PMath 334 Notes - Rings and Fields

Laurent W. Marcoux

April 14, 2024

Preface

Preface to the Second Edition - April 14, 2024

A number of typos from the first edition have now been corrected. Presumably, many others remain, and there may even be new ones! Please read the preface below, and bring any remaining typos/errors to my attention.

In particular, I would like to thank the following people for already having brought some typos and/or errors to my attention: T. Dieudonné, W. Dong, M. Hürlimann, Z. Kun, M. Nguyen, J. Nordmann, A. Okell, N. Saloojee, S. Steel, J. Tzoganakis, Z. Wang, and X. Yin.

Preface to the First Edition - April 19, 2021

The following is a set of class notes for the PMath 334 course I am currently teaching at the University of Waterloo in January, 2021. They are a work in progress, and – this being the "first edition" – they are replete with typos. A student should approach these notes with the same caution he or she would approach buzz saws; *they can be very useful, but you should be alert and thinking the whole time you have them in your hands.* Enjoy.

Just one short comment about the Exercises at the end of each chapter. These are of varying degrees of difficulty. Some are much easier than Assignment Questions, and some are of a comparable level of difficulty. Those few Exercises that marked by three asterisks are definitely worth doing, as they are crucial to understanding the underlying concepts. The marked exercises are also of varying levels of difficulty, but it is better for the reader to discover some things on his/her own, since the reader will then understand and retain those things better. Besides, the only way to learn mathematics is to do mathematics.

In our humble opinion, an excellent approach to reading these notes is as follows.

- One first gathers the examples of rings from the second and third chapters. One then reads the statements of the theorems/propositions/corollaries, etc., and interprets those results for each of those examples. The purpose of the theory is to understand and unify the examples.

- To learn the proofs, we recommend that one read the statement of a given theorem or proposition, and tries to prove the result oneself. If one gets stuck at a certain point in the proof, one reads the proof until one gets past that point, and then one resumes the process of proving the result oneself.

Also, one should keep in mind that *if one doesn't know where to start, one can always start with the definition,* which means that one always knows where to start. Just saying.

I strongly recommend that the reader consult other textbooks as well as these notes. As ridiculous as this may sound, there are other people who can write as well as if not better than your humble author, and it is important that the reader find the source which best suits the reader. Moreover, by consulting multiple sources, the reader will discover results not covered in any single reference. I shall only mention two namely the book of I.N. Herstein [**Her69**], and a wonderful little book on Galois (and Field Theory) by I. Stewart [**Ste04**].

I would like to thank the following people for bringing some typos to my attention: T. Cooper, J. Demetrioff, R. Gambhir, A. Murphy (more than once), K. Na, and B. Virsik. Any remaining typos and mistakes are the fault of my colleagues... well.... ok, maybe they're partly my fault too.

Finally, I would like to thank Mr. Nic Banks for providing me with a short list of examples of Euclidean domains beyond the much shorter and apparently universally standard list which appears in so many elementary ring theory textbooks I consulted.

April 14, 2024

The reviews are in!

He is a writer for the ages, the ages of four to eight.

Dorothy Parker

This paperback is very interesting, but I find it will never replace a hardcover book - it makes a very poor doorstop.

Alfred Hitchcock

It was a book to kill time for those who like it better dead.

Rose Macaulay

That's not writing, that's typing.

Truman Capote

Only the mediocre are always at their best.

Jean Giraudoux

Contents

CHAPTER 1

A brief overview

Somewhere on this globe, every ten seconds, there is a woman giving birth to a child. She must be found and stopped.

Sam Levenson

1. Groups

1.1. Pure Mathematics is, we would argue, the study of mathematical objects and structures, and of the relationships between these. We seek to understand and classify these structures. This is a very vague statement, of course. What kinds of objects and structures are we dealing with, and what do we mean by "relationships"?

The objects and structures are many and varied. Vector spaces, continuous functions, integers, real and complex numbers, groups, rings, fields, algebras, topological spaces and manifolds are just a very tiny sample of the cornucopia of objects which come under the purview of modern mathematics. In dealing with various members of a single category – and here we are not using *category* in the strict mathematical sense, for categories are also a structure unto themselves – we often use functions or *morphisms* between them to see how they are related. For example, an injective function f from a non-empty set A to a non-empty set B suggests that B must have at least as many elements as A. (In fact, we take this as the definition of "at least as many elements" when A and B are infinite!)

1.2. The two principal structures we shall examine in this course are *rings* and *fields*. The typical approach in most books begins with the definition of a ring, for example, and then produces a (hopefully long) list of examples of rings to demonstrate their importance. This process is, however, essentially the inverse of how such a concept develops in the first place. In practice, one normally starts with a long list of examples of objects, each of which has its own intrinsic value. Through inspiration and hard work, one may discover certain commonalities between those objects, and one may observe that any other object which shares that commonality will – by necessity – behave in a certain way as suggested by the initial objects of interest. The advantage of doing this is that if one can prove that certain behaviour is a result of the commonality, as opposed to being specific to one of the examples,

then one may deduce that such behaviour will apply to *all* objects in this collection and beyond.

But enough of generalities. Let us illustrate our point through an example. Since both rings and fields exhibit a *group* structure under addition, let us momentarily digress to examine these.

1.3. Example. Consider the following list of structures.

(a) $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, 3, \ldots\}$, the set of **integers**. Note that $\mathbb{Z}$ admits a binary operation, called **addition**, which is denoted by $+$. To say that addition is a binary operation on $\mathbb{Z}$ means that $+$ is actually a *function* on the set of ordered pairs of integers $\mathbb{Z} \times \mathbb{Z}$ with range included in $\mathbb{Z}$. That is,

$$\begin{array}{rccc} + : & \mathbb{Z} \times \mathbb{Z} & \to & \mathbb{Z} \\ & (a, b) & \mapsto & a + b. \end{array}$$

We also say that $\mathbb{Z}$ is **closed** under addition.

(b) Consider next the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of **natural numbers**. Again, addition, also denoted by $+$, is a binary operation on $\mathbb{N}$: that is, we have a function

$$\begin{array}{rccc} + : & \mathbb{N} \times \mathbb{N} & \to & \mathbb{N} \\ & (a, b) & \mapsto & a + b. \end{array}$$

(c) Let $n \in \mathbb{N}$, and consider next the set of $n \times n$ complex matrices $\mathbb{M}_n(\mathbb{C})$. Given $A = [a_{i,j}]$ and $B = [b_{i,j}]$ in $\mathbb{M}_n(\mathbb{C})$, we can define the *sum* of A and B to be

$$A + B := [a_{i,j} + b_{i,j}].$$

In a similar manner to the first two examples, we find that $\mathbb{M}_n(\mathbb{C})$ is closed under addition.

(d) Let

$$\mathcal{C}([0, 1], \mathbb{R}) = \{f : [0, 1] \to \mathbb{R} : f \text{ is continuous}\}.$$

Given $f, g \in \mathcal{C}([0, 1], \mathbb{R})$, we may define their *sum* $f + g$ via

$$(f + g)(x) := f(x) + g(x) \text{ for all } x \in [0, 1].$$

From first-year calculus, we know that since f and g are each continuous on $[0, 1]$, so is $f + g$. Once again, we say that $\mathcal{C}([0, 1], \mathbb{R})$ is closed under addition.

(e) Let $\mathbb{Q} := \{\frac{p}{q} : p, q \in \mathbb{Z}, q \neq 0\}$ denote the set of rational numbers, and set $\mathbb{Q}^* := \mathbb{Q} \setminus \{0\}$. Given $r_1 = \frac{p_1}{q_1}, r_2 = \frac{p_1}{q_2} \in \mathbb{Q}^*$, the *product* $r_1 \cdot r_2 = \frac{p_1 p_2}{q_1 q_2} \in \mathbb{Q}^*$. Thus $\mathbb{Q}^*$ is closed under *multiiplication.*

(f) If we set $\mathbb{Q}^+ := \{r \in \mathbb{Q} : r > 0\}$, then $\mathbb{Q}^+$ is again closed under multiplication.

Examples (a), (c), (d), (e) and (f) have (at least the following) two things in common:

(i) in each case - there is a **neutral element**; that is, an element z in the set which satisfies $a + z = a = z + a$ for all a in the original set.

For example, in the case of $\mathbb{Z}$, we simply set $z = 0$, whereas for $\mathbb{M}_n(\mathbb{C})$, we need to set $z = [0_{i,j}]$, the zero matrix. In the case of $\mathcal{C}([0,1],\mathbb{R})$, we would set z to be the *zero function*, that is, the function $z(x) = 0$, $x \in [0,1]$.

In the case of $\mathbb{Q}^*$ and $\mathbb{Q}^+$, the number $1 = \frac{1}{1}$ serves as a neutral element under multiplication.

Note that $\mathbb{N}$ does not possess a neutral element. Regardless of which $z \in \mathbb{N}$ one chooses, $n + z \neq n$ for any $n \in \mathbb{N}$.

(ii) Note also that for each element a in each of the first, third and fourth examples, there is an element b such that $a + b = z = b + a$, where z is the neutral element from item (i) above. This coincides with what we normally refer to as the **negative** or the **additive inverse** of a. For example, in the case of $f \in \mathcal{C}([0,1],\mathbb{R})$, we set $-f$ to be the function $(-f)(x) = -(f(x))$, $x \in [0,1]$. Again – that $-f$ is continuous when f is continuous is proven in a first-year Calculus course.

Given $r = p/q \in \mathbb{Q}^*$ (in using this notation we are assuming that $p, q \in \mathbb{Z}$), we note that $p \neq 0 \neq q$, and so $s := q/p \in \mathbb{Q}^*$, and $sr = 1 = rs$. That is, the product of r and s yields the neutral element from above. Of course, we normally refer to s as the **(multiplicative) inverse** of r.

Once again, $\mathbb{N}$ fails to have this property.

1.4. Example. The next example appears, at least on the surface, rather different from those above. Let Δ denote an equilateral triangle, with vertices labelled 1, 2 and 3.

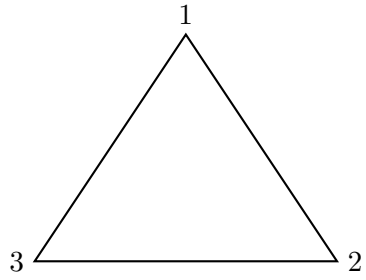

FIGURE 1. Δ

Let us think of Δ as sitting in the x,y-plane sitting inside the 3-dimensional real vector space $\mathbb{R}^3$. We equip $\mathbb{R}^3$ with its usual Euclidean distance (or **metric**) as follows: if $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3) \in \mathbb{R}^3$, the distance between x and y is defined as

$$d(x,y) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + |x_3 - y_3|^2}.$$

We first consider **isometries** of the triangle. By an isometry, we mean a map $\varphi : \Delta \to \mathbb{R}^3$ such that $d(x,y) = d(\varphi(x), \varphi(y))$. That is, φ **preserves distance**.

We then define a **symmetry** of Δ to be an isometry of Δ with the property that

$$\varphi(\Delta) = \{\varphi(x) : x \in \Delta\} = \Delta.$$

For example, we might rotate the triangle Δ *clockwise* through 0 radians, $2\pi/3$ radians, or through $4\pi/3$ radians about its centre. Let us call these three symmetries ϱ_0, ϱ_1 and ϱ_2 respectively.

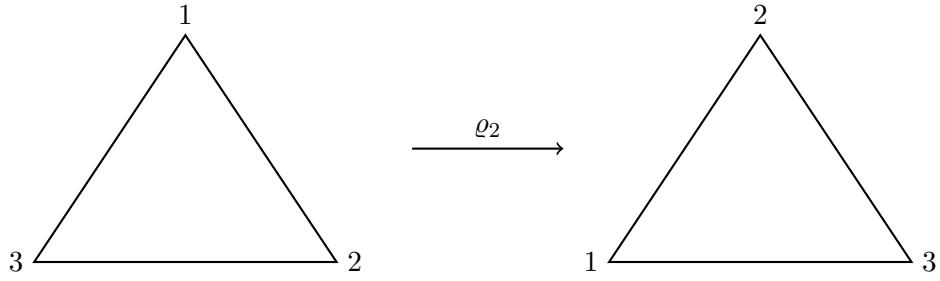

FIGURE 2. THE ACTION OF ϱ_2

We might also fix one of the vertices, for example the top vertex (labelled 1 in Figure 1), and reflect Δ along the vertical line from the top vertex to the midpoint of the horizontal line segment connecting the leftmost vertex (labelled 3 in Figure 1) to the rightmost vertex (labelled 2 in Figure 1). Let us call this symmetry φ_1, and the symmetries similarly obtained by fixing vertex 2 by φ_2, while that similarly obtained by fixing vertex 3 by φ_3.

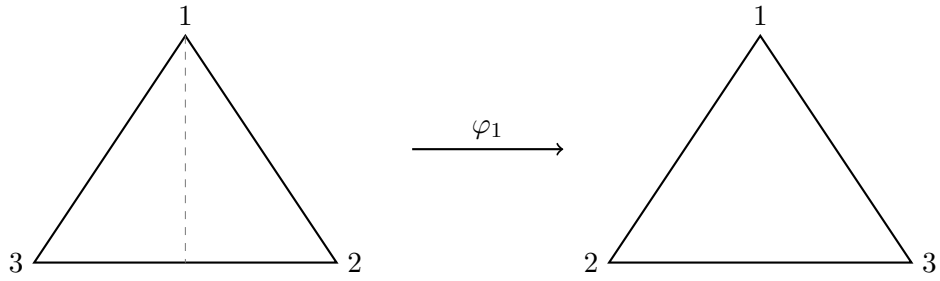

FIGURE 3. THE ACTION OF φ_1

We have so far discovered six distinct symmetries of Δ. Are there others?

A moment's thought will hopefully convince the reader that any symmetry must take distinct vertices of Δ to distinct vertices of Δ. There are only six ways of doing this: once we have chosen where to send vertex 1, there remain two choices for where to send vertex 2, and then vertex 3 must be sent to the remaining vertex. The total number of choices is therefore $6 = 3 \cdot 2 \cdot 1$. Since we already have produced six distinct symmetries above, we indeed have them all. Let us denote by $\textsc{symm}(\Delta)$ the set

$$\textsc{symm}(\Delta) = \{\varrho_0, \varrho_1, \varrho_2, \varphi_1, \varphi_2, \varphi_3\}$$

of all symmetries of the equilateral triangle Δ.

Here is where it gets interesting. Suppose that we choose to *compose* two of these symmetries, that is: first we perform one symmetry, and then another. For example, we might choose to first do ϱ_2, and then φ_1. What would the result look like?

We can answer this by considering what the maps do to each of the vertices. Note that ϱ_2 sends vertex 1 to vertex 3, vertex 2 to vertex 1 and vertex 3 to vertex 2. We abbreviate this to

$$\varrho_2(1) = 3, \qquad \varrho_2(2) = 1, \qquad \varrho_2(3) = 2.$$

Similarly, φ_1 takes vertex 1 to vertex 1, vertex 2 to vertex 3, and vertex 3 to vertex 2, which we may write as

$$\varphi_1(1) = 1, \qquad \varphi_1(2) = 3, \qquad \varphi_1(3) = 2.$$

Thus

$$\varphi_1 \circ \varrho_2(1) = \varphi_1(\varrho_2(1)) = \varphi_1(3) = 2,$$

and similiarly $\varphi_1 \circ \varrho_2(2) = 1$, $\varphi_1 \circ \varrho_2(3) = 3$.

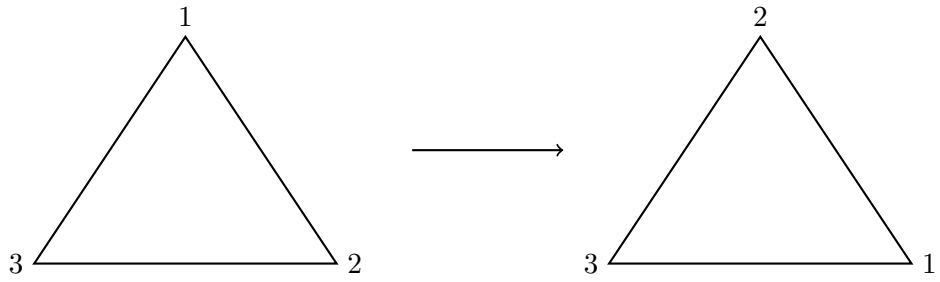

FIGURE 4. THE ACTION OF $\varphi_1 \circ \varrho_2$.

But this is precisely what happens if we simply apply φ_3 to Δ! In other words,

$$\varphi_1 \circ \varrho_2 = \varphi_3.$$

In analogy to the three examples from Example 1.3, a quick calculation shows that ϱ_0 serves as a neutral element for $\textsc{symm}(\Delta)$. Also, given any symmetry α in $\textsc{symm}(\Delta)$, there exists a second symmetry β such that $\beta \circ \alpha = \varrho_0 = \alpha \circ \beta$.

For example, if $\alpha = \varrho_1$, then we may set $\beta = \varrho_2$, while if $\alpha = \varphi_1$, then we choose $\beta = \varphi_1$ as well.

So, in terms of its behaviour using composition as the binary operation, $\textsc{symm}(\Delta)$ behaves more like $(\mathbb{Z}, +)$ than $(\mathbb{N}, +)$ does!

1.5. Each of the sets we have defined in Example 1.3 and Example 1.4 is interesting in its own right. But now we have discovered that, except for the example of the natural numbers, they all share a few properties in common. It is this kind of analysis that leads mathematicians to make definitions such as the following.

1.6. Definition. *A non-empty set G equipped with a binary operation*

$$\cdot : G \times G \to G$$

is said to be a ***group*** *if the following conditions hold:*

(a) *the operation $\cdot$ is* ***associative****: that is,*

$$(g \cdot h) \cdot k = g \cdot (h \cdot k) \text{ for all } g, h, k \in G.$$

(b) *There exists an element $e \in G$, called a* ***neutral element*** *or an* ***identity element*** *for G such that*

$$e \cdot g = g = g \cdot e \text{ for all } g \in G.$$

(c) *Given $g \in G$, there exists an element $h \in G$ such that*

$$g \cdot h = e = h \cdot g,$$

where e is the neutral element of G defined above. The element h is referred to as an ***inverse*** *of g.*

The group G is said to be ***abelian*** *if $g \cdot h = h \cdot g$ for all $g, h \in G$. It is customary to use "+" to denote the binary operation for abelian groups (instead of "$\cdot$").*

1.7. Remark. Since a group consists of the non-empty set G along with the binary operation $\cdot$, we normally write $(G, \cdot)$ to denote a group. *Informally*, when the binary operation is understood and there is no risk of confusion, we also refer to a *group* G, and we even drop the $\cdot$ from the notation, writing gh to denote $g \cdot h$.

