective since ϕ is, and is a homomorphism since for all r and s,

$$\psi(\bar{r} + \bar{s}) = \psi(\overline{r+s}) = \phi(r+s) = (\phi r)(\phi s) = (\psi\bar{r})(\psi\bar{s}).$$

Hence ψ is an isomorphism, as required. □

If S is an arbitrary set then the number of elements of S (sometimes by $\#S$. In particular, the statement that S has only finitely many elements is conveniently written as "$|S| < \infty$".

Definition (12.6): If G is a group then $|G|$ is called the *order* of G. If $x \in G$ then the order of the cyclic subgroup $\langle x \rangle$ is called the *order* or *period* of the element x.†

Note that $g \in G$ has finite order if and only if there is a positive integer n such that $g^n = i_G$, in which case the order is the least such positive integer, and the integers n such that $g^n = i_G$ are precisely the multiples of the order. Thus if g has order n then the order of any power of g will be a divisor of n, since $(g^r)n = (g^n)^r = i_G$. Furthermore, if $n = de$ then g^e has order d, since $(g^e)^d = g^n = i_G$, and if r is a positive integer less than d then $(g^e)^r = g^{er} \neq i_G$ (since er is a positive integer less than ed, which is the order of g).

Exercise 47. Show that if G is a cyclic group of order n and H a subgroup of G then H is also cyclic, and the order of H is a divisor of n. Show furthermore that for each divisor d of n there is exactly one subgroup of G of order d. (Hint: If g is a generator of G then a subgroup H of G will be generated by g^k, where k is the least positive integer such that $g^k \in H$. Now $n = qk + r$ for some r with $0 \le r < k$, and since $g^n = 1$ this gives $g^r = (g^k)^{-q} \in H$, forcing $r = 0$.)

Lemma (12.7): *Let G be a group and $x, y \in G$ such that $xy = yx$. Let m be the order of x and n the order of y, and suppose that* $\gcd(m,n) = 1$. *Then the order of xy is mn.*

Proof. Since $xy = yx$ we see that in any product $xyxy\cdots xy$ the x's and y's can be collected together. That is, $(xy)^k = x^k y^k$. Now if $(xy)^k = i_G$ then $x^k = y^{-k}$, and since this element is both a power of x and also a power of y it follows that its order is both a divisor of m and a divisor of n. But $\gcd(m,n) = 1$, and it follows that $x^k = y^{-k} = i_G$, since this is the only element of order 1. But now $x^k = i_G$ yields that $m \mid k$, since the order of x is m, and similarly $y^{-k} = i_G$ yields that $n \mid k$. By Exercise 27 we conclude that k is a multiple of lcm(m,n), which equals mn since $\gcd(m,n) = 1$. Thus we have shown that if $(xy)^k = i_G$ then k is a multiple of mn, and since also $(xy)^{mn} = (x^m)^n (y^n)^m = i_G$ it follows that the order of xy is mn, as required. □

It is a non-obvious fact that if p is a prime number then the multiplicative group of (nonzero elements of) $\mathbb{Z}_p$ is cyclic. Indeed, the multiplicative group of any field which has only finitely many elements is cyclic. For example, the successive powers of 3 in $\mathbb{Z}_{17}$ are 3, 9, 10, 13, 5, 15, 11, 16, 14, 8, 7, 4, 12, 2, 6 and 1, exhausting all nonzero elements of $\mathbb{Z}_{17}$. As a first step towards proving this theorem about finite fields we need the following lemma.

Lemma (12.8): *Let G be an Abelian group, and $x, y \in G$ elements of finite orders m, n respectively. Then G contains an element whose order is* $\operatorname{lcm}(m,n)$.

† "Order" is almost universally used in preference to "period", which is regrettable since "order" is an over-used word in mathematics, and "period" under-used.

Proof. For each prime number p let $a(p)$ be the largest integer such that $p^{a(p)} \mid m$ and let $b(p)$ be the largest integer such that $p^{b(p)} \mid n$. Let $p_1, p_2, \ldots, p_k$ be all the primes p such that $p \mid m$ and $a(p) \geq b(p)$, and let $p_{k+1}, p_{k+2}, \ldots, p_l$ be the primes p such that $a(p) < b(p)$. So, writing $a_i = a(p_i)$ and $b_i = b(p_i)$, we have

$$\begin{aligned} m &= p_1^{a_1} p_2^{a_2} \cdots p_l^{a_l} \\ n &= p_1^{b_1} p_2^{b_2} \cdots p_l^{b_l} \end{aligned}$$

with $a_i \geq b_i$ for $1 \leq i \leq k$ and $a_i < b_i$ for $k < i \leq l$. If we now define

$$\begin{aligned} m_1 &= p_1^{a_1} p_2^{a_2} \cdots p_k^{a_k}, & m_2 &= p_{k+1}^{a_{k+1}} p_{k+2}^{a_{k+2}} \cdots p_l^{a_l}, \\ n_1 &= p_1^{b_1} p_2^{b_2} \cdots p_k^{b_k}, & n_2 &= p_{k+1}^{b_{k+1}} p_{k+2}^{b_{k+2}} \cdots p_l^{b_l}, \end{aligned}$$

then as well as $n = n_1 n_2$ and $m = m_1 m_2$ we have that $m_1 n_2 = \operatorname{lcm}(m, n)$ and that $\gcd(m_1, n_2) = 1$. Since the order of x^{m_2} is m_1 and the order of y^{n_1} is n_2 we conclude by Lemma (12.7) that $x^{m_2} y^{n_1}$ is an element of G of order $m_1 n_2$, as sought. □

Corollary (12.9): *Let G be an Abelian group and let $x_1, x_2, \ldots, x_k$ be elements of G of orders $n_1, n_2, \ldots, n_k$. Then there is an element $g \in G$ whose order is a multiple of n_i, for each i from 1 to k.*

Proof. This is trivial for $k = 1$. Proceeding inductively we may assume that $g' \in G$ has order divisible by all n_i for $1 \leq i \leq k-1$, and now Lemma (12.8) applied to g' and x_k yields an element g whose order is divisible by the order of g' and by n_k. Then the order of g is divisible by all of $n_1, n_2, \ldots, n_k$, as required. □

In particular, this result shows that if G is a finite Abelian group then there is an element $g \in G$ whose order n is a multiple of the order of every other element of G, so that $h^n = i_G$ for all $h \in G$. Suppose now that F is a finite field, with identity element 1_F, and choose such an element g in the multiplicative group of F. Thus, letting n be the order of g, we deduce that every nonzero element of the field is a root of the polynomial $x^n - 1$. Now a polynomial in $F[x]$ of degree n can have at most n roots (since every root t yields a factor $x - t$, and since the degree of a product is the sum of the degrees of the factors there can be at most n factors of degree 1), and so the number of nonzero elements of F is at most n. But on the other hand the elements g^i (for $0 \leq i < n$) are all nonzero, and there are n of them (since they are in bijective correspondence with the elements of $\mathbb{Z}_n$, by Proposition (12.5)). So these are all the nonzero elements of G, and we have proved the following theorem (as foreshadowed).

Theorem (12.10): *The multiplicative group of a finite field is cyclic.*

Regrettably, apart from trial and error, there is no simple method known for finding a generator of the multiplicative group of a given finite field.

The next two exercises on finite fields should really have been included earlier, but they were accidentally forgotten.

Exercise 48. Observe that 2 and 3 are zero divisors in $\mathbb{Z}_6$. Show that if R is an integral domain of characteristic $m \neq 0$ then $\mathbb{Z}_m$ is a subring of R, and deduce that m must be prime.

Exercise 49. Let F be a field. The *prime field* F_0 of F is the smallest subfield of F. Show that if the characteristic of F is zero then F_0 is isomorphic to $\mathbb{Q}$, while if the characteristic is $p \neq 0$ then p is prime and F_0 is isomorphic to $\mathbb{Z}_p$. Deduce that if F is finite then the number of elements in F is p^d for some integer d. (Hint: d is the dimension of F considered as a vector space over F_0.)

Exercise 50. Let F be a field of odd finite order q. Show that there is an element $t \in F$ with $t^2 = -1$ if and only if $q - 1$ is divisible by 4. (Hint: Let g be a generator of the multiplicative group of F, so that $g^{q-1} = 1_F$, and the distinct nonzero elements of F are $1, g, g^2, \ldots, g^{q-2}$. Check that the two roots of $x^2 - 1 = 0$ are 1 and $g^{(q-1)/2}$, and deduce that $g^{(q-1)/2} = -1$. Observe that if $(q-1)/2$ is even then $t = g^{(q-1)/4}$ satisfies $t^2 = -1$, but if $(q-1)/2$ is odd then there is no i such that $2i \equiv (q-1)/2 \pmod{q-1}$.)

Exercise 51. Show that if p is a prime and $p \equiv 3 \pmod 4$ then p cannot be expressed as a sum of two squares. (Hint: If $p = a^2 + b^2$ then $(\overline{a})^2 + (\overline{b})^2 = 0$ in the field $\mathbb{Z}_p$, and hence $t = \overline{a}(\overline{b})^{-1}$ is a solution of $t^2 = -1$, contradicting Exercise 50.)

Exercise 52. Show that if p is a prime and $p \equiv 1 \pmod 4$ then p can be expressed as a sum of two squares. (Hint: By Exercise 50 there is an integer t such that $t^2 + 1 \equiv 0 \pmod p$. Working now in the ring $\mathbb{G}$ of Gaussian integers, observe that p is a divisor of $t^2 + 1 = (t+i)(t-i)$, so that if p were irreducible then, by Theorem (9.18), p would have to be a divisor of either $t + i$ or $t - i$. But neither $\frac{1}{p}t + \frac{1}{p}i$ nor $\frac{1}{p}t - \frac{1}{p}i$ is a Gaussian integer; so p is not irreducible in $\mathbb{G}$, and so (as we saw in our previous discussion of irreducible elements of $\mathbb{G}$) there exist integers a and b with $p = (a + bi)(a - bi)$.)

We return now to our discussion the aspects of group theory which will be of relevance for our introduction to Galois theory. If X is any set then a *permutation* of X is by definition a bijective function from X to itself. We define $\mathrm{Sym}(X)$ to be the set of all permutations of X. If $\sigma, \tau \in \mathrm{Sym}(X)$ then the composite function $\alpha\beta\colon X \to X$ (defined by $(\alpha\beta)x = \alpha(\beta x)$ for all $x \in X$) is also a permutation, and so composition can be regarded as a multiplication operation on $\mathrm{Sym}(X)$. This operation is associative, since if α, β and γ are arbitrary elements of $\mathrm{Sym}(X)$ then $((\alpha\beta)\gamma)x$ and $(\alpha(\beta\gamma))x$ are both equal to $\alpha(\beta(\gamma x))$ (for each $x \in X$). The identity function $i\colon X \to X$ (defined by $ix = x$ for all $x \in X$) is clearly an identity element for this multiplication, and all elements of $\mathrm{Sym}(X)$ have inverses (since every bijective function has an inverse which is also bijective). Hence $\mathrm{Sym}(X)$ is a group. It is known as the *symmetric group* on the set X.

In the case $X = \{1, 2, \ldots, n\}$ we shall for brevity denote the symmetric group on X by S_n. We shall also use the cycle notation for permutations: if $i_1, i_2, \ldots, i_k$ are distinct elements of the set $X = \{1, 2, \ldots, n\}$ then $(i_1, i_2, \ldots, i_k)$ denotes the function $\sigma\colon X \to X$ satisfying

$$\begin{aligned} \sigma i_j &= i_{j+1} \quad \text{for } j = 1, 2, \ldots, k-1, \\ \sigma i_k &= i_1, \end{aligned}$$

and

$$\sigma i = i \quad \text{for all } i \in X \text{ such that } i \notin \{i_1, i_2, \ldots, i_k\}.$$

That is, σ permutes $i_1, i_2, \ldots, i_k$ cyclically and fixes the other elements of X. We call such a permutation a *cycle*, and the set $\{i_1, i_2, \ldots, i_k\}$ is called the *support* of $(i_1, i_2, \ldots, i_k)$. A *k-cycle* is a cycle with k elements in its support. It is easily seen that if $\tau \in \mathrm{Sym}(X)$ is arbitrary then X is a union of disjoint subsets which are each permuted cyclically by τ, and τ is correspondingly a product of cycles with disjoint supports. Thus, for example, the element $\tau \in S_6$ given by

$$1 \mapsto 4,\ 2 \mapsto 6,\ 3 \mapsto 1,\ 4 \mapsto 3,\ 5 \mapsto 5, \text{and } 6 \mapsto 2,$$

is $(1,4,3)(2,6)$. Note that disjoint cycles commute: we can equally well write $\tau = (2,6)(1,4,3)$. Furthermore, within the individual cycles it is only the cyclic order of the terms which matters, and so τ above can also be written as $(6,2)(3,1,4)$.

Exercise 53. Check that the 24 elements of S_4 comprise the identity, six transpositions (2-cycles), eight 3-cycles, six 4-cycles, and three elements which are products of two disjoint transpositions. Similarly check that the elements of S_5 comprise the identity, ten transpositions, twenty 3-cycles, thirty 4-cycles, twenty-four 5-cycles, fifteen elements which are products of two disjoint transpositions, and twenty elements which are products of a 3-cycle and a transposition with disjoint supports.

If G is a group and H a subgroup of G then we can define relations $\sim_R$ and $\sim_L$ on G as follows: for all $x, y \in G$,

$$\begin{aligned} x \sim_R y &\text{ if } x = hy \text{ for some } h \in H,\\ x \sim_L y &\text{ if } x = yh \text{ for some } h \in H. \end{aligned}$$

It is not hard to show that these are both equivalence relations on G. The equivalence classes for $\sim_R$ are called the *right cosets* of H in G, and the equivalence classes for $\sim_L$ are called the left cosets of H in G. We see that the right coset of H that contains the element x of G is the set

$$Hx = \{\, hx \mid h \in H\,\}.$$

By the general properties of equivalence classes, each element of G lies in exactly one of these right cosets; thus if x and y are arbitrary elements of G, then $x \in Hy$ if and only if $Hx = Hy$, and $x \notin Hy$ if and only if $Hx \cap Hy = \emptyset$. The analogous properties hold for left cosets: the left coset containing x is

$$xH = \{\, xh \mid h \in H\,\},$$

and for all $x, y \in G$ we have $xH = yH$ if and only if $x \in yH$, and $xH \cap yH = \emptyset$ if and only if $x \notin yH$. Note that because multiplication in G need not be commutative, the left coset xH and the right coset Hx will not in general be equal.

Consider for example the group $G = S_4$, consisting of all permutations of the set $\{1, 2, 3, 4\}$. Let $H = \{\, \sigma \in G \mid \sigma 4 = 4\,\}$. Then H is a subgroup of G. It is clear that H is isomorphic to the group S_3, since the restriction to $\{1, 2, 3\}$ of a permutation of $\{1, 2, 3, 4\}$ that fixes 4 is a permutation of $\{1, 2, 3\}$, and this gives us a function from H to S_3 which is easily shown to be bijective and preserve composition. We find that

$$\begin{aligned} iH = H &= \{\, i,\ (1,2),\ (1,3),\ (2,3),\ (1,2,3),\ (1,3,2)\,\}\\ (1,4)H &= \{\, (1,4),\ (1,2,4),\ (1,3,4),\ (1,4)(2,3),\ (1,2,3,4),\ (1,3,2,4)\,\}\\ (2,4)H &= \{\, (2,4),\ (1,4,2),\ (2,4)(1,3),\ (2,3,4),\ (1,4,2,3),\ (1,3,4,2)\,\}\\ (3,4)H &= \{\, (3,4),\ (3,4)(1,2),\ (1,4,3),\ (2,4,3),\ (1,2,4,3),\ (1,4,3,2)\,\}, \end{aligned}$$

and each element of G occurs in exactly one of these. The elements i, $(1,4)$, $(2,4)$, $(3,4)$ constitute a *left transversal*, or *system of left coset representatives* for the subgroup H of G, in the sense that

$$G = iH \mathbin{\dot\cup} (1,4)H \mathbin{\dot\cup} (2,4)H \mathbin{\dot\cup} (3,4)H$$

(where the symbol $\dot\cup$ indicates a union of disjoint sets). Of course there are many other left transversals, since any element of a coset can be chosen as the representative of that coset.

It is a general fact that a right transversal for a subgroup H of G can be obtained form a left transversal for H by taking the inverses of the elements. In the above example, by a coincidence, the elements i, $(1,4)$, $(2,4)$, $(3,4)$ of our chosen left transversal are all self-inverse, and so they also form a right transversal. This is easily checked, with a little calculation. We find that

$$\begin{aligned} Hi = H &= \{\, i,\ (1,2),\ (1,3),\ (2,3),\ (1,3,2),\ (1,2,3)\,\}\\ H(1,4) &= \{\, (1,4),\ (1,4,2),\ (1,4,3),\ (2,3)(1,4),\ (1,4,3,2),\ (1,4,2,3)\,\}\\ H(2,4) &= \{\, (2,4),\ (1,2,4),\ (1,3)(2,4),\ (2,4,3),\ (1,3,2,4),\ (1,2,4,3)\,\}\\ H(3,4) &= \{\, (3,4),\ (1,2)(3,4),\ (1,3,4),\ (2,3,4),\ (1,3,4,2),\ (1,2,3,4)\,\}. \end{aligned}$$

Note that the right cosets are not the same as the left cosets, and that many left transversals are not also right transversals. (For example, i, $(1,2,4)$, $(1,4,2)$, $(1,4,3)$ is a left transversal but not a right transversal).

Proposition (12.11): *Let H be a subgroup of the group G and let $x \in G$ be arbitrary. Then there is a bijection $F\colon H \to Hx$ given by $Fh = hx$ for all $h \in H$. Similarly, $h \mapsto xh$ is a bijection $H \to xH$.*

Proof. By definition every element of Hx has the form $hx = Fh$ for some $h \in H$; so F is surjective. And if h, $h' \in H$ with $Fh = Fh'$ then

$$h = hi_G = h(xx^{-1}) = (hx)x^{-1} = (Fh)x^{-1} = (Fh')x^{-1} = (h'x)x^{-1} = h'(xx^{-1}) = h'i = h',$$

so that F is also injective. □

Assume now that H is a subgroup of G and $|G| < \infty$. Then

$$G = x_1 H \,\dot{\cup}\, x_2 H \,\dot{\cup} \cdots \dot{\cup}\, x_n H,$$

where $x_1, x_2, \ldots, x_n$ is any left transversal for H, and we see that

$$|G| = |x_1 H| + |x_2 H| + \cdots + |x_n H|.$$

But Proposition (12.11) tells us that $|x_i H| = |H|$ for all i, and so we conclude that $|G| = n|H|$, where n is the number of left cosets of H in G. Exactly the same reasoning shows that $|G| = n'|H|$ where n' is the number of right cosets of H in G; so we conclude that $n = n'$. (One can also readily check that the mapping $x \to x^{-1}$ from G to G induces a bijection from the set of left cosets of H in G to the set of right cosets of H in G.)

Definition (12.12): Let H be a subgroup of the group G. The number of left cosets of H in G (which equals the number of right cosets of H in G) is called the *index* of H in G. The index is denoted by $[G : H]$.

Our discussion above proved the following proposition.

Proposition (12.13): *If H is a subgroup of G then $|G| = [G : H]|H|$. In particular, if $|G| < \infty$ then $|H|$ and $[G : H]$ are both divisors of $|G|$.*

We remark in passing that it is possible to define arithmetic of infinite numbers so that Proposition (12.13) remains true when G is infinite.

Definition (12.14): Let G_1 and G_2 be groups and $\phi\colon G_1 \to G_2$ a homomorphism. The *kernel* of ϕ is the set

$$\ker\phi = \{\, x \in G_1 \mid \phi x = i_{G_2} \,\}.$$

Just as kernels of ring homomorphisms are subrings, so kernels of group homomorphisms are subgroups. And just as kernels of ring homomorphisms possess an extra closure property which is not possessed by all subrings—kernels of ring homomorphisms are ideals, rather than merely subgroups—so kernels of group homomorphisms possess an extra property, not possessed by all subgroups. They are normal subgroups, in the sense of the following definition.

Definition (12.15): A subgroup K of the group G is said to be *normal* in G if $x^{-1}kx \in K$ for all $k \in K$ and $x \in G$.

Equivalently, the subgroup K is normal in G if $x^{-1}Kx = K$ for all $x \in G$, where of course by definition $x^{-1}Kx = \{\, x^{-1}kx \mid k \in K \,\}$. Equivalently also, K is normal if $xK = Kx$ for all $x \in G$.

Exercise 54. Prove that the kernel of a group homomorphism $G_1 \to G_2$ is a normal subgroup of G_1, and its image is a subgroup of G_2.

If S and T are arbitrary subsets of the group G then it is natural to define the product, ST, by

$$ST = \{\, st \mid s \in S \text{ and } t \in T\,\}.$$

It is easily checked that this multiplication of subsets is associative (but not usually commutative, of course). If H is a subgroup of G then the fact that H is closed under multiplication and contains the identity element makes it easy to show that $HH = H$. Suppose now that K is a normal subgroup of the group G, and let x, $y \in G$ be arbitrary. Then since $y^{-1}Ky = K$ we find that

$$(xK)(yK) = xy(y^{-1}Ky)K = xyKK = (xy)K.$$

So multiplication of subsets of G defined above yields an operation on the set $\{\, xK \mid x \in G\,\}$ of cosets in G of the normal subgroup K. Note that left cosets and right cosets of K coincide, since K is normal.