We shall leave it as an exercise for the reader to show that if G is a group, then it admits exactly one identity element e, and that given $g \in G$, the element $h \in G$ satisfying $g \cdot h = e = h \cdot g$ is also uniquely defined. For this reason, we speak of *the* identity element of a group G, as well as *the* inverse of $g \in G$. We also adopt the notation g^{-1} to denote the unique inverse of g.

1.8. Example. Although we have not verified the associativity of addition in each of the examples from Example 1.3, this can indeed be done, and so it follows that $(\mathbb{Z}, +)$, $(\mathbb{M}_n(\mathbb{C}), +)$ and $(\mathcal{C}([0,1], \mathbb{R}), +)$ are all (abelian) groups under addition.

Furthermore, $(\textsc{symm}(\Delta), \circ)$ from Example 1.4 is a group under composition. The fact that $\varrho_2 \circ \varphi_1 = \varphi_2 \neq \varphi_3 = \varphi_1 \circ \varrho_2$ means that $\textsc{symm}(\Delta)$ is not abelian.

The set $\mathbb{N}$ of natural numbers fails to be a group under addition, since it admits neither a neutral element, nor inverses.

1.9. Definition. *A* ***subgroup*** *H of a group $(G, \cdot)$ is a non-empty subset of G which is a group using the binary operation $\cdot$ inherited from G.*

1.10. Example. Let $H = 2\mathbb{Z} = \{2m : m \in \mathbb{Z}\}$. Then H is a subgroup of $(\mathbb{Z}, +)$.

1.11. Example. Recall that $\textsc{symm}(\Delta) = \{\varrho_0, \varrho_1, \varrho_2, \varphi_1, \varphi_2, \varphi_3\}$ as defined in Example 1.4 is a group using composition $\circ$.

We leave it to the reader to verify that $H := \{\varrho_0, \varrho_1, \varrho_2\}$ is a group under $\circ$, and thus H is a subgroup of $\textsc{symm}(\Delta)$.

1.12. Example. The set $\mathbb{Q}^+ := \{q \in \mathbb{Q} : q > 0\}$ is a group under multiplication. While $\mathbb{Q}^+$ is a subset of $\mathbb{Q}$, it is not a subgroup of $(\mathbb{Q}, +)$, since we are not using the same binary operation. That is, in the first case we are using multiplication as our operation, while in the case of $(\mathbb{Q}, +)$, we are using addition. Since $1 \in \mathbb{Q}^+$ but $-1 \notin \mathbb{Q}^+$, we see that $\mathbb{Q}^+$ is not closed under addition, and so $(\mathbb{Q}^+, +)$ is *not* a group.

On the other hand, $(\mathbb{Q}^+, \cdot)$ *is* a subgroup of $(\mathbb{Q}^*, \cdot)$, where $\mathbb{Q}^* = \{q \in \mathbb{Q} : q \neq 0\}$.

Supplementary Examples

S1.1. Example. Let $n \in \mathbb{N}$. Let $\mathrm{GL}_n(\mathbb{C}) := \{T \in \mathbb{M}_n(\mathbb{C}) : T \text{ is invertible}\}$, and let $\cdot$ represent usual matrix multiplication.

Then $(\mathrm{GL}_n(\mathbb{C}), \cdot)$ is a group, and $I_n = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \in \mathbb{M}_n(\mathbb{C})$ is the identity of this group. We refer to $\mathrm{GL}_n(\mathbb{C})$ as the **general linear group** in $\mathbb{M}_n(\mathbb{C})$.

Also,

$$\mathrm{U}_n(\mathbb{C}) := \{U \in \mathbb{M}_n(\mathbb{C}) : U \text{ is unitary}\}$$

is a group, called the **unitary group** of $\mathbb{M}_n(\mathbb{C})$.

S1.2. Example. Let $n \in \mathbb{N}$. Then $\mathrm{SL}_n(\mathbb{C}) := \{T \in \mathbb{M}_n(\mathbb{C}) : \det T = 1\}$ is a group, again using usual matrix multiplication as the group operation.

This group is referred to as the **special linear group** of $\mathbb{M}_n(\mathbb{C})$.

S1.3. Example. More generally, let $\Omega \subseteq \mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$ be a multiplicative group. Let

$$\mathrm{D}_n(\Omega) := \{T \in \mathbb{M}_n(\mathbb{C}) : \det T \in \Omega\}.$$

Then $\mathrm{D}_n(\Omega)$ is a group, using usual matrix multiplication as the group operation.

S1.4. Example. The set of rational numbers $\mathbb{Q}$ is an abelian group under addition, while the set $\mathbb{Q}^*$ of non-zero rational numbers is an abelian group under multiplication. As we have seen, the set $\mathbb{Q}^+$ of *strictly positive* rational numbers is also an abelian group under multiplication.

S1.5. Example. Let $(\mathcal{V}, +, \cdot)$ be a vector space over $\mathbb{R}$. (Here, $\cdot$ refers to scalar multiplication.) Then $(\mathcal{V}, +)$ is an abelian group.

In particular, given a natural number n, $(\mathbb{R}^n, +)$ is an additive, abelian group.

This explains why the examples given in Example 1.3 are (abelian) groups.

S1.6. Example. Amongst the most important classes of groups are the so-called **permutation groups**.

Let $\varnothing \neq X$ be a non-empty set. A **permutation** of X is a bijective function $f : X \to X$. Typically, if $X = \{1, 2, 3, \ldots, n\}$ for some natural number n, then we use symbols such as σ or ϱ to denote permutations. We refer to the set of *all* permutations of X as the **symmetric group** of X and (here we shall) denote it by $\Sigma(X)$. The group operation is composition of functions.

Note that the identity map $\iota : X \to X$ given by $\iota(x) = x$, $x \in X$ is a bijection.

Given two bijections f, g of X, their composition $f \circ g$ is again a bijection, and $f \circ \iota = f = \iota \circ f$. Thus ι serves as the identity of $\Sigma(X)$ under composition. Also, since f is a bijection, given $y \in X$, we may write $y = f(x)$ for a unique choice of $x \in X$. Thus we may define the bijection $f^{-1} : X \to X$ via $f^{-1}(y) = x$. Clearly $f^{-1} \circ f = \iota = f \circ f^{-1}$, so that f^{-1} is the inverse for f under composition.

In general, $\Sigma(X)$ is non-abelian. (For which sets X is it abelian?)

A **permutation group** is any subset of $\Sigma(X)$ which is itself a group under composition.

The fact that we refer to the set of permutations as the *symmetric group*, and that we refer to the group $\textsc{symm}(\Delta)$ from Example 1.4 as a group of *symmetries* is not merely a coincidence. As we observed in the discussion of that Example, any symmetry of Δ is really just a permutation of the vertices, and any permutation of the vertices leads to a symmetry of Δ.

S1.7. Example. The set $\mathbb{Z}_n = \{0, 1, 2, \ldots, n-1\}$ is an abelian group under *addition modulo n*. It is an interesting exercise to prove that if $p \in \mathbb{N}$ is a prime number, then every group G with p elements behaves exactly like $\mathbb{Z}_p$ under addition. Technically speaking, $(\mathbb{Z}_p, +)$ is the unique group *of order p* (i.e. with p elements) *up to group isomorphism.*

To be explicit, if $p \in \mathbb{N}$ is a prime number and $G = \{g_1, g_2, \ldots, g_p\}$ is a group using the binary operation $\cdot$, then there exists a bijective map $\varrho : \mathbb{Z}_p \to G$ such that $\varrho(a+b) = \varrho(a) \cdot \varrho(b)$ for all $a, b \in \mathbb{Z}_p$. Such a map is referred to as a **group isomorphism**. The proof of this for $n = 2, 3$ or $n = 5$ is certainly within reach.

We shall have more to say about so-called *homomorphisms* of groups in Chapter 4.

S1.8. Example. If $(G, \cdot)$ and $(H, *)$ are groups, then so is $G \times H := \{(g, h) : g \in G, h \in H\}$ under the operation $\bullet$, where

$$(g_1, h_1) \bullet (g_2, h_2) := (g_1 \cdot g_2, h_1 * h_2).$$

We leave the verification of this to the reader.

S1.9. Example. The **discrete Heisenberg group** $H_3(\mathbb{Z})$ is the set of 3×3 matrices of the form

$$\begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix},$$

where $a, b, c \in \mathbb{Z}$, and where

$$\begin{bmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & x & y \\ 0 & 1 & z \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & a+x & y+az+b \\ 0 & 1 & c+z \\ 0 & 0 & 1 \end{bmatrix}.$$

If one replaces $\mathbb{Z}$ by $\mathbb{R}$, one obtains the **continuous Heisenberg group**.

S1.10. Example. The set $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$ is a group, using usual complex multiplication as the operation.

Appendix

A1.1. Niels Henrik Abel (5 August 1802 - 6 April 1829) is known for a wide variety of results that bear his name. Perhaps his most famous result is the proof that given a polynomial p of degree five over the real numbers, it is not possible to solve the equation $p(x) = 0$ in radicals. (The well-known **quadratic equation** does exactly this for polynomials of degree two. In the case of polynomials of degree three or four, solutions in radicals also exist, but they are not as simple, and are subsequently less known.)

Abel lived in poverty and died of tuberculosis. The **Abel Prize** in mathematics is named in his honour. (There is no Nobel Prize in mathematics.) Given how prestigious and merited the Nobel Peace Prize always is, a prize in the world of mathematics is... (we leave it to the reader to complete this sentence).

Exercises for Chapter 1

Exercise 1.1.
Let $(G,\cdot)$ be a group. Prove that if $e, f \in G$ and $gf = ge = g = eg = fg$ for all $g \in G$, then $e = f$. That is, prove that the identity element of a group is unique.

Exercise 1.2.
Let $(G,\cdot)$ be a group. Prove that if g, h_1, and $h_2 \in G$ and if $gh_2 = gh_1 = e = h_1g = h_2g$, then $h_1 = h_2$. This shows that the inverse of g is unique, allowing us to refer to **the** inverse of g, which we then denote by g^{-1}.

Exercise 1.3.
Let $(G,\cdot)$ be a group. Suppose that $g^2 := g \cdot g = e$ for all $g \in G$. Prove that G is abelian.

Exercise 1.4.
In Example 1.4, we determined that $(\textsc{symm}(\Delta), \circ)$ is a non-abelian group with six elements, and we determined all of its elements by observing what each symmetry does to the vertices of the triangle Δ.

(a) Determine $(\textsc{symm}(\square), \circ)$, the group of symmetries of a square – in particular, how many elements must it have?
(b) Is this group abelian?

Exercise 1.5.
Let Γ denote an isosceles triangle which is not equilateral. Determine

$$(\textsc{symm}(\Gamma), \circ).$$

Exercise 1.6.
A **permutation** of a non-empty set X is a bijective function $f : X \to X$. Denote by $\Sigma(X)$ the set of all permutations of X. As we saw in Example S1.6, using composition as the operation: $f \circ g(x) := f(g(x))$ for all $x \in X$, we have that $(\Sigma(X), \circ)$ is a group.

In the case where $X = \{1, 2, \ldots, n\}$, we may write a permutation by specifying its value at each of the elements of X as a $2 \times n$ matrix: for example,

$$\sigma = \begin{bmatrix} 1 & 2 & 3 & \cdots & n-1 & n \\ a_1 & a_2 & a_3 & \cdots & a_{n-1} & a_n \end{bmatrix}$$

represents the permutation $\sigma \in \Sigma(X)$ that satisfies $\sigma(j) = a_j$, $1 \le j \le n$.

(a) What is the relationship (if any) between $\Sigma(\{1,2,3\})$ and $\textsc{symm}(\Delta)$?
(b) What is the relationship (if any) between $\Sigma(\{1,2,3,4\})$ and $\textsc{symm}(\square)$)?

Exercise 1.7.

Let $\mathbb{Q}^* := \{q \in \mathbb{Q} : q \neq 0\}$. Let $\cdot$ represent the usual product of rational numbers. Prove that $(\mathbb{Q}^*, \cdot)$ is an abelian group.

Exercise 1.8.

For those of you with a background in linear algebra: let $\mathcal{V}$ be a vector space over $\mathbb{C}$ and let

$$\mathcal{L}(\mathcal{V}) := \{T : \mathcal{V} \to \mathcal{V} : T \text{ is linear}\}.$$

Set $\mathcal{G}(\mathcal{V}) := \{T \in \mathcal{L}(\mathcal{V}) : T \text{ is invertible}\}$. Using composition of linear maps $\circ$ as the binary operation, prove that $(\mathcal{G}(\mathcal{V}), \circ)$ is a group.

Exercise 1.9.

Let $(G, \cdot)$ be a group and g, h and $k \in G$. Suppose that

$$g \cdot h = g \cdot k.$$

Prove that $h = k$.

We say that the **cancellation law** holds for groups.

Exercise 1.10.

Let $n \in \mathbb{N}$ and let $\omega_n := e^{2\pi i/n} \in \mathbb{C}$. Let $\Omega_n := \{1, \omega, \omega^2, \ldots, \omega^{n-1}\}$. Prove that Ω_n is an abelian group under complex multiplication, and that if n is prime, then the only subgroups of Ω_n are $\{1\}$ and Ω_n itself.

CHAPTER 2

An introduction to Rings

I can speak Esperanto like a native.

Spike Milligan

1. Definitions and basic properties

1.1. As we indicated in the first Chapter, the method of writing down a definition of a mathematical structure followed by examples is not the way that such structures are typically discovered. The reason for teaching the material in this order, however, is expediency. It is simply that much faster.

Granted an unlimited amount of time, it would be instructive to approach an entire course through carefully selected examples which gradually lead the students to discover on their own the desired definitions, propositions and theorems. Over a period of twelve weeks, however, this is not practical, and so we submit to the common, if more prosaic practice of definition, example, theorem and proof.

The student whose primary goal is to learn the material (as opposed the student whose primary goal is to obtain a good grade) would be well-served to spend as much time as their other courses allow to understand what each of the results we shall prove means in the examples we give, as well as in examples that they should endeavour to discover on their own.

Let us keep in mind that the purpose of the theory is to try to unite and understand the examples, which should be important in their own right. We are not trying to motivate the examples using the definitions!

1.2. Definition. *A **ring** R is a non-empty set equipped with two binary operations, known as **addition** (denoted by $+$), and **multiplication** (denoted by $\cdot$, or simply by juxtaposition of two elements). The operations must satisfy (A1) - (A5) and (M1)-(M4) below:*

(A1) $a + b \in R$ *for all* $a, b \in R$.

(A2) $(a + b) + c = a + (b + c)$ *for all* $a, b, c \in R$.

(A3) *There exists a neutral element* $0 \in R$ *such that*

$$a + 0 = a = 0 + a \text{ for all } a \in R.$$

(A4) *Given $a \in R$, there exists an element $b \in R$ such that*

$$a + b = 0 = b + a.$$

(A5) *$a + b = b + a$ for all $a, b \in R$.*

Taken together, conditions (A1) - (A4) are the statement that $(R, +)$ is a group, and condition (A5) is the statement that $(R, +)$ is in fact abelian.

We also require that

(M1) *$a \cdot b \in R$ for all $a, b \in R$.*
(M2) *$a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all $a, b, c \in R$.*
(M3) *$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$ for all $a, b, c \in R$.*
(M4) *$(a + b) \cdot c = (a \cdot c) + (b \cdot c)$ for all $a, b, c \in R$.*

*We say that R is **unital** if there exists an element $1 \neq 0$ in R such that $1 \cdot a = a = a \cdot 1$ for all $a \in R$. This element 1 is also referred to as a **multiplicative identity**.*

1.3. Remark. Condition (M1) above states that R is closed under multiplication, (M2) says that multiplication is associative on R, while conditions (M3) and (M4) state that multiplication is left and right distributive with respect to addition.

We point out that some authors require that all rings be unital. It is true that given a non-unital ring, there is a way of attaching a multiplicative identity to it (this will eventually appear as an Assignment question), however we shall refrain from doing so because the abstract construction is not always natural in terms of the (non-unital) rings we shall consider.

We point out that the decision to require that $1 \neq 0$ is to ensure that $R = \{0\}$ is not considered a unital ring.

1.4. Example. Given an abelian group $(G, +)$, we can always turn it into a ring by setting defining $g_1 \cdot g_2 = 0$ for all $g_1, g_2 \in G$.

1.5. Examples. In some cases, the examples of (abelian) groups we studied in Chapter 1 were actually examples of rings, as we shall now see.

(a) In Example 1.1.3, we saw that $(\mathbb{Z}, +)$ is an abelian group under addition. That $(\mathbb{Z}, \cdot)$ satisfies (M1)-(M4) is a standard result. Thus $(\mathbb{Z}, +, \cdot)$ is a ring. Since $1 \in \mathbb{Z}$ and $1 \cdot a = a = a \cdot 1$ for all $a \in \mathbb{Z}$, $(\mathbb{Z}, +, \cdot)$ is a unital ring.

(b) Again, from Example 1.1.3, $\mathbb{M}_n(\mathbb{C})$ is an abelian group under addition. Given $A = [a_{i,j}]$ and $B = [b_{i,j}] \in \mathbb{M}_n(\mathbb{C})$, recall from linear algebra that we define

$$A \cdot B := [c_{i,j}],$$

where for each $1 \leq i, j \leq n$, we set $c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}$. That $(\mathbb{M}_n(\mathbb{C}), +, \cdot)$ satisfies (M1)-(M4) is a standard exercise from linear algebra. Note that $1 := I_n$, the $n \times n$ identity matrix, serves as a multiplicative identity for $(\mathbb{M}_n(\mathbb{C}), +, \cdot)$.

(c) Yet again, as in Example 1.1.3, we set

$$\mathcal{C}([0,1],\mathbb{R}) = \{f : [0,1] \to \mathbb{R} : f \text{ is continuous}\}.$$

As we saw there, $(\mathcal{C}[0,1],\mathbb{R}),+)$ is an abelian group using (point-wise) defined addition. Given $f, g \in \mathcal{C}([0,1],\mathbb{R})$, we may define their *product* to be fg, where

$$(fg)(x) := f(x)g(x) \text{ for all } x \in [0,1].$$

That $(\mathcal{C}([0,1],\mathbb{R}),+,\cdot)$ is a (unital) ring with these operations is left as an exercise for the reader.

1.6. Example. Let $n \in \mathbb{N}$. Recall from Math 135 that we define

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

equipped with addition $\dot{+}$ and multiplication $\otimes$, modulo n. That is, given $a, b \in \mathbb{Z}_n$, we define

$$a \dot{+} b = (a + b) \bmod n$$

and

$$a \otimes b := ab \bmod n.$$

For example, if $n = 12$, then $\mathbb{Z}_{12} = \{0, 1, 2, \ldots, 10, 11\}$ and

$$8 \dot{+} 9 = 17 \bmod 12 = 5,$$

while

$$8 \otimes 10 = 80 \bmod 12 = 8.$$

Of course, in practice, we keep the "$\bmod n$" in our heads, and we use the usual notation + for addition and $\cdot$ (or juxtaposition) for multiplication. Thus we typically write

$$8 + 9 = 5$$

and

$$8(10) = 8 \cdot 10 = 8$$

in $\mathbb{Z}_{12}$.

We leave it to the reader to verify that $(\mathbb{Z}_n,+,\cdot)$ is a ring. Note that it is commutative.

1.7. Examples. The following are also rings, and the verification of this is left to the reader.

(a) $(\mathbb{C},+,\cdot)$. The complex numbers.
(b) $(\mathbb{R},+,\cdot)$. The real numbers.
(c) $(\mathbb{Q},+,\cdot)$. The rational numbers.

(d) $(7\mathbb{Z}, +, \cdot)$. Here, you will recall that

$$7\mathbb{Z} := \{7x : x \in \mathbb{Z}\} = \{\ldots, -14, -7, 0, 7, 14, \ldots\}.$$

The addition and multiplication that we define on $7\mathbb{Z}$ are those inherited from the fact that $7\mathbb{Z} \subseteq \mathbb{Z}$.

More generally, for any $k \in \mathbb{Z}$, $k\mathbb{Z} := \{km : m \in \mathbb{Z}\}$ is a ring. This ring is unital if and only if $|k| = 1$; i.e. if and only if $k \in \{-1, 1\}$.

(e) $(\mathbb{Z}[\sqrt{2}], +, \cdot)$, where $\mathbb{Z}[\sqrt{2}] := \{a + b\sqrt{2} : a, b \in \mathbb{Z}\}$.

(f) $(\mathbb{Q}[i], +, \cdot)$, where $\mathbb{Q}[i] := \{a + bi : a, b \in \mathbb{Q}\}$, and where $i^2 = -1$.

1.8. Example. A similar construction allows us to define an example of a **group ring**.

Let $(G, \cdot)$ be a group, and R be a ring. We define the group ring $R\langle G\rangle$ as follows: the elements of $R\langle G\rangle$ are formal *finite* sums of the form

$$\sum_{j=1}^{m} r_j g_j,$$

where $r_j \in R$ and $g_j \in G$, $1 \leq j \leq m$.

The sum of two elements of $R\langle G\rangle$ is performed as follows: let $a = \sum_{j=1}^{p} r_j g_j$ and $b = \sum_{i=1}^{q} s_i h_i$ be two elements of $R\langle G\rangle$. Set $\mathcal{E} := \{g_1, g_2, \ldots, g_p\} \cup \{h_1, h_2, \ldots, h_q\}$. After a change of notation, we may rewrite $\mathcal{E} = \{y_1, y_2, \ldots, y_k\}$ where all of the y_j's are distinct (and $k \leq p + q$). We may then write $a = \sum_{j=1}^{k} a_j y_j$ and $b = \sum_{j=1}^{k} b_j y_j$, noting that it is entirely possible that $a_j = 0$ or that $b_j = 0$ for some of the j's.

The purpose of this was to write both a and b as a combination of the *same* set of group elements.

Then

$$a + b = \left(\sum_{j=1}^{k} a_j y_j\right) + \left(\sum_{j=1}^{k} b_j y_j\right) := \sum_{j=1}^{k} (a_j + b_j) y_j.$$

Multiplication is performed by allowing elements of R to commute with elements of G:

$$\left(\sum_{j=1}^{p} r_j g_j\right)\left(\sum_{i=1}^{q} s_i h_i\right) = \sum_{j=1}^{p}\sum_{i=1}^{q} r_j s_i (g_j h_i).$$

Let's have a look at this in action through two concrete examples of group rings.

(a) Let $G = \{e, \omega, \omega^2\}$, and equip G with the multiplication determined by the following table:

	e	ω	ω^2
e	e	ω	ω^2
ω	ω	ω^2	e
ω^2	ω^2	e	ω

The verification that G is a group is left to the reader. Suppose that $R = \mathbb{Q}$, the rational numbers. To see what addition and multiplication look like in $\mathbb{Q}\langle G\rangle$, consider the following. A general element of $\mathbb{Q}\langle G\rangle$ looks like $a = r_1 \cdot e + r_2\omega + r_3\omega^2$, where $r_1, r_2, r_3 \in \mathbb{Q}$. Suppose also that $b = s_1 e + s_2\omega + s_3\omega^2 \in \mathbb{Q}\langle G\rangle$. Then

$$\begin{aligned} a + b &= (r_1\, e + r_2\ \omega + r_3\ \omega^2) + (s_1 e + s_2\ \omega + s_3\ \omega^2) \\ &= (r_1 + s_1)\ e + (r_2 + s_2)\ \omega + (r_3 + s_3)\ \omega^2, \end{aligned}$$

and

$$\begin{aligned} a \cdot b &= (r_1\ e + r_2\ \omega + r_3\ \omega^2)(s_1 e + s_2\ \omega + s_3\ \omega^2) \\ &= r_1 s_1\ e^2 + r_1 s_2\ e\cdot\omega + r_1 s_3\ e\cdot\omega^2 + r_2 s_1\ \omega\cdot e + r_2 s_2\ \omega^2 + r_2 s_3\ \omega^3 \\ &\qquad + r_3 s_1 \omega^2 \cdot e + r_3 s_2 \omega^3 + r_3 s_3 \omega^4 \\ &= (r_1 s_1 + r_2 s_3 + r_3 s_2)\ e + (r_1 s_2 + r_2 s_1 + r_3 s_3)\ \omega + (r_1 s_3 + r_2 s_2 + r_3 s_1)\ \omega^2. \end{aligned}$$

(b) As a second example, let $n \in \mathbb{N}$. A matrix $P \in \mathbb{M}_n(\mathbb{C})$ is said to be a **permutation matrix** if every row and every column of P contains exactly one "1", and zeroes everywhere else. The reason for this nomenclature is that if we view elements of $\mathbb{M}_n(\mathbb{C})$ as linear transformations on $\mathbb{C}^n$ written with respect to the standard orthonormal basis $\mathcal{B} := \{e_1, e_2, \ldots, e_n\}$, then a permutation matrix corresponds to a linear transformation which permutes the basis $\mathcal{B}$; i.e. $Pe_j = e_{\sigma(j)}$, where $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ is a bijection.

Consider the set $\mathcal{P}_n$ of all $n \times n$ permutation matrices. We leave it to the reader to prove that $\mathcal{P}_n$ is a group. In the case where $n = 3$, we find that

$$\begin{aligned} \mathcal{P}_3 &= \{P_1, P_2, P_3, P_4, P_5, P_6\} \\ &= \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right\}. \end{aligned}$$

A general element of the group ring $\mathbb{Z}\langle\mathcal{P}_3\rangle$ looks like

$$a = \sum_{j=1}^{6} a_j P_j,$$

where $a_j \in \mathbb{Z},\ 1 \le j \le 6$. As an example of a product in this ring, consider

$$\begin{aligned} &(2P_2 + 4P_5)(3P_3 + 1P_4 - 7P_5) = \\ &\qquad = 6P_2P_3 + 2P_2P_4 - 14P_2P_5 + 12P_5P_3 + 4P_5P_4 - 28P_5P_5 \\ &\qquad = 6P_1 + 2P_5 - 14P_6 + 12P_6 + 4P_2 - 28P_1 \\ &\qquad = -22P_1 + 4P_2 + 2P_5 - 2P_6. \end{aligned}$$

1.9. Example. Let $\mathcal{V}$ be a vector space over $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Let

$$\mathcal{L}(\mathcal{V}) := \{T : \mathcal{V} \to \mathcal{V} : T \text{ is } \mathbb{K}\text{-linear}\}.$$

Then $(\mathcal{L}(\mathcal{V}), +, \circ)$ is a ring. That is, given $T, R \in \mathcal{L}(\mathcal{V})$, we define $T + R$ to be the map satisfying $(T + R)(v) = Tv + Rv$ for all $v \in \mathcal{V}$, while $T \circ R$ is defined through composition, namely $T \circ R(v) = T(Rv)$ for all $v \in \mathcal{V}$.

1.10. Remark. Notice that in the example above, $\mathcal{L}(\mathcal{V})$ is not only a ring, but it is in fact a vector space over $\mathbb{K}$. We say that it as an **algebra over** $\mathbb{K}$.

As a second example, note that $(\mathcal{C}([0,1], \mathbb{R}), +, \cdot)$ is an algebra over $\mathbb{R}$.

2. Polynomial Rings – a first look

2.1. One particular class of rings which will be of special importance to us is the class of polynomial rings. Later, we shall devote an entire chapter to studying factorisation in polynomial rings with coefficients in a field.

2.2. Definition.

(a) *Let $n \in \mathbb{N}$ and $x_1, x_2, \ldots, x_n$ be a finite collection of abstract symbols (also called **indeterminates**). The set $\{x_1, x_2, \ldots, x_n\}$ is often called an **alphabet**. A **word** in $\{x_1, x_2, \ldots, x_n\}$ is a formal product (i.e. a concatenation)*

$$W := x_{i_1} x_{i_2} \cdots x_{i_k}$$

*where $k \in \mathbb{N}$ and $i_j \in \{1, 2, \ldots, n\}$ for all $1 \leq j \leq k$. That is, it is just a concatenation of finitely many elements of the alphabet. The **length** of the word W above is defined to be k. By convention, we define W_0 to be the empty word, with the understanding that $W_0 W = W = W W_0$ for any word $W \in \mathcal{W}$.*

Let us denote by $\mathcal{W}_k$ the set of all words of length k in $\{x_1, x_2, \ldots, x_n\}$, and by $\mathcal{W}$ the set

$$\mathcal{W} := \cup_{k=0}^{\infty} \mathcal{W}_k.$$

Given a commutative, unital *ring R, we then define*

$$R\langle x_1, x_2, \ldots, x_n \rangle := \{r_0 \cdot 1 + \sum_{j=1}^{m} r_j W_j : m \in \mathbb{N}, r_j \in R \text{ and } W_j \in \mathcal{W}, 1 \leq j \leq m\},$$

*and refer to this as the **ring of polynomials in n non-commuting variables $x_1, x_2, \ldots, x_n$ over the commutative, unital ring R**. Alternatively, the terminology $R\langle x_1, x_2, \ldots, x_n \rangle$ is the **ring of polynomials in n non-commuting variables with coefficients in R** is often used.*

Given $A := \sum_{j=1}^{p} r_j W_j$ *and* $B := \sum_{i=1}^{q} s_i V_i \in R\langle x_1, x_2, \ldots, x_n\rangle$, *we set*

$$\begin{aligned} A \cdot B &= (\sum_{j=1}^{p} r_j W_j) \cdot (\sum_{i=1}^{q} s_i V_i) \\ &= \sum_{j=1}^{p} \sum_{i=1}^{q} r_j s_i W_j V_i. \end{aligned}$$

For example, if $n = 3$, *then* $W_1 := x_3x_1x_1x_2x_3x_2x_2 = x_3x_1^2x_2x_3x_2^2$ *defines an element of* $\mathcal{W}_7$. *Also,*

$$\begin{aligned} &(5x_1x_3x_2 + 3x_2x_3)(-x_2 + 3x_3x_1) \\ &\qquad = -5x_1x_3x_2^2 + 15x_1x_3x_2x_3x_1 - 3x_2x_3x_2 + 9x_2x_3^2x_1. \end{aligned}$$

Observe that $1 \cdot W_0$ *serves as the multiplicative identity element of* $R\langle x_1, x_2, \ldots, x_n\rangle$, *which means that* $R\langle x_1, x_2, \ldots, x_n\rangle$ *is unital.*

(b) *We can also insist as part of our hypotheses above that the indeterminates* $x_1, x_2, \ldots, x_n$ *must commute with one another – i.e. that* $x_ix_j = x_jx_i$ *for all* $1 \le i, j \le n$. *In this case, we write*

$$R[x_1, x_2, \ldots, x_n]$$

and call this the ***ring of polynomials in*** n ***commuting variables*** $x_1, x_2, \ldots, x_n$ ***over the commutative, unital ring*** R.

Note that if $n = 1$, then $R\langle x_1\rangle = R[x_1]$ is just the ring of polynomials in x_1. In this setting, the expression "*commuting variables*" is extraneous.

2.3. The case where $n = 1$ will be very close to our hearts. Given a commutative, unital ring R, we have defined the ring of polynomials in x with coefficients in R to be the set of formal symbols

$$R[x] = R\langle x\rangle = \{p_0W_0 + p_1x + \cdots + p_mx^m : m \ge 1, p_k \in R, 0 \le k \le m\}.$$

In practice, we typically do one of two things:

- either we drop the W_0 notation and simply write $p_0 + p_1x + \cdots + p_mx^m$, or
- we denote W_0 by x^0 and write $p_0x^0 + p_1x + \cdots + p_mx^m$.

In these notes, we will tend to drop the W_0 altogether. Hopefully, this should not cause any confusion, as the notation agrees with the usual notation (that we've all come to know and love) for polynomials in x.

Two elements $p_0 + p_1x + \cdots + p_mx^m$ (where $p_m \ne 0$) and $q_0 + q_1x + \cdots + q_nx^n$ (where $q_n \ne 0$) are considered equal if and only if $n = m$ and $p_k = q_k$, $0 \le k \le n$.

As we have just seen, $R[x]$ is a ring using the operations (say $m \le n$)

$$\begin{aligned} &(p_0 + p_1x + \cdots + p_mx^m) + (q_0 + q_1x + \cdots + q_nx^n) \\ &\qquad = (p_0 + q_0) + (p_1 + q_1)x + \cdots + (p_m + q_m)x^m + q_{m+1}x^{m+1} + \cdots + q_nx^n, \end{aligned}$$

and

$$(p_0 + p_1x + \cdots + p_mx^m)(q_0 + q_1x + \cdots + q_nx^n)$$
$$= p_0q_0 + (p_0q_1 + p_1q_0)x^1 + (p_0q_2 + p_1q_1 + p_2q_0)x^2 + \cdots + p_mq_nx^{n+m}.$$

2.4. Terminology. Given $0 \neq p(x) := p_0 + p_1x + p_2x^2 + \cdots + p_mx^m \in R[x]$, we define the **degree** of $p(x)$ to be

$$\deg p(x) = \max\{k : p_k \neq 0\}.$$

If $p(x) = 0$, then the degree of $p(x)$ is *undefined*.

Note that any polynomial $r(x)$ of degree at most $m \geq 1$ can be written as $r(x) = r_0 + r_1x + \cdots + r_mx^m$ by adding terms whose coefficients are equal to zero if necessary. Thus, for example, given a polynomial $r(x)$ of degree 2, we may always write $r(x) = r_0 + r_1x + r_2x^2 + 0x^3 + 0x^4$.

(1) If $\deg p(x) = m$ and $p_m = 1$, we say that $p(x)$ is a **monic polynomial**;
(2) if $p(x) = p_0$, we say that $p(x)$ is a **constant polynomial**.

Thus if $p(x)$ and $q(x)$ are polynomials of degree $0 \leq m$ and $0 \leq n$ respectively, we find that either $p(x) + q(x) = 0$ or

$$\deg(p(x) + q(x)) \leq \max(\deg(p(x)), \deg(q(x))),$$

while

$$\deg(p(x)q(x)) \leq \deg(p(x)) + \deg(q(x)).$$

The reason that we need not have equality in either of the two inequalities above is that

- if $n = m$ and $p_m = -q_m$, then $p(x) + q(x)$ has degree strictly less than m (or is undefined when $p(x) + q(x) = 0$), while
- if $p_mq_n = 0$, then $p(x)q(x)$ has degree strictly less than $m + n$.

3. New rings from old

3.1. We have seen a number of examples of rings. Let us now examine a few constructions that allow us to build new rings from these.

3.2. Example. Let R be a ring.

(a) Let $n \in \mathbb{N}$. Then

$$\mathbb{M}_n(R) = \{[r_{i,j}] : r_{i,j} \in R, 1 \leq i, j \leq n\}$$

is a ring, using usual matrix addition and multiplication (and the product and sum from R).

(b) Let $n \in \mathbb{N}$. Then

$$\mathbb{T}_n(R) = \{T = [t_{i,j}] : t_{i,j} \in R, 1 \leq i, j \leq n, t_{i,j} = 0 \text{ if } j < i\}$$

is a ring, again using usual matrix addition and multiplication.

(c) Let $\varnothing \neq X$ be a set. Then

$$R^X := \{f : X \to R : f \text{ a function}\}$$

is a ring, using point-wise addition and multiplication.

3.3. Example.

(a) DIRECT PRODUCTS. Let Λ be a non-empty set, which may be finite or infinite, countable or uncountable. Suppose that for each $\lambda \in \Lambda$, we are given a ring R_λ. We define the **direct product** of the family $\{R_\lambda : \lambda \in \Lambda\}$ to be the set

$$\prod_{\lambda\in\Lambda} R_\lambda := \{(r_\lambda)_\lambda : r_\lambda \in R_\lambda, \lambda \in \Lambda\}.$$

Thus, if $\Lambda = \{1, 2\}$, then

$$\prod_{\lambda\in\{1,2\}} R_\lambda = R_1 \times R_2 = \{(r_1, r_2) : r_1 \in R_1, r_2 \in R_2\}$$

is the usual crossed product of R_1 and R_2.

We define addition and multiplication co-ordinate-wise:

$$(r_\lambda)_\lambda + (s_\lambda)_\lambda := (r_\lambda + s_\lambda)_\lambda,$$

and

$$(r_\lambda)_\lambda \cdot (s_\lambda)_\lambda := (r_\lambda \cdot s_\lambda)_\lambda.$$

An equivalent formulation of this definition is to first let $R := \cup_{\lambda\in\Lambda} R_\lambda$ (as a set, with no operations). We then define

$$\prod_{\lambda\in\Lambda} R_\lambda := \{f : \Lambda \to R : f(\lambda) \in R_\lambda \text{ for all } \lambda \in \Lambda\}.$$

Any such function f is referred to as a *choice function* for Λ. Writing $r_\lambda := f(\lambda)$, $\lambda \in \Lambda$ and $(r_\lambda)_\lambda$ for f, we arrive at the equivalence of the two formulations.