Theorem (12.16): *Let K be a normal subgroup of the group G. Then the set of cosets of K in G forms a group, with multiplication satisfying $(xK)(yK) = xyK$ for all x, $y \in G$.*

Proof. Our discussion above showed that there is a well-defined associative operation on the set of cosets satisfying $(xK)(yK) = xyK$. An arbitrary coset α has the form xK for some $x \in G$, and so

$$\alpha(i_G K) = (xK)(i_G K) = (xi_G)K = xK = (i_G x)K = (i_G K)(xK) = (i_G K)\alpha,$$

which shows that the coset $i_G K = K$ is an identity element. And if $\alpha = xK$ is an arbitrary coset, then $x^{-1}K$ is an inverse for α, since

$$(xK)(x^{-1}K) = (xx^{-1})K = i_G K = (x^{-1}x)K = (x^{-1}K(xK).$$

Thus all the group axioms are satisfied. □

The group of cosets of a normal subgroup K of G is called the *quotient group* (of G modulo K), and it is denoted by G/K. It is immediate from the definition of multiplication in G/K that $x \mapsto xK$ is a homomorphism from G to G/K. We leave it to the reader to check the truth the next result, which is the First Isomorphism Theorem for groups.

Theorem (12.17): *Let $\phi\colon G \to H$ be a group homomorphism. Then $K = \ker\phi$ is a normal subgroup of G, and there is an isomorphism $\psi\colon G/K \to \operatorname{im}\phi$ such that $\psi(xK) = \phi x$ for all $x \in G$.*

We shall show in the next section how a group can be associated to a polynomial equation. The group is known as the *Galois group* of the equation. One can then give a group theoretic criterion for a polynomial equation to be soluble by radicals. Specifically, an equation is soluble by radicals if and only if its Galois group is a soluble group, in the sense of the following definition.

Definition (12.18): A group G is said to be *soluble* if there is a chain of subgroups

$$G = G_0 \geq G_1 \geq G_2 \geq \cdots \geq G_n = \{i_G\}$$

such that G_i is a normal subgroup of G_{i-1} and the quotient group G_i/G_{i-1} is Abelian, for each $i = 1, 2, \ldots, k$.

Our strategy for showing that not all equations are soluble by radicals is to show that there exists an equation whose Galois group is the symmetric group S_5, and to show that S_5 is not soluble. The remainder of this section is devoted an investigation of S_5, including a proof that it is not soluble.

Let n be any positive integer and let $\sigma \in S_n$. Recall that σ can be written as a product of disjoint cycles; let $c_i(\sigma)$ be the number of i-cycles occurring in this expression for σ. For the purposes of this definition it is intended that a 1-cycle (k) should appear in the disjoint cycle expression for σ for every $k \in \{1, 2, \ldots, n\}$ such that $\sigma k = k$, so that $c_1(\sigma)$ is the number of fixed points of σ. Thus the element $\sigma \in S_{10}$ which would normally be written as $(7, 9)(2, 4, 6)(3, 5, 10)$ becomes $(1)(8)(7, 9)(2, 4, 6)(3, 5, 10)$ when the 1-cycles are explicitly written in, and we see that $c_1(\sigma) = 2$, $c_2(\sigma) = 1$ and $c_3(\sigma) = 2$, while $c_i(\sigma) = 0$ for all $i > 3$. We call the n-tuple $(c_1(\sigma), c_2(\sigma), \ldots, c_n(\sigma))$ the *cycle type* of the element σ of S_n.

Proposition (12.19): *Elements σ, $\tau \in S_n$ have the same cycle type if and only if there exists $\rho \in S_n$ with $\tau = \rho\sigma\rho^{-1}$.*

Before proving this, we give an example which should clarify things somewhat. Consider the element $\sigma = (2, 5)(3, 4, 7, 6) \in S_7$. According to Proposition (12.19), if $\rho \in S_7$ is arbitrary then $\rho\sigma\rho^{-1}$ should have the same cycle type as σ; that is, it should be the product of a 2-cycle and a 4-cycle with disjoint supports. In fact, it can readily be seen that $\rho\sigma\rho^{-1} = (\rho 2, \rho 5)(\rho 3, \rho 4, \rho 7, \rho 6)$. For example,

$$(\rho\sigma\rho^{-1})(\rho 2) = \rho(\sigma(\rho^{-1}(\rho 2))) = \rho(\sigma 2) = \rho 5,$$

and

$$(\rho\sigma\rho^{-1})(\rho 5) = \rho(\sigma(\rho^{-1}(\rho 5))) = \rho(\sigma 5) = \rho 2.$$

Proposition (12.19) also says, conversely, that given any $\tau \in S_7$ with the same cycle type as σ—such as $\tau = (1, 4)(3, 2, 5, 6)$ for example—there is a $\rho \in S_7$ such that $\rho\sigma\rho^{-1} = \tau$. We can find a suitable ρ as follows. First write down σ, including the 1-cycles, and underneath it write down τ in such a way that each cycle of τ is below a cycle of σ of the same length:

$$\begin{gathered}(1)(2, 5)(3, 4, 7, 6,)\\ (7)(1, 4)(3, 2, 5, 6,).\end{gathered}$$

Then define ρ to be the permutation which maps each number to the one written directly below it. Thus, in this example, $\rho 1 = 7$, $\rho 2 = 1$, $\rho 5 = 4$, $\rho 3 = 3$, $\rho 4 = 2$, $\rho 7 = 5$ and $\rho 6 = 6$. We see that

$$\rho\sigma\rho^{-1} = (\rho 2, \rho 5)(\rho 3, \rho 4, \rho 7, \rho 6) = (1, 4)(3, 2, 5, 6) = \tau,$$

as required.

Proof of (12.19). Let σ_1, σ_2, …, σ_k be all the cycles (including those of length 1) that appear in the expression for σ as a product of disjoint cycles. Suppose that σ_i is an n_i-cycle (so that $n = sum_{i=1}^{k} n_i$), and let $\sigma_i = (a_{i1}, a_{i2}, \ldots, a_{in_i})$. If $\rho \in S_n$ is arbitrary, then for each i we have that the permutation $\tau_i = \rho\sigma_i\rho^{-1}$ is also an n_i-cycle; specifically,

$$\tau_i = (\rho a_{i1}, \rho a_{i2}, \ldots, \rho a_{in_i}),$$

since for each $j < n_i$ we have

$$\tau_i(\rho a_{ij}) = (\rho\sigma_i\rho^{-1})(\rho a_{ij}) = \rho(\sigma_i a_{ij}) = \rho a_{i\,j+1},$$

and similarly $\tau_i(\rho a_{in_i}) = \rho a_{i1}$. Furthermore, the τ_i have pairwise disjoint supports, since the permutation ρ maps the support of σ_i to the support of τ_i, and the supports of the σ_i are pairwise disjoint. Thus it follows that τ_1, τ_2, ..., τ_k are the cycles that appear in the expression for $\tau = \rho\sigma\rho^{-1}$ as a product of disjoint cycles, and so τ has the same cycle type as σ.

Conversely, suppose that τ is any element of S_n with the same cycle type as σ. Then we can write $\tau = \tau_1\tau_2\cdots\tau_k$, where τ_i is an n_i-cycle, and the supports of the τ_i are pairwise disjoint. Then we have

$$\tau_i = (b_{i1}, b_{i2}, \ldots, b_{in_i})$$

for some b_{ij}, and furthermore

$$\{1, 2, \ldots, n\} = \bigcup_{i=1}^{k}\{a_{i1}, a_{i2}, \ldots, a_{in_i}\} = \bigcup_{i=1}^{k}\{b_{i1}, b_{i2}, \ldots, b_{in_i}\}.$$

It can readily be seen that there is a permutation ρ of $\{1, 2, \ldots, n\}$ such that $\rho a_{ij} = b_{ij}$ for all i and j, and since this gives

$$\tau_i = (\rho a_{i1}, \rho a_{i2}, \ldots, \rho_{in_i}) = \rho\sigma_i\rho^{-1},$$

we conclude that

$$\tau = \tau_1\tau_2\cdots\tau_k = (\rho\sigma_1\rho^{-1})(\rho\sigma_2\rho^{-1})\cdots(\rho\sigma_k\rho^{-1}) = \rho\sigma_1\sigma_2\cdots\sigma_k\rho^{-1} = \rho\sigma\rho^{-1},$$

as required. □

Recall that, by the definition, if a normal subgroup K of a group G contains an element g then it also contains $x^{-1}gx$ for all $x \in G$. Hence Proposition (12.19) shows that if K is a normal subgroup of S_n and if $\sigma \in K$ then all permutations of the same cycle type as σ are also contained in K. This makes it a relatively easy task to determine all the normal subgroups of S_n, for all values of n. However, we shall content ourselves with doing so for the case $n = 5$, since this is all that is needed for this course.

A subset of S_n consisting of all elements of a given cycle type is called a *conjugacy class*, or simply *class*, of S_n. In S_5 there are exactly seven classes, and for ease of reference we label them (a) to (g) in accordance with the following table.

Class	Size	Description	Sample element
(a)	1	Five 1-cycles	Identity
(b)	10	Three 1-cycles and one 2-cycle	$(1, 2)$
(c)	20	Two 1-cycles and one 3-cycle	$(1, 2, 3)$
(d)	30	One 1-cycle and one 4-cycle	$(1, 2, 3, 4)$
(e)	24	One 5-cycle	$(1, 2, 3, 4, 5)$
(f)	15	One 1-cycle and two 2-cycle	$(1, 2)(3, 4)$
(g)	20	One 2-cycle and one 3-cycle	$(1, 2)(3, 4, 5)$

Suppose now that K is a normal subgroup of S_5. Suppose first of all that K contains an element of Class (b) (the transpositions). Then since K is normal it must contain all elements Class (b), and since K is closed under multiplication it follows that it contains all elements which can be expressed as products of elements of Class (b). But short calculations yield

$$\begin{aligned}(1, 2, 3) &= (1, 2)(2, 3)\\ (1, 2, 3, 4) &= (1, 2)(2, 3)(3, 4)\\ (1, 2, 3, 4, 5) &= (1, 2)(2, 3)(3, 4)(4, 5)\\ (1, 2)(3, 4, 5) &= (1, 2)(3, 4)(4, 5);\end{aligned}$$

so K contains $(1,2,3)$, $(1,2,3,4)$, $(1,2,3,4,5)$ and $(1,2)(3,4,5)$, and hence all elements of Classes (c), (d), (e) and (g). It also contains the elements of Class (f), since trivially these are products of transpositions, as well as the identity element (which is in every subgroup). We conclude that every element of S_5 is in K.

Proposition (12.20): *The only normal subgroup of S_5 that contains a transposition is the group S_5 itself.*

Suppose next that K is a normal subgroup which contains an element of Class (c) (the 3-cycles). Then K contains all 3-cycles, and hence contains $(1,2,3)(3,4,5) = (1,2,3,4,5)$ and $(1,2,3)(2,3,4) = (1,2)(3,4)$. So K must contain all elements of Classes (e) and (f), as well as Class (c) and the identity. Note that these four classes together have sixty elements, which is exactly half the total number of elements of S_5. It is in fact true that these sixty elements constitute a subgroup of S_5, known as the *alternating group* of degree 5, and denoted by A_5. We leave the proof that A_5 is a subgroup for later. We now come to our main theorem concerning the group S_5.

Theorem (12.21): *The group S_5 has no normal subgroups other than $\{i\}$, A_5 and S_5.*

Proof. Let K be a normal subgroup of S_5, and suppose that $K \neq \{i\}$. It suffices to prove that K contains all the elements of Classes (c),(e) and (f), for then K either consists of these 59 elements together with i, and thus equals A_5, or else contains more than 60 elements. But $|K|$ must be a divisor of $120 = |G|$ (by Proposition (12.13)); so $|K| > 60$ forces $|K| = 120$, so that $K = S_5$.

We have already seen in our discussion above that if K contains a transposition or a 3-cycle then K must include all of A_5. If K contains a 4-cycle then it must also contain a 3-cycle, since $(1,2,3,4)(5,4,3,2) = (1,2,5)$. If K contains a 5-cycle then it must also contain a 3-cycle, since $(1,2,3,4,5)(1,5,4,2,3) = (2,4,3)$. If K contains an element of Class (f) then it contains a 3-cycle since $((1,2)(3,4))((1,2)(4,5)) = (3,4,5)$. And if K contains an element of Class (g) then it contains a 3-cycle, since $((1,2)(3,4,5))^2 = (3,5,4)$. So in all cases K contains a 3-cycle, and hence, by our earlier arguments, includes all of A_5, as required. □

Before we can complete the proof that S_5 is not soluble, we need some general information concerning normal subgroups such that the corresponding quotient group is Abelian. We define the *commutator* of two group elements x and y to be the element $x^{-1}y^{-1}xy$. The commutator of x and y is commonly denoted by $[x,y]$. The *commutator subgroup*, or *derived group*, of a group G is defined to be the smallest subgroup of G which contains all commutators. By Exercise 46 this is the intersection of all those subgroups of G which contain all the commutators.

Theorem (12.22): *Let G be a group and G' the derived group of G. If H is any subgroup of G such that $G' \leq H$ then H is normal in G and G/H is Abelian. Conversely, if H is any normal subgroup of G such that G/H is abelian then $G' \leq H$.*

Proof. Suppose that H is a subgroup of G containing G'. Then for all $h \in H$ and $g \in G$ we have that $h^{-1}g^{-1}hg = [h,g] \in G' \leq H$, and so (by closure of H under multiplication) $g^{-1}hg = h(h^{-1}g^{-1}hg) \in H$. Since this holds for all $h \in H$ and $g \in G$ it follows that H is normal in G. Now for all x, $y \in G$ we have that $(x^{-1}y^{-1}xy)H = H$ (since $x^{-1}y^{-1}xy = [x,y] \in G' \leq H$), and so

$$(yH)(xH) = yxH = yx(x^{-1}y^{-1}xyH) = xyH = (xH)(yH),$$

and hence G/H is Abelian.

Conversely, suppose that H is normal in G and G/H is abelian. Then for all x, $y \in G$ we have that $(xH)(yH) = (yH)(xH)$, and so

$$(x^{-1}y^{-1}xy)H = (xH)^{-1}(yH)^{-1}(xH)(yH) = H,$$

which shows that $x^{-1}y^{-1}xy \in H$. So H is a subgroup which contains all the commutators, and hence contains G' (which is the intersection of all subgroups that contain all the commutators). □

Observe that the derived group G' of G can be described explicitly as the set of all elements of G that can be expressed as products of commutators (allowing any number of factors). That is, if we define

$$S = \{\, [x_1, y_1][x_2, y_2] \cdots [x_n, y_n] \mid 0 \le n \in \mathbb{Z} \text{ and } x_i,\ y_i \in G \text{ for all } i \in \{1, 2, \ldots, i\} \,\}$$

then $G' = S$. To prove this, observe first that since G' contains all commutators and is closed under multiplication, it must contain all products $[x_1, y_1][x_2, y_2] \cdots [x_n, y_n]$. So $S \subseteq G'$. The reverse inclusion will follow once we have proved that S is a subgroup of G, since S certainly contains all commutators (for these are elements of the form $[x_1, y_1][x_2, y_2] \cdots [x_n, y_n]$ with $n = 1$). Clearly $S \neq \emptyset$; indeed $i_G \in S$ since by the standard convention for empty products, $[x_1, y_1][x_2, y_2] \cdots [x_n, y_n] = i_G$ when $n = 0$. It is trivial that S is closed under multiplication, since concatenating two products of commutators gives a (longer) product of commutators. Finally, the inverse of an element of S is an element of S, since the inverse of $[x_1, y_1][x_2, y_2] \cdots [x_n, y_n]$ is $[y_n, x_n] \cdots [y_2, x_2][y_1, x_1]$. (Recall the general fact that $(xy)^{-1} = y^{-1}x^{-1}$, whence $[x, y]^{-1} = (x^{-1}y^{-1}xy)^{-1} = y^{-1}x^{-1}yx = [y, x]$.)

Proposition (12.23): *If H is a normal subgroup of the group G then the derived group H' of H is also a normal subgroup of G.*

Proof. Let $g \in G$ and $t \in H'$ be arbitrary. Then we have

$$t = [x_1, y_1][x_2, y_2] \cdots [x_n, y_n]$$

for some integer $n \ge 0$ and some elements $x_i,\ y_i \in H$. Now

$$g^{-1}tg = (g^{-1}[x_1, y_1]g)(g^{-1}[x_2, y_2]g) \cdots (g^{-1}[x_n, y_n]g),$$

since all the internal g's and g^{-1}'s cancel out. Similarly,

$$\begin{aligned} g^{-1}[x, y]g &= (g^{-1}x^{-1}g)(g^{-1}y^{-1}g)(g^{-1}xg)(g^{-1}yg) \\ &= (g^{-1}xg)^{-1}(g^{-1}yg)^{-1}(g^{-1}xg)(g^{-1}yg) \\ &= [g^{-1}xg, g^{-1}yg] \end{aligned}$$

for any x and y; so if we define $x_i' = g^{-1}x_ig$ and $y_i' = g^{-1}y_ig$ then we have that

$$g^{-1}tg = [x_1', y_1'][x_2', y_2'] \cdots [x_n', y_n'],$$

which is in H' since the fact that H is normal in G yields that $x_i',\ y_i' \in H$ for all i. Thus we have shown that $g^{-1}tg \in H'$ whenever $t \in H'$ and $g \in G$, as required. □

Proposition (12.24): *The group S_5 is not soluble.*

Proof. Suppose to the contrary that it is soluble, so that there is a descending chain of subgroups $S_5 = G_0 > G_1 > \cdots > G_n = \{i\}$, each normal in the preceding term in the chain, and such that G_{i-1}/G_i is Abelian for all i. Since G_1 is normal in S_5 it follows from Theorem (12.21) that G_1 is either A_5 or $\{i\}$. But since G_0/G_1 Abelian it follows that G_1 contains the derived group of G_0, and hence contains all commutators $[\sigma, \tau]$ for $\sigma,\ \tau \in G_0 = S_5$. Since, for example, $[(1, 2), (2, 3)] \neq i$, we conclude that $G_1 \neq \{i\}$. Hence $G_1 = A_5$. By Proposition (12.23) the derived group G_1' of $G_1 = A_5$ is a normal subgroup of S_5 contained in A_5. By Proposition (12.21) it follows that $G_1' = \{i\}$ or A_5. But since $(1, 2, 3)$, $(2, 3, 4) \in G_1$ and $[(1, 2, 3), (2, 3, 4)] \neq i$ it follows that $G_1' \neq \{i\}$, and so we conclude that $G_1' = A_5 = G_1$. However, since G_1/G_2 is Abelian we must have $G_2 \ge G_1'$, and this contradicts the fact that every nonidentity term in the series $S_5 = G_0 > G_1 > \cdots > G_n = \{i\}$ is strictly larger than the following term. This contradiction shows that our original assumption that S_5 is soluble must have been false. □

13. Introduction to Galois theory

In this final section we shall prove that there are equations of degree 5 over the field Q which are not soluble by radicals. In accordance with Definition (11.3) and the subsequent discussion, we need to investigate field extensions. As a first step we have the following proposition, which is really just a Corollary of Theorem (10.5).

Theorem (13.1): *If F is a field and $f(x) \in F[x]$ a nonconstant polynomial, then there exists an extension of F in which $f(x)$ has a root.*

Proof. Let $p(x)$ be any irreducible factor of $f(x)$. By Theorem (10.5) (or Corollary (10.6)) $E = F[x]/p(x)F[x]$ is a field which can be regarded as an extension of F, and $p(x)$ has a root in E. Hence $f(x)$ has a root in E (since $p(x) \mid f(x)$). □

If $\alpha \in E$, where E is an extension of F, then as in Exercise 38 we shall write $F(\alpha)$ for the extension of F generated by α. If α is algebraic over F we shall write $m_{\alpha,F}(x)$ for the minimal polynomial of α over F. Recall that $m_{\alpha,F}(x)$ is irreducible, and $F(\alpha) \cong F[x]/m_{\alpha,F}(x)F[x]$.

As a corollary of Theorem (13.1) we obtain the following result.

Corollary (13.2): *If F is a field and $f(x) \in F[x]$ a nonconstant polynomial then there is a field E that is an extension of F and has the property that $f(x)$ splits into linear factors in $E[x]$. That is, $f(x) = c(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_d)$ for some $\alpha_i \in E$, where c is the leading coefficient of $f(x)$ and d its degree.*

Proof. Choose an extension of F in which $f(x)$ has a root α_1, and let $E_1 = F(\alpha_1)$. Since α_1 is a root of $f(x)$ in $E_1[x]$ we have $f(x) = (x - \alpha_1)f_1(x)$ for some $f_1(x) \in E_1[x]$ (by Theorem (9.3)), where the degree of $f_1(x)$ is $\deg(f(x)) - 1$. If $f_1(x)$ is not constant we may repeat the argument with $f_1(x)$ in place of $f(x)$ and E_1 in place of F, and obtain an extension $E_2 = E_1(\alpha_2)$ such that $f_1(x) = (x - \alpha_2)f_2(x)$, and hence $f(x) = (x - \alpha_1)(x - \alpha_2)f_2(x)$ for some $f_2(x) \in E_2[x]$. If $f_2(x)$ is not constant we repeat the argument again, and so on, constructing in this way an increasing chain of fields

$$F = E_0 \subseteq E_1 = E_0(\alpha_1) \subseteq E_2 = E_1(\alpha_2) \subseteq \cdots \subseteq E_d = E_{d-1}(\alpha_d) = E$$

such that, in $E[x]$,

$$f(x) = (x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_d)c(x)$$

with $\deg(c(x)) = 0$. Comparing the degrees and leading coefficients on either side of this equation yields $c(x) = c$, the leading coefficient of $f(x)$, and $d = \deg(f(x))$. □

Definition (13.3): If $f(x) \in F[x]$ (where F is a field) then a field K which is an extension of F is called a *splitting field* for $f(x)$ if

(i) $f(x)$ splits into linear factors over K, and

(ii) if L is any subfield of K with $F \subseteq L \subsetneq K$ then $f(x)$ does not split into linear factors over L.