(For those with a knowledge of the **Axiom of Choice**, we point out that since $0_\lambda \in R_\lambda$ for each λ, the zero function $f(\lambda) = 0_\lambda$ lies in $\prod_{\lambda\in\Lambda} R_\lambda \neq \varnothing$ without appealing to this axiom. In other words, we are not dealing with **Bertrand Russell's** socks here, but rather with his shoes.)

(b) DIRECT SUMS. As before, we let Λ be a non-empty set, which may be finite or infinite, countable or uncountable. Suppose that for each $\lambda \in \Lambda$, we are give a ring R_λ. We define the **direct sum** of the family $\{R_\lambda : \lambda \in \Lambda\}$ to be the set

$$\bigoplus_{\lambda\in\Lambda} R_\lambda := \{(r_\lambda)_\lambda \in \prod_{\lambda\in\Lambda} R_\lambda : r_\lambda = 0 \text{ for all but finitely many } \lambda \in \Lambda\}.$$

Again, we define addition and multiplication co-ordinate-wise.

Observe that if Λ is finite (for example if $\Lambda = \{1, 2, \ldots, n\}$ for some $n \in \mathbb{N}$), then

$$\prod_{\lambda\in\Lambda} R_\lambda = \prod_{j=1}^{n} R_j = \bigoplus_{j=1}^{n} R_n = \{(r_1, r_2, \ldots, r_n) : r_j \in R_j,\ 1 \le j \le n\}.$$

Also, if $R_\lambda = R$ for all $\lambda \in \Lambda$, then

$$\prod_{\lambda\in\Lambda} R_\lambda = \prod_{\lambda\in\Lambda} R = R^\Lambda,$$

as defined in Example 2.3.2(c).

3.4. Example. Suppose that $R_1 = \mathbb{R}$, $R_2 = \mathbb{Q}$, $R_3 = \mathbb{Z}_4$ and $R_4 = \mathbb{M}_2(\mathbb{Z})$. For each of these rings, we shall use the usual operations of addition and multiplication. Let $\Lambda := \{1, 2, 3, 4\}$. Then

$$\begin{aligned}\prod_{\lambda\in\Lambda} R_\lambda &:= R_1 \times R_2 \times R_3 \times R_4\\ &= \mathbb{R} \times \mathbb{Q} \times \mathbb{Z}_4 \times \mathbb{M}_2(\mathbb{Z})\\ &= \{(x, q, a, T) : x \in \mathbb{R}, q \in \mathbb{Q}, a \in \mathbb{Z}_4, T \in \mathbb{M}_2(\mathbb{T})\}.\end{aligned}$$

Since the set Λ is finite,

$$\prod_{\lambda\in\Lambda} R_\lambda = \oplus_{\lambda\in\Lambda} R_\lambda.$$

It is important to keep in mind that Λ is an *ordered set*. That is, if we define $\Omega = \{2, 4, 1, 3\}$, then

$$\begin{aligned}\prod_{\omega\in\Omega} R_\omega &:= R_2 \times R_4 \times R_1 \times R_3\\ &= \mathbb{Q} \times \mathbb{M}_2(\mathbb{Z}) \times \mathbb{R} \times \mathbb{Z}_4\\ &= \{(q, T, x, a) : x \in \mathbb{R}, q \in \mathbb{Q}, a \in \mathbb{Z}_4, T \in \mathbb{M}_2(\mathbb{T})\}\\ &= \oplus_{\omega\in\Omega} R_\omega,\end{aligned}$$

but

$$\prod_{\lambda\in\Lambda} R_\lambda \ne \prod_{\omega\in\Omega} R_\omega.$$

3.5. Example. Suppose that $R = (\mathbb{Z}, +, \cdot)$. Then

$$\prod_{n\in\mathbb{N}} R = \prod_{n\in\mathbb{N}} \mathbb{Z} = \{(z_1, z_2, z_3, \ldots) : z_n \in \mathbb{Z}, n \ge 1\},$$

and

$$\oplus_{n\in\mathbb{N}} R = \oplus_{n\in\mathbb{N}} \mathbb{Z} = \{(z_1, z_2, z_3, \ldots) : z_n \in \mathbb{Z}, n \ge 1, \{k \in \mathbb{N} : z_k \ne 0\} \text{ is finite}\}.$$

Thus $e := (1, 1, 1, \ldots) \in \prod_{n\in\mathbb{N}} \mathbb{Z}$, but $e \notin \oplus_{n\in\mathbb{N}} \mathbb{Z}$.

We add and multiply terms coordinate-wise. Thus if $z = (z_1, z_2, z_3, \ldots)$ and $w = (w_1, w_2, w_3, \ldots) \in \prod_{n\in\mathbb{N}} \mathbb{Z}$, then

$$z + w = (z_1 + w_1, z_2 + w_2, z_3 + w_3, \ldots),$$

and

$$z \cdot w = (z_1 w_1, z_2 w_2, z_3 w_3, \ldots).$$

It follows that the element $e = (1, 1, 1, \ldots)$ defined above is the identity element of $\prod_{n\in\mathbb{N}} \mathbb{Z}$. We leave it as an exercise for the reader to prove that $\oplus_{n\in\mathbb{N}}\mathbb{Z}$ is non-unital.

3.6. Example. Let R be a ring and $X \subseteq R$. Let us define

$$R_X := \cap\{S \subseteq R : X \subseteq S \text{ and } S \text{ is a ring}\}.$$

Then R_X is a ring, and in fact it is the smallest ring in R that contains X, in the sense that if $\mathcal{T}$ is any ring satisfying $X \subseteq \mathcal{T} \subseteq R$, then $R_X \subseteq \mathcal{T}$.

We say that R_X is the **ring generated by** X.

3.7. Example. Let $R = \mathbb{M}_2(\mathbb{R})$. Let $Y = \{T_1 := \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, T_2 := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\}$. If $S \subseteq \mathbb{M}_2(\mathbb{R})$ is a ring and $T_1 \in S$, then $T_1^2 \in S$, and therefore $X_1 := 2T_1 - T_1^2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \in S$. From this we see that $X_2 := T_1 - X_1 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \in S$ as well. This in turn implies that

$$\begin{bmatrix} m & 0 \\ 0 & 0 \end{bmatrix} = mX_1 := X_1 + X_1 + \cdots + X_1 \quad m \text{ times}$$

and $nX_2, pT_2 \in S$ for all $m, n, p \in \mathbb{N}$. Since (additive) inverses are also in S, we see that

$$\mathcal{W} := \left\{ \begin{bmatrix} m & p \\ 0 & 2n \end{bmatrix} : m, n, p \in \mathbb{Z} \right\} \subseteq S.$$

Thus

$$\mathcal{W} \subseteq \cap\{V \subseteq R : Y \subseteq V \text{ and } V \text{ is a ring}\}.$$

By definition, $\mathcal{W}$ consists of all upper-triangular 2×2 matrices whose $(1,1)$ and $(1,2)$ entries are arbitrary integers, and whose $(2,2)$ entry is an arbitrary *even* integer.

To see that $\mathcal{W} = R_Y$, it suffices to know that $\mathcal{W}$ is a ring. We could go through the entire list of conditions from Definition 2.2 above, but it will be much easier to apply the Subring Test below (see Proposition 4.10).

That $\mathcal{W} \neq \varnothing$ is obvious from its definition. If $W_1 := \begin{bmatrix} m_1 & p_1 \\ 0 & 2n_1 \end{bmatrix}$ and $W_2 := \begin{bmatrix} m_2 & p_2 \\ 0 & 2n_2 \end{bmatrix} \in \mathcal{W}$, then

$$W_1 - W_2 = \begin{bmatrix} m_1 - m_2 & p_1 - p_2 \\ 0 & 2(n_1 - n_2) \end{bmatrix} \in \mathcal{W},$$

and

$$W_1 W_2 = \begin{bmatrix} m_1 m_2 & m_1 p_2 + p_1(2n_2) \\ 0 & 2(2n_1 n_2) \end{bmatrix} \in \mathcal{W}.$$

By the Subring Test, $\mathcal{W}$ is a ring, and therefore it must be the smallest ring that contains Y, i.e., $\mathcal{W} = R_Y$.

4. Basic results

4.1. We now have at hand a large number of rings. It is time that we prove our first results. We shall adopt the common notation whereby we express multiplication through juxtaposition, that is: we write ab to mean $a \cdot b$. We also introduce the notation $a - b$ to mean $a + (-b)$.

4.2. Proposition. *Let $(R, +, \cdot)$ be a ring and $a, b, c \in R$. Then*

(a) *If $a + b = a + c$, then $b = c$.*
(b) $a0 = 0 = 0a$.
(c) $a(-b) = -(ab) = (-a)b$.
(d) $(-a)(-b) = ab$.
(e) $a(b - c) = ab - ac$, *and* $(a - b)c = ac - bc$.

If R is unital, then

(f) $(-1)a = -a$, *and*
(g) $(-1)(-1) = 1$.

Proof.

(a) This is the cancellation law that was stated as Exercise 1.9.

Let's prove this. Indeed, recall that $(R, +)$ is an abelian group. Since $a \in R$, there exists an element $x \in R$ such that $x + a = 0$, the neutral element of $(R, +)$.

Then $a + b = a + c$ implies that

$$b = 0 + b = (x + a) + b = x + (a + b) = x + (a + c) = (x + a) + c = 0 + c = c.$$

Observe that associativity of addition was crucial to this proof.

(b) Note that $a0 + 0 = a0 = a(0 + 0) = a0 + a0$. By the cancellation law, $a0 = 0$. The proof that $0a = 0$ is similar, and is left as an exercise.

(c) Now $0 = a0 = a(b + (-b)) = ab + a(-b)$. Since the additive inverse of ab is unique (Exercise 1.2), it follows that $a(-b) = -(ab)$. Again, the proof that $(-a)b = -(ab)$ is similar and is left as an exercise.

(d) By item (c) above,

$$(-a)(-b) = -(a(-b)) = -(-(ab)).$$

That is, $(-a)(-b)$ is the additive inverse of $-(ab)$. But ab is also the additive inverse of $-(ab)$, and so by uniqueness,

$$(-a)(-b) = ab.$$

(e) Now, using item (c) above,

$$a(b-c) = a(b+(-c)) = ab + a(-c) = ab + (-(ac)) = ab - ac.$$

Also,

$$(a-b)c = (a+(-b))c = ac + (-b)c = ac - bc.$$

(f) We see from item (c) above that

$$(-1)a = -(1a) = -a.$$

(g) From item (d),

$$(-1)(-1) = (1)(1) = 1.$$

□

4.3. Proposition. *Let $(R, +, \cdot)$ be a ring.*

(a) *If R has a multiplicative identity, then it is unique.*
(b) *If $x \in R$ has a multiplicative inverse, then it is unique.*

Proof.

(a) Suppose that $e, f \in R$ and that $er = fr = r = rf = re$ for all $r \in R$, so that both e and f are multiplicative identities for R.

Then

$$e = ef = f.$$

(b) Suppose that $r \in R$ and that $x, y \in R$ satisfy $ry = yr = 1 = xr = rx$. Then

$$y = (1)y = (xr)y = x(ry) = x(1) = x.$$

□

4.4. Remark. Proposition 4.3 is what allows us to adopt the notation 1 to denote the unique multiplicative identity of R (when it exists), and to adopt the notation x^{-1} for the unique multiplicative inverse of x (when it exists).

4.5. Definition. *Let $(R, +, \cdot)$ be a unital ring. An element $r \in R$ is said to be* ***invertible*** *if it admits a multiplicative inverse. That is, r is invertible if there exists $s \in R$ such that $rs = 1 = sr$.*

We denote the set of invertible elements of R by $\text{INV}(R)$.

By Remark 4.4, when r is invertible, the element s satisfying $sr = 1 = rs$ is unique and is denoted by r^{-1}.

4.6. Examples.

(a) The only invertible elements of $(\mathbb{Z}, +, \cdot)$ are 1 and -1, with inverses 1 and -1 respectively.
(b) Every non-zero element of $\mathbb{R}$ is invertible (with respect to the usual multiplication of real numbers).
(c) From linear algebra, we recall that if $n \geq 1$ is an integer, then a matrix $T \in \mathbb{M}_n(\mathbb{R})$ is invertible if and only if its determinant is non-zero.

4.7. Definition. *A non-empty subset S of a ring $(R,+,\cdot)$ is called a* ***subring*** *of R if $(S,+,\cdot)$ is a ring* ***using the addition and multiplication operations inherited from*** *R.*

4.8. Example. Let $S = 2\mathbb{Z} := \{\ldots,-4,-2,0,2,4,\ldots\}$ denote the set of all even integers. We leave it as an exercise for the reader to show that $2\mathbb{Z}$ is a subring of $\mathbb{Z}$.

4.9. Remark. In defining a subring S of a ring R, it is crucial that we maintain the same operations as those used in R. For example, we know that $(\mathbb{Z},+,\cdot)$ is a ring using the usual multiplication and addition operations. By Example 2.3.2(a), $\mathbb{M}_2(\mathbb{Z})$ is also a ring.

Let $S := \left\{\begin{bmatrix} 0 & r \\ s & 0 \end{bmatrix} : r,s \in \mathbb{Z}\right\}$, and for $A = \begin{bmatrix} 0 & r_1 \\ s_1 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & r_2 \\ s_2 & 0 \end{bmatrix}$, define

$$A + B = \begin{bmatrix} 0 & r_1 + r_2 \\ s_1 + s_2 & 0 \end{bmatrix}$$

(this is the usual addition) and

$$A * B := \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It is not difficult to verify that $(S,+,*)$ is a ring, but it is not a subring of $(\mathbb{M}_2(\mathbb{Z}),+,\cdot)$, because we have changed the definition of multiplication in S.

Note that under the usual definition of multiplication (say $\cdot$) in $\mathbb{M}_2(\mathbb{Z})$, $S_1 := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $S_2 := \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \in S$, but $S_1 \cdot S_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \notin S$. Thus $(S,+,\cdot)$ is not a ring, and therefore it cannot be a subring of $(\mathbb{M}_2(\mathbb{Z}),+,\cdot)$.

4.10. Proposition. (The Subring Test) *Let $\varnothing \neq S \subseteq R$, where $(R,+,\cdot)$ is a ring. Then S is a subring of R if and only if $a-b \in S$ and $ab \in S$ for all $a,b \in S$.*

Proof.

Suppose first that S is a subring of R. If $a,b \in S$, then, since S is a ring, $-b \in S$ and $a-b = a+(-b) \in S$. Also, $a,b \in S$ implies that $ab \in S$.

Next, suppose that $S \subseteq R$, and that $a-b, ab \in S$ for all $a,b \in S$. Let $a,b,c \in S$. Then – since the addition and multiplication on S are those inherited from R, it follows that $(a+b)+c = a+(b+c)$ and $a(bc) = (ab)c$, since $a,b,c \in R$ and addition and multiplication are associative in R.

Similarly, multiplication distributes over addition in S because it does in R, and we are using the same addition and multiplication. Furthermore, $a+b = b+a$ since this holds in R.

Now $0 = a+(-a) \in S$ by hypothesis, so S has a neutral element under addition. Also, $0, a \in S$ implies that $0+(-a) = -a \in S$, so S includes the additive inverses of each of its elements. Finally, $a+b = a-(-b) \in S$, so that S is closed under addition.

Thus $(S,+)$ is an abelian group, closed under multiplication, which distributes over addition. That is, $(S,+,\cdot)$ is a ring under the operations inherited from R, and so it is a subring of $(R,+,\cdot)$.

□

4.11. Remark. We point out that in the Subring Test, it is crucial that we show that $a-b \in S$ (and $ab \in S$) rather than $a+b$ (and $ab \in S$) for all $a,b \in S$.

Consider $\mathbb{N} = \{1,2,3,\dots,\}$, equipped with the usual addition and multiplication it inherits as a subset of the ring $\mathbb{Z}$. For $m,n \in \mathbb{N}$, clearly $m+n \in \mathbb{N}$ and $mn \in \mathbb{N}$. But $(\mathbb{N},+,\cdot)$ is not a ring because – for example – $2 \in \mathbb{N}$, but $-2 \notin \mathbb{N}$. Alternatively, $(\mathbb{N},+,\cdot)$ is not a ring because it does not have a neutral element under addition - i.e. $0 \notin \mathbb{N}$.

4.12. Examples.

(a) Let $R := (\mathcal{C}([0,1],\mathbb{R}),+,\cdot)$ and $S = \{f \in R : f(x) = 0 \text{ for all } x \in [1/2, 8/9]\}$. Then S is a subring of R.
(b) For any ring R and any natural number n, $\mathbb{T}_n(R)$ is a subring of $\mathbb{M}_n(R)$.
(c) $\mathbb{Q}$ is a subring of $\mathbb{R}$, and $\mathbb{Z}$ is a subring of $\mathbb{Q}$ and of $\mathbb{R}$.
(d) $\mathbb{Z}_5$ is *not* a subring of $\mathbb{Z}_{10}$. (Why not?)

4.13. Example. Let $(R,+,\cdot)$ be a ring. We define the **centre** of R to be the set

$$Z(R) := \{z \in R : zr = rz \text{ for all } r \in R\}.$$

Suppose that $w,z \in Z(R)$, and $r \in R$. Then

- $(wz)r = w(zr) = w(rz) = (wr)z = (rw)z = r(wz)$.
  Since $r \in R$ was arbitrary, we conclude that $wz \in Z(R)$.
- $(w-z)r = wr - zr = rw - rz = r(w-z)$.
  Again, since $r \in R$ was arbitrary, we conclude that $w - z \in Z(R)$.

By the Subring Test, $Z(R)$ is a subring of R.

Supplementary Examples

S2.1. Example. Let $R = \mathbb{Z}$. Then $\mathbb{M}_2(\mathbb{Z}) = \left\{\begin{bmatrix} a & b \\ c & d \end{bmatrix} : a, b, c, d \in \mathbb{Z}\right\}$ is a ring, where

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} + \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{bmatrix},$$

and

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}\begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{bmatrix}.$$

S2.2. Example. Let $\mathcal{D}((0,1),\mathbb{R}) = \{f : (0,1) \to \mathbb{R} : f \text{ is differentiable on } (0,1)\}$. Define $(f+g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$ for all $x \in (0,1)$.