Corollary (13.2) shows that splitting fields always exist; indeed, the field E constructed in the proof of (13.2) is a splitting field for the polynomial $f(x)$. This can be seen as follows. Suppose that L is any field with $F \subseteq L \subseteq E$, and such that in $L[x]$ there is a factorization

$$f(x) = (a_1x - b_1)(a_2x - b_2)\cdots(a_dx - b_d)$$

expressing $f(x)$ as a product of factors of degree 1. Then by uniqueness of factorization in $E[x]$ the elements $b_ia_i^{-1} \in L$ (for $1 \le i \le d$) coincide with the roots α_i (in some order). So L contains

$F = E_0$ and all the α_i, and we successively deduce that L contains $E_1 = E_0(\alpha_1)$, and $E_2 = E_1(\alpha_2)$, and so on, giving finally that L contains $E_d = E$. So there is no field L with $F \subseteq L \subsetneq E$ over which $f(x)$ splits into linear factors.

The above discussion shows that if E is a splitting field for the polynomial $f(x) \in F[x]$ then E contains a complete set of roots of $f(x)$, and, moreover, E is the extension of F generated by the roots. Note that the degree $[E : F]$ must be finite, since, in the notation used above,

$$[E : F] = [E_d : E_0] = [E_d : E_{d-1}][E_{d-1} : E_{d-2}] \cdots [E_1 : E_0]$$

and $[E_i : E_{i-1}] = [E_{i-1}(\alpha_i) : E_{i-1}]$ is finite for each i by Theorem (10.20).

The following lemma is a triviality, but certainly important.

Lemma (13.4): *Suppose that F and K are fields and $\theta\colon F \to K$ an isomorphism. Then there is an isomorphism $F[x] \to K[x]$ given by $f(x) \mapsto (\theta f)(x)$ for all $f(x) \in F[x]$, where if $f(x) = \sum_i a_i x^i$ then by definition $(\theta f)(x) = \sum_i (\theta a_i) x^i$.*

In fact, this result was included in Exercise 7. We shall refer to the map $f(x) \mapsto (\theta f)(x)$ from $F[x]$ to $K[x]$ as the map *induced* by θ.

Of course we often like to regard isomorphic fields as being equal, and since identifying F with K would surely mean identifying $F[x]$ with $K[x]$, it is just as well that Lemma (13.4) holds! (However, we shall also encounter situations in which we shall not want to identify the isomorphic fields F and K.)

The next theorem is of fundamental importance to our cause, and to highlight this fact we give it a name as well as a number: we shall call it the **Isomorphism Extension Theorem**.

Theorem (13.5): *Suppose that $\theta\colon F \to K$ is an isomorphism of fields, and let $F' = F(\alpha)$ and $K' = K(\beta)$ be extensions of F and K. Let $a(x) \in F[x]$ be the minimal polynomial of α over F, and $b(x) \in K[x]$ the minimal polynomial of β over K. Then if $b(x) = (\theta a)(x)$ the isomorphism $\theta\colon F \to K$ extends to an isomorphism $F' \to K'$ such that $\alpha \mapsto \beta$.*

Proof. Assume $b(x) = (\theta a)(x)$. Then the isomorphism $F[x] \to K[x]$ (given by $f(x) \mapsto (\theta f)(x)$ for all $f(x) \in F[x]$) takes the ideal $I = a(x)F[x]$ to $J = (\theta a)(x)K[x] = b(x)K[x]$. Hence

$$f(x) + I \mapsto (\theta f)(x) + J \qquad \text{(for all } f(x) \in F[x]\text{)}$$

defines an isomorphism $F[x]/I \to K[x]/J$. But by Theorem (10.15) there is an isomorphism $F[x]/I \to F(\alpha)$ such that

$$f(x) + I \mapsto f(\alpha)$$

for all $f(x) \in F[x]$, and an isomorphism $K[x]/J \to K(\beta)$ such that

$$g(x) + J \mapsto g(\beta)$$

for all $g(x) \in K[x]$. Combining these isomorphisms gives

$$\begin{array}{ccccccc} F' = F(\alpha) & \cong & F[x]/I & \cong & K[x]/J & \cong & K(\beta) = K' \\ f(\alpha) & \mapsto & f(x) + I & \mapsto & (\theta f)(x) + J & \mapsto & (\theta f)(\beta). \end{array}$$

That is, there is an isomorphism $F' \to K'$ such that $f(\alpha) \mapsto (\theta f)(\beta)$ for all $f(x) \in F[x]$. In particular, applying this with $f(x) = x$ gives $\alpha \mapsto \beta$. □

The following diagram illustrates the situation that Theorem (13.5) deals with:

$$\begin{tikzcd}
a(x) = m_{\alpha,F}(x) \in F[x] \arrow[d, mapsto] & F \arrow[r, hook] \arrow[d, "\theta"'] & F' = F(\alpha) \arrow[d, dotted] & \alpha \arrow[d, mapsto, dotted] \\
b(x) = m_{\beta,K}(x) \in K[x] & K \arrow[r, hook] & K' = K(\beta) & \beta
\end{tikzcd}$$

Assuming that θ is an isomorphism $F \to K$ which carries the minimal polynomial of α to the minimal polynomial of β, the conclusion of the theorem is that the broken arrows in the diagram above can be completed so that the resulting square commutes (meaning that the two routes from F to K' give the same function). Moreover, the map from $F' \to K'$ which completes the square is an isomorphism mapping α to β.

The horizontal arrows in the above diagram are drawn with hooks on the left hand end to indicate that the corresponding mappings are embeddings: the point is that $\hookrightarrow$ resembles to some extent a cross between an arrow and a subset sign.

Just as iteration of Theorem (13.1) gave us Corollary (13.2), so Theorem (13.5) yields, on iteration, the following corollary, which tells us that splitting fields are essentially unique. Since this result is also rather important, we call it a theorem rather than a corollary.

Theorem (13.6): *Let $\theta\colon F \to K$ be an isomorphism and $f(x) \in F[x]$ a polynomial. Let E be a splitting field for $f(x)$ over F and L a splitting field for $(\theta f)(x)$ over K. Then θ extends to an isomorphism $E \to L$.*

Proof. We use induction on $[E : F]$, the degree of the extension. This is a finite number because, as we observed previously, E is the extension of F generated by a finite number of algebraic elements (namely, the roots of $f(x)$). In the case $[E : F] = 1$ we have $E = F$, which implies that $f(x)$ splits into linear factors in $F[x]$. Hence, applying the isomorphism $F[x] \to K[x]$ induced by θ (see Lemma (13.4)) we deduce that $(\theta f)(x)$ splits into linear factors over K. Thus $L = K$, and the desired isomorphism $E \to L$ extending θ is just θ itself.

Now suppose that $[E : F] > 1$. Let $\alpha \in E$ be any root of $f(x)$ which is not in F, and let $F' = F(\alpha)$. Let $p(x) = m_{\alpha,F}(x) \in F[x]$, the minimal polynomial of α over F. Since $f(\alpha) = 0$ it follows by Proposition (10.12) that $p(x) \mid f(x)$, and applying the isomorphism $F[x] \to K[x]$ induced by θ we deduce that $(\theta p)(x) \mid (\theta f)(x)$. Now since $(\theta f)(x)$ splits into linear factors over L it follows that $(\theta p)(x)$ also splits into linear factors over L; thus

$$(\theta p)(x) = (x - \beta_1)(x - \beta_2) \cdots (x - \beta_k)$$

for some $\beta_i \in L$. Since $p(x)$ is an irreducible element of $F[x]$ (being the minimal polynomial of an element) it follows that $(\theta p)(x) \in K[x]$ must also be irreducible, as any nontrivial factorization of $(\theta p)(x)$ would yield (upon application of the isomorphism $K[x] \to F[x]$ induced by θ^{-1}) a nontrivial factorization of $p(x)$. We deduce that $(\theta p)(x)$ must be the minimal polynomial of β_1 over L (by Proposition (10.14)). Now by the Isomorphism Extension Theorem, the isomorphism

$$\theta\colon F \to K$$

extends to an isomorphism

$$\theta'\colon F' = F(\alpha) \to K' = K(\beta_1).$$

We have now achieved a partial extension of θ. Diagrammatically, we have

$$\begin{tikzcd}
F \arrow[r, hook] \arrow[d, "\theta"'] & F' \arrow[r, hook] \arrow[d, "\theta'"'] & E \\
K \arrow[r, hook] & K' \arrow[r, hook] & L
\end{tikzcd}$$

and our objective is to now extend θ' to an isomorphism $E \to L$. But since $[E : F] = [E : F'][F' : F]$ and $[F' : F] > 1$ (since $\alpha \notin F$) we see that $[E : F'] < [E : F]$, and this enables us to use the inductive hypothesis to complete the proof. All we need to do is check that the hypotheses of the theorem are satisfied with F' and K' replacing F and K and θ' replacing θ.

Since $x - \alpha \mid f(x)$ in $F'[x]$, there is a polynomial $f_1(x) \in F'[x]$ such that $f(x) = (x - \alpha)f_1(x)$. Since $E = F(\alpha_1, \alpha_2, \ldots, \alpha_d)$ (the extension of F generated by the α_i) where $\alpha_1 = \alpha$, α_2, …, α_d are the roots of $f(x)$, and since $F' = F(\alpha_1)$, we see that $E = F'(\alpha_2, \ldots, \alpha_d)$ is the extension of F' generated by the roots of $f_1(x)$. Thus E is a splitting field for $f_1(x)$ over F'. Applying the isomorphism $F'[x] \to K'[x]$ induced by θ' we see that $(\theta f)(x) = (x - \beta_1)(\theta' f_1)(x)$, and reasoning as above we deduce that L is a splitting field for $(\theta' f_1)(x)$ over K'. So the theorem hypotheses are indeed satisfied with F', K', θ', $f_1(x)$ in place of F, K, θ, $f(x)$, and the inductive hypothesis yields that there is an isomorphism $E \to L$ extending θ' (and hence extending θ). □

To illustrate the above ideas, consider the polynomial $x^4 - 2 \in \mathbb{Q}[x]$. Eisenstein's Criterion shows that this is irreducible. Elementary calculus can be used to prove that $x^4 - 2$ has a positive root in $\mathbb{R}$; we denote this number by α. (That is, $\alpha = \sqrt[4]{2}$.) Let K be the extension of $\mathbb{Q}$ generated by α, so that

$$K = \mathbb{Q}(\alpha) = \{\, a + b\alpha + c\alpha^2 + d\alpha^3 \mid a, b, c, d \in \mathbb{Q}\,\},$$

a degree 4 extension of $\mathbb{Q}$. Observe that K is not a splitting field for $x^4 - 2$; indeed, $x^4 - 2$ has only two roots in K, and in $K[x]$ the factorization of $x^4 - 2$ into irreducibles is

$$x^4 - 2 = (x - \alpha)(x + \alpha)(x^2 + \alpha^2).$$

To obtain a splitting field we would have to make a further degree 2 extension to split the irreducible quadratic factor $x^2 + \alpha^2$ into factors of degree 1. In other words, we would have to adjoin a root of $x^2 + \alpha^2$ to K.

Considering K as a subfield of the complex field $\mathbb{C} = \mathbb{R}(i)$, put $L = K(i\alpha)$. (Observe that this equals $K(i)$, since $\alpha \in K$.) Then L is a degree 2 extension of K, and hence a degree 8 extension of $\mathbb{Q}$. Observe that $x^4 - 2$ is the minimal polynomial of α over $\mathbb{Q}$ (by Proposition (10.14)) and also the minimal polynomial of $-\alpha$ over $\mathbb{Q}$. Hence The Isomorphism Extension Theorem can be applied to the identity isomorphism $\mathbb{Q} \to \mathbb{Q}$, observing that the induced isomorphism $\mathbb{Q}[x] \to \mathbb{Q}[x]$ carries the minimal polynomial of α to the minimal polynomial of $-\alpha$, to conclude that there is an isomorphism $\phi\colon \mathbb{Q}(\alpha) \to \mathbb{Q}(-\alpha)$ which takes α to $-\alpha$ and acts as the identity on $\mathbb{Q}$. (It is worthwhile, at this point, to look carefully back at the steps that have led us to this conclusion, just to make certain that we have not cheated.) Since $-\alpha \in \mathbb{Q}(\alpha)$, it is clear that in fact $\mathbb{Q}(-\alpha) = \mathbb{Q}(\alpha) = K$.

What we have shown above is really not very surprising. The isomorphism $\phi\colon K \to K$ satisfies $\phi\alpha = -\alpha$, and it follows that

$$\phi(\alpha^2) = (\phi\alpha)(\phi\alpha) = (-\alpha)(-\alpha) = \alpha^2,$$

and similarly

$$\phi(\alpha^3) = (\phi\alpha)(\phi(\alpha^2)) = (-\alpha)(\alpha^2) = -\alpha^3.$$

Since also $\phi a = a$ for all $a \in \mathbb{Q}$ we conclude that the action of ϕ on a general element of K is given by

$$a + b\alpha + c\alpha^2 + d\alpha^3 \xmapsto{\phi} a - b\alpha + c\alpha^2 - d\alpha^3 \tag{21}$$

(where a, b, c, $d \in \mathbb{Q}$). This is somewhat similar to the isomorphism $\mathbb{C} \to \mathbb{C}$ given by complex conjugation: the mapping from $\mathbb{C}$ to itself defined by

$$a + bi \mapsto a - bi \qquad \text{(for all } a,\ b \in \mathbb{R})$$

is well known to be bijective and to preserve addition and multiplication. Philosophically, the idea is that from the point of view of the real numbers, $-i$ is just as good a square root of -1 as i is, so that swapping the two should not change anything much. Similarly, from the point of view of the rational numbers one root of $x^4 - 2$ should be as good as any other, and consequently there should be an isomorphism which fixes $\mathbb{Q}$ and takes any given root of $x^4 - 2$ to any other given root of $x^4 - 2$. One can also check directly, by a routine calculation, that Eq.(21) does define an isomorphism $K \to K$.

There is more to be said yet in this situation, since $x^4 - 2$ has two other roots in L that we have not considered yet, namely, $i\alpha$ and $-i\alpha$. Since $x^4 - 2$ is irreducible in $\mathbb{Q}[x]$ it must be the minimal polynomial over $\mathbb{Q}$ for $i\alpha$ and for $-i\alpha$. Since the identity isomorphism $\mathbb{Q} \to \mathbb{Q}$ takes the minimal polynomial of α to the minimal polynomial of $i\alpha$ it can be extended to an isomorphism $\mathbb{Q}(\alpha) \to \mathbb{Q}(i\alpha)$ such that $\alpha \mapsto i\alpha$. And likewise, since the identity isomorphism $\mathbb{Q} \to \mathbb{Q}$ takes the minimal polynomial of α to the minimal polynomial of $-i\alpha$ it can be extended to an isomorphism $\mathbb{Q}(\alpha) \to \mathbb{Q}(i\alpha)$ such that $\alpha \mapsto -i\alpha$. Note that $\mathbb{Q}(i\alpha)$ and $\mathbb{Q}(-i\alpha)$, are not equal to K, although they are equal to each other. We denote this subfield of L by K'.

We can now identify four different isomorphisms from K to subfields of L, as follows:

(i) $\phi_1 \colon K \to K$ such that $\alpha \mapsto \alpha$,

(ii) $\phi_2 \colon K \to K$ such that $\alpha \mapsto -\alpha$,

(iii) $\phi_3 \colon K \to K'$ such that $\alpha \mapsto i\alpha$,

(iv) $\phi_4 \colon K \to K'$ such that $\alpha \mapsto -i\alpha$.

(Of course, ϕ_1 is simply the identity isomorphism.) We can now use the Isomorphism Extension Theorem again to show that each of these isomorphisms has two extensions to isomorphism $L \to L$. For example, consider the isomorphism $\phi_4 \colon K \to K'$. For the general element of of K we find that

$$a + b\alpha + c\alpha^2 + d\alpha^3 \overset{\phi_4}{\longmapsto} a - bi\alpha - c\alpha^2 + di\alpha^3, \tag{22}$$

and in particular the induced isomorphism $K[x] \to K'[x]$ takes $x^2 + 1$ to $x^2 + 1$. That is, it takes the minimal polynomial of i over K to the minimal polynomial of i over K'. Consequently there is an isomorphism from $K(i)$ to $K'(i)$ which extends ϕ_4 and takes i to i. Similarly, since we can also regard $x^2 + 1$ as the minimal polynomial of $-i$ over K', there is also an isomorphism $K(i) \to K'(-i) = K'(i)$ extending ϕ_4 and taking i to $-i$. It is easily checked that $K(i)$ and $K'(i)$ are both equal to L.

Exactly similar reasoning applies to ϕ_1, ϕ_2 and ϕ_3. The upshot is that we find eight different isomorphisms from L to L which extend the identity isomorphism from $\mathbb{Q}$ to $\mathbb{Q}$:

(i) $\psi_1 \colon \alpha \mapsto \alpha,\ i \mapsto i$.

(ii) $\psi_2 \colon \alpha \mapsto \alpha,\ i \mapsto -i$.

(iii) $\psi_3 \colon \alpha \mapsto i\alpha,\ i \mapsto i$.

(iv) $\psi_4 \colon \alpha \mapsto i\alpha,\ i \mapsto -i$.

(v) $\psi_5 \colon \alpha \mapsto -\alpha,\ i \mapsto i$.

(vi) $\psi_6 \colon \alpha \mapsto -\alpha,\ i \mapsto -i$.

(vii) $\psi_7 \colon \alpha \mapsto -i\alpha,\ i \mapsto i$.

(viii) $\psi_8 \colon \alpha \mapsto -i\alpha,\ i \mapsto -i$.

These eight are in fact the only isomorphisms $L \to L$, and it is no coincidence that eight is also equal to $[L : \mathbb{Q}]$, the degree of the extension. However, it is convenient to temporarily postpone further investigation of this, since we have not yet introduced all the relevant concepts.

Definition (13.7): (i) An injective homomorphism is called a *monomorphism*.

(ii) An isomorphism from a field to itself is called an *automorphism* of the field.

(iii) If F is a subfield of E then an automorphism ϕ of E is called an F*-automorphism* of E if $\phi t = t$ for all $t \in F$. Similarly, if L is another field containing F as a subfield then a monomorphism $\phi\colon E \to L$ is called an F-monomorphism if $\phi t = t$ for all $t \in F$.

We use the notation $\mathrm{Aut}_F(E)$ for the set of all F-automorphisms of E. An F-automorphism of E is a symmetry of E, in the sense that it is a bijective transformation which preserves the essential structure of E as an extension field of F. In accordance with our general remarks at the beginning of the section on symmetry above, $\mathrm{Aut}_F(E)$ should be a group under the operation of composition of maps. And indeed it is easy to check the truth of this. If $\phi, \psi \in \mathrm{Aut}_F(E)$ then $\phi\psi$ is certainly an automorphism of E (since the composite of two isomorphisms is an isomorphism), and for all $t \in E$ we have $(\phi\psi)t = \phi(\psi t) = \phi t = t$, whence $\phi\psi \in \mathrm{Aut}_F(E)$. Hence composition does define an operation on $\mathrm{Aut}_F(E)$. Checking that the group axioms are satisfied is straightforward.

Suppose, for example, that ϕ is an $\mathbb{R}$-automorphism of $\mathbb{C}$. Then $(\phi i)^2 = \phi(i^2) = \phi(-1) = -1$, since ϕ preserves multiplication and fixes elements of $\mathbb{R}$. So $\phi i = \pm i$. Furthermore, for all $a, b \in \mathbb{R}$ we have that

$$\phi(a + bi) = (\phi a) + (\phi b)(\phi i) = a + b(\phi i),$$

and so ϕ is either the identity ($a+bi \mapsto a+bi$ for all $a, b \in \mathbb{R}$) or complex conjugation ($a+bi \mapsto a-bi$ for all $a, b \in \mathbb{R}$). We conclude that the group $\mathrm{Aut}_{\mathbb{C}}(\mathbb{R})$ has order 2.

Note that all the elements of $\mathrm{Aut}_R(\mathbb{C})$ have the property that they permute the roots $i, -i$ of the polynomial $x^2 + 1$. This is an instance of the following general result.

Proposition (13.8): *Suppose that F and E are fields, with E an extension of F, and let $\alpha \in \mathrm{Aut}_F(E)$. Let $p(x) \in F[x]$ be arbitrary, and let $S = \{\, t \in E \mid p(t) = 0 \,\}$, the set of roots of $p(x)$ in E. Then $\alpha t \in S$ whenever $t \in S$, and, moreover, the map $S \to S$ given by $t \mapsto \alpha t$ (for all $t \in S$) is a permutation of S.*

Proof. Let $p(x) = a_0 + a_1 x + \cdots + a_d x^d$ and let $t \in S$. Since α preserves addition and multiplication, and $\alpha a = a$ for all $a \in F$, we have

$$\begin{aligned} 0 = \alpha 0 = \alpha(p(t)) = \alpha(a_0 + a_1 t + \cdots + a_d t^d) &= \alpha a_0 + (\alpha a_1)(\alpha t) + \cdots + (\alpha a_d)(\alpha t)^d \\ &= a_0 + a_1(\alpha t) + \cdots + a_d(\alpha t)^d = p(\alpha t), \end{aligned}$$

so that $\alpha t \in S$. Since α is injective and S is a finite set it follows that $t \mapsto \alpha t$ is a permutation of S, as required. □

The principal theme of Galois theory is that each polynomial $f(x) \in F[x]$ has a splitting field E, which is an essentially uniquely determined extension of the field F, and from this extension we obtain the group $\mathrm{Aut}_F(E)$. The Main Theorem of Galois Theory states (amongst other things) that there is a one to one correspondence between the subgroups of $\mathrm{Aut}_F(E)$ and the subfields of E that contain F. Information about the group—in particular, information about its subgroups—then yields information about the equation $f(x) = 0$.

Proposition (13.8) above enables us to regard $\mathrm{Aut}_F(E)$ as a group of permutations.

Proposition (13.9): *Assume that E be a splitting field for the polynomial $f(x) \in F[x]$, and let $f(x) = (x - t_1)^{n_1}(x - t_2)^{n_2} \cdots (x - t_d)^{n_d}$, where the n_i are positive integers and the t_i are distinct elements of E. Each $\alpha \in \mathrm{Aut}_F(E)$ yields a permutation $\sigma = \sigma_\alpha \in S_d$ such that $\alpha t_i = t_{\sigma i}$. Furthermore, the mapping $\mathrm{Aut}_F(E) \to S_d$ given by $\alpha \to \sigma_\alpha$ is an injective homomorphism (permitting $\mathrm{Aut}_F(E)$ to be identified with a subgroup of S_d.*

Proof. We have already seen in Proposition (13.8) that if $\alpha \in \mathrm{Aut}_F(E)$ then $t_i \mapsto \alpha t_i$ is a permutation of $\{t_1, t_2, \dots, t_d\}$. The obvious one to one correspondence between $\{t_1, t_2, \dots, t_d\}$ and $\{1, 2, \dots, d\}$ permits us to convert this into a permutation of $\{1, 2, \dots, d\}$; that is, an element of S_d. Specifically, the permutation $\sigma \in S_d$ that we get maps i to j whenever α takes t_i to t_j. That is, $\alpha t_i = t_{\sigma i}$.

Let $f\colon \mathrm{Aut}_F(E) \to S_d$ be defined by $f\alpha = \sigma$, where $\alpha t_i = t_{\sigma i}$ for all $i \in \{1, 2, \dots, d\}$. That is, in the notation used in the proposition statement above, $f\alpha = \sigma_\alpha$.

Let us show firstly that $f(\alpha\beta) = (f\alpha)(f\beta)$ for all α, $\beta \in \mathrm{Aut}_F(E)$. Given α and β, let $\sigma = f\alpha$ and $\tau = f\beta$. Then $\alpha t_i = t_{\sigma i}$ and $\beta t_j = t_{\tau j}$ for all i and j. So for all j,

$$(\alpha\beta)t_j = \alpha(\beta t_j) = \alpha(t_{\tau j}) = t_{\sigma(\tau j)} = t_{(\sigma\tau)j},$$

which shows that $f(\alpha\beta) = \sigma\tau = (f\alpha)(f\beta)$, as required. Thus f is a homomorphism from $\mathrm{Aut_F(E)}$ to S_d.

It remains to show that f is injective. The key to this is the observation that E is the extension of F generated by t_1, t_2, …, t_d (since E is a splitting field for $f(x)$ over F), and hence all elements of E can be expressed in terms of the t_i. This means that an automorphism of E is completely determined by its effect on the roots t_i. To make this more precise, let $E_0 = F$ and then define E_i recursively for $i > 0$ by $E_i = E_{i-1}(t_i)$, so that $E = E_d$. For all $i \in \{1, 2, \dots, d\}$ each element $a \in E_i$ can be expressed in the form $a_0 + a_1 t_i + a_2 t_i^2 + \cdots + a_m t_i^m$ for some integer m and some elements $a_j \in E_{i-1}$. It follows readily that every element of E_d can be expressed as a sum of terms of the form

$$ct_1^{n_1} t_2^{n_2} \cdots t_d^{n_d}$$

where n_1, n_2, …, n_d are nonnegative integers and $c \in F$. Now suppose that α, $\beta \in \mathrm{Aut}_F(E)$ are such that $f\alpha = f\beta$. Then for each i we have $\alpha t_i = t_{\sigma i} = \beta t_i$, where $\sigma = f\alpha = f\beta$. Furthermore, $\alpha c = c = \beta c$ for all $c \in F$, since α and β are F-automorphisms. Hence

$$\begin{aligned}\alpha(ct_1^{n_1} t_2^{n_2} \cdots t_d^{n_d}) &= (\alpha c)(\alpha t_1)^{n_1}(\alpha t_2)^{n_2} \cdots (\alpha t_d)^{n_d}\\ &= (\beta c)(\beta t_1)^{n_1}(\beta t_2)^{n_2} \cdots (\beta t_d)^{n_d} = \beta(ct_1^{n_1} t_2^{n_2} \cdots t_d^{n_d})\end{aligned}$$

for all choices of the element $c \in F$ and the nonnegative integers n_i. Now for an arbitrary $a \in E$ we can find a finite collection of elements u_1, u_2, …, u_k such that $a = u_1 + u_2 + \cdots + u_k$ and each u_i is expressible in the form $ct_1^{n_1} t_2^{n_2} \cdots t_d^{n_d}$, and it follows that

$$\begin{aligned}\alpha a = \alpha(u_1 + u_2 + \cdots + u_k) &= \alpha u_1 + \alpha u_2 + \cdots + \alpha u_k\\ &= \beta u_1 + \beta u_2 + \cdots + \beta u_k = \beta(u_1 + u_2 + \cdots + u_k) = \beta a,\end{aligned}$$

and hence $\alpha = \beta$. Thus f is injective, as required. □

Let us relate this result to the example we considered above. The polynomial $x^4 - 2 \in \mathbb{Q}[x]$ has four roots in $\mathbb{C}$, namely $t_1 = \alpha$, $t_2 = i\alpha$, $t_3 = -\alpha$ and $t_4 = -i\alpha$, where $\alpha = \sqrt[4]{2}$. Let $L = \mathbb{Q}(t_1, t_2, t_3, t_4)$, the extension of $\mathbb{Q}$ generated by these roots. Then L is a splitting field for $x^4 - 2$ over $\mathbb{Q}$, and by Proposition (13.9) we can identify each element of the group $\mathrm{Aut}_{\mathbb{Q}}(L)$ with a permutation in S_4. Recall that we found eight $\mathbb{Q}$-automorphisms of L, which we labelled ψ_1 to ψ_8. It is straightforward to determine the corresponding permutations. For example, since $\psi_3\alpha = i\alpha$ and $\psi_3 i = i$ we find that

$$\psi_3(i\alpha) = (\psi_3 i)(\psi_3\alpha) = i(i\alpha) = -\alpha,$$

and thus $\psi_3(-\alpha) = -i\alpha$ and $\psi_3(-i\alpha) = \alpha$. In other words, we have

$$\psi_3 t_1 = t_2, \qquad \psi_3 t_2 = t_3, \qquad \psi_3 t_3 = t_4, \qquad \psi_3 t_4 = t_1,$$

and thus psi_3 corresponds to the permutation $(1,2,3,4) \in S_4$. Carrying this kind of analysis out for the other ψ_i, we find that

$$\begin{array}{llll} \psi_1 = \text{identity} & \psi_2 = (2,4) & \psi_3 = (1,2,3,4) & \psi_4 = (1,2)(3,4) \\ \psi_5 = (1,3)(2,4) & \psi_6 = (1,3) & \psi_7 = (1,4,2,3) & \psi_8 = (1,4)(2,3). \end{array}$$

It is perhaps dubious to say that the automorphisms ϕ_j are actually equal to the corresponding permutations, but the permutations do at least determine the corresponding automorphisms uniquely, and in practice it is often convenient to think of an automorphism of the splitting field as a permutation of the roots. But beware that there may be permutations which do not give rise to automorphisms. For example, the permutation $(1,2,3)$ does not correspond to any $\mathbb{Q}$-automorphism of L. This is easily seen, for if there were such an automorphism ψ then it would satisfy $\psi t_1 = t_2$ and $\psi t_3 = t_1$, which is impossible since $\psi t_3 = \psi(-t_1) = -\psi t_1 = -t_2$, which is not equal to t_1.

In fact, it is easily seen that the above eight permutations in S_4 are the only ones that yield automorphisms of L. If $\psi \in \mathrm{Aut}_{\mathbb{Q}}(L)$ is arbitrary, then we must have $\psi t_1 = t_j$ for some j, for which there are four possible choices. Since $t_3 = -t_1$ it follows that $\psi t_3 = -\psi t_1 = -t_j$. Of course, $-t_j$ is some t_k, but the main point is that once ψt_1 is chosen then ψt_3 is also fixed. There remain only two possible choices for ψt_2, and once that choice is made then ψt_4 will also be determined. So there can be at most $4 \times 2 = 8$ possibilities altogether, and hence the eight we found are the only ones.

Returning to theoretical matters, the next proposition explains the correspondence we mentioned earlier between subgroups of $\mathrm{Aut}_F(E)$ and fields K such that $F \subseteq K \subseteq E$.

Proposition (13.10): *Let E be a field and F a subfield of E, and let $\mathcal{G}$ be the group* $\mathrm{Aut}_F(E)$.

(i) For each subgroup $\mathcal{H}$ of $\mathcal{G}$ the set

$$\mathrm{Fix}(\mathcal{H}) = \{\, t \in E \mid \alpha t = t \textit{ for all } \alpha \in \mathcal{H} \,\}$$

is a subfield of E such that $F \subseteq K \subseteq E$.

(ii) For each subfield K of E with $F \subseteq K \subseteq E$ the set

$$\mathrm{Aut}_K(E) = \{\, \alpha \in \mathcal{G} \mid \alpha t = t \textit{ for all } t \in K \,\}$$

is a subgroup of $\mathcal{G}$.

This is a relatively straightforward application of Theorems (5.9) and (12.3), and we leave most of the details to the reader. The main task is to check that various closure properties are satisfied. For example, in Part (i) it is necessary to check that $\mathrm{Fix}(\mathcal{H})$ is closed under addition. So suppose that $t, u \in \mathrm{Fix}(\mathcal{H})$. Then $\alpha t = t$ and $\alpha u = u$ for all $\alpha \in \mathrm{Fix}(\mathcal{H})$, and so for all $\alpha \in \mathrm{Fix}(\mathcal{H})$ we have

$$\alpha(t + u) = \alpha t + \alpha u = t + u$$

and it follows that $t + u \in \mathrm{Fix}(\mathcal{H})$, as required.

Proposition (13.10) gives us functions from the set of subgroups of $\mathcal{G}$ to the set of intermediate fields (fields lying between E and F) and vice versa. Ideally, these functions would be inverse to each other. This is not in fact always true, but we shall see that it is true when certain very reasonable extra hypotheses are assumed.

A slightly complicating feature of the correspondence between subgroups and intermediate fields is that it is inclusion reversing. If $\mathcal{H}$ and $\mathcal{K}$ are subgroups of $\mathcal{G}$ with $\mathcal{H} \leq \mathcal{K}$ then the reverse inclusion holds for their fixed fields: $\mathrm{Fix}(\mathcal{K}) \subseteq \mathrm{Fix}(\mathcal{H})$. Similarly, if H and K are intermediate fields with $H \subseteq K$ then $\mathrm{Aut}_K(E) \leq \mathrm{Aut}_H(E)$. These facts are both immediate from the respective definitions. Note also that if K is any intermediate fields then $K \subseteq \mathrm{Fix}(\mathrm{Aut}_K(E))$ (for this simply says that elements of K are fixed by K-automorphisms of E), and if $\mathcal{H}$ is any subgroup of $\mathcal{G}$ then $\mathcal{H} \leq \mathrm{Aut}_{\mathrm{Fix}(\mathcal{H})}(E)$ (which says that elements of $\mathcal{H}$ are automorphisms of E which fix all the elements of the fixed field of $\mathcal{H}$).

Definition (13.11): An irreducible polynomial $f(x) \in F[x]$ is said to be *separable* if there is no extension field E of F in which $f(x)$ has a repeated root.

Irreducible polynomials that are not separable are harder to analyse than those that are, and hence we intend to restrict our discussion to the separable case. Fortunately, inseparability is a relatively rare phenomenon, and in particular it cannot occur when F has characteristic zero. We shall be quite content to consider only the case $F = \mathbb{Q}$.

Let us sketch briefly the proof that inseparability can only occur in nonzero characteristic. Suppose that $f(x) = \sum_{i=0}^{d} a_i x^i \in F[x]$ is a monic irreducible polynomial, and that E is an extension of F over which $f(x)$ has a factorization $f(x) = (x-t)^2 g(x)$ for some $g(x) \in E[x]$ and $t \in E$. Note that $f(x)$ is the minimal polynomial of t over F. But if we define the *formal derivative* $f'(x)$ of $f(x)$ by the usual formula, $f'(x) = \sum_{i=1}^{d} i a_i x^{i-1}$, and define $g'(x)$ similarly, then by a direct calculation of both sides we can prove that $f'(x) = (x-t)((x-t)g'(x) + 2g(x)$, and we conclude that $f'(t) = 0$. This appears to contradict the fact that $f(x)$ is the minimal polynomial of t over F, since $f'(x) \in F[x]$ and $\deg(f'(x)) < \deg(f(x))$. There is, however, a circumstance in which this contradiction is avoided. If the polynomial $f'(x)$ is the zero polynomial then $f'(t) = 0$ does not contradict the fact that $f(x)$ is the minimal polynomial of t. But how can $f'(x)$ be zero? After all, $f(x)$ is a polynomial of degree at least 1, and every student of calculus knows that the only polynomial whose derivative is zero are the constants. This familiar fact from calculus does not remain true for polynomials over fields of nonzero characteristic. In characteristic zero it is true: if $f(x)$ has degree d and leading coefficient a_d then $f'(x)$ has degree $d-1$ and leading coefficient da_d, since $da_d \neq 0$. But if the characteristic of F is nonzero and a divisor of d then $da_d = 0$ in F; moreover, it is possible for all the terms in the derivative of $f(x)$ to disappear in a similar fashion. For example, working over the field $\mathbb{Z}_7$ we find that the derivative of the polynomial $x^{14} - 3x^7 + 2$ is zero.

It is natural to also apply the term "separable" to elements and to field extensions, in accordance with the following definition.

Definition (13.12): If E is an algebraic extension of F then an element $\alpha \in E$ is said to be separable over F if its minimal polynomial over F is separable. The extension is said to be separable if every element of E is separable over F.

From now on we shall concern ourselves only with subfields of the complex field $\mathbb{C}$, so that all fields under discussion will have characteristic 0. As we have seen, this ensures that all extensions are separable. Only minor modifications to the proofs would be required to deal with arbitrary fields of characteristic 0.

One useful feature of the complex field is that every polynomial in $\mathbb{C}[x]$ splits into linear factors in $\mathbb{C}[x]$. This is the famous "Fundamental Theorem of Algebra". Most proofs make use of complex analysis; a particularly simple one, given by R. P. Boas ("Yet another proof of the Fundamental Theorem of Algebra", *American Mathematical Monthly* no. 71, p. 180 (1964)) goes as follows. Suppose, for a contradiction, that $f(x) \in \mathbb{C}[x]$ has positive degree and satisfies $f(z) \neq 0$ for all $z \in \mathbb{C}$. For each $z \in \mathbb{C}$ let $\overline{z}$ be the complex conjugate of z, and let $\overline{f}(x)$ be the polynomial obtained from $f(x)$ by replacing the coefficients by their complex conjugates. If we put $g(x) = f(x)\overline{f}(x)$, then it is an easy calculation to check that the coefficients of $f(x)\overline{f}(x)$ are all self-conjugate, so that $g(x) \in \mathbb{R}[x]$. In particular, $g(t) \in \mathbb{R}$ for all $t \in \mathbb{R}$. Furthermore, for each $z \in \mathbb{C}$,

$$g(z) = f(z)\overline{f}(z) = f(z)\overline{f(\overline{z})} \neq 0$$

since $f(z)$ and $f(\overline{z})$ are both nonzero. It follows that

$$\int_0^{2\pi} \frac{d\theta}{g(2\cos(\theta))} \neq 0,$$

since $g(2\cos(\theta))$ can never change sign.

Write $g(x) = a_0 + a_1x + \cdots + a_nx^n$, where $a_n \neq 0$, and define

$$h(x) = x^n g(x + x^{-1}) = a_0x^n + a_1x^n(x + x^{-1}) + \cdots + a_nx^n(x + x^{-1})^n.$$

Observe that the negative powers of x cancel out, so that $h(x)$ is a polynomial. Moreover, the constant term of $h(x)$ is a_n. Since $g(z + z^{-1}) \neq 0$ for all nonzero $z \in \mathbb{C}$, and $h(0) = a_n \neq 0$, we see that $1/h(z)$ is analytic in all of $\mathbb{C}$. So we have

$$\begin{aligned}\int_0^{2\pi} \frac{d\theta}{g(2\cos(\theta))} &= \frac{1}{i}\int_{|z|=1} \frac{dz}{zg(z + z^{-1})} \\ &= \frac{1}{i}\int_{|z|=1} \frac{z^{n-1}dz}{h(z)} \\ &= 0,\end{aligned}$$

where the last equality is by Cauchy's Integral Theorem. So we have a contradiction. This shows that every $f(x) \in \mathbb{C}[x]$ with $\deg(f(x)) > 0$ has a root in $\mathbb{C}$, and hence a factorization $(x - \alpha)f_1(x)$ for some $\alpha \in \mathbb{C}$ and some $f_1(x) \in \mathbb{C}[x]$ of degree $\deg(f(x)) - 1$. Repeating the argument with $f_1(x)$ in place of $f(x)$, and continuing in this way, yields an expression for $f(x)$ as a product of linear factors.

Recall that in our investigation of $\mathbb{Q}$-automorphisms of the splitting field of $x^4 - 2$ over $\mathbb{Q}$ we discovered that the number of $\mathbb{Q}$-automorphisms equals the degree of the extension. Our next objective is to prove a general result of this kind.

Proposition (13.13): *Let $F \subseteq E \subseteq \mathbb{C}$ with $[E : F]$ finite. Then the number of F-monomorphisms $E \to \mathbb{C}$ is exactly $[E : F]$. More generally, if $\theta\colon F \to F'$ is an isomorphism, where F' is also a subfield of $\mathbb{C}$, then there are exactly $[E : F]$ monomorphisms $E \to \mathbb{C}$ extending θ.*

Proof. Observe that the first assertion follows from the second by taking θ to be the identity. We prove the second assertion by induction on $[E : F]$, observing that the result is trivial if $[E : F] = 1$ (since $E = F$ implies that the only map $E \to \mathbb{C}$ extending θ is θ itself.

Suppose now that $[E : F] > 1$, and choose a field H with $F \subseteq H \subsetneq E$ and $[E : H]$ as small as possible. If t be any element of E that is not in H then $H \subsetneq H(t) \subseteq E$, and since this gives $[E : H] = [E : H(t)][H(t) : H]$ and $[H(t) : H] > 1$ we deduce that $[E : H(t)] < [E : H]$. In view of the way H was chosen, this forces $H(t) = E$. Furthermore, since $[H : F] < [E : F]$ the inductive hypothesis tells us that the number of monomorphisms $H \to \mathbb{C}$ extending θ is precisely $[H : F]$.

Let ϕ be any one of the monomorphisms $H \to \mathbb{C}$ which extends θ, and let K be the image of ϕ. Thus $u \mapsto \phi u$ is an isomorphism $H \to K$ extending θ. Let $p(x) \in H[x]$ be the minimal polynomial of t over H, and let $q(x) = (\phi p)(x) \in K[x]$. Then $p(x)$ is irreducible in $H[x]$ (by Theorem (10.11)) and so $q(x)$ is irreducible in $K[x]$ (since the map $H[x] \to K[x]$ induced by ϕ is an isomorphism). Furthermore, by Proposition (10.19), $\deg(p(x)) = [H(t) : H] = [E : H]$. By the Fundamental Theorem of Algebra combined with the fact that $q(x)$ is separable we know that $q(x)$ has d distinct roots $t_1, t_2, \ldots, t_d$ in $\mathbb{C}$, where $d = \deg(q(x)) = \deg(p(x)) = [E : H]$. By the Isomorphism Extension Theorem (13.5), for each $i \in \{1, 2, \ldots, d\}$ the isomorphism $\phi\colon H \to K$ extends to an isomorphism $E = H(t) \to K(t_i)$ such that $t \mapsto t_i$. Since $K(t_i)$ is a subfield of $\mathbb{C}$ an isomorphism $E \to K(t_i)$ can be regarded as a monomorphism $E \to \mathbb{C}$. Thus we have constructed d monomorphisms $E \to \mathbb{C}$ extending the given monomorphism $\phi\colon H \to \mathbb{C}$. We know that the monomorphisms we have constructed are all distinct from each other: they are distinguished by their effect on the element t since the elements $t_1, t_2, \ldots, t_d$ are all distinct from each other.