Then $\mathcal{D}((0,1),\mathbb{R})$ is a ring. The details are left to the reader.

S2.3. Example. Most readers are familiar with the ring of complex numbers $\mathbb{C} := \{a + bi : a, b \in \mathbb{R}, i^2 = -1\}$. Perhaps less well-known is the ring of **real quaternions**:

Define $\mathbb{H} := \{a\mathbf{1} + b\,\mathbf{i} + c\,\mathbf{j} + d\,\mathbf{k} : a, b, c, d \in \mathbb{R}\}$, where $\mathbf{i}, \mathbf{j}, \mathbf{k}$ satisfy the follow multiplication table:

$\times$	$\mathbf{1}$	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\mathbf{1}$	$\mathbf{1}$	$\mathbf{i}$	$\mathbf{j}$	$\mathbf{k}$
$\mathbf{i}$	$\mathbf{i}$	$\mathbf{-1}$	$\mathbf{k}$	$\mathbf{-j}$
$\mathbf{j}$	$\mathbf{j}$	$\mathbf{-k}$	$\mathbf{-1}$	$\mathbf{i}$
$\mathbf{k}$	$\mathbf{k}$	$\mathbf{j}$	$\mathbf{-i}$	$\mathbf{-1}$

We define

$$(a_1\mathbf{1}+b_1\,\mathbf{i}+c_1\,\mathbf{j}+d_2\,\mathbf{k})+(a_2\mathbf{1}+b_2\,\mathbf{i}+c_2\,\mathbf{j}+d_2\,\mathbf{k}) = (a_1+a_2)\mathbf{1}+(b_1+b_2)\,\mathbf{i}+(c_1+c_2)\,\mathbf{j}+(d_1+d_2)\,\mathbf{k}.$$

Multiplication behaves analogously to complex multiplication, only following the multiplication table above. Thus, for example,

$$\begin{aligned}(\pi + 2\mathbf{j} + 3\mathbf{k})(e + 2\mathbf{i}) &= \pi e + 2e\,\mathbf{j} + 3e\,\mathbf{k} + 2\pi\,\mathbf{i} + 4\,\mathbf{j}\,\mathbf{i} + 6\,\mathbf{k}\,\mathbf{i} \\ &= \pi\, e + 2\pi\,\mathbf{i} + (2e + 6)\,\mathbf{j} + (3e - 4)\,\mathbf{k}.\end{aligned}$$

We leave it to the reader to find the (long) formula for multiplying two arbitrary real quaternions.

Note that by replacing $\mathbb{R}$ by $\mathbb{Q}$ above, we obtain the **rational quaternions**.

S2.4. Example. Let

$$\mathcal{C}_0(\mathbb{R},\mathbb{R}) := \{f : \mathbb{R} \to \mathbb{R} : f \text{ is continuous and } \lim_{x\to\infty} f(x) = 0 = \lim_{x\to-\infty} f(x)\}.$$

Then $\mathcal{C}_0(\mathbb{R},\mathbb{R})$ is a ring under pointwise addition and multiplication: that is, $(fg)(x) = f(x)g(x)$ and $(f+g)(x) = f(x) + g(x)$ for all $x \in \mathbb{R}$.

It is non-unital.

S2.5. Example. Let $R := \mathbb{Z}_2 \times \mathbb{Z}_3$ be the direct product of $\mathbb{Z}_2$ and $\mathbb{Z}_3$. Given $(a, b) \in R$, we have that $a \in \mathbb{Z}_2$ and $b \in \mathbb{Z}_3$. We define

$$(a_1, b_1) + (a_2, b_2) := (a_1 + a_2, b_1 + b_2)$$

and

$$(a_1, b_1)(a_2, b_2) := (a_1 a_2, b_1 b_2),$$

where $a_1 + a_2, a_1 a_2$ are determined in $\mathbb{Z}_2$ while $b_1 + b_2, b_1 b_2$ are determined by the operations in $\mathbb{Z}_3$.

Note that R has 6 elements, as does $\mathbb{Z}_6$. We invite the reader to write out the addition and multiplication tables for R and $\mathbb{Z}_6$ and to look for similarities. The relation will be made more explicit in Chapter 4.

S2.6. Example. Let $R := \mathbb{Z}_2 \times \mathbb{Z}_2$ be the direct product of $\mathbb{Z}_2$ and $\mathbb{Z}_2$. Given $(a, b) \in R$, we have that $a \in \mathbb{Z}_2$ and $b \in \mathbb{Z}_2$. We define

$$(a_1, b_1) + (a_2, b_2) := (a_1 + a_2, b_1 + b_2)$$

and

$$(a_1, b_1)(a_2, b_2) := (a_1 a_2, b_1 b_2),$$

where $a_1 + a_2, b_1 + b_2, a_1 a_2$ and $b_1 b_2$ are determined by the operations in $\mathbb{Z}_2$.

This ring has 4 elements, as does $\mathbb{Z}_4$. Note, however, that if $(a, b) \in R$, then $(a, b) + (a, b) = (0, 0)$ is the zero element of R. On the other hand, $1 \in \mathbb{Z}_4$ and $1 + 1 = 2 \neq 0$ in $\mathbb{Z}_4$.

There is something fundamentally different about these two rings. That too will be made explicit below.

S2.7. Example. Let

$$R = \left\{ \begin{bmatrix} w & z \\ -\overline{z} & \overline{w} \end{bmatrix} : w, z \in \mathbb{C} \right\} \subseteq \mathbb{M}_2(\mathbb{C}).$$

Then R is a ring.

Can you see any relationship between R and the ring $\mathbb{H}$ of real quaternions from Example 4.1?

S2.8. Example. Let $\varnothing \neq X$ be a non-empty set, and denote by $\mathcal{P}(X)$ the **power set** of X, that is:

$$\mathcal{P}(X) = \{Y : Y \subseteq X\}.$$

Given $Y, Z \in \mathcal{P}(X)$, set $Y \oplus Z := (Y \setminus Z) \cup (Z \setminus Y)$, and $Y \otimes Z := Y \cap Z$.

Then $(\mathcal{P}(X), \oplus, \otimes)$ is a (commutative) ring. The additive identity is $\varnothing$, while the multiplicative identity is X. We leave it to the reader to verify the details.

Suppose that $X = \{a, b\}$. Is there any relationship between $(\mathcal{P}(X), \oplus, \otimes)$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$? Or between $(\mathcal{P}(X), \oplus, \otimes)$ and $\mathbb{Z}_4$? If so, what is the relationship? If not, why does this fail?

S2.9. Example. Let $(G, +)$ denote an abelian group. An **endomorphism** of G is a map

$$\varphi : G \to G$$

that satisfies $\varphi(g + h) = \varphi(g) + \varphi(h)$ for all $g, h \in G$.

Let $\text{END}(G) := \{\varphi : \varphi \text{ is an endomorphism of } G\}$. For $\varphi, \psi \in \text{END}(G)$, define

$$(\varphi \dot{+} \psi)(g) = \varphi(g) + \psi(g),$$

and

$$(\varphi * \psi)(g) = \varphi(\psi(g)), \quad g \in G.$$

We leave it to the reader to verify that $(\text{END}(G), \dot{+}, *)$ is a ring.

S2.10. Example. We can also consider the so-called **ring of formal power series** in x, which extends the notion of a polynomial ring over a ring.

Let R be a commutative ring. Denote by $R[[x]]$ the ring

$$R[[x]] := \{\sum_{n=0}^{\infty} r_n x^n : r_n \in R \text{ for all } n\}.$$

(*There is no notion of convergence here – this is just notation!*)

We define addition and multiplication on $R[[x]]$ by setting:

$$\sum_{n=0}^{\infty} r_n x^n + \sum_{n=0}^{\infty} s_n x^n := \sum_{n=0}^{\infty} (r_n + s_n) x^n,$$

and

$$\left(\sum_{n=0}^{\infty} r_n x^n\right)\left(\sum_{n=0}^{\infty} s_n x^n\right) := \sum_{n=0}^{\infty} t_n x^n,$$

where for all $n \geq 0$,

$$t_n := \sum_{j=0}^{n} r_j s_{n-j}.$$

We leave it to the reader to verify that this is indeed a ring.

Note that when we say that there is *no notion of convergence here*, we mean that we could just have easily defined an element of $R[[x]]$ to be a sequence $(r_n)_{n=0}^{\infty}$ with $r_n \in R$ for all $n \geq 0$, and $(r_n)_n + (s_n)_n = (r_n + s_n)_n$.

The reason for preferring the "series" notation is that helps to "explain" how one might come up with the multiplication

$$(r_n)_n * (s_n)_n := (t_n)_n,$$

where

$$t_n := \sum_{j=0}^{n} r_j s_{n-j}.$$

In other words, the x^n, $n \geq 0$ is really just a placeholder for the coefficient $r_n \in R$.

Appendix

A2.1. As mentioned in the introductions to each of the first two chapters, mathematics typically evolves by considering a large number of objects that one is interested in, and having one or more clever individuals notice a common link which relates those objects to one another.

In the present case, examples arose out of number theory (integers, rational numbers, real numbers, complex numbers), linear algebra (linear transformations, matrices), set theory (Boolean rings), analysis (continuous functions, differentiable functions), amongst other areas.

The formulation of the notion of a ring began in the 1870's with the study of polynomials and algebraic integers, in part by **Richard Dedekind**. The term "**Zahlring**" (German for "number ring") was introduced by **David Hilbert** in the 1890s. **Adolf Fraenkel** gave the first axiomatic definition of a ring in 1914, but his definition was stricter than the current definition. The modern axiomatic definition of a (commutative) ring was given by **Emmy Noether**.

A2.2. Some authors – especially algebraists – require a ring to have a multiplicative identity. For example, Fraenkel required this, although Noether did not, despite the fact that she was an algebraist (and an excellent one at that). There is a sickening tendency of some people to refer to rings without identity as *rngs*. Let us agree that this *vox nihili* should never be mentioned in public again. (Not that the current author has a strong opinion on this, mind you.)

It is true that it is always possible to "add" an identity to a ring in order to make it unital, but this construction is not always natural. (For example, the set

$$\mathcal{C}_0(\mathbb{R},\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous and } \lim_{x\to\infty} f(x) = 0 = \lim_{x\to-\infty} f(x)\}$$

defined in Example 4.1 above becomes a ring using pointwise addition, but with pointwise multiplication replaced by **convolution**:

$$(f * g)(x) := \int_{-\infty}^{\infty} f(t)g(x-t)\,dt.$$

There is no reasonable function $h : \mathbb{R} \to \mathbb{R}$ which could serve as an identity for this function under convolution. Instead, one has to artificially introduce an abstract symbol and declare it to be an identity by convention.

On the other hand, it is possible to find a *sequence* $(\delta_n)_n$ of functions in $\mathcal{C}_0(\mathbb{R},\mathbb{R})$ satisfying

$$\lim_n (f * \delta_n) = f,$$

but this requires us to understand what we mean by "*limits of functions*", which in this setting is a bit beyond the scope of this course. We mention in passing that the sequence is referred to as an **approximate identity** for the ring. These notions tend to be popular in Analysis, especially Harmonic analysis (commutative and non-commutative) and Operator Algebras.

A2.3. The ring $\mathbb{H}$ of real quaternions was first described by the Irish mathematician **William Rowan Hamilton**. Note that if $\omega_1, \omega_2 \in \mathbb{H}$, then $\omega_1\omega_2 \neq \omega_2\omega_1$ in general, but if $0 \neq \omega \in \mathbb{H}$, then ω is invertible. A non-commutative ring in which every non-zero element is invertible is referred to as a **division ring** or a **skew field**.

A2.4. In some textbooks, the degree of the zero polynomial is defined to be "$-\infty$", with the understanding that $-\infty + (-\infty) = -\infty = -\infty + n = n + (-\infty)$ for all $n \in \mathbb{N}$. In this way, we see that

$$\deg(p(x) + q(x)) \leq \max(\deg(p(x)), \deg(q(x))$$

for all polynomials $p(x), q(x)$ – that is, we don't have to separately argue that $p(x) + q(x)$ *might* be equal to zero.

We have chosen not to use this terminology, because there is a tendency for people to think of "$-\infty$" as a number as opposed to a symbol, and to think of the above "sums" as being actual addition, instead of mere convention.

A2.5 The ring of polynomials revisited. Recall that if R is a commutative ring, then we defined the ring of polynomials $R[x]$ to be the set of formal symbols:

$$R[x] = R\langle x\rangle = \{p_0 + p_1 x + \cdots + p_m x^m : m \geq 1, p_k \in R, 0 \leq k \leq m\}.$$

Let us now give an equivalent, but slightly different and very useful definition of the ring of polynomials $R[x]$.

Let $(R, +, \cdot)$ be a commutative, unital ring and consider the ring

$$\begin{aligned}\mathbb{P}_R &:= \oplus_{n=0}^{\infty} R \\ &:= \{(p_0, p_1, p_2, \ldots) : p_n \in R \text{ for all } n \text{ and } p_n = 0 \text{ except for finitely many } n \geq 0\}.\end{aligned}$$

Given $p = (p_0, p_1, p_2, \ldots)$ and $q = (q_0, q_1, q_2, \ldots) \in \mathbb{P}_R$, we define

$$p + q := (p_0 + q_0, p_1 + q_1, p_2 + q_2, \ldots)$$

and

$$p * q := (r_0, r_1, r_2, \ldots),$$

where

$$r_m := \sum_{k=0}^{m} p_k q_{m-k}, \qquad m \geq 0.$$

That is,

$$p * q := (p_0q_0, p_0q_1 + p_1q_0, p_0q_2 + p_1q_1 + p_2q_0, \ldots).$$

(If this multiplication reminds you of the multiplication of polynomials, that's a good thing!)

Exercise. We leave it as an exercise for the reader to verify that $(\mathbb{P}_R, +, *)$ is a ring with these two operations.

Note that if R is unital, say $1_R \in R$, then

$$e := (1_R, 0, 0, 0, \ldots) \in \mathbb{P}_R$$

satisfies $e * p = p = p * e$ for all $p \in \mathbb{P}_R$, and thus e is the multiplicative identity of $\mathbb{P}_R$.

Define $x := (0, 1_R, 0, 0, 0, \ldots) \in \mathbb{P}_R$. If $m \geq 1$, then

$$x^m = (0, 0, \ldots, 0, 1_R, 0, 0, 0, \ldots),$$

where the 1_R term occurs at the $m+1$-st position. In other words,

$$x^m = (a_0, a_1, a_2, \ldots, a_m, a_{m+1}, a_{m+2}, \ldots),$$

where $a_k = 0$ if $k \neq m+1$ and $a_{m+1} = 1_R$.

Given $r \in R$ and $p = (p_0, p_1, p_2, \ldots) \in \mathbb{P}_R$, we define

$$r \cdot p := (rp_0, rp_1, rp_2, \ldots) \in \mathbb{P}_R.$$

In other words, if we identify $r \in R$ with the element $\widehat{r} := (r, 0, 0, 0, \ldots) \in \mathbb{P}_R$, then we have that

$$r \cdot p = \widehat{r} * p.$$

This allows us to write a general element $p = (p_0, p_1, p_2, \ldots) \in \mathbb{P}_R$ (where $n \in \mathbb{N}$ is chosen such that $p_m = 0$ if $m \geq n$) in the form

$$p = p_0 e + p_1 \cdot x + p_2 \cdot x^2 + \cdots + p_N \cdot x^n.$$

It is clear that $p = q$ if and only if $p_k = q_k$ for all $k \geq 0$, and we have specifically chosen the addition and multiplication in $\mathbb{P}_R$ to mimic the addition and multiplication of polynomials.

In the language of Chapter Four, $R[x]$ and $\mathbb{P}_R$ are **isomorphic**. They are precisely the same ring, just expressed with different notation.

We invite the reader to think of a way of expressing $R[x_1, x_2]$ in a similar form.

Exercises for Chapter 2

Exercise 2.1.

Consider the set $\mathcal{P}_n$ of $n \times n$ permutation matrices in $\mathbb{M}_n(\mathbb{C})$. Thinking of $\mathbb{M}_n(\mathbb{C})$ as a vector space over $\mathbb{C}$, is

$$\operatorname{span}\{\mathcal{P}_n\} = \mathbb{M}_n(\mathbb{C})?$$

Exercise 2.2.

Let

$$H = \{H_1, H_2, H_3\} := \left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \right\}.$$

Prove that H is a group and write out its multiplication table (as we did with G in Example 2.1.8).

(a) Is there a relationship between H and the group G from that example?
(b) Is there a relationship between $\mathbb{Q}\langle H \rangle$ and $\mathbb{Q}\langle G \rangle$?

If relationships do exist, how would you describe them?

Exercise 2.3.

Find all subrings of $(\mathbb{Z}, +, \cdot)$.

Exercise 2.4.

Let R and S be subrings of a *commutative* ring T.

(a) Define

$$RS := \{\sum_{i=1}^{n} r_i s_i : n \geq 1, r_i \in R, s_i \in S, 1 \leq i \leq n\}.$$

Prove that RS is a subring of T. Is it always true that $R \subseteq RS$? Either prove that it is, or find a counterexample to show that it is not always true.

(b) Did we need to take all finite sums in the expression for RS above? In other words, if

$$A := \{rs : r \in R, s \in S\},$$

is A a subring of T? Again, you must justify your answer!

(c) Prove that $R \cap S$ is a subring of T. Is the commutativity of T required here?

(d) Was commutativity of T required in part (a) above? If not, prove part (a) without assuming that T is commutative. Otherwise, provide an example of a non-commutative ring T where RS is not a subring of T.

Exercise 2.5.

Let R and S be rings. Describe all subrings of $R \oplus S$ in terms of the subrings of R and S.

Exercise 2.6.

Let $G = \mathbb{Z} \oplus \mathbb{Z}$, equipped with the operation $(g_1, h_1) + (g_2, h_2) := (g_1 + g_2, h_1 + h_2)$.

(a) Prove that G is an abelian group.

(b) Describe $\textsc{End}(G)$, the endomorphism ring of G as defined in Example 2.9.

Exercise 2.7.

Let R be a ring and $a, b \in R$. Show that in general,

$$(a+b)^2 := (a+b)(a+b) \neq a^2 + 2ab + b^2.$$

State the correct formula, and find the formula for $(a+b)^3$.

Exercise 2.8.