There are $[H : F]$ distinct monomorphisms $\phi\colon H \to \mathbb{C}$ extending θ, and each of them extends in d distinct ways to a monomorphism $E \to \mathbb{C}$. Note that extensions of distinct mappings are still distinct; so we have definitely obtained $d[H : F] = [E : H][H : F] = [E : F]$ distinct monomorphisms $E \to \mathbb{C}$ extending θ. It remains to prove that these are the only monomorphisms $E \to \mathbb{C}$ extending θ. But if $\psi\colon E \to \mathbb{C}$ is an arbitrary monomorphism extending θ then the restriction of ψ to H is an monomorphism $H \to \mathbb{C}$ extending θ, which by the inductive hypothesis must be one of the $[H : F]$ monomorphisms ϕ considered in our calculations. Writing $q(x) = (\phi p)(x)$ as above, and noting that $(\phi p)(x)$ is the same as $(\psi p)(x)$, we see that $q(\psi t) = \psi(p(t)) = \psi 0 = 0$. Hence $\psi t = t_i$ for some i. So ψ is an extension of ϕ such that $t \mapsto t_i$. On such mapping is included among the $[E : F]$ monomorphisms that we counted above; so to prove that ψ is one of the maps we have counted it remains to prove that the extension ψ of ϕ satisfying $\psi t = t_i$ is unique. But every element $u \in E$ can be expressed in the form $u = a_0 + a_1 t + \cdots + a_k t^k$ for some nonnegative integer k and some elements $a_j \in H$, and thus

$$\psi u = \psi a_0 + (\psi a_1)(\psi t) + \cdots + (\psi a_k)(\psi t)^k = \phi a_0 + (\phi a_1)t_i + \cdots + (\phi a_k)(t_i)^d$$

is uniquely determined by the conditions that ψ extends ϕ and takes t to t_i, as required. □

As we have seen, if $L \subseteq \mathbb{C}$ is the splitting field for $x^4 - 2$ over $\mathbb{Q}$ then $[L : \mathbb{Q}] = 8$, and so Proposition (13.13) says that there are exactly eight $\mathbb{Q}$-monomorphisms $L \to \mathbb{C}$. However, we discovered that in fact these eight monomorphisms all map L to itself, and are therefore, in effect, $\mathbb{Q}$-automorphisms of L. In contrast, we also saw that if $\alpha = \sqrt[4]{2}$ then the four $\mathbb{Q}$-monomorphisms $\mathbb{Q}(\alpha) \to \mathbb{C}$ do not all map $\mathbb{Q}(\alpha)$ to itself.

Definition (13.14): Let F and E be subfields of $\mathbb{C}$ with E an extension of F. We say that E is a *normal extension* of F if $\phi E = E$ for every F-monomorphism $E \to \mathbb{C}$.

It is fairly easy to see that splitting field have to be normal extensions. For suppose that $f(x) = a_0 + a_1 x + \cdots + a_d x^d \in F[x]$ and let $E \subseteq \mathbb{C}$ be the splitting field of $f(x)$ over F. That is, $E = F(t_1, t_2, \ldots, t_k)$ (the extension of F generated by $t_1, t_2, \ldots, t_k$), where $t_1, t_2, \ldots, t_k$ are the distinct roots of $f(x)$ in $\mathbb{C}$. Now if $\phi\colon E \to \mathbb{C}$ is a F-monomorphism we have for all i from 1 to k that

$$0 = \phi 0 = \phi(a_0 + a_1 t_i + \cdots + a_k t^k) = \phi a_0 + (\phi a_1)(\phi t_i) + \cdots + (\phi a_k)(\phi t_i)^k = a_0 + a_1(\phi t_i) + \cdots + a_k(\phi t_i)^k,$$

whence $\phi t_i = t_j$ for some j. Since ϕ is injective it follows that ϕ permutes the set $t_1, t_2, \ldots, t_k$, and hence

$$\phi E = \phi(F(t_1, \ldots, t_k)) = (\phi F)(\phi t_1, \ldots, \phi t_k) = F(t_1, \ldots, t_k) = E,$$

and this shows, as required, that all F-monomorphisms $E \to \mathbb{C}$ take E to itself.

The following proposition gives an alternative characterization of normality, which we could have used as the definition.

Proposition (13.15): *An finite extension E of the field F is normal if and only if every irreducible $f(x) \in F[x]$ that has a root in E splits into linear factors over E.*

Proof. Suppose that E is a normal extension of F, and let $f(x) \in F[x]$ be an irreducible polynomial with a root $t \in E$. Let $t' \in \mathbb{C}$ be any other root of $f(x)$. By the Isomorphism Extension Theorem there is an F-monomorphism $F(t) \to \mathbb{C}$ taking t to t'. Furthermore, by Proposition (13.13) any such mapping extends to an F-monomorphism $\theta\colon E \to \mathbb{C}$, and it follows that $t' = \theta t \in \theta E = E$, since E is a normal extension of F. So E contains all the roots of $f(x)$ in $\mathbb{C}$, and hence $f(x)$ splits over E.

Suppose, conversely, that E is a finite extension of F having the property that every irreducible $f(x) \in F[x]$ with a root in E splits over E. We show that if $\theta\colon E \to \mathbb{C}$ is an arbitrary F-monomorphism and $t \in E$ is arbitrary then $\theta t \in E$. Indeed, suppose that $f(x) \in F[x]$ is the minimal polynomial of t over F. Then $f(t) = 0$, and applying the homomorphism θ it follows that $(\theta f)(\theta t) = 0$. But $(\theta f)(x) = f(x)$ since θ fixes all elements of F. So θt is a root of $f(x)$. But since $f(x)$ splits over E it follows that E contains all the roots of $f(x)$, and so $\theta t \in E$. Since t was an arbitrary element of E we conclude that $\theta E \subseteq E$. But $[\theta E : F] = [E : F]$ is finite, and thus $[E : \theta E] = 1$. So $\theta E = E$, and we have shown, as required, that every F-monomorphism $E \to \mathbb{C}$ fixes E. □

Definition (13.16): Let E be a field and F a subfield of E. We say that E is a *Galois extension* of F if it is a finite, normal separable extension of F. Under these circumstances the group $\mathrm{Aut}_F(E)$ is called the *Galois group* of the extension, and it is denoted by $\mathrm{Gal}(E : F)$.

We need one further preliminary result before we can prove the Main Theorem of Galois Theory.

Proposition (13.17): *Let E be a finite extension of F and let $\mathcal{G}$ be any subgroup of* $\mathrm{Aut}_\mathrm{F}(\mathrm{E})$. *Then* $|\mathcal{G}| \geq [E : \mathrm{Fix}(\mathcal{G})]$.

Proof. Let $K = \mathrm{Fix}(\mathcal{G})$, and suppose, for a contradiction, that $|\mathcal{G}| = n < [E : K]$. Then we may choose $n + 1$ elements, $t_1, t_2, \ldots, t_{n+1}$ in E which are linearly independent over K. Write the elements of $\mathcal{G}$ as $g_1, g_2, \ldots, g_n$, where $g_1 = i_G$. Now

$$\begin{pmatrix} g_1t_1 \\ g_2t_1 \\ \vdots \\ g_nt_1 \end{pmatrix}, \begin{pmatrix} g_1t_2 \\ g_2t_2 \\ \vdots \\ g_nt_2 \end{pmatrix}, \ldots, \begin{pmatrix} g_1t_{n+1} \\ g_2t_{n+1} \\ \vdots \\ g_nt_{n+1} \end{pmatrix},$$

are $n + 1$ vectors in the n-dimensional space E^n, and so they must be linearly dependent. Thus we can find $s_1, s_2, \ldots, s_{n+1} \in E$ which are not all zero and which satisfy

$$s_1 \begin{pmatrix} g_1t_1 \\ g_2t_1 \\ \vdots \\ g_nt_1 \end{pmatrix} + s_2 \begin{pmatrix} g_1t_2 \\ g_2t_2 \\ \vdots \\ g_nt_2 \end{pmatrix} + \cdots + s_{n+1} \begin{pmatrix} g_1t_{n+1} \\ g_2t_{n+1} \\ \vdots \\ g_{n+1}t_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \qquad (23)$$

Amongst all possible choices for coefficients $s_1, s_2, \ldots, s_{n+1}$ satisfying Eq.(23), with at least one of the s_i being nonzero, let us choose one for which the number of nonzero s_i is as small as possible.

Observe that the lth component of the above vector equation is

$$s_1(g_l t_1) + s_2(g_l t_2) + \cdots + s_{n+1}(g_l t_{n+1}) = 0. \tag{24}$$

If $g \in \mathcal{G}$ is arbitrary, then applying g to this gives

$$(gs_1)(gg_l t_1) + (gs_2)(gg_l t_2) + \cdots + (gs_{n+1})(gg_l t_{n+1}) = 0 \tag{25}$$

(where we have used the fact that g preserves addition and multiplication and takes 0 to 0). Here g_l is any element of $\mathcal{G}$; so for a given $g \in \mathcal{G}$, Eq.(25) remains valid if g_l is replaced by any element of $\mathcal{G}$. In particular, for each $j \in \{1, 2, \ldots, n\}$ Eq.(25) holds with g_l replaced by $g^{-1}g_j$. Thus

$$(gs_1)(g_j t_1) + (gs_2)(g_j t_2) + \cdots + (gs_{n+1})(g_j t_{n+1}) = 0 \tag{26}$$

holds for all j. Now choose $k \in \{1, 2, \ldots, n+1\}$ such that $s_k \neq 0$, and multiply both sides of Eq.(24) gs_k. Replacing l by j this gives

$$(gs_k)s_1(g_j t_1) + (gs_k)s_2(g_j t_2) + \cdots + (gs_k)s_{n+1}(g_j t_{n+1}) = 0. \tag{27}$$

Multiplying Eq.(26) through by s_k gives

$$s_k(gs_1)(g_j t_1) + s_k(gs_2)(g_j t_2) + \cdots + s_k(gs_{n+1})(g_j t_{n+1}) = 0, \tag{28}$$

and subtracting Eq.(28) from Eq.(27) gives

$$((gs_k)s_1 - s_k(gs_1))(g_j t_1) + ((gs_k)s_2 - s_k(gs_2))(g_j t_2) + \cdots + ((gs_k)s_{n+1} - s_k(gs_{n+1}))(g_j t_{n+1}) = 0.$$

Writing $s_i' = (gs_k)s_i - s_k(gs_i)$ for each i, this is the jth component of the vector equation

$$s_1' \begin{pmatrix} g_1 t_1 \\ g_2 t_1 \\ \vdots \\ g_n t_1 \end{pmatrix} + s_2' \begin{pmatrix} g_1 t_2 \\ g_2 t_2 \\ \vdots \\ g_n t_2 \end{pmatrix} + \cdots + s_{n+1}' \begin{pmatrix} g_1 t_{n+1} \\ g_2 t_{n+1} \\ \vdots \\ g_{n+1} t_{n+1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Note that if $s_i = 0$ then $gs_i = 0$ and so $s_i' = 0$; moreover, $s_k' = (gs_k)s_k - s_k(gs_k) = 0$, while $s_k \neq 0$ by our choice of k. So the number of i such that s_i' is nonzero is strictly less than the number of i such that s_i is nonzero. This forces all the s_i' to be zero, since the alternative would contradict our original choice of the coefficients s_i. Thus

$$(gs_k)s_i = s_k(gs_i)$$

for all i, and whenever $s_i \neq 0$ this can be rewritten as

$$s_i s_k^{-1} = g(s_i s_k^{-1}).$$

But $g \in \mathcal{G}$ was arbitrary, and so we conclude that $s_i s_k^{-1} \in \mathrm{Fix}(\mathcal{G})$ for each i such that $s_i \neq 0$. Writing $c_i = s_i s_k^{-1}$ we have $c_i \in \mathrm{Fix}(\mathcal{G})$ for each i, and multiplying Eq.(24) through by s_k^{-1} we obtain

$$c_1(g_l t_1) + c_2(g_l t_2) + \cdots + c_{n+1}(g_l t_{n+1}) = 0$$

where the c_i are elements of Fix($\mathcal{G}$) which are not all zero. So $g_l t_1, g_l t_2, \ldots, g_l t_{n+1}$ are linearly dependent over Fix($\mathcal{G}$). In particular, taking $l = 1$, so that g_l is the identity, this says that $t_1, t_2, \ldots, t_{n+1}$ are linearly dependent over Fix($\mathcal{G}$), contrary to our original choice of these elements. This final contradiction shows that the assumption that $|\mathcal{G}| < [E : K]$ is unsustainable; that is, $|\mathcal{G}| \geq [E : K]$. □

Our next result is known as the **Main Theorem of Galois Theory**.

Theorem (13.18): *Let E be a Galois extension of F and let $\mathcal{G} = \mathrm{Gal}(E : F)$. Let S be the set of all subgroups of $\mathcal{G}$ and T the set of all subfields K of E such that $F \subseteq K \subseteq E$. Then then function* $\mathrm{Fix}\colon \mathsf{S} \to \mathsf{T}$ *which takes each $\mathcal{H} \leq \mathcal{G}$ to its fixed field* $\mathrm{Fix}(\mathcal{H}) = \{\, t \in E \mid gt = t \text{ for all } g \in \mathcal{H} \,\}$ *is a bijection, and the inverse bijection $\mathsf{T} \to \mathsf{S}$ is given by $K \mapsto \mathrm{Gal}(E : K)$ (for each intermediate field K). Furthermore, a subgroup $\mathcal{H}$ of $\mathcal{G}$ is normal if and only if the corresponding field $K = \mathrm{Fix}(\mathcal{H})$ is a normal extension of F, and in this situation* $\mathrm{Gal}(K : F)$ *is isomorphic to the quotient group* $\mathrm{Gal}(E : F)/\mathrm{Gal}(E : K)$.

Proof. Recall first that $\mathrm{Gal}(E : K) = \mathrm{Aut}_K(E)$, which is a subgroup of $\mathrm{Gal}(E : F) = \mathrm{Aut}_F(E)$, since every K-automorphism of E is certainly also an F-automorphism of E.

To show that $\mathcal{H} \mapsto \mathrm{Fix}(\mathcal{H})$ and $K \mapsto \mathrm{Aut}_K(E)$ are mutually inverse bijections it suffices to show that $K = \mathrm{Fix}(\mathrm{Aut}_K(E))$ for each field K with $F \subseteq K \subseteq E$ and $\mathcal{H} = \mathrm{Aut}_{\mathrm{Fix}(\mathcal{H})}(E)$ for each subgroup $\mathcal{H}$ of $\mathcal{G}$.

Let K be an arbitrary subfield of E containing F, and let $\mathcal{K} = \mathrm{Aut}_K(E)$. Let $K' = \mathrm{Fix}(\mathcal{K})$. Our aim is to prove that $K = K'$. It is clear that $K \subseteq K'$, because every element of K is fixed by every K-automorphism (and K' is by definition the set of points fixed by every K-automorphism). Now $K \subseteq K'$ implies that every K'-automorphism of E is also a K-automorphism of E; that is, $\mathrm{Aut}_{K'}(E) \subseteq \mathrm{Aut}_K(E)$. But, by the definition of K', every element of $\mathcal{K}$ is a K'-automorphism of E. So $\mathrm{Aut}_K(E) \subseteq \mathrm{Aut}_{K'}(E)$, and thus $\mathrm{Aut}_K(E) = \mathrm{Aut}_{K'}(E)$.

Since E is a normal extension of F we know that every F-monomorphism $E \to \mathbb{C}$ fixes E. Every K-monomorphisms is an F-monomorphism, since $F \subseteq K$, and the same applies to K'-monomorphisms. So the number of K-monomorphisms $E \to \mathbb{C}$ equals $|\mathrm{Aut}_K(E)|$, and the number of K'-monomorphisms $E \to \mathbb{C}$ equals $|\mathrm{Aut}_{K'}(E)|$. But by Proposition (13.13) these numbers are also equal to $[E : K]$ and $[E : K']$ respectively. As $\mathrm{Aut}_K(E) = \mathrm{Aut}_{K'}(E)$, it follows that $[E : K] = [E : K']$, and combined with $K \subseteq K'$ this yields $K = K'$, as required.

Now let $\mathcal{H}$ be an arbitrary subgroup of $\mathcal{G}$, and let $H = \mathrm{Fix}(\mathcal{H})$. Let $\mathcal{H}' = \mathrm{Aut}_H(E)$. We aim to prove that $\mathcal{H} = \mathcal{H}'$. It is clear that $\mathcal{H} \subseteq \mathcal{H}'$, since every automorphism in $\mathcal{H}$ fixes all the points of H (and $\mathcal{H}'$ is by definition the set of all automorphisms that fix every point of H. Now $\mathcal{H} \subseteq \mathcal{H}'$ implies that anything fixed by all elements of $\mathcal{H}'$ is fixed by all elements of $\mathcal{H}$; that is $\mathrm{Fix}(\mathcal{H}') \subseteq \mathrm{Fix}(\mathcal{H})$. But, by the definition of $\mathcal{H}'$, every element of H is a fixed point of $\mathcal{H}'$. So $\mathrm{Fix}(\mathcal{H}) \subseteq \mathrm{Fix}(\mathcal{H}')$, and thus $\mathrm{Fix}(\mathcal{H}') = \mathrm{Fix}(\mathcal{H})$.

Since $\mathrm{Aut}_H(E)$ equals the number of H-monomorphisms $E \to \mathbb{C}$, Proposition (13.13) yields $|\mathcal{H}'| = |\mathrm{Aut}_H(E)| = [E : H]$. But Proposition (13.17) tells us that $|\mathcal{H}| \geq [E : H]$. So $|\mathcal{H}| \geq |\mathcal{H}'|$, and combined with $\mathcal{H} \subseteq \mathcal{H}'$ this yields $\mathcal{H} \subseteq \mathcal{H}'$, as required.

Let us now show that normal subgroups of $\mathcal{G}$ correspond to intermediate fields that are normal extensions of F. Suppose that $\mathcal{H}$ is a normal subgroup of $\mathcal{G}$, and let $H = \mathrm{Fix}(\mathcal{H})$. Suppose that $\theta\colon H \to \mathbb{C}$ is an F-monomorphism, and let $H' = \theta H$. By Proposition (13.13) there is an F-monomorphism $E \to \mathbb{C}$ extending θ, and any such monomorphism must take E to itself since E is a normal extension of F. So there is a $\phi \in \mathcal{G}$ whose restriction to H coincides with θ. Now let $t \in H'$ and $\psi \in \mathcal{H}$ be arbitrary. Since $H' = \theta H = \phi H$ there exists $u \in H$ with $\phi u = t$. Since $\mathcal{H}$ is normal in $\mathcal{G}$ and $\psi \in \mathcal{H}$ it follows that $\phi^{-1}\psi\phi \in \mathcal{H}$. Since $u \in H = \mathrm{Fix}(\mathcal{H})$ it follows that $(\phi^{-1}\psi\phi)u = u$, and thus

$$\psi t = \psi(\phi u) = (\phi\phi^{-1})(\psi(\phi u)) = \phi((\phi^{-1}\psi\phi)u) = \phi u = t.$$

Since ψ was an arbitrary element of $\mathcal{H}$, this shows that $t \in \mathrm{Fix}(\mathcal{H})$. Hence $H' \subseteq H$, and since $[H' : E] = [H : E]$ we deduce that $H' = H$. We have shown that an arbitrary F-monomorphism $\theta: H \to \mathbb{C}$ maps H to itself; that is, H is a normal extension of F.