Consider $\mathbb{M}_2(\mathbb{R})$ equipped with usual matrix addition and multiplication. If $L \subseteq R$ is a subring with the additional property that $R \in R$ and $M \in L$ implies that $RM \in L$, we say that L is a **left ideal** of $\mathbb{M}_2(\mathbb{R})$.

Find all left ideals of $\mathbb{M}_2(\mathbb{R})$. Can you generalise this to $\mathbb{M}_n(\mathbb{R})$ for $n \geq 3$?

Note. As you might imagine, one can also define **right ideals** in a similar manner. What is the relationship between left and right ideals of $\mathbb{M}_n(\mathbb{R})$?

Exercise 2.9.

Let $(R, +, \cdot)$ be a ring. Define a new multiplication $*$ on R by setting

$$a * b := b \cdot a \text{ for all } a, b \in R.$$

Prove that $(R, +, *)$ is also a ring.

This ring is referred to as the **opposite ring** of R.

Exercise 2.10.

Let R be a ring and suppose that $a^2 = a$ for all $a \in R$. Prove that R is commutative; that is, prove that $ab = ba$ for all $a, b \in R$.

A ring R for which $a^2 = a$ for all $a \in R$ is called a **Boolean ring**. Provide an example of such a ring.

CHAPTER 3

Integral Domains and Fields

If you want to know what God thinks of money, just look at the people he gave it to.

Dorothy Parker

1. Integral domains - definitions and basic properties

1.1. Recall from Exercise 1.9 that one of the nicer properties that holds in a group $(G, \cdot)$ is the **cancellation law**, namely:

if g, h and $k \in G$ and

$$g \cdot h = g \cdot k,$$

then $h = k$.

If $(R, +, \cdot)$ is a ring, then $(R, +)$ is an abelian group, and so the cancellation law holds for $(R, +)$. That is, given $a, b, c \in R$, the equation

$$a + b = a + c$$

implies that $b = c$.

The cancellation law is the group analogue of a familiar technique we employ when *solving equations*: for example, if we are working with real numbers, we solve the equation

$$5x = 35$$

by multiplying both sides of the equation by $\frac{1}{5} = 5^{-1}$ to get

$$x = 5^{-1} \cdot 5x = 5^{-1}35 = 7.$$

In other words,

$$5x = 5 \cdot 7$$

implies that $x = 7$.

Unfortunately, the cancellation law is not some universal truth that holds in all algebraic structures at all times. In particular, the cancellation law does not have

to hold for multiplication in a ring. For example, consider the matrices

$$A := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B := \begin{bmatrix} 0 & 0 \\ -100 & \pi \end{bmatrix}, \quad C := \begin{bmatrix} 0 & 0 \\ 12 & e^3 + 19 \end{bmatrix},$$

all of which lie in the ring $\mathbb{M}_2(\mathbb{R})$.

A simple calculation shows that

$$AC = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = AB,$$

but clearly $B \neq C$.

In this Chapter, we shall isolate this "bad behaviour", and we shall give a special name to those rings which avoid it.

1.2. Definition. *Let R be a ring. An element $0 \neq r \in R$ is said to be a* ***left divisor of zero*** *if there exists $0 \neq x \in R$ such that $rx = 0$.*

Similarly, $0 \neq s \in R$ is said to be a ***right divisor of zero*** *if there exists $0 \neq y \in R$ such that $ys = 0$.*

Finally, $0 \neq t \in R$ is said to be a ***(joint) divisor of zero*** *if it is both a left and a right divisor of 0.*

We remark that if R is a commutative ring, then every left (or right) divisor of zero is automatically a joint divisor of zero.

1.3. Examples.

(a) Let $R = (\mathcal{C}([0,1], \mathbb{R}), +, \cdot)$. Let $f \in R$ be the function

$$f(x) = \begin{cases} 0 & 0 \leq x \leq \frac{1}{2} \\ x - \frac{1}{2} & \frac{1}{2} \leq x \leq 1. \end{cases}$$

Let $g \in R$ be the function

$$g(x) = \begin{cases} \frac{1}{2} - x & 0 \leq x \leq \frac{1}{2} \\ 0 & \frac{1}{2} \leq x \leq 1. \end{cases}$$

Then $f \neq 0 \neq g$, but $fg = 0 = gf$, and thus f is a joint divisor of zero (as is g).

(b) Let $\mathcal{V} = \{x = (x_n)_{n=1}^{\infty} \in \mathbb{R}^{\mathbb{N}} : \lim_n x_n = 0\}$. We leave it as an exercise for the reader to prove that $\mathcal{V}$ is a vector space over $\mathbb{R}$ under the operations

$$(x_n)_n + (y_n)_n := (x_n + y_n)_n,$$
$$\lambda(x_n)_n := (\lambda x_n)_n,$$

for all $(x_n)_n, (y_n)_n \in \mathcal{V}$ and $\lambda \in \mathbb{R}$.

Define the linear map

$$\begin{array}{rccc} S: & \mathcal{V} & \to & \mathcal{V} \\ & (x_1, x_2, x_3, \ldots) & \mapsto & (0, x_1, x_2, x_3, x_4, \ldots). \end{array}$$

Again - we leave it to the reader to verify that $S \in \mathcal{L}(\mathcal{V})$. Note that $x \in \mathcal{V}$ and $Sx = 0$ implies that $x = 0$; that is, S is injective. Clearly $S \neq 0$.

Suppose that $T \in \mathcal{L}(\mathcal{V})$ and that $ST = 0$. Then for all $y \in \mathcal{V}$, $STy = S(Ty) = 0$ implies that $Ty = 0$. But then $Ty = 0$ for all $y \in \mathcal{V}$ implies that $T = 0$. This shows that S is not a left divisor of zero.

On the other hand, let $R \in \mathcal{L}(\mathcal{V})$ be the map

$$\begin{array}{rccc} R: & \mathcal{V} & \to & \mathcal{V} \\ & (x_1, x_2, x_3, \ldots) & \mapsto & (x_1, 0, 0, 0, \ldots). \end{array}$$

(One should verify that this is indeed a linear map on $\mathcal{V}$!)

Then $R \neq 0$ since $R(1, 0, 0, \ldots) = (1, 0, 0, \ldots)$, but for any $x = (x_n)_n \in \mathcal{V}$,

$$RSx = R(Sx) = R(0, x_1, x_2, \ldots) = (0, 0, 0, \ldots) = 0.$$

Thus S is a right divisor of zero.

(c) If $(R, +, \cdot)$ is a unital ring and $0 \neq r \in R$ is invertible, then it is neither a left nor a right divisor of zero. Indeed, denoting by r^{-1} the multiplicative inverse of r, we see that the equation $r \cdot x = 0$ implies that

$$x = 1 \cdot x = (r^{-1} \cdot r) \cdot x = r^{-1} \cdot (r \cdot x) = r^{-1} \cdot 0 = 0,$$

proving that r is not a left divisor of zero. The argument that r is not a right divisor of zero is similar and is left to the reader.

(d) Suppose that $R = (\mathbb{Z}_{24}, +, \cdot)$. Then, with addition and multiplication done "MOD 24", as described in Example 2.1.6, we see that $6 \cdot 8 = 48 \text{ MOD } 24 = 0$. Since $\mathbb{Z}_{24}$ is a commutative ring, we find that 6 is a joint divisor of zero.

1.4. Definition. *A ring $(D, +, \cdot)$ is said to be an **integral domain** if*

- *D is unital.*
- *D is commutative, and*
- *if $a, b \in D$ and $a \cdot b = 0$, then either $a = 0$ or $b = 0$. In other words, D has no divisors of zero.*

(Note that since D is commutative, if it were to admit any kind of divisor of zero, it would have to be a joint divisor of zero.)

1.5. Examples.

(a) The motivating example of an integral domain, and probably the origin of the terminology, is the ring $(\mathbb{Z}, +, \cdot)$ of integers.

(b) Each of $(\mathbb{Q}, +, \cdot)$, $(\mathbb{R}, +, \cdot)$ and $(\mathbb{C}, +, \cdot)$ is also an integral domain.

(c) From Example 1.3 above,

$$(\mathcal{C}([0, 1], \mathbb{R}), +, \cdot)$$

is *not* an integral domain.

(d) $\mathbb{M}_2(\mathbb{R})$ is not commutative - it has no chance of being an integral domain.

1.6. Examples.

(a) The **ring of Gaussian integers**

$$\mathbb{Z}[i] := \{a + bi : a, b \in \mathbb{Z}, i^2 = -1\}$$

is an integral domain.

(b) The subring $(2\mathbb{Z}, +, \cdot)$ of $\mathbb{Z}$ is *not* an integral domain, because it is not unital!

(c) The ring $(\mathbb{R}[x], +, \cdot)$ of polynomials in one variable with coefficients in $\mathbb{R}$ is an integral domain. We leave it to the reader to verify this.

Again - we remind the reader that polynomial rings will be of special importance later. Despite the simplicity of the proof, the following result is crucial, and hence is designated a "theorem".

1.7. Theorem. *Suppose that D is an integral domain. Then so is the ring $D[x]$ of polynomials with coefficients in D.*

Proof. Let $p(x) = p_0 + p_1x + p_2x^2 + \cdots + p_mx^m$ and $q(x) = q_0 + q_1x + q_2x^2 + \cdots + q_nx^n$ be polynomials of degree m and n respectively. (Note that the hypothesis that their degrees are defined means that the polynomials are non-zero.)

Then $p_m \neq 0 \neq q_n$, and

$$r(x) := p(x)q(x) = r_0 + r_1x + r_2x^2 + \cdots + r_{n+m}x^{n+m}$$

is a polynomial of degree $n + m$, given that $r_{n+m} = p_mq_n \neq 0$ since D is an integral domain.

Thus $D[x]$ is an integral domain.

□

1.8. Remark. The following useful fact was embedded in the above proof. For ease of reference, we isolate it as a remark.

If D is an integral domain and $0 \neq p(x), q(x) \in D[x]$, then

$$\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x)).$$

1.9. Proposition. *Let $(R, +, \cdot)$ be a unital, commutative ring. The following statements are equivalent.*

(a) *R is an integral domain.*

(b) *If $a, b, c \in R$ with $0 \neq a$, and if $ab = ac$, then $b = c$.*

Proof.

(a) implies (b). Suppose that $(R, +, \cdot)$ is an integral domain, and let $a, b, c \in R$. If $a \neq 0$ and $ab = ac$, then $a(b - c) = 0$ implies that $b - c = 0$, i.e. $b = c$.

(b) implies (a). Now suppose that $(R, +, \cdot)$ is a unital commutative ring and that condition (b) holds. If there exists $a \neq 0$ and $b \in R$ with $ab = 0$, then $ab = 0 = a0$. By hypothesis, $b = 0$. But then a is not a left divisor of zero. Since R is commutative, a is not a right divisor of zero either. In other words, R has no divisors of zero, and as such, it is an integral domain.

□

1.10. Proposition 1.9 identifies integral domains as precisely those commutative, unital rings where the cancellation law (for multiplication) holds. As pointed out at the beginning of this section, this is useful in trying to solve equations.

Suppose, for example, that one is given a polynomial equation

$$p_0 + p_1 x + p_2 x^2 + \cdots + p_n x^n = 0$$

with coefficients $p_k \in D$, where $(D, +, \cdot)$ is an integral domain and $p_n \neq 0$. Suppose furthermore that the polynomial factors into linear terms. (This hypothesis is non-trivial, and need not hold in general. For example, the polynomial $q(x) = x^2 + 1$ with coefficients in $\mathbb{R}$ does not factor into linear terms.)

We may then write

$$p_n(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n) = 0,$$

with each $\alpha_j \in D$. The fact that we are in an integral domain D means that the only solutions to this equation are found by taking $x = \alpha_j$ for some j.

When the ring of coefficients is not an integral domain, it is possible that we may have many other solutions. For example, consider the polynomial equation

$$q(x) = \left(x - \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right)\left(x - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

in one variable with coefficients in $\mathbb{M}_2(\mathbb{R})$ (which is not an integral domain). Then

Then

$$q\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

even though

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \notin \left\{\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right\}.$$

1.11. Proposition. *Let $2 \leq n \in \mathbb{N}$. Then $(\mathbb{Z}_n, +, \cdot)$ is an integral domain if and only if n is prime.*

Proof. Suppose first that n is prime. Suppose that $a, b \in \mathbb{Z}_n$ and that $a \cdot b = 0$. Then n divides ab. But n prime then implies that n either divides a or that n divides b. That is, $a = 0$ or $b = 0$ in $\mathbb{Z}_n$. Thus $\mathbb{Z}_n$ has no divisors of zero.

Next, suppose that n is not prime, say that $n = m_1 m_2$, where $1 < m_1, m_2 < n$. Then $m_1 \neq 0 \neq m_2 \in \mathbb{Z}_n$, and $m_1 m_2 = n \bmod n = 0$ in $\mathbb{Z}_n$, proving that m_1 and m_2 are joint divisors of zero in $\mathbb{Z}_n$. Thus $\mathbb{Z}_n$ is not an integral domain.

□

1.12. Exercise. Let $p \in \mathbb{N}$ be a prime number. Then $\mathbb{Q}(\sqrt{p}) = \{a + b\sqrt{p} : a, b \in \mathbb{Q}\}$ is an integral domain, using the usual addition and multiplication of real numbers.

2. The character of a ring

2.1. In dealing with integers, rational numbers and real numbers, we know that if we keep adding a (non-zero) number to itself many times, we can never end up getting the value 0. It might also seem counterintuitive that such a thing could ever happen under any circumstance (which I believe is sufficiently vague to classify as a true statement). Ah, but then we look at the watch that grandmaman gave us, and realise that if it is one o'clock now, and if we keep moving forward in time (as we are wont to do), then – twelve hours from now, and twelve hours from then – it will be one o'clock again. Somehow, adding twelve times one hour is, insofar as the face of our grandmaman's watch is concerned, the same as adding zero hours.

We shall have more to say about grandmaman's watch once we begin to examine quotient rings.

In the meantime, let us observe that the phenomenon observed on the face of our grandmaman's watch occurs in a number of rings, and it is an extremely useful phenomenon to keep in mind. (Just one more reason to appreciate your grandmaman.)

2.2. Notation. Let $(R, +, \cdot)$ be a ring and $n \in \mathbb{N}$. Given $r \in R$, we denote by $n\,r$ the element

$$n\,r := \sum_{j=1}^{n} r = r + r + \cdots + r \quad (n \text{ times}).$$

2.3. Definition. *Let R be a ring. We define the* ***character*** *of R as*

$$\operatorname{CHAR}(R) := \min\{n \in \mathbb{N} : n\,r = 0 \text{ for all } r \in R\},$$

if such a natural number $n \in \mathbb{N}$ exists. Otherwise, we say that $\operatorname{CHAR}(R) = 0$.

2.4. Examples.

(a) Each of the following rings has characteristic zero: $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$.
(b) If $n \in \mathbb{N}$, then $\operatorname{CHAR}(\mathbb{Z}_n, +, \cdot) = n$. In particular, $\operatorname{CHAR}(\mathbb{Z}_{12}) = 12$, which explains in part what is going on with grandmaman's watch.

2.5. Proposition. *Let R be a unital ring. If* $\text{CHAR}(R) \neq 0$*, then*

$$\text{CHAR}(R) = \min\{n \in \mathbb{N} : n\,1 = 0\}.$$

Proof. Suppose first that $\gamma := \text{CHAR}(R) \neq 0$. Then $\gamma\, r = 0$ for all $r \in R$, and so $\gamma\, 1 = 0$. In particular, $\gamma \in \{n \in \mathbb{N} : n\,1 = 0\}$. It follows that $\{n \in \mathbb{N} : n\,1 = 0\} \neq \varnothing$ and that

$$\min\{n \in \mathbb{N} : n\,1 = 0\} \leq \gamma = \text{CHAR}(R).$$

Next, if $\kappa := \min\{n \in \mathbb{N} : n\,1 = 0\}$, then for any $r \in R$,

$$\kappa\, r = r + r + \cdots r = (1 + 1 + \cdots + 1)r = (\kappa\, 1)r = 0r = 0,$$

and so $\text{CHAR}(R) \leq \kappa$. That is,

$$\text{CHAR}(R) \leq \min\{n \in \mathbb{N} : n\,1 = 0\}.$$

Combining these inequalities yields the desired result.

□

The above Proposition will make our work *significantly* easier when it comes to computing the characteristic of a *unital* ring, even those that on the surface appear rather complicated.

2.6. Examples.

(a) Consider $R = \mathbb{Z}_2 \oplus \mathbb{Z}_3 = \{(a, b) : a \in \mathbb{Z}_2, b \in \mathbb{Z}_3\}$. We claim that $\text{CHAR}(R) = 6$.

Indeed, the identity of R is $e := (1, 1)$. Note that as an ordered list, we have that

$$(1e, 2e, 3e, 4e, 5e, 6e) = ((1, 1), (0, 2), (1, 0), (0, 1), (1, 2), (0, 0)).$$

That is, $6 = \min\{n \in \mathbb{N} : n\, e = 0\}$, and so by the above Proposition 2.5,

$$\text{CHAR}(R) = 6.$$

(b) Consider next $R = \mathbb{Z}_2 \oplus \mathbb{Z}_4 = \{(a, b) : a \in \mathbb{Z}_2, b \in \mathbb{Z}_4\}$. Let $e := (1, 1)$ denote the multiplicative identity of R.

Then

$$(1e, 2e, 3e, 4e) = ((1, 1), (0, 2), (1, 3), (0, 0)).$$

By Proposition 2.5, we deduce that

$$\text{CHAR}(R) = 4.$$

(c) Let $R = \mathbb{Z}_5\langle x, y, z\rangle$ denote the ring of polynomials in three non-commuting variables x, y and z with coefficients in $\mathbb{Z}_5$. Then R is a unital ring with identity $e = 1 \in \mathbb{Z}_5$. Thus

$$(1e, 2e, 3e, 4e, 5e) = (1, 2, 3, 4, 0),$$

and so by Proposition 2.5,

$$\text{CHAR}(\mathbb{Z}_5\langle x, y, z\rangle) = 5.$$

2.7. Remark. To learn mathematics, it is not sufficient to merely read mathematics, although reading mathematics is certainly an important step that cannot be overlooked. In addition to reading, one should always be asking oneself questions about what one is reading.

- What was the key step in the proof of a result? Where was it used?
- Can we eliminate any of the hypotheses? Can we generalise the proof?

In the example above, we saw that $\text{CHAR}(\mathbb{Z}_2 \oplus \mathbb{Z}_3) = 6 = 2 \cdot 3$, whereas $\text{CHAR}(\mathbb{Z}_2 \oplus \mathbb{Z}_4) = 4 = \max\{2, 4\}$.