Suppose, on the other hand, that K is a normal extension of F, and let $\mathcal{H} = \mathrm{Gal}(E : K)$. We must show that $\mathcal{H}$ is normal in $\mathcal{G}$, and also that $\mathrm{Gal}(K : F)$ is isomorphic to $\mathrm{Gal}(E : F)/\mathrm{Gal}(E : K)$. For each $\phi \in \mathrm{Gal}(E : F) = \mathrm{Aut}_F(E)$ let $f\phi: K \to \phi K$ be the restriction of ϕ. Observe that since K is a normal extension of F we must have $\phi K = K$ in all cases, and so $\phi \mapsto f\phi$ defines a map $f: \mathrm{Gal}(E : F) \to \mathrm{Aut}_F(K) = \mathrm{Gal}(K : F)$. It is clear that this map is a group homomorphism, and it is surjective since by Proposition (13.13) every F-automorphism of K can be extended to an F-automorphism of E (since any F-monomorphism $E \to \mathbb{C}$ must map E to itself, and hence correspond to an F-automorphism of E, since E is a normal extension of F). The kernel of f is the set of all F-automorphisms of E whose restriction to K is the identity; that is, the kernel of f is $\mathrm{Aut}_K(E) = \mathrm{Gal}(E : K)$. The First Isomorphism Theorem for groups (Theorem (12.17)) now tells us that $\mathrm{Gal}(E : K)$ is normal in $\mathrm{Gal}(E : F)$, and that $\mathrm{Gal}(K : F)$ is isomorphic to $\mathrm{Gal}(E : F)/\mathrm{Gal}(E : K)$, as required. □

Returning to an example we considered earlier, let $L \subseteq \mathbb{C}$ be the splitting field for $x^4 - 2$ over $\mathbb{Q}$. We found that $\mathcal{G} = \mathrm{Gal}(L : \mathbb{Q})$ has eight elements ψ_j, where $1 \leq j \leq 8$, and we identified these with permutations of the roots t_1, t_2, t_3, t_4 of $x^4 - 2$ as follows:

$$\begin{array}{llll} \psi_1 = \mathrm{id} & \psi_2 = (2,4) & \psi_3 = (1,2,3,4) & \psi_4 = (1,2)(3,4) \\ \psi_5 = (1,3)(2,4) & \psi_6 = (1,3) & \psi_7 = (1,4,2,3) & \psi_8 = (1,4)(2,3) \end{array}$$

(where we have denoted the identity automorphism by id to avoid confusion with the complex number i.) It is now not difficult to determine all the subgroups of $\mathcal{G}$. Firstly, each element generates a cyclic subgroup. This yields a subgroup of order 1, five subgroups of order 2 and a subgroup of order 4:

$$\begin{array}{ll} \mathcal{H}_1 = \{\mathrm{id}\} & \mathcal{H}_5 = \{\mathrm{id},\ (1,3)(2,4)\} \\ \mathcal{H}_2 = \{\mathrm{id},\ (2,4)\} & \mathcal{H}_6 = \{\mathrm{id},\ (1,3)\} \\ \mathcal{H}_3 = \{\mathrm{id},\ (1,2,3,4),\ (1,3)(2,4),\ (1,4,3,2)\} & \mathcal{H}_7 = \{\mathrm{id}, (1,4)(2,3)\}. \\ \mathcal{H}_4 = \{\mathrm{id},\ (1,2)(3,4)\} & \end{array}$$

Observe that $(1,2,3,4)$ and $(1,4,3,2)$ generate the same subgroup. Clearly a group of order 2 has to be cyclic, and so $\mathcal{G}$ has no other subgroups of order 2. Since the order of any subgroup has to be a divisor of $|\mathcal{G}| = 8$ (by Proposition (12.13)) all other subgroups of $\mathcal{G}$ have order 4 or order 8. The nonidentity elements of a noncyclic group of order 4 can only have order 2 (since if an element had order 4 the group would be cyclic, and only the identity has order 1). So a noncyclic group of order 4 must have the form $\{\mathrm{id}, a, b, c\}$, where a, b and c all have order 2. Since ab cannot equal a (since $b \neq \mathrm{id}$) or b (since $a \neq \mathrm{id}$) or id (since $b \neq a^{-1} = a$) it follows that $ab = c$, and similarly we see also that $ba = c$. Looking through the elements of $\mathcal{G}$ for two elements a and b of order 2 such that $ab = ba$, we readily discover that $\mathcal{G}$ has just two noncyclic subgroups of order 4:

$$\begin{aligned} \mathcal{H}_8 &= \{\mathrm{id},\ (1,3),\ (2,4),\ (1,3)(2,4)\} \\ \mathcal{H}_9 &= \{\mathrm{id}, (1,2)(3,4),\ (1,3)(2,4),\ (1,4)(2,3)\}. \end{aligned}$$

Together with $\mathcal{H}_{10} = \mathcal{G}$, we have now obtained all the subgroups. If we define $K_j = \text{Fix}(\mathcal{H}_j)$ then the Main Theorem of Galois Theory tells us that the ten fields K_j are all the subfields of L containing $\mathbb{Q}$.

The extension of $\mathbb{Q}$ generated by $\alpha = \sqrt[4]{2}$ $(= t_1)$ has degree 4, and 1, α, α^2, α^3 form a basis for $\mathbb{Q}(\alpha)$ as a vector space over $\mathbb{Q}$. Since $L = \mathbb{Q}(\alpha, i)$ is a degree 2 extension of $\mathbb{Q}(\alpha)$ for which 1, i is a basis, we deduce that every element of L is uniquely expressible in the form

$$\lambda_1 + \lambda_2\alpha + \lambda_3\alpha^2 + \lambda_4\alpha^3 + \lambda_5 i + \lambda_6 i\alpha + \lambda_7 i\alpha^2 + \lambda_8 i\alpha^3 \tag{29}$$

where the coefficients λ_j are elements of $\mathbb{Q}$.

Recall that ψ_2 fixes α and takes $i\alpha$ to $-i\alpha$, which means that it takes i to $-i$. So ψ_2 takes the general element (29) of L to

$$\lambda_1 + \lambda_2\alpha + \lambda_3\alpha^2 + \lambda_4\alpha^3 - \lambda_5 i - \lambda_6 i\alpha - \lambda_7 i\alpha^2 - \lambda_8 i\alpha^3,$$

and so the element is fixed by ψ_2 if and only if $\lambda_5 = \lambda_6 = \lambda_7 = \lambda_8 = 0$. In other words,

$$K_2 = \text{Fix}(\mathcal{H}_2) = \text{Fix}(\{\text{id}, \psi_2\}) = \mathbb{Q}(\alpha).$$

(Observe that $[E : K_2] = 2 = |\mathcal{H}_2| = |\text{Aut}_{K_2}(E)|$, which is as it should be!) Similarly, ψ_6 takes α to $-\alpha$ and i to $-i$, and we find that the element (29) is fixed by ψ_6 if and only if $\lambda_2 = \lambda_4 = \lambda_5 = \lambda_7 = 0$. Thus $K_6 = \mathbb{Q}(i\alpha)$. Observe that α and $i\alpha$ are different roots of the same polynomial $x^4 - 2 \in \mathbb{Q}[x]$, and so the extensions K_2 and K_6 of $\mathbb{Q}$, though distinct, are isomorphic to each other. These are not normal extensions of $\mathbb{Q}$ since there is a monomorphism $K_2 \to \mathbb{C}$ taking K_2 to $K_6 \neq K_2$, and similarly there is a monomorphism taking K_6 to $K_2 \neq K_6$.

We can determine the fixed fields of the other subgroups in the same kind of way. The automorphism ψ_5 takes α to $-\alpha$ and fixes i; so it takes the general element (29) to

$$\lambda_1 - \lambda_2\alpha + \lambda_3\alpha^2 - \lambda_4\alpha^3 + \lambda_5 i - \lambda_6 i\alpha + \lambda_7 i\alpha^2 - \lambda_8 i\alpha^3.$$

We conclude that K_5 consists of those elements such that $\lambda_2 = \lambda_4 = \lambda_6 = \lambda_8 = 0$. Thus $K_5 = \mathbb{Q}(i, \alpha^2)$, which can be seen to be the splitting field of $(x^2 - 2)(x^2 + 1)$. Since this is a normal extension of $\mathbb{Q}$ the group $\mathcal{H}_5 = \{\text{id}, (1,3)(2,4)\}$ must be a normal subgroup of $\mathcal{G}$. This is easily checked. Indeed, the permutation $(1,3)(2,4)$ commutes with every element of $\mathcal{G}$, and so for all $\phi \in \mathcal{G}$ and $\psi \in \mathcal{H}_5$ we have that $\phi^{-1}\psi\phi = \psi \in \mathcal{H}_5$.

None of the other subgroups of $\mathcal{G}$ of order 2 are normal. For example, $(1,2)(3,4) = \psi_4 \in \mathcal{H}_4$, but

$$\psi_2^{-1}\psi_4\psi_2 = (2,4)(1,2)(3,4)(2,4) = (1,4)(2,3) \notin \mathcal{H}_4.$$

Since ψ_4 takes α to $i\alpha$ and $i\alpha$ to α, it takes i to $-i$ and α^2 to $-\alpha^2$. So it takes the general element (29) to

$$\lambda_1 + \lambda_6\alpha - \lambda_3\alpha^2 - \lambda_8\alpha^3 + -\lambda_5 i + \lambda_2 i\alpha + \lambda_7 i\alpha^2 - \lambda_4 i\alpha^3,$$

and we conclude that K_5 consists of all those elements such that $\lambda_3 = \lambda_5 = 0$, $\lambda_2 = \lambda_6$ and $\lambda_4 = -\lambda_8$. So $K_5 = \mathbb{Q}(\gamma)$, where $\gamma = \alpha + i\alpha$. Observe that $\gamma^2 = 2i\alpha^2$, and so $\gamma^4 = -8$. That is, γ is a root of the polynomial $x^4 + 8 \in \mathbb{Q}[x]$. In a similar fashion we find that ψ_7 takes the element (29) to

$$\lambda_1 - \lambda_6\alpha - \lambda_3\alpha^2 + \lambda_8\alpha^3 + -\lambda_5 i - \lambda_1 i\alpha + \lambda_7 i\alpha^2 + \lambda_4 i\alpha^3,$$

so that K_7 consists of those elements of L such that $\lambda_3 = \lambda_5 = 0$, $\lambda_4 = \lambda_8$ and $\lambda_2 = -\lambda_6$. We deduce that $K_7 = \mathbb{Q}(\gamma')$, where $\gamma' = \alpha - i\alpha$ is another root of the polynomial $x^4 - 8$.

It remains for us to determine K_3, K_8 and K_9, the fixed fields of the subgroups of order 4. These will all be extensions of $\mathbb{Q}$ of degree 2, since $[E : K] = |\mathrm{Aut}_K(E)| = 4$ means that $[K : F] = 2$. Thus each of these fields is the result of adjoining to $\mathbb{Q}$ a root of an appropriate quadratic polynomial, and since a quadratic factorizes completely if it has one factor of degree 1, we deduce that these fields are splitting fields for the relevant polynomials. Thus they are normal extensions of $\mathbb{Q}$, which tells us that the groups $\mathcal{H}_3$, $\mathcal{H}_8$ and $\mathcal{H}_9$ must be normal subgroups of $\mathcal{G}$. This is a fact which we could have easily checked directly (and indeed it is a general theorem of group theory that a subgroup of index 2 in any group is necessarily normal).

The 4-cycle $(1, 2, 3, 4)$ takes α to $i\alpha$ and $i\alpha$ to $-\alpha$, and so we see that i is fixed. The general element (29) is taken to

$$\lambda_1 - \lambda_6\alpha - \lambda_3\alpha^2 + \lambda_8\alpha^3 + \lambda_5 i + \lambda_2 i\alpha - \lambda_7 i\alpha^2 - \lambda_4 i\alpha^3,$$

and so is fixed if and only if $\lambda_2 = \lambda_3 = \lambda_4 = \lambda_6 = \lambda_7 = \lambda_8 = 0$. So $K_3 = \mathbb{Q}(i)$ is the splitting field of $x^2 + 1$. The field K_8 consists of all elements of L that are fixed by both ψ_2 and ψ_6; so

$$K_8 = \{\lambda + \mu\alpha^2 \mid \lambda, \mu \in \mathbb{Q}\} = \mathbb{Q}(\sqrt{2}),$$

the splitting field of $x^2 - 2$. And K_9 consists of those elements of L that are fixed by both ψ_5 and ψ_7; we find that

$$K_8 = \{\lambda + \mu i\alpha^2 \mid \lambda, \mu \in \mathbb{Q}\} = \mathbb{Q}(i\sqrt{2}),$$

the splitting field of $x^2 + 2$.

As a second example, let $\zeta = \cos(2\pi/30) + i\sin(2\pi/30)$, a complex thirtieth root of 1, and let $E = \mathbb{Q}(\zeta)$. Although ζ is a root of $x^{30} - 1$, this is not the minimal polynomial of ζ over $\mathbb{Q}$, which, as we shall see, actually has degree 8. Note, however, the thirty complex roots of $x^{30} - 1$ are the powers ζ^k of ζ, where $0 \le k \le 29$. (Recall that $\zeta^k = \cos(2k\pi/30) + i\sin(2k\pi/30)$.) Since these all lie in $\mathbb{Q}(\zeta)$, and no proper subfield of $Q(\zeta)$ contains them all, we conclude that $E = \mathbb{Q}(\zeta)$ is the splitting field for $x^{30} - 1$ over $\mathbb{Q}$. Hence it is a normal extension of $\mathbb{Q}$. To determine the Galois group $\mathrm{Gal}(E : \mathbb{Q})$, consider the possible effect on ζ of an arbitrary $\mathbb{Q}$-automorphism ϕ of E. Certainly $\phi\zeta$ must be another root of $x^{30} - 1$, and so we must have $\phi\zeta = \zeta^k$ for some k with $0 \le k \le 29$. But since $\zeta^m \ne 1$ if m is not divisible by thirty, we must also have that $\zeta^{km} \ne 1$ if m is not divisible by thirty. In fact this means that $\gcd(k, 30) = 1$, for if $\gcd(k, 30) = d > 1$ then $(\zeta^k)30/d = (\zeta^{30})^{k/d} = 1$ despite the fact that $30/d$ is not a multiple of 30. The conclusion is that the only possible values for m are 1, 7, 11, 13, 17, 19, 23 and 29. Since a $\mathbb{Q}$-automorphism of $\mathbb{Q}(\zeta)$ is completely determined by its effect on ζ we deduce that $\mathrm{Gal}(E : F)$ has at most eight elements, and so $[E : \mathbb{Q}] \le 8$.

To show that $[E : \mathbb{Q}]$ is not less than 8 needs a little subtlety. Observe that ζ^6 is a fifth root of 1, and hence (since it is not 1) a root of the polynomial $x^4 + x^3 + x^2 + x + 1 \in \mathbb{Q}[x]$. We saw earlier that for all primes p the polynomial $\sum_{j=1}^{p-1} x^j$ is irreducible in $\mathbb{Q}[x]$; this was done by a change of variable followed by an application of Eisenstein's Criterion. So $x^4 + x^3 + x^2 + x + 1$ is the minimal polynomial of ζ^6 over $\mathbb{Q}$, and hence ζ^6 generates an extension of $\mathbb{Q}$ of degree 4. It remains for us to check that ζ itself is not in the field generated by $\omega = \zeta^6$, for this will show that $E : \mathbb{Q}(\omega)] \ge 2$ and hence that $[E : \mathbb{Q}] \ge 2[\mathbb{Q}(\omega) : \mathbb{Q}] = 8$.

The Galois group $\mathrm{Gal}(\mathbb{Q}(\omega) : \mathbb{Q})$ has order 4, consisting of automorphisms id, σ, ρ and τ, defined by

$$\mathrm{id}\omega = \omega, \quad \sigma\omega = \omega^2, \quad \rho\omega = \omega^3, \quad \tau\omega = \omega^4.$$

We see that $\sigma^2 = \tau$, and we deduce that the group is cyclic. Hence its only subgroup of order 2 is $\{\mathrm{id}, \tau\}$, and the fixed field of this is the only degree 2 extension of $\mathbb{Q}$ contained in $\mathbb{Q}(\omega)$. Now

$(\omega + \omega^4)^2 = 2 + \omega^2 + \omega^3 = 1 - \omega - \omega^4$, and we deduce that $\omega + \omega^4$, which is in the fixed field of τ, is a root of $x^2 + x - 1$. So $\mathbb{Q}(\omega + \omega^4) = \mathbb{Q}(i\sqrt{5})$ is the only degree 2 extension of Q contained in $\mathbb{Q}(\omega)$. Suppose now, for a contradiction, that $\zeta\mathbb{Q}(\omega)$. Then $\zeta^5 \in \mathbb{Q}(\omega)$. Now $\zeta^5 = \cos(\pi/3) + i\sin(\pi/3) = (1 + i\sqrt{3})/2$ is a root of the polynomial $x^2 - x + 1$, and so generates a degree 2 extension of $\mathbb{Q}$, which must therefore equal $\mathbb{Q}(i\sqrt{5})$. Thus there exist rational numbers a and b such that $a + bi\sqrt{5} = 1 + i\sqrt{3}$. Equating real and imaginary parts we conclude that $\sqrt{(}3/5) = b \in \mathbb{Q}$. But, of course, $\sqrt{(}3/5)$ is irrational (since, for instance, it is a root of $5x^2 - 3 \in \mathbb{Q}[x]$, which is irreducible by Eisenstein's Criterion).

After all this, we are able to say definitely that for each $j \in S = \{1, 7, 11, 13, 17, 19, 23, 29\}$ there is an element $\phi_j \in \mathrm{Gal}(E : \mathbb{Q})$ (where $E = \mathbb{Q}(\zeta)$) such that $\phi_j\zeta = \zeta^j$. Let j, $k \in S$ and choose $l \in S$ such that $l \equiv jk \pmod{30}$ (which is possible since $\gcd(jk, 30) = 1$). Then

$$(\phi_j\phi_k)\zeta = \phi_j(\phi_k\zeta) = \phi_j(\zeta^k) = (\phi_j\zeta)^k = (\zeta^j)^k = \zeta^{kj} = \zeta^l = \phi_l\zeta$$

and so we deduce that $\phi_j\phi_k = \phi_l$. Since $jk = kj$ the same argument shows that $\phi_k\phi_j = \phi_l$, and, in particular, $\mathcal{G} = \mathrm{Gal}(E : \mathbb{Q})$ is Abelian. If we write $g = \phi_7$ and $h = \phi_{11}$ then it is readily checked that g generates a cyclic subgroup of order 4 and h generates a cyclic subgroup of order 2, and each element of $\mathcal{G}$ is uniquely expressible as a product xy where x is a power of g and y a power of h. In the terminology of group theory, $\mathcal{G}$ is the *direct product* of its subgroups $\langle g\rangle$ and $\langle h\rangle$. Although we shall not go into this further, it is a relatively easy exercise to determine the subgroups of $\mathcal{G}$ and hence the subfields of E containing Q.

The hardest part of the example we have just done was the proof that the degre of the extension was 8 and not 4. But even without doing this, the proof that the Galois group is Abelian would still work. Indeed, the same idea works for nth roots of 1 for any positive integer n. If we let $\zeta_n = e^{2i\pi/n} \in \mathbb{C}$, then $\mathbb{Q}(\zeta_n)$ is a splitting field for $x^n - 1$ over $\mathbb{Q}$. Every $\mathbb{Q}$-automorphism ϕ of $\mathbb{Q}(\zeta_n)$ is uniquely determined by $\phi\zeta_n = \zeta_n^k$, where k is an integer such that $\gcd(k, n) = 1$. If j, k are two such integers and ϕ_j, ϕ_k corresponding $\mathbb{Q}$-automorphisms of $\mathbb{Q}(\zeta_n)$, then $\phi_j\phi_k = \phi_k\phi_j = \phi_l$, where $l \equiv jk \pmod{n}$ and $\phi_l\zeta_n = \zeta_n^l$. Thus we have the following proposition.

Proposition (13.19): *Let n be a positive integer and $\zeta_n = e^{2i\pi/n)} \in \mathbb{C}$. Then $E_n = \mathbb{Q}(\zeta_n)$ is a normal extension of $\mathbb{Q}$ such that* $\mathrm{Gal}(E_n : \mathbb{Q})$ *is Abelian. The degree $[E_n : \mathbb{Q}]$ is at most equal to the number of positive integers k such that $k \leq n$ and* $\gcd(k, n) = 1$.

In fact the degree is always equal to the number of positive integers k such that $k \leq n$ and $\gcd(k, n) = 1$, but we shall not prove this.

Solution by radicals

We aim to prove that there exists a polynomial equation of degree 5 over $\mathbb{Q}$ which is not soluble by radicals, and the strategy is to relate the problem to Galois groups. We start with a lemma which deals with Galois groups related to equations of the form $x^k - a = 0$.

Lemma (13.20): *Let F be a subfield of $\mathbb{C}$ which contains all the complex kth roots of 1, and let $t \in \mathbb{C}$ be a root of the polynomial $x^k - a \in F[x]$ (where a is any element of F). Then $F(t)$ is a normal extension of F and* $\mathrm{Gal}(F(t) : F)$ *is cyclic.*

Proof. If t' is another root of $x^k - a$ then $(t^{-1}t')^k = a^{-1}a = 1$, and so $t' = t\zeta$ where ζ is a kth root of 1. Since $\zeta \in F$ it follows that $t' \in F(t)$, and hence $x^k - a$ splits into linear factors over $F(t)$. On the other hand, no proper subfield of $F(t)$ contains the root t of $x^k - a$, and so $F(t)$ is the splitting field of $x^k - a$ over F. Hence it is a normal extension of F.

Now let $\phi \in \mathrm{Gal}(F(t) : F)$ be arbitrary. Then ϕt must be root of $x^k - a$, and so $\phi t = \zeta t$ for some kth root ζ of 1. Since t generates $F(t)$ as an extension of F, the automorphism ϕ is uniquely determined by ϕt, and hence by ζ. Let G be the set of all those kth roots of 1 for which there is an element $\phi = \phi_\zeta$ in $\mathrm{Gal}(F(t) : F)$ such that $\phi_\zeta t = \zeta t$. If η, $\zeta \in G$ then since $\zeta \in F$ we have $\phi_\eta \zeta = \zeta$, and thus

$$(\phi_\eta \phi_\zeta)t = \phi_\eta(\phi_\zeta t) = \phi_\eta(\zeta t) = (\phi_\eta \zeta)(\phi_\eta t) = \zeta(\eta t) = \phi_{\eta\zeta} t.$$

Hence $\phi_\eta \phi_\zeta = \phi_{\eta\zeta}$. It follows that there is a bijective mapping $G \to \mathrm{Gal}(F(t) : F)$ given by $\zeta \mapsto \phi_\zeta$ for all $\zeta \in G$, and this map preserves multiplication. So G is a group isomorphic to $\mathrm{Gal}(F(t) : F)$. So G is a subgroup of the cyclic group generated by $e^{2i\pi/k}$, and is also cyclic by Exercise 47. Thus $\mathrm{Gal}(F(t) : F)$ is cyclic. □

We now come to the key theorem, giving the group-theoretic condition for the solubility of a polynomial equation.