Why do these results appear to be different? What (if anything) is really going on?

These are the kinds of questions one should be asking oneself throughout the learning process. Yes, it is time-consuming, but the better one understands mathematics, the less one has to memorise. In a perfect world, one would only need to memorise definitions. (I've yet to figure out how to get around that!)

For example, rather than trying to memorise a proof line-by-line, one would only need to remember the key step, relying upon oneself to fill in the routine steps of the proof. The time invested in learning these things at the beginning is time saved later on.

2.8. Proposition. *Let R be a commutative ring, and suppose that $p := \text{CHAR}(R)$ is prime. If $r, s \in R$, then*

$$(r+s)^p = r^p + s^p.$$

Proof. Consider

$$(r+s)^p = r^p + \binom{p}{1} r^{p-1} s^p + \binom{p}{2} r^{p-2} s^2 + \cdots + \binom{p}{p-1} r^1 s^{p-1} + s^p.$$

Here, for $1 \leq k \leq p-1$,

$$\binom{p}{k} = \frac{p!}{k!(p-k)!}$$

Now, p is prime and $1 \leq k, (p-k) < p$ implies that p does not divide $k!$ nor $(p-k)!$. Since p divides the numerator, it follows that p divides $\binom{p}{k}$, $1 \leq k \leq p-1$. But then

$$\binom{p}{k} r^{p-k} s^k = 0, \quad 1 \leq k \leq p-1,$$

and so

$$(r+s)^p = r^p + 0 + 0 + \cdots + 0 + s^p = r^p + s^p.$$

□

Again, the reader should be asking (and hopefully answering) the question: where was the commutativity of R used in this proof?

3. Fields - an introduction

3.1. In Section 1 of this Chapter, we identified a property of rings which is equivalent to the ring satisfying the cancellation law for multiplication, namely: we asked that the ring be an integral domain. We argued that this is a requirement if we hope to solve equations that involve elements of that ring.

Note that $(\mathbb{Z}, +, \cdot)$ is an integral domain, and yet if we only knew about integers and we didn't know anything about rational numbers, we could not solve the equation

$$2x = 1024.$$

That is, we could not do it by simply multiplying 2 by $\frac{1}{2}$, because $\frac{1}{2} \notin \mathbb{Z}$. (Trial and error would still work, I suppose.)

This leads us to consider rings which are even more special than integral domains. In some ways, these new rings will behave almost as well as the real numbers themselves.

3.2. Definition. *A field $(\mathbb{F}, +, \cdot)$ is a commutative, unital ring in which every non-zero element admits a multiplicative inverse. That is, if $0 \neq x \in \mathbb{F}$, then there exists $y \in \mathbb{F}$ such that $xy = 1 = yx$. As we have seen, such an element y is unique and denoted by x^{-1}.*

3.3. Remark. We remark that by Example 1.3(d), every field $\mathbb{F}$ is automatically an integral domain. Since $\mathbb{F}$ is unital, it must have at least two elements, namely 0 and 1.

3.4. Examples.

(a) $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ are all fields, using the usual notions of addition and multiplication.

(b) $(\mathbb{Z}, +, \cdot)$ is not a field, since $0 \neq 2 \in \mathbb{Z}$ does not admit a multiplicative inverse in $\mathbb{Z}$.

(c) Let $\mathbb{F} := \{0, 1, x, y\}$, equipped with the following addition and multiplication operations.

+	0	1	x	y
0	0	1	x	y
1	1	0	y	x
x	x	y	0	1
y	y	x	1	0

and

·	0	1	x	y
0	0	0	0	0
1	0	1	x	y
x	0	x	y	1
y	0	y	1	x

.

Then one can verify that $(\mathbb{F}, +, \cdot)$ is a field. Note that $\mathbb{F}$ has $4 = 2^2$ elements. Coincidence? We shall see.

3.5. Theorem. *If $(D,+,\cdot)$ is an integral domain and D is finite, then it is a field.*

Proof. Suppose that $(D,+,\cdot)$ is an integral domain with $m \in \mathbb{N}$ elements, say $D = \{0, 1, d_3, d_4, \ldots, d_m\}$.

Clearly $1 = 1^{-1}$ is invertible. Now suppose that $3 \leq k \leq m$. Then

$$d_k D := \{d_k d : d \in D\} \subseteq D.$$

Moreover, if there exist $x, y \in D$ with $d_k x = d_k y$, then $x = y$ by virtue of the fact that D is an integral domain.

Thus $d_k D$ has exactly m elements, and from this we conclude that $d_k D = D$. In particular, there exists $x \in D$ such that $d_k x = 1$. Since D is commutative, $x d_k = 1$ as well.

That is, every non-zero element of D is invertible, and so D is a field.

□

3.6. Corollary. *Let $2 \leq n \in \mathbb{N}$. Then $(\mathbb{Z}_n,+,\cdot)$ is a field if and only if n is prime.*

Proof. Suppose first that $(\mathbb{Z}_n,+,\cdot)$ is a field. Then, by Remark 3.3, $\mathbb{Z}_n$ is an integral domain. By Proposition 1.11, we see that n is prime.

Conversely, suppose that $2 \leq n \in \mathbb{N}$ is prime. By Proposition 1.11, $(\mathbb{Z}_n,+,\cdot)$ is an integral domain (with $n < \infty$ elements). By Theorem 3.5 above, it is a field.

□

3.7. Example. If $p \in \mathbb{N}$ is prime, then $(\mathbb{Q}(\sqrt{p}),+,\cdot)$ is a field.

Recall that $\mathbb{Q}(\sqrt{p}) = \{a + b\sqrt{p} : a, b \in \mathbb{Q}\}$. Clearly $\mathbb{Q}(\sqrt{p}) \subseteq \mathbb{R}$. To prove that it is a ring, it suffices to apply the Subring Test, Proposition 2.4.10.

Note that if $r_1 := a_1 + b_1\sqrt{p}, r_2 := a_2 + b_2\sqrt{p} \in \mathbb{Q}$, then

$$r_1 - r_2 = (a_1 - a_2) + (b_1 - b_2)\sqrt{p} \in \mathbb{Q}(\sqrt{p}),$$

since $\mathbb{Q}$ is a ring which implies that $a_1 - a_2, b_1 - b_2 \in \mathbb{Q}$.

Moreover,

$$r_1 \cdot r_2 = (a_1 + b_1\sqrt{p})(a_2 + b_2\sqrt{p}) = (a_1 b_1 + p a_2 b_2) + (a_1 b_2 + a_2 b_1)\sqrt{p} \in \mathbb{Q}(\sqrt{p}),$$

since $a_1 b_1 + p a_2 b_2, a_1 b_2 + a_2 b_1 \in \mathbb{Q}$.

By the Subring Test, $\mathbb{Q}(\sqrt{p})$ is a subring of $\mathbb{R}$. Since $\mathbb{R}$ is commutative, so is $\mathbb{Q}(\sqrt{p})$, and $e = 1 + 0\sqrt{p} = 1$ serves as the identity element of $\mathbb{Q}(\sqrt{p})$ under multiplication.

Recall that $\mathbb{R}$ is a field, and so it is an integral domain. As such, $\mathbb{R}$ has no divisors of zero. From this it follows that $\mathbb{Q}(\sqrt{p})$ has no divisors of zero, for otherwise they would be divisors of zero in $\mathbb{R}$.

So far we have that $\mathbb{Q}(\sqrt{p})$ is an integral domain. There remains to show that every non-zero element of $\mathbb{Q}(\sqrt{p})$ is invertible.

Suppose that $0 \neq r := a + b\sqrt{p} \in \mathbb{Q}(\sqrt{p})$. By Exercise 7 below, $a^2 - pb^2 \neq 0$. Thinking of r as an element of $\mathbb{R}$, where we know it is invertible, we may write

$$r^{-1} = \frac{1}{a + b\sqrt{p}} = \frac{1}{a + b\sqrt{p}} \left(\frac{a - b\sqrt{p}}{a - b\sqrt{p}} \right) = \frac{a - b\sqrt{p}}{a^2 - pb^2}.$$

That is,

$$r^{-1} = \frac{a}{a^2 - pb^2} - \frac{b}{a^2 - pb^2}\sqrt{p} \in \mathbb{Q}(\sqrt{p}).$$

Thus $\mathbb{Q}(\sqrt{p})$ is a field.

Supplementary Examples.

S3.1. Example. We invite the student to verify that

$$\mathbb{Q}(\sqrt{2},\sqrt{3}) := \{a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mid a, b, c, d \in \mathbb{Q}\}$$

is a subfield of $\mathbb{R}$. That is, it is a subset of $\mathbb{R}$ which is a field using the addition and multiplication it inherits from $\mathbb{R}$.

S3.2. Example. Let $R = \prod_{n=1}^{\infty} \mathbb{Z}_2 = \{(a_n)_{n=1}^{\infty} : a_n \in \mathbb{Z}_2, n \geq 1\}$ be the product ring as defined in the previous Chapter.

If $a := (a_n)_{n=1}^{\infty} \in R$, then

$$a + a = (a_n)_n + (a_n)_n := (a_n + a_n)_n.$$

But $a_n \in \mathbb{Z}_2$, so $a_n + a_n = 0$ for all $n \geq 1$, and therefore $a + a = 0$.

This shows that R has finite characteristic (equal to 2), despite the fact that it is an infinite ring.

S3.3. Example. Let D be an integral domain. If $e \in D$ is an **idempotent**, i.e. if $e = e^2$, then $e \in \{0, 1\}$. Indeed,

$$0 = e^2 - e = e(e - 1)$$

implies that $e = 0$ or $e = 1$.

On the other hand, for any real number $r \in \mathbb{R}$,

$$E := \begin{bmatrix} 1 & r \\ 0 & 0 \end{bmatrix} \in \mathbb{M}_2(\mathbb{R})$$

satisfies $E^2 = E$. This is a very serpentine way of showing that $\mathbb{M}_2(\mathbb{R})$ is not an integral domain.

S3.4. Example. Let $\mathbb{D}_4$ denote the set of all 2×2 matrices of the form

$$\begin{bmatrix} a & b \\ b & a + b \end{bmatrix},$$

where a and $b \in \mathbb{Z}_2$. We add and multiply elements of $\mathbb{D}_4$ using the usual matrix operations.

We invite the reader to prove that $\mathbb{D}_4$ is a field under these operations.

Note. This example is due to R.A. Dean, and thus our choice of notation.

S3.5. Example. If R is a commutative ring, then $R[x]$ is not a field. After all, $q(x) := x \in R[x]$, but if $p(x) := (q(x))^{-1} \in R[x]$, then as we have seen (see Remark 1.8)

$$\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x)) = \deg(p(x)) + 1 \geq 1,$$

so that $p(x)q(x) \neq 1$.

S3.6. Example. Let R be a commutative ring, and let $R[[x]]$ denote the ring of formal power series defined in Example 2.4.1.

We leave it as an exercise for the reader to prove that $R[[x]]$ is an integral domain. (Hint. if $0 \neq p(x) = \sum_{n=0}^{\infty} p_n x^n \in R[[x]]$, let $\xi(p(x)) := \min\{k \geq 0 : p_k \neq 0\}$. Show that if $0 \neq q(x) \in R[[x]]$, then $\xi(p(x)q(x)) = \xi(p(x)) + \xi(q(x))$. Why is this helpful?)

S3.7. Example. Let R be a commutative ring and $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0 \in R[x]$. We define the **derivative** of $q(x)$ to be

$$q'(x) := nq_n x^{n-1} + (n-1)q_{n-1}x^{n-2} + \cdots + 2q_2 x + 1q_1 + 0q_0 \in R[x].$$

We leave it as an exercise for the reader to check that the usual rules of differentiation of real functions holds in this setting:

- $(p(x) + q(x))' = p'(x) + q'(x)$;
- $(r_0 q(x))' = r_0 q'(x)$ for all $r_0 \in R$;
- $(p(x)q(x))' = p'(x)q(x) + p(x)q'(x)$.

Let $a \in R$. If $(x-a)^2 \mid q(x)$ in the sense that $q(x) = (x-a)^2 g(x)$ for some $g(x) \in R[x]$, then $(x-a) \mid q'(x)$. To see this, we use the third property above to write

$$\begin{aligned} q'(x) &= ((x-a)^2 g(x))' \\ &= [(x-a)^2]' g(x) + (x-a)^2 g'(x) \\ &= (x-a)'(x-a) + (x-a)(x-a)' + (x-a)^2 g'(x) \\ &= (x-a)((x-a)' + (x-a)' + (x-a)g'(x)) \\ &= (x-a)(2 + (x-a)g'(x)). \end{aligned}$$

Thus $(x-a) \mid q'(x)$.

We emphasise that elements of $R[x]$ are *not* functions acting on a set (e.g. the real line). As such, this is an abstract notion of *derivative*, which we have named "*derivative*" because it behaves like the derivative of real-valued functions on $\mathbb{R}$.

S3.8. Example. Let R be a unital ring. An element $m \in R$ is said to be **nilpotent of order** $k \geq 1$ if $m^k = 0 \neq m^{k-1}$. (Note that by convention, $m^0 := 1$. Using this convention, we see that 0 is always nilpotent of order 1.)

Observe that if $0 \neq m \in R$ is a nilpotent, then (even if R is commutative) R is not an integral domain. Indeed, m is nilpotent of order $k \geq 1$, then $m^1 m^{k-1} = m^k = 0$, although $m \neq 0 \neq m^{k-1}$.

Suppose that R is commutative and consider the set $\text{NIL}(R) := \{m \in R : m = 0 \text{ or } m \text{ is nilpotent of order } k \geq 1\}$. We claim that $\text{NIL}(R)$ is a subring of R. Indeed,

if $m_1, m_2 \in \text{NIL}(R)$ with orders k_1 and k_2 respectively, then

$$(m_1 - m_2)^{k_1+k_2} = \binom{k_1+k_2}{0} m_1^{k_1+k_2} + \binom{k_1+k_2}{1} m_1^{k_1+k_2-1} m_2^1 + \\ \cdots + \binom{k_1+k_2}{k_1+k_1-1} m_1^1 m_2^{k_1+k_2-1} + \binom{k_1+k_2}{k_1+k_2} m_2^{k_1+k_2}.$$

Note that in each term we have that $m_1^i m_j^j$ appears where either $i \geq k_1$ or $j \geq k_2$. In the first case, $m_1^i = 0$ and in the second case, $m_2^j = 0$. Thus $m_1 - m_2 \in \mathcal{N}$ and it has order at most $k_1 + k_2$.

Also, if $k := \min(k_1, k_2)$, then

$$(m_1 m_2)^k = m_1^k m_2^k = 0,$$

since at least one of these terms is zero. (Note that the fact that R is commutative was important to both of these proofs.)

In fact, if $r \in R$ and $m \in \text{NIL}(R)$ is nilpotent of order k, then $(rm)^k = r^k m^k = r^k 0 = 0$, so $rm \in \text{NIL}(R)$. We shall examine such subrings in greater detail below, calling them *ideals* of R.

S3.9. Example. Let $\Lambda \neq \varnothing$ be a set with at least two elements, and for each $\lambda \in \Lambda$, suppose that D_λ is an integral domain. Then $\prod_{\lambda \in \Lambda} D_\lambda$ is not an integral domain.

Indeed, let $\lambda_1 \neq \lambda_2 \in \Lambda$, and let $d := (d_\lambda)_\lambda$, $e = (e_\lambda)_\lambda$ be elements of $\prod_\lambda D_\lambda$, where $d_\lambda = 0$ if $\lambda \neq \lambda_1$, otherwise $d_{\lambda_1} = 1$, while $e_\lambda = 0$ if $\lambda \neq \lambda_2$, otherwise $e_{\lambda_2} = 1$. Then $d \neq 0 \neq e$ in $\prod_\lambda D_\lambda$, and yet $de = 0$.

Since $d, e \in \oplus_\lambda D_\lambda$, we see that $\oplus_\lambda D_\lambda$ is not an integral domain either when Λ has more than one element.

S3.10. Example. We invite the reader to verify that

$$D := \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} \subseteq \mathbb{M}_2(\mathbb{Z}_2)$$

is an integral domain, using usual matrix addition and multiplication (with entries added and multiplied in $\mathbb{Z}_2$!!!).

Appendix

A3.1. Fields play a central role in algebra and number theory, in particular in the search for solutions to equations. It can be shown that they have a strong relation to Euclidean geometry by using field theory to demonstrate that one cannot, using a compass and straightedge, trisect an (arbitrary) angle or find a square whose area is the same as that as a circle of radius 1.

The solutions to a quadratic equation $ax^2 + bx + c = 0$ where a, b and $c \in \mathbb{R}$ and $0 \neq a$ are given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

There exist formulae, albeit more complicated than the one above, to identify the solutions of cubic and quartic equations. In 1799, the Italian mathematician **Paolo Ruffini** claimed that no such formula exists for a quintic equation (i.e. one of degree 5), although his argument contained errors which were finally resolved by the Norwegian **Niels Henrik Abel**, this being the same *Abel* whose name is used to describe *abelian groups*.

The crowning result in this area is due to **Évariste Galois**, who derived necessary and sufficient conditions for a polynomial equation to be solvable algebraically. Galois' contributions to algebra are extensive and quite amazing, especially considering that he died at the age of 20 of peritonitis, caused by a gunshot wound suffered in a dual. Go figure.

The original word for a field was the German word *Körper*, introduced by Richard Dedekind in 1871 (for subfields of the complex numbers), while the expression **field** was introduced much later (1893) by **E. Hastings Moore**. We mention in passing that *Körper* is actually the German word for "body", and the equivalent word *corps* is used to describe a field in French, while *cuerpo* is used to describe a field in Spanish. So English is the outlier, pretty much. How surprised the reader chooses to be by this is really up to the reader, frankly.

Exercises for Chapter 3

Exercise 3.1.
Our definition of a joint divisor of zero $0 \neq r$ in a ring R says that there exist $0 \neq x, y \in R$ such that

$$xr = 0 = ry.$$

If r is a joint divisor of zero in R, is it always the case that there exists $0 \neq z \in R$ such that

$$zr = 0 = rz?$$

Exercise 3.2.
Consider the ring of polynomials in two commuting variables $(\mathbb{Z}[x, y], +, \cdot)$ with coefficients in (the commutative, unital ring) $\mathbb{Z}$, as defined in Example 2.1.8(b).

Prove that $(\mathbb{Z}[x, y], +, \cdot)$ is an integral domain.