Theorem (13.21): *If $f(x) \in \mathbb{Q}[x]$ and the equation $f(x) = 0$ is soluble by radicals, then* $\mathrm{Gal}(E : \mathbb{Q})$ *is a soluble group, where E is the splitting field for $f(x)$ over $\mathbb{Q}$.*

Proof. Suppose that $f(x) = 0$ is soluble by radicals. Then by Definition (11.3) there is a chain of fields $\mathbb{Q} = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n = K$ such that $f(x)$ splits into linear factors over K and, for each j from 1 to n, there is an element $t_j \in F_j$ such that $F_j = F_{j-1}(t_j)$ and $t_j^{k_j} \in F_{j-1}$ for some positive integer k_j. We start by defining $k = k_1 k_2 \cdots k_n$ and $L_0 = \mathbb{Q}(\zeta)$, where $\zeta = e^{2i\pi/k}$. By Proposition (13.19) the group L_0 is a normal extension of $\mathbb{Q}$ and $\mathrm{Gal}(L_0 : \mathbb{Q})$ is Abelian. Since $e^{2i\pi/k_j}$ is a power of ζ we see that L_0 contains all the complex k_jth roots of 1, for all j from 1 to n. Now recursively define $L_j = L_{j-1}(t_j)$ for $j = 1, 2, \ldots, n$. Lemma (13.20) yields that L_j is a normal extension of L_{j-1} and $\mathrm{Gal}(L_j : L_{j-1})$ is cyclic for each j. Furthermore, an easy induction shows that $F_j \subseteq L_j$ for each $j \in \{0, 1, \ldots, n\}$: this holds for $j = 0$ since $F_0 = \mathbb{Q}$ and $L_0 = \mathbb{Q}(\zeta)$, and for $j > 1$ if we have $F_{j-1} \subseteq L_{j-1}$ then it follows that $F_j = F_{j-1}(t_j) \subseteq L_{j-1}(t_j) = L_j$. Hence $K \subseteq L_n$, and so $f(x)$ splits into linear factors over L_n. In particular, the splitting field E is contained in E_n.

Let $E_{-1} = \mathbb{Q}$, and for each $j \in \{0, 1, \ldots, n\}$ let $E_j = L_j \cap E$. Observe that $E_n = E$. We seek to prove that E_j is a normal extension of E_{j-1} (for each $j \in \{0, 1, \ldots, n\}$). So suppose that $p(x) \in E_{j-1}[x]$ is irreducible and has a root $u \in E_j$. Since E is the splitting field for $f(x)$ over E_{j-1} we know that E is a normal extension of E_{j-1}, and hence all the roots of $p(x)$ are contained in E. Now since $p(x) \in L_{j-1}[x]$, and u is a root of $p(x)$, it follows that the minimal polynomial $q(x)$ of u over L_{j-1} is a divisor of $p(x)$. In particular, all the roots of $q(x)$ are roots of $p(x)$, and so lie in E. So $q(x) = (x - u_1)(x - u_2) \cdots (x - u_r)$ for some $u_1, u_2, \ldots, u_r \in E$. It follows that $q(x) \in E[x]$. But by definition the coefficients of $q(x)$ lie in the field L_{j-1}, and so it follows that these coefficients are in $L_{j-1} \cap E = E_{j-1}$. So $q(x)$ is a divisor of $p(x)$ in $E_{j-1}(x)$, and since $p(x)$ is irreducible in $E_{j-1}[x]$ it follows that $q(x)$ and $p(x)$ are associates. As $q(x)$ is the minimal polynomial of u over L_{j-1} it is irreducible in $L_{j-1}(x)$, and since it has a root u in L_j, which is a normal extension of L_{j-1}, it must split over L_j. So all the roots of $p(x)$—which are the same as the roots of $q(x)$—are in L_j. But we have already seen that they are all in E; so they are all in $L_j \cap E = E_j$. So we have shown that every irreducible $p(x) \in E_{j-1}[x]$ with a root in E_j splits over E_j, and thus by Proposition (13.15) it follows that E_j is a normal extension of E_{j-1}.

If $\phi \in \mathrm{Gal}(L_j : L_{j-1}) = \mathrm{Aut}_{L_{j-1}}(L_j)$ then the restriction of ϕ to E_j yields an E_{j-1}-monomorphism $E_j \to \mathbb{C}$, which must take E_j to E_j since E_j is a normal extension of E_{j-1}. So there is an element $\phi' \in \mathrm{Aut}_{E_{j-1}}(E_j)$ such that $\phi' u = \phi u$ for all $u \in E_j$. We seek to prove that every element of $\mathrm{Aut}_{E_{j-1}}(E_j)$ is obtained in this fashion from some element ϕ of $\mathrm{Aut}_{L_{j-1}}(L_j)$. Clearly if we set $\mathcal{H}$ to be the set of all elements of $\mathcal{G} = \mathrm{Aut}_{E_{j-1}}(E_j)$ which are so obtained, then $\mathcal{H}$ is a subgroup of $\mathcal{G}$. Now let

$$M = \mathrm{Fix}(\mathcal{H}) = \{\, u \in E_j \mid \phi u = u \text{ for all } \phi \in \mathrm{Gal}(L_j : L_{j-1}) \,\}.$$

But the Main Theorem of Galois Theory tells us that the only elements of L_j that are fixed by all elements of the Galois group $\mathrm{Gal}(L_j : L_{j-1})$ are the elements of L_{j-1}, as $\mathrm{Fix}(\mathrm{Gal}(E : K) = K$ for all the intermediate fields K, including $K = L_{j-1}$. So $M \subseteq L_{j-1}$. But by definition $M \subseteq E$, and so we conclude that $M \subseteq L_{j-1} \cap E = E_{j-1}$. Since $\mathcal{H}$ is by definition a group of E_{j-1}-automorphisms of E_j it follows that $M = \mathrm{Fix}(\mathcal{H}) = E_{j-1}$, and hence, by the Main Theorem, $\mathcal{H} = \mathrm{Gal}(E_j : E_{j-1})$. That is, every element of $\mathrm{Gal}(E_j : E_{j-1})$ is obtained as the restriction of an element of $\mathrm{Gal}(L_j : L_{j-1})$, as claimed.

Suppose now that $\theta_1, \theta_2 \in \mathrm{Gal}(E_j : E_{j-1})$. Then we may choose $\phi_1, \phi_2 \in \mathrm{Gal}(L_j : L_{j-1})$ whose restrictions are θ_1, θ_2. Since $\phi_1\phi_2 = \phi_2\phi_1$ we deduce that for all $u \in E_j$,

$$(\theta_1\theta_2)u = \theta_1(\theta_2 u) = \phi_1(\phi_2 u) = (\phi_1\phi_2)u = (\phi_2\phi_1)u = \phi_2(\phi_1 u) = \theta_2(\theta_1)u = (\theta_2\theta_1)u,$$

and so $\theta_1\theta_2 = \theta_2\theta_1$. So $\mathrm{Gal}(E_j : E_{j-1})$ is Abelian.

Thus we have obtained a chain of fields

$$\mathbb{Q} = E_{-1} \subseteq E_0 \subseteq E_1 \subseteq \cdots \subseteq E_n = E$$

such that each is a normal extension of the preceding and the Galois groups are all Abelian. Note also that E is the splitting field for $f(x)$ over each E_j, and hence is a normal extension of each E_j. We define $\mathcal{G} = \mathrm{Gal}(E : \mathbb{Q})$ and $\mathcal{G}_j = \mathrm{Gal}(E : E_j)$ for each j. Then each $\mathcal{G}_j$ is a subgroup of $\mathcal{G}$, and these subgroups form a decreasing chain

$$\mathcal{G} = \mathcal{G}_{-1} \geq \mathcal{G}_0 \geq \mathcal{G}_1 \geq \cdots \geq \mathcal{G}_n = \{\mathrm{id}\}. \qquad (30)$$

(For example, $\mathcal{G}_n$ consists of all E_n-automorphisms of E_n, and the identity is the only such thing. And $\mathcal{G}_j \leq \mathcal{G}_{j-1}$ since every automorphism of E that fixes all elements of E_j fixes all elements of E_{j-1}.) Now the Main Theorem of Galois Theory, applied to the situation

$$E_{j-1} \subseteq E_j \subseteq E$$

in which E_j is a normal extension of E_{j-1}, yields that $\mathrm{Gal}(E : E_j)$ is a normal subgroup of $\mathrm{Gal}(E : E_{j-1})$ and the quotient group $\mathrm{Gal}(E : E_{j-1})/\mathrm{Gal}(E : E_j)$ is isomorphic to $\mathrm{Gal}(E_j : E_{j-1})$. That is, $\mathcal{G}_{j-1}/\mathcal{G}_j$ is isomorphic to $\mathrm{Gal}(E_j : E_{j-1})$, which we have shown to be Abelian. So in the chain of subgroups Eq.(30) each successive term is normal in the preceding, and the quotient groups are Abelian. Hence, by Definition (12.18), the group $\mathcal{G}$ is soluble, as required. □

We shall now give an example of a polynomial $f(x) \in \mathbb{Q}[x]$ such that $\mathrm{Gal}(E : \mathbb{Q}) \cong S_5$, where E is the splitting field of $f(x)$ over $\mathbb{Q}$. Since S_5 is not a soluble group, it follows from Theorem (13.21) that the equation $f(x) = 0$ is not soluble by radicals.

Let $f(x) = 2x^5 - 10x + 5 \in \mathbb{Q}[x]$. Eisenstein's Criterion tells us that $f(x)$ is irreducible over $\mathbb{Q}$. By first year calculus we find that $f(x)$ has exactly three real roots: the derivative $f'(x) = 10x^4 - 10$ has only two real roots (at ± 1), and since the two turning points of $f(x)$ are at $(1, -3)$ and $(-1, 13)$ it follows that there is exactly one root greater than 1, exactly one between -1 and 1, and exactly one less than -1. So

$$f(x) = (x - t_1)(x - t_2)(x - t_3)(2x^2 + rx + s)$$

for some $t_1, t_2, t_3, r, s \in \mathbb{R}$, and $2x^2 + rx + s$ has two non-real complex roots t_4 and t_5 (which are conjugates of each other). Let $E = \mathbb{Q}(t_1, t_2, t_3, t_4, t_5) \subseteq \mathbb{C}$, the splitting field for $f(x)$ over $\mathbb{Q}$.

Each $\mathbb{Q}$-automorphism of E permutes the roots of $f(x)$, and the automorphism is uniquely determined by the permutation since the roots generate E as an extension of $\mathbb{Q}$. Let G be the subgroup of S_5 consisting of all those permutations $\sigma \in S_5$ such that there is a Q-automorphism of E satisfying $t_j \mapsto t_{\sigma j}$ (for $1 \leq j \leq 5$). The mapping from $\mathrm{Gal}(E : \mathbb{Q}) = \mathrm{Aut}_{\mathbb{Q}}(E)$ to G which takes each automorphism to the corresponding permutation is an isomorphism.

Step 1. $(4,5) \in G$.

Proof. The assertion is that there exists $\phi \in \mathrm{Aut}_{\mathbb{Q}}(E)$ such that $\alpha t_4 = t_5$ and $\alpha t_5 = t_4$. Defining $\alpha t = \bar{t}$ (the complex conjugate of t) does the trick. Complex conjugation is a $\mathbb{Q}$-automorphism of $\mathbb{C}$, and it takes E to itself since it fixes t_1, t_2 and t_3 and interchanges t_4 and t_5. □

Step 2. *Suppose that $(h,j) \in G$ and $\tau \in G$ satisfies $\tau h = k$ and $\tau j = l$. Then $(k,l) \in G$.*

Proof. Let $\gamma \in \mathrm{Aut}_{\mathbb{Q}}(E)$ correspond to $(h,j) \in G$, so that γ swaps t_h and t_j and fixes the other roots, and let $\beta \in \mathrm{Aut}_{\mathbb{Q}}(E)$ correspond to $\tau \in G$, so that $\beta t_h = t_k$ and $\beta t_j = t_l$. Then

$$(\beta\gamma\beta^{-1})t_k = (\beta\gamma)t_h = \beta t_j = t_l,$$
$$(\beta\gamma\beta^{-1})t_l = (\beta\gamma)t_j = \beta t_h = t_k.$$

Moreover, if $m \notin \{k,l\}$ then $\beta^{-1}t_m \notin \{\beta^{-1}t_k, \beta^{-1}t_l\} = \{t_h, t_j\}$; so γ fixes $\beta^{-1}t_m$, and

$$(\beta\gamma\beta^{-1})t_m = \beta(\gamma(\beta^{-1}t_m)) = \beta(\beta^{-1}t_m) = t_m.$$

So $\beta\gamma\beta^{-1}$ swaps t_k and t_l and fixes the others. □

Step 3. *For each j, $k \in \{1,2,3,4,5\}$ there is a $\sigma \in G$ such that $\sigma k = j$.*

Proof. There exists a $\mathbb{Q}$-isomorphism $\phi \colon \mathbb{Q}(t_k) \to \mathbb{Q}(t_j)$ with $\phi t_k = t_j$ since t_k and t_j have the same minimal polynomial over $\mathbb{Q}$. By Proposition (13.13) there is an extension of ϕ to a $\mathbb{Q}$-monomorphism $E \to \mathbb{C}$, and since E is a normal extension of $\mathbb{Q}$ the monomorphism in question must take E to itself. That is, we obtain a $\mathbb{Q}$-automorphism of E with $t_k \mapsto t_j$. □

Fix $j \in \{1,2,3,4,5\}$. In view of Step 3, we may choose a permutation $\sigma_j \in G$ with $\sigma_j 5 = j$. Let $\sigma_j 4 = k$, and note that $k \neq j$. By Steps 1 and 3 we see that $(j,k) \in G$ (since $(4,5) \in G$ and $\sigma_j \in G$ satisfies $\sigma_j 5 = j$ and $\sigma_j 4 = k$). Now j was arbitrary; so we have shown that for each $j \in \{1,2,3,4,5\}$ there exists $k \neq j$ such that $(j,k) \in G$. Thus, for some a, b, c we have that $(1,a)$, $(2,b)$, $(3,c)$, $(4,5) \in G$. Our aim is to show that $G = S_5$; that is, we aim to show that all elements of S_5 are in G.

Step 4. *There exist j, k, l, distinct from each other, such that both (j,k) and (j,l) are in G.*

Proof. If $a = 4$ then we can take $j = 4$, $k = 1$ and $l = 5$. If $a = 5$ we can take $j = 5$, $k = 4$ and $l = 1$. If either b or c is 4 or 5 then the same applies with 2 in place of 1: for example, if $b = 4$ then we can take $j = 4$, $k = 2$ and $l = 5$. So we are left with the possibility that $a, b, c \in \{1,2,3\}$. Then $(1,a)$, $(2,b)$ and $(3,c)$ must overlap, so to speak. For example, suppose that $a = 2$. Then if $c = 1$ we have that $(1,3)$, $(1,2) \in G$, and so we may take j, k, l to be 1, 3, 2 respectively. On the other hand, if $c = 2$ then $(2,3)$, $(2,1) \in G$, and we may take j, k, l to be 2, 3, 1. Suppose alternatively that $a = 3$. Then if $b = 1$ we may take j, k, l to be 1, 3, 2, and if $b = 3$ then we may take j, k, l to be 3, 1, 2. □

Step 5. *With j, k, l as in Step 4, all six permutations in S_5 which permute j, k, l amongst themselves and fix the other two elements of $\{1,2,3,4,5\}$ are in G.*

Proof. We are given that (j,k), $(j,l) \in G$; so since G is closed under multiplication of permutations, we deduce that $(j,l,k) = (j,k)(j,l) \in G$ and $(j,k,l) = (j,l)(j,k) \in G$, and also $(j,l)(j,k)(j,l) = (k,l) \in G$. It is trivial that the identity is in G, and this now accounts for all the six. □

Let us now renumber the roots so that $\{j,k,l\} = \{3,4,5\}$. Choose $\tau \in G$ with $\tau 5 = 2$ (possible by Step 3 above). Then $\tau 4 = m$ and $\tau 3 = h$ (for some m, h). By Step 2, since $(5,4)$, $(5,3) \in G$, we obtain $(2,m)$, $(2,h) \in G$. But m and h cannot both be 1 (since $\tau 4 \neq \tau 3$), and so $(2,j) \in G$ for some $j \in \{3,4,5\}$. And by exactly the same argument, using 1 in place of 2, we deduce that $(1,k) \in G$ for some $k \in \{3,4,5\}$.

It is now easy to deduce that every transposition (r,s) is in G. Suppose first that $j = 3$. Then $(2,5) = (3,5)(2,3)(3,5) \in G$, and $(2,4) = (3,4)(2,3)(3,4) \in G$. So $(2,3)$, $(2,4)$, $(2,5)$ are all in G. The same works if $j = 4$ or 5. By the same reasoning with 1 in place of 2 we deduce that $(1,3)$, $(1,4)$, $(1,5)$ are all in G. This gives us six of the ten transpositions, and we already knew that $(3,4)$, $(4,5)$, $(3,5) \in G$, making nine. The last one is $(1,2)$, and it too must be in G since $(1,2) = (1,3)(2,3)(1,3)$.

Since G contains all the transpositions it must be the whole of S_5, since every permutation can be expressed as a product of transpositions. (This is simply a statement of the fact that a row of n objects can be rearranged into any order by a sequence of operations consisting of interchanging pairs of objects.) So $\mathrm{Gal}(E : \mathbb{Q})$ is isomorphic to $S_5 = G$, and hence is not soluble. By Theorem (13.21) the equation $2x^5 - 10x + 5$ is not soluble by radicals.

The converse of Theorem (13.21) is also true: if the Galois group $\mathrm{Gal}(E : F)$ of the splitting field E of $f(x) \in F[x]$ is a soluble group, then the equation $f(x) = 0$ is soluble by radicals. To see this, first adjoin to F all the kth roots of 1, for $k = 3, 4, \ldots, \deg(f(x))$. It is fairly easy to show that if the Galois group of the splitting field was soluble before doing this, it will still be soluble after. In other words, we may as well assume to begin with that F contains all these roots of unity. (We are looking for a formula for the roots of $f(x)$ which uses only field operations and radicals; we do not mind if $\sqrt[k]{1}$ appears in the formula.)

Solubility of the Galois group tells us that there is a decreasing chain of subgroups of the Galois group such that each is a normal subgroup of the preceding one and the quotient groups of successive terms in the chain are all Abelian. It is a (not difficult) theorem of group theory that the chain of subgroups can then be refined, by the insertion of extra subgroups, so that the quotient groups of successive terms are cyclic of prime order. The fixed fields of these subgroups form an increasing chain

$$F = F_0 \subseteq F_1 \subseteq F_2 \subseteq \cdots \subseteq \cdots F_n = E$$

such that each is a normal extension of the preceding and $\mathrm{Gal}(F_j : F_{j-1})$ is cyclic of prime order for each j. Our next proposition shows that in this situation $F_j = F_{j-1}(\sqrt[p]{t})$ for some p and some $t \in F_{j-1}$. It follows that the splitting field E is obtained by successively adjoining to F the pth roots of elements, for various values of p, which is what it means to say that $f(x) = 0$ is soluble by radicals.

Proposition (13.22): *Suppose that E, F are subfields of $\mathbb{C}$ such that E is a normal extension of F. Suppose that* $\mathrm{Gal}(E : F)$ *is cyclic of order p, where p is a prime number, and suppose that F contains all the complex pth roots of 1. Then $E = F(a)$ for some a such that $a^p = t \in F$. Furthermore, E is the splitting field over F of the polynomial $x^p - t$.*

Proof. Let $\omega = e^{2\pi i/k} \in F$, and let $\alpha \in \mathrm{Aut}_F(E)$ be a generator of the Galois group. That is, the powers α^j of α, for $0 \leq j < p$, are all the F-automorphisms of E. We prove first that the functions

α^j are linearly independent over E; that is, there is no nontrivial solution of

$$\lambda_0\alpha^0 + \lambda_1\alpha^1 + \lambda_2\alpha^2 + \cdots + \lambda_{p-1}\alpha^{p-1} = 0$$

with $\lambda_0, \lambda_1, \ldots, \lambda_{p-1} \in E$.

Suppose for a contradiction that a nontrivial solution exists, and choose one such that the number of nonzero coefficients λ_j is as small as possible. Leaving out the zero terms, we have

$$\mu_1(\beta_1 t) + \mu_2(\beta_2 t) + \cdots + \mu_k(\beta_k t) = 0 \tag{31}$$

for all $t \in E$, where $\beta_1, \beta_2, \ldots, \beta_k$ are distinct elements of $\{\, \alpha^j \mid 0 \leq j < p \,\}$, and the coefficients $\mu_j \in E$ are all nonzero. Clearly $k \geq 2$, since β_1 is not the zero function. Since $\beta_2 \neq \beta_1$ we may choose $u \in E$ such that $\beta_2 u \neq \beta_1 u$, and now replacing t by tu in Eq.(31), and using the fact that the β_j are all automorphisms, we obtain

$$\mu_1(\beta_1 u)\beta_1 t) + \mu_2(\beta_2 u)(\beta_2 t) + \cdots + \mu_k(\beta_u)(\beta_k t) = 0. \tag{32}$$

If we multiply both sides of Eq.(31) by $\beta_1 u$ and subtract from Eq.(32) we get

$$\mu_2(\beta_2 u - \beta_1 u)(\beta_2 t) + \mu_3(\beta_3 u - \beta_1 u)(\beta_3 t) + \cdots + \mu_{p-1}(\beta_{p-1} u - \beta_1 u)(\beta_{p-1} t) = 0 \tag{33}$$

for all $t \in E$. This is an equation of the same form as Eq.(31), but with fewer terms. It is still a nontrivial relationship since the coefficient $\mu_2(\beta_2 u - \beta_1 u)$ is nonzero. This contradicts our choice of Eq.(31).

In particular the above shows that $\mathrm{id} + \omega\alpha + \omega^2\alpha^2 + \cdots + \omega^{p-1}\alpha^{p-1}$ is not the zero function on E, and so we may choose $b \in E$ such that $a \neq 0$, where

$$a = b + \omega(\alpha b) + \omega^2(\alpha^2 b) + \cdots + \omega^{p-1}(\alpha^{p-1} b).$$

Since α fixes ω we see that

$$\alpha a = \alpha b + \omega(\alpha^2 b) + \cdots + \omega^{p-1} b = \omega^{-1} a$$

(since α^p is the identity) and $\omega^p = 1$. So

$$a(\alpha a)(\alpha^2 a)\cdots(\alpha^{p-1} a) = a(\omega^{-1} a)(\omega^{-2} a)\cdots(\omega^{-p+1} a) = \omega^{p(1-p)/2} a^p,$$

and since the left hand side of this equation is fixed by α, and hence by all powers of α, we deduce that $\omega^{p(1-p)/2} a^p$ is in F (the fixed field of $\mathrm{Gal}(E : F)$). If $p = 2$ then $\omega^{p(1-p)/2} = -1$, and if $p > 2$ then $\omega^{p(1-p)/2} = 1$; so in either case it follows that $a^p \in F$. On the other hand a itself is not in F since $\alpha a \neq a$. Hence $F(a)$ is an extension of F with $F \subsetneq F(a) \subseteq E$. Now $\mathrm{Gal}(E : F)$, being cyclic of prime order, has no proper subgroups, and so there are no fields lying strictly between F and E. Hence we must have $F(a) = E$. Obviously a is a root of $x^p - a^p$, which is in $F[x]$ since $a^p \in F$, and furthermore E splits this polynomial since its other roots are all of the form $\omega^j a$, and these all lie in E. □

The reason why polynomial equations of degree 2, 3 and 4 are soluble by radicals, while those of degree 5 or more may not be, is because the groups S_2, S_3 and S_4, unlike S_n for higher values of n, are soluble. Let us review the process for solving cubic equations that we described previously, in the light of the theory we have just been through, and then investigate quartics.

For the group S_3 the subset $A = \{\mathrm{id}, (1,2,3), (1,3,2)\}$ is a normal subgroup such that S_3/A is cyclic of order 2 and A is cyclic of order 3. Suppose that F is a subfield of $\mathbb{C}$ containing $\omega = e^{2\pi i/3}$, and $f(x) \in F[x]$ has degree 3. Let $E \subseteq \mathbb{C}$ be the splitting field for $f(x)$ over F. From the point of view of solving the equation $f(x) = 0$, the worst case scenario is that $\mathrm{Gal}(E : F) = S_3$. Any process which solves the equation if the Galois group is S_3 will also work if the Galois group is a proper subgroup of S_3, the only difference being that obtaining the various cube roots or square roots that are required may not necessitate extending the field. So let us proceed under the assumption that the Galois group is S_3.

Let $t_1 \in \mathbb{C}$ be a root of $f(x)$. Then 1, t_1, t_1^2 form a basis for $F(t_1)$ as a vector space over F, and $f(x) = (x - t_1)g(x)$ for some quadratic polynomial $g(x)$ with coefficients in $F(t_1)$. If t_2 is a root of $g(x)$ then 1, t_2 form a basis of $F(t_1, t_2)$ over $F(t_1)$, and since the remaining root t_3 is automatically in $F(t_1, t_2)$ (since $x - t_3 = f(x)/(x - t_1)(x - t_2)$) we see that $F(t_1, t_2) = E$, the splitting field. Furthermore,

$$1,\ t_1,\ t_1^2,\ t_2,\ t_2t_1,\ t_2t_1^2 \tag{34}$$

form a basis for E over F.

The theory indicates that the first step in the process of solving the equation should be to construct the fixed field of the subgroup $\mathcal{A}$. This will be a field K such that $F \subseteq K \subseteq E$, with $\mathrm{Gal}(E : K) \cong \mathcal{A}$ and $\mathrm{Gal}(K : F) \cong S_3/\mathcal{A}$. Write $\alpha = (1,2,3)$, a generator of the group $\mathcal{A}$, and observe that if $u \in E$ is arbitrary then $u + \alpha u + \alpha^2 u$ is fixed by α, and hence is an element of K. To find elements which span K as a vector space over F it suffices to apply this for each of the basis elements in Eq.(34) above. In this way we discover that $K = F(v)$, where $v = t_2t_1^2 + t_3t_2^2 + t_1t_3^2$. It is a simple task to square v and thus find a quadratic equation over F of which v is a root.

By Proposition (13.22) we know that $E = K(\sqrt[3]{t})$ for some $t \in K$. The proof of (13.22) gives a method for finding a suitable t: if we choose any $b \in E$ and put $a = b + \omega(\alpha b) + \omega^2(\alpha^2 b)$, then $t = a^3$ will be in the field K. If we put $b = t_1$ then $\alpha b = t_2$ and $\alpha^2 b = t_3$, giving

$$t = a^3 = (t_1 + \omega t_2 + \omega^2 t_3)^3,$$

and it is straightforward to express this in terms of v, which we have already found. Now since 1, a and $a^2 = t/a$ form a basis for E as a vector space over K, the roots t_1, t_2 and t_3 can be expressed as linear combinations of 1, a and a^2, where the coefficients are elements of K. Finding the coefficients is a matter of solving simultaneous linear equations, and we essentially did this in our previous calculations.

Consider now the quartic equation $x^4 - S_1x^3 + S_2x^2 - S_3x + S_4 = 0$, and suppose that its roots are t_1, t_2, t_3 and t_4. The key observation is that the group S_4 has a normal subgroup $\mathcal{H} = \{\mathrm{id}, (12)(34), (13)(24), (14)(23)\}$. The group $\mathcal{H}$ is Abelian, and $S_4/\mathcal{H}$ is isomorphic to S_3. We start by finding the fixed field H of the group $\mathcal{H}$, which we do by observing that if u is an arbitrary element of the splitting field $E = F(t_1, t_2, t_3)$ then

$$u + (12)(34)u + (13)(24)u + (14)(23)u \in H.$$

Applying this first with $u = t_2t_3$, then with $u = t_1t_3$, and then with $u = t_1t_2$, yields that

$$\begin{aligned} r_1 &= t_1t_4 + t_2t_3 \\ r_2 &= t_1t_3 + t_2t_4 \\ r_3 &= t_1t_2 + t_3t_4 \end{aligned}$$

are elements of H. It can be seen that each permutation of t_1, t_2, t_3, t_4 gives rise to a permutation of r_1, r_2, r_3; for example, interchanging t_3 and t_4 fixes r_3 and interchanges r_1 and r_2. (This mapping from permutations of 1, 2, 3, 4 to permutations of 1, 2, 3 is in fact a homomorphism from S_4 to S_3 whose kernel is the normal subgroup $\mathcal{H}$.) It follows that the elements of the Galois group $\mathrm{Gal}(E : F)$ fix the polynomial $(x - r_1)(x - r_2)(x - r_3)$. Thus the coefficients of this polynomial must lie in the field F, and with a little calculation we find that indeed

$$\begin{aligned} r_1 r_2 r_3 &= S_3^2 - 4S_4 S_2 + S_1^2 S_4, \\ r_1 r_2 + r_1 r_3 + r_2 r_3 &= S_1 S_3 - 4S_4, \\ r_1 + r_2 + r_3 &= S_2. \end{aligned}$$

Since we already know how to solve cubics, we can solve

$$x^3 - S_2 x^2 + (S_1 S_3 - 4S_4)x - (S_3^2 - 4S_4 S_2 + S_1^2 S_4)$$

and thus find r_1, r_2 and r_3.

Now let K be the fixed field of the group $\mathcal{K} = \{\mathrm{id}, (12)(34)\}$, a subgroup of index 2 in $\mathcal{H}$. Elements of K will be roots of quadratic equations over H. The element $t_1 + (12)(34)t_1 = t_1 + t_2$ is in K, and $(13)(24) \in \mathcal{H}$ swaps $t_1 + t_2$ and $t_3 + t_4$, which is also in K. So $t_1 + t_2$ and $t_3 + t_4$ are the roots of a quadratic polynomial with coefficients in H. In fact, $(x - t_1 - t_2)(x - t_3 - t_4) = x^2 - S_1 x + (r_1 + r_2)$. We have found r_1 and r_2; so we can solve this quadratic to find $t_1 + t_2$ and $t_3 + t_4$. Since we can similarly find $t_1 + t_3$ and $t_2 + t_4$, and also $t_1 + t_4$ and $t_2 + t_3$, we can now easily get the roots t_i, since (for instance) $t_1 = (1/2)((t_1 + t_2) + (t_1 + t_3) + (t_1 + t_4) - S_1)$.

Examples

We close with two examples which are beyond the scope of the Maths 392F course, and are included only for interested readers who may wish to pursue the subject further. The examples show how to find polynomials in $\mathbb{Q}[x]$ with whose galois groups are, respectively, the quaternion group of order 8 and the nonabelian group of order 21.

If a group contains elements g' and f' which both have order 4 and which satisfy $(f')^2 = (g')^2$ and $f'g'f' = g'$ then it is not hard to show that f' and g' generate a subgroup of order 8. The complex matrices

$$\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

provide an example of such a pair of elements. The term “quaternion group of order 8” means any group isomorphic to this. We proceed to construct an extension of $\mathbb{Q}$ whose Galois group is generated by two elements g' and f' satisfying the specified relations.

Let $F = \mathbb{Q}(\sqrt{10}, \sqrt{26})$, a degree 4 normal extension of $\mathbb{Q}$. Let f be the $\mathbb{Q}$-automorphism of F given by

$$f(\sqrt{10}) = \sqrt{10} \qquad f(\sqrt{26}) = -\sqrt{26} \qquad f(\sqrt{65}) = -\sqrt{65},$$

let g be the $\mathbb{Q}$-automorphism of F given by

$$g(\sqrt{10}) = -\sqrt{10} \qquad g(\sqrt{26}) = \sqrt{26} \qquad g(\sqrt{65}) = -\sqrt{65},$$

and let $h = fg$, so that

$$h(\sqrt{10}) = -\sqrt{10} \qquad h(\sqrt{26}) = -\sqrt{26} \qquad h(\sqrt{65}) = \sqrt{65}.$$

Let $T = -(10 + 3\sqrt{10})(65 + 5\sqrt{65}) \in F$, and observe that T has no square root in F since $T < 0$ and F is real. Let $E = F(u)$, where u is a root of $x^2 - T$, so that E is a degree 8 extension of $\mathbb{Q}$. Define elements $u_1, u_2, u_3, u_4 \in E$ by

$$\begin{aligned}
u_1 &= u \\
u_2 &= \frac{10\sqrt{26}}{65 + 5\sqrt{65}}u \\
u_3 &= \frac{10\sqrt{10}\sqrt{26}}{(10 + 3\sqrt{10})(65 + 5\sqrt{65})}u \\
u_4 &= \frac{\sqrt{10}}{10 + 3\sqrt{10}}u.
\end{aligned}$$

Then we find that

$$\begin{aligned}
u_2^2 &= -\frac{2600(10 + 3\sqrt{10})}{65 + 5\sqrt{65}} = -(10 + 3\sqrt{10})(65 - 5\sqrt{65}) = f(T), \\
u_3^2 &= -\frac{26000}{(10 + 3\sqrt{10})(65 + 5\sqrt{65})} = -(10 - 3\sqrt{10})(65 - 5\sqrt{65}) = g(T), \\
u_4^2 &= -\frac{10(65 + 5\sqrt{65})}{10 + 3\sqrt{10}} = -(10 - 3\sqrt{10})(65 + 5\sqrt{65}) = h(T).
\end{aligned}$$

Hence $\pm u_1, \pm u_2, \pm u_3, \pm u_4$ are the eight roots of $p(x) = (x^2 - T)(x^2 - f(T))(x^2 - g(T))(x^2 - h(T))$, and this polynomial has coefficients in $\mathbb{Q}$ since it is fixed by all elements of the Galois group of F over $\mathbb{Q}$. Hence E is a splitting field for $p(x)$ over $\mathbb{Q}$, and in particular it is a normal extension of $\mathbb{Q}$. Since $F = \mathbb{Q}(T)$, clearly $E = \mathbb{Q}(u)$; furthermore, $p(x)$ is the minimal polynomial of u over $\mathbb{Q}$.

Let $f' \colon E \to E$ be the $\mathbb{Q}$-automorphism defined by $f'(u) = u_2$, and let $g' \colon E \to E$ be the $\mathbb{Q}$-automorphism defined by $g'(u) = u_3$. Then $f'(T) = f'(u^2) = u_2^2 = f(T)$, and similarly $g'(T) = g'(u^2) = u_3^2 = g(T)$; so f' and g' are extensions of f and g respectively. Now

$$\begin{aligned}
f'(u_1) &= u_2 \\
f'(u_2) &= \frac{10f(\sqrt{26})}{65 + 5f(\sqrt{65})}f'(u) = \left(\frac{-10\sqrt{26}}{65 - 5\sqrt{65}}\right)\left(\frac{10\sqrt{26}}{65 + 5\sqrt{65}}\right)u = -u_1 \\
f'(u_3) &= \frac{10f(\sqrt{10})f(\sqrt{26}}{(10+3f(\sqrt{10}))(65+5f(\sqrt{65}))}f'(u) = \frac{-10\sqrt{10}\sqrt{26}10\sqrt{26}}{(10+3\sqrt{10})(65-5\sqrt{65})(65+5\sqrt{65})}u = -u_4 \\
f'(u_4) &= \frac{f(\sqrt{10})}{10 + 3f(\sqrt{10})}f'(u) = \frac{\sqrt{10}10\sqrt{26}}{(10 + 3\sqrt{10})(65 + 5\sqrt{65})}u = u_3,
\end{aligned}$$

and similarly

$$\begin{aligned}
g'(u_1) &= u_3 \\
g'(u_2) &= \frac{10g(\sqrt{26})}{65 + 5g(\sqrt{65})}g'(u) = \frac{10\sqrt{26}}{65 - 5\sqrt{65}}u_3 = u_4 \\
g'(u_3) &= \frac{10g(\sqrt{10})g(\sqrt{26})}{(10 + 3g(\sqrt{10}))(65 + 5g(\sqrt{65}))}g'(u) = \frac{-10\sqrt{10}\sqrt{26}}{(10 - 3\sqrt{10})(65 - 5\sqrt{65})}u_3 = -u_1 \\
g'(u_4) &= \frac{g(\sqrt{10})}{10 + 3g(\sqrt{10})}g'(u) = \frac{-\sqrt{10}}{(10 - 3\sqrt{10})}u_3 = -u_2.
\end{aligned}$$

It is now easily checked that $(f')^2 = (g')^2$ satisfies $u_i \mapsto -u_i$ for all i, and that $f'g'f' = g'$, as required.

There is a nonabelian group of order 21, which can be described as follows: it is generated by two elements f and g such that f has order 7, g has order 3, and $g^{-1}fg = f^2$. If $\gamma = e^{2\pi i/7}$ then the following complex matrices provide an example of a pair of elements satisfying these relations:

$$\begin{pmatrix} \gamma & 0 & 0 \\ 0 & \gamma^2 & 0 \\ 0 & 0 & \gamma^4 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

We proceed to find an extension of $\mathbb{Q}$ whose Galois group is generated by two elements satisfying the specified relations.

Let F be a splitting field over $\mathbb{Q}$ for the polynomial $f(x) = x^{14} - 8x^7 + 128$. Modulo 7 we find that $(y+4)^{14} - 8(y+4)^7 + 128$ is congruent to $(y^7+4)^2 - 8(y^7+4) + 2 \equiv y^{14}$, and modulo 49 the constant coefficient is $4^{14} - 8 \times 4^7 + 128 = 2^7(2^{21} - 2^{10} + 1) \equiv 2^7(2 \times (-5)^2 + 5 + 1) \not\equiv 0$. So, by Eisenstein's Criterion, $(y+4)^{14} - 8(y+4)^7 + 128$ is irreducible over $\mathbb{Q}$, and hence so is $f(x)$. Let $\gamma = 4 + 4\sqrt{7}i$, a root of $x^2 - 8x + 128$, and let α be a 7th root of γ, so that α is a root of $f(x)$. Observe that $(\alpha\bar{\alpha})^7 = (4 + 4\sqrt{7}i)(4 - 4\sqrt{7}i) = 128$, and so $\bar{\alpha} = 2\alpha^{-1}$. Now if ω is a primitive complex 7th root of 1 then $\omega^i\alpha$ and $2\omega^i\alpha^{-1}$ are roots of $f(x)$ for each i from 1 to 7. Note that these are 14 distinct roots, since $\omega^i\alpha = 2\omega^j\alpha^{-1}$ would give $\alpha^{14} \in \mathbb{Q}$, contradicting the fact that $f(x)$ is the minimal polynomial of α. Since $\sqrt{7}i \in \mathbb{Q}(\omega)$ it follows that the minimal polynomial of ω over $\mathbb{Q}(\sqrt{7}i)$ has degree 3; furthermore since $\sqrt{7}i \in \mathbb{Q}(\alpha)$, which is a degree 14 extension of $\mathbb{Q}$, it follows that ω generates a degree 3 extension of $\mathbb{Q}(\alpha)$. Hence $[F : \mathbb{Q}] = 42$.

Let us now investigate the Galois group of this extension. The mapping $\mathbb{Q}(\alpha) \to \mathbb{Q}(\omega\alpha)$ given by $\alpha \mapsto \omega\alpha$ extends to an automorphism f of F which fixes ω, since the minimal polynomial of ω over $\mathbb{Q}(\alpha)$ has its coefficients in $\mathbb{Q}(\sqrt{7}i) = \mathbb{Q}(\alpha^7)$, and is therefore fixed by $\alpha \mapsto \omega\alpha$. Since $\bar{\alpha} \in \mathbb{Q}(\alpha)$ there is an automorphism of $\mathbb{Q}(\alpha)$ with $\alpha \mapsto \bar{\alpha}$. (Indeed, this automorphism is simply complex conjugation.) Since it takes $(x-\omega)(x-\omega^2)(x-\omega^4)$, the minimal polynomial of ω over $\mathbb{Q}(\alpha)$, to $(x-\omega^3)(x-\omega^5)(x-\omega^6)$, it extends to an automorphism g of F such that $g(\omega) = \omega^3$. Now $g^2(\omega) = \omega^2$ and $g^2(\alpha) = \alpha$, and it follows easily that g^2 has order 3 and g has order 6. Clearly f has order 7. Now

$$(gfg^{-1})(\alpha) = (gf)(\bar{\alpha}) = (gf)(2\alpha^{-1}) = g(2(\omega\alpha)^{-1}) = g(\omega^{-1}\bar{\alpha}) = \omega^{-3}\alpha = f^4(\alpha),$$
$$(gfg^{-1})(\omega) = (gf)(\omega^5) = g(\omega^5) = \omega = f^4(\omega),$$

so that conjugation by g induces an automorphism of order 3 on the cyclic group generated by f. The Galois group is thus the direct product of a group of order 2 with the nonabelian group of order 21. The fixed field of the element of order 2, which is the intersection of F with $\mathbb{R}$ since the element of order 2 is complex conjugation, will be a normal extension of $\mathbb{Q}$ whose Galois group is nonabelian of order 21.

It remains to find a polynomial which has the field in question as its splitting field. Since $(x-\alpha)(x-\bar{\alpha})$ has coefficients in $\mathbb{Q}(\alpha+\bar{\alpha})$ (since $\alpha\bar{\alpha} \in \mathbb{Q}$), it follows that $\mathbb{Q}(\alpha)$ is a degree 2 extension of $\mathbb{Q}(\alpha+\bar{\alpha})$, which is therefore a degree 7 extension of $\mathbb{Q}$. If E is a splitting field for the minimal polynomial of $\alpha+\bar{\alpha}$, then $E(\alpha)$ is a degree 2 extension of E (the minimal polynomial again being $(x-\alpha)(x-\bar{\alpha})$, and hence a normal extension of $\mathbb{Q}$ since E is. So $E(\alpha)$ contains all the algebraic conjugates of α, and it follows that $E(\alpha) = F$. Hence E is the requisite degree 21 extension of $\mathbb{Q}$. Now using $\alpha\bar{\alpha} = 2$ and $\alpha^7 = 4 + \sqrt{7}i$ we find that

$$\begin{aligned}
(\alpha+\bar{\alpha})^7 &= 8 + 14(\alpha^5+\bar{\alpha}^5) + 84(\alpha^3+\bar{\alpha}^3) + 280(\alpha+\bar{\alpha}),\\
(\alpha+\bar{\alpha})^5 &= (\alpha^5+\bar{\alpha}^5) + 10(\alpha^3+\bar{\alpha}^3)40(\alpha+\bar{\alpha}),\\
(\alpha+\bar{\alpha})^3 &= (\alpha^3+\bar{\alpha}^3) + 6(\alpha+\bar{\alpha}),
\end{aligned}$$

and therefore $\alpha + \bar{\alpha}$ is a root of

$$x^7 - 14x^5 + 56x^3 - 56x - 8.$$

So this polynomial has the required Galois group.