Exercise 3.3.
Prove that the *ring* of Gaussian integers $\mathbb{Z}[i]$ is indeed a ring, as the name suggests. (Hint: use the Subring Test. Which ring should $\mathbb{Z}[i]$ be contained in?)

Exercise 3.4.
For which rings $(R, +, \cdot)$ is the ring of polynomials $R[x]$ in one variable with coefficients in R an integral domain? Formulate and prove your conjecture.

Exercise 3.5.
Let $n_1, n_2, n_3 \in \mathbb{N}$ and suppose that $R = \mathbb{Z}_{n_1} \oplus \mathbb{Z}_{n_2} \oplus \mathbb{Z}_{n_3}$. Find

$$\text{CHAR}(R),$$

and formulate a conjecture of what the character of the ring

$$S := \oplus_{j=1}^{k} \mathbb{Z}_{n_j}$$

should be, given $n_1, n_2, \ldots, n_k \in \mathbb{N}$. Better yet, try to prove your conjecture!

Exercise 3.6.
Find a unital ring R and a polynomial $p(x) = p_0 + p_1 x + \cdots + p_n x^n$ of degree at least two such that the equation

$$p(x) = 0$$

admits infinitely many solutions in R.

Exercise 3.7.
Let $a, b \in \mathbb{Q}$ and suppose that $p \in \mathbb{N}$ is prime. Prove that $a^2 - pb^2 \neq 0$. This was used in Example 3.7 above.

Exercise 3.8. ***

Let D be an integral domain, and suppose that $\text{CHAR}(D) = 0$. Given $d \in D$ and $n \geq 1$, recall that $n\,d = d + d + \cdots + d$, the sum of n copies of d. If $n = 0$, let us define $n\,d = 0$, and for an integer $n \leq -1$, we shall define $n\,d = -((-n)d) = -(d+d+d+\cdots+d)$ (n times).

Prove that if $0 \neq d$, then $nd = 0$ if and only if $n = 0$.

We note that in some texts (e.g. [**Her75**]), this is taken as the *definition* of characteristic zero for integral domains. While this works for integral domains, this fails for more general commutative, unital rings.

Exercise 3.9. ***

Let $(D, +, \cdot)$ be an integral domain. Prove that if $\text{CHAR}(D) \neq 0$, then $\text{CHAR}(D)$ is a prime number.

Exercise 3.10.

Let $\mathbb{H}$ denote the ring of real Hamiltonians, defined in Example 2.4.1. Prove that $\mathbb{H}$ satisfies all of the properties of a field, except that multiplication need not be commutive. As pointed out in the Appendix to Chapter 2, we say that $\mathbb{H}$ is a *division ring* or a *skew field*.

CHAPTER 4

Homomorphisms, ideals and quotient rings

If at first you don't succeed... so much for skydiving.

Henny Youngman

1. Homomorphisms and ideals

1.1. Given two mathematical objects sharing a common structure – be they both sets, vector spaces, groups, rings or whatever – one is interested in studying the relationship between them. A common theme in mathematics is the notion of a *homomorphism*. The nature of a homomorphism depends upon the nature of the mathematical structure which our objects of interest share in common.

In the case of (arbitrary) sets A and B, there is essentially no algebraic structure, and so the **set homomorphisms** from A to B then consist of arbitrary maps whose domain is A and whose codomain is B. The kind of information one might glean from these relates to the only structure the sets have: their size. For example, if one knows that there exists an injective function $f : A \to B$ (i.e. a one-to-one function), then the size of B must be at least as great as the size of A. Indeed, this is how we compare the sizes of infinite sets – through the existence or non-existence of injections from the first to the second.

In the case of two vector spaces $\mathcal{V}$ and $\mathcal{W}$ over the same field (e.g. both vector spaces over $\mathbb{R}$), the morphisms are precisely the functions from $\mathcal{V}$ to $\mathcal{W}$ which preserve the vector space structure. What is the "vector space structure"? The two basic properties of a vector space $\mathcal{V}$ over $\mathbb{R}$ are that given two vectors x and y in $\mathcal{V}$ and a scalar $\kappa \in \mathbb{R}$, we may define their sum $x+y$ and scalar multiple κx as elements of $\mathcal{V}$. As such, we ask that our **vector space homomorphisms** $T : \mathcal{V} \to \mathcal{W}$ satisfy $T(x + y) = Tx + Ty$ and $T(\kappa x) = \kappa T(x)$ for all $x, y \in \mathcal{V}$ and $\kappa \in \mathbb{R}$. You will recall from your first linear algebra course that these are precisely the functions we refer to as **linear maps**.

A group $(G, \cdot)$ admits a single binary operation, namely multiplication. Given two groups $(G, \cdot)$ and $(H, *)$, a **group homomorphism** $\varphi : G \to H$ is a map that satisfies

$$\varphi(g_1 \cdot g_2) = \varphi(g_1) * \varphi(g_2) \quad \text{for all } g_1, g_2 \in G.$$

Observe that the first multiplication, namely $g_1 \cdot g_2$, is taking place in $(G, \cdot)$, while the second product $\varphi(g_1) * \varphi(g_2)$ is taking place in $(H, *)$. (Something similar happened with linear maps above: the sum of x and y occurred in $\mathcal{V}$, whereas the sum of Tx and Ty occurred in $\mathcal{W}$.) More often than not, multiplication in both G and H is denoted simply by juxtaposition, and the reader will often see a group homomorphism defined as a map $\varphi : G \to H$ which satisfies $\varphi(g_1 g_2) = \varphi(g_1)\varphi(g_2)$ for all $g_1, g_2 \in G$. The reader is expected to keep track of which multiplication is being used where.

Which brings us to rings. Rings $(R, +, \cdot)$ admit two operations, as did vector spaces. It is not surprising, therefore, that – like vector space homomorphisms – our notion of a homomorphism for a ring should require that we respect both of these operations.

1.2. Definition. *Let $(R, +, \cdot)$ and $(S, \dot{+}, *)$ be two rings. A map $\varphi : R \to S$ is said to be a* ***ring homomorphism*** *if*

$$\varphi(r_1 + r_2) = \varphi(r_1) \dot{+} \varphi(r_2)$$

and

$$\varphi(r_1 \cdot r_2) = \varphi(r_1) * \varphi(r_2)$$

for all $r_1, r_2 \in R$.

1.3. Similar to the case when dealing with group homomorphisms, the reader is often expected to keep track of the two different notions of addition and of multiplication arising in R and S, and a ring homomorphism from R to S is (informally) defined as a map $\varphi : R \to S$ that satisfies $\varphi(r_1 + r_2) = \varphi(r_1) + \varphi(r_2)$ and $\varphi(r_1 r_2) = \varphi(r_1)\varphi(r_2)$ for all $r_1, r_2 \in R$.

If there is a risk of confusion, we shall adopt the very strict notation from Definition 1.2. In general, however, the notation given in the previous paragraph should not cause misunderstanding, and we shall bow to tradition and use it with almost reckless abandon.

1.4. The homomorphisms which appear in different branches of algebra themselves share a great many properties. If one has seen them at work in the theory of groups, for example, then one may already have a very good idea of what to expect from ring homomorphisms. In particular, the so-called *First, Second and Third Isomorphism Theorems* each have their variants for groups, rings, fields, algebras and even vector spaces. (Unsurprisingly, our main focus in this course will be the ring-theoretic versions of these results.)

In particular, we remind the reader of the *First Homomorphism Theorem for (real) Vector spaces*, which says that if $\mathcal{V}$ and $\mathcal{W}$ are vector spaces over $\mathbb{R}$, and if $T : \mathcal{V} \to \mathcal{W}$ is an $\mathbb{R}$-linear map, then

$$\operatorname{ran} T := T(\mathcal{V}) = \{Tv : v \in \mathcal{V}\} \simeq \mathcal{V}/\ker T,$$

where $\ker T = \{v \in \mathcal{V} : Tv = 0\}$ is the **kernel** of the linear map T.

Here $\simeq$ refers to a **vector space isomorphism**; that is, a map $\varphi : \operatorname{ran} T \to \mathcal{V}/\ker T$ which is linear and bijective.

The First Homomorphism Theorem for Vector spaces requires the notion of a *quotient vector space* of a vector space with one of its subspaces. As we shall soon see, a similar result holds for ring homomorphisms between two rings R and S. The notion of a *quotient ring* will also make its appearance. Unlike the vector space setting, however, it will not be fruitful to consider the quotient of a ring by an arbitrary subring. Instead, we shall see that the kernel of a ring homomorphism is a subring which has special properties, and these are the subrings which will be of most interest to us.

1.5. Definition. *Let R and S be rings and let $\varphi : R \to S$ be a (ring) homomorphism. The **kernel** of φ is the set*

$$\ker \varphi := \{r \in R : \varphi(r) = 0\}.$$

1.6. Remark. Note that if R, S, and φ are as above, then $K := \ker \varphi$ is a subring of R. To show this, we shall use the Subring Test, Proposition 2.4.10.

Note first that $0_R \in \ker \varphi \neq \varnothing$. Indeed,

$$\varphi(0_R) + 0_S = \varphi(0_R) = \varphi(0_R + 0_R) = \varphi(0_R) + \varphi(0_R).$$

By Proposition 2.4.2, $0_S = \varphi(0_R)$, and therefore $0_R \in \ker \varphi$.

Let $k_1, k_2 \in K$, Then

$$\varphi(k_1 - k_2) = \varphi(k_1) - \varphi(k_2) = 0 - 0 = 0,$$

and

$$\varphi(k_1 k_2) = \varphi(k_1)\varphi(k_2) = 0 \cdot 0 = 0.$$

Thus $k_1 - k_2$ and $k_1 k_2 \in K$, showing that K is a subring of R.

1.7. Example. Let R and S be arbitrary rings and define $\varphi : R \to S$ via $\varphi(r) = 0$ for all $r \in R$. Then φ is a ring homomorphism, known as the *trivial homomorphism* from R to S. As the reader may have already surmised, not much information about R and S can be gleaned from studying this map.

1.8. Example. Let $p \in \mathbb{N}$ be a prime number and consider $R = \mathbb{Q}(\sqrt{p}) = S$. Define

$$\varphi : \begin{array}{ll} \mathbb{Q}(\sqrt{p}) & \to \mathbb{Q}(\sqrt{p}) \\ a + b\sqrt{p} & \mapsto a - b\sqrt{p} \end{array}.$$

Then

$$\begin{aligned} \varphi((a_1 + b_1\sqrt{p}) + (a_2 + b_2\sqrt{p})) &= \varphi((a_1 + a_2) + (b_1 + b_2)\sqrt{p}) \\ &= (a_1 + a_2) - (b_1 + b_2)\sqrt{p} \\ &= (a_1 - b_1\sqrt{p}) + (a_2 - b_2\sqrt{p}) \\ &= \varphi(a_1 + b_1\sqrt{p}) + \varphi(a_2 + b_2\sqrt{p}). \end{aligned}$$

Similarly,

$$\begin{aligned}\varphi((a_1+b_1\sqrt{p})(a_2+b_2\sqrt{p})) &= \varphi((a_1a_2+pb_1b_2)+(a_1b_2+b_1a_2)\sqrt{p})\\ &= (a_1a_2+pb_1b_2)-(a_1b_2+b_1a_2)\sqrt{p}\\ &= (a_1-b_1\sqrt{p})(a_2-b_2\sqrt{p})\\ &= \varphi(a_1+b_1\sqrt{p})\varphi(a_2+b_2\sqrt{p}).\end{aligned}$$

Thus φ is a ring homomorphism of $\mathbb{Q}(\sqrt{p})$ into (in fact *onto*) itself.

1.9. Example. Let $n \in \mathbb{N}$. The map

$$\begin{array}{rccc}\varphi: & \mathbb{Z} & \to & \mathbb{Z}_n\\ & a & \mapsto & a \bmod n\end{array}$$

is called the **canonical homomorphism** of $\mathbb{Z}$ onto $\mathbb{Z}_n$. That is is indeed a ring homomorphism is left as an exercise for the reader.

1.10. Example. Let R, S and T be rings, and set $\mathcal{A} = R \oplus S \oplus T$, so that $\mathcal{A}$ is also a ring. The map $\varphi_1 : \mathcal{A} \to R$ defined by $\varphi(r,s,t) = r$ for all $(r,s,t) \in \mathcal{A}$ is a ring homomorphism, as is the map $\varphi_2 : \mathcal{A} \to S$ defined by $\varphi_2(r,s,t) = s$ and the map $\varphi_3 : \mathcal{A} \to T$ defined by $\varphi_3(r,s,t) = t$ for all $(r,s,t) \in \mathcal{A}$.

This is again a good time to make sure that we are thinking while reading these notes: what is so special about having three rings R, S and T? Would it have worked with four rings? With twenty? With infinitely many? If not, what goes wrong?

Is the map $\psi : \mathcal{A} \to S \oplus R$ defined by $\psi(r,s,t) = (s,r)$ a ring homomorphism? So many questions! That is indeed the nature of mathematics.

1.11. Example. Consider the map

$$\begin{array}{rccc}\varphi: & \mathbb{M}_2(\mathbb{R}) & \to & \mathbb{R}\\ & \begin{bmatrix}a & b\\ c & d\end{bmatrix} & \mapsto & a.\end{array}$$

Then

$$\varphi\left(\begin{bmatrix}a_1 & b_1\\ c_1 & d_1\end{bmatrix}+\begin{bmatrix}a_2 & b_2\\ c_2 & d_2\end{bmatrix}\right) = a_1+a_2,$$

but

$$\begin{aligned}\varphi\left(\begin{bmatrix}a_1 & b_1\\ c_1 & d_1\end{bmatrix}\begin{bmatrix}a_2 & b_2\\ c_2 & d_2\end{bmatrix}\right) &= \varphi\left(\begin{bmatrix}a_1a_2+b_1c_2 & a_1b_2+b_1d_2\\ c_1a_2+d_1c_2 & c_1b_2+d_1d_2\end{bmatrix}\right)\\ &= a_1a_2+b_1c_2\\ &\neq a_1a_2\\ &= \varphi\left(\begin{bmatrix}a_1 & b_1\\ c_1 & d_1\end{bmatrix}\right)\varphi\left(\begin{bmatrix}a_2 & b_2\\ c_2 & d_2\end{bmatrix}\right)\end{aligned}$$

in general. Yes, it can happen for *some matrices*, but it is not always true. For example, if $A := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B := \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$, then $\varphi(A) = 0 = \varphi(B)$, but

$$\varphi(AB) = \varphi(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}) = 1 \neq 0 = \varphi(A)\varphi(B).$$

Thus φ is *not* a ring homomorphism.

1.12. Definition. *Let R and S be rings and let $\varphi : R \to S$ be a homomorphism.*

- *If φ is injective, then we say that φ is a* ***monomorphism****.*
- *If φ is surjective, we say that φ is a* ***epimorphism****.*
- *If φ is bijective, we say that it is an* ***isomorphism****. In this case, we say that R and S are* ***isomorphic****, and we write $R \simeq S$.*
- *Finally, a homomorphism of a ring R into itself is referred to as an* ***endomorphism****, while a bijective endomorphism is called an* ***automorphism*** *of R.*

1.13. Remark. From the point of view of set theory, there is not much difference between the sets $A = \{a, b, c\}$ and $B = \{1, 2, 3\}$. True, one consists of letters and the other of numbers, but those are not set-theoretic properties. That fact that each set has three elements is a set-theoretic property, as is the fact that each set admits a total of 8 subsets. In fact, from the point of view of Set Theory, B is just the set A (after relabelling the entries of A). If we wish to impress people, we might say that A and B are *isomorphic as sets*. Why we would want to impress people, especially the kinds of people who would be impressed by such a statement, is another matter altogether.

In much the same way, however, one should think of two isomorphic rings R and S are being essentially the same, although presented differently. While one ring might consist of numbers and another of functions, or matrices, or polynomials with coefficients in some ring, any *ring-theoretic* property of R will be possessed by S and vice-versa. That is, *as rings*, S is the just the ring R, with the elements and the operations relabelled.

1.14. Examples. The proofs of the following facts are computations which we leave to the reader.

(a) Consider the map

$$\begin{array}{rccc} \alpha : & \mathbb{T}_2(\mathbb{C}) & \to & \mathbb{T}_3(\mathbb{C}) \\ & \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} & \mapsto & \begin{bmatrix} t_{11} & 0 & t_{12} \\ 0 & 0 & 0 \\ 0 & 0 & t_{22} \end{bmatrix}. \end{array}$$

Then α is a monomorphism of $\mathbb{T}_2(\mathbb{C})$ into $\mathbb{T}_3(\mathbb{C})$. Clearly it is not surjective, and so it is not an isomorphism.

(b) Consider the map

$$\begin{array}{rccc} \beta: & \mathbb{T}_3(\mathbb{C}) & \to & \mathbb{T}_2(\mathbb{C}) \\ & \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ 0 & t_{22} & t_{23} \\ 0 & 0 & t_{33} \end{bmatrix} & \mapsto & \begin{bmatrix} t_{22} & t_{23} \\ 0 & t_{33} \end{bmatrix}. \end{array}$$

Then φ is an epimorphism of $\mathbb{T}_3(\mathbb{C})$ onto $\mathbb{T}_2(\mathbb{C})$. Note that

$$\beta(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & -9 \\ 0 & 0 & 6 \end{bmatrix}) = \begin{bmatrix} 2 & -9 \\ 0 & 6 \end{bmatrix} = \beta(\begin{bmatrix} -\pi & 2 & 43 \\ 0 & 2 & -9 \\ 0 & 0 & 6 \end{bmatrix}),$$

and thus β is not injective. In particular, β is not an isomorphism.

(c) Let $S \in \mathbb{M}_n(\mathbb{C})$ be an invertible matrix. Consider the map

$$\begin{array}{rccc} \gamma: & \mathbb{M}_n(\mathbb{C}) & \mapsto & \mathbb{M}_n(\mathbb{C}) \\ & A & \mapsto & S^{-1}AS. \end{array}$$

Then

$$\gamma(A+B) = S^{-1}(A+B)S = S^{-1}AS + S^{-1}BS = \gamma(A) + \gamma(B),$$

and

$$\gamma(AB) = S^{-1}(AB)S = (S^{-1}AS)(S^{-1}BS) = \gamma(A)\ \gamma(B).$$

Thus γ is a homomorphism.

If $\gamma(A) = \gamma(B)$, then $S^{-1}AS = S^{-1}BS$, whence

$$A = S(S^{-1}AS)S^{-1} = S(S^{-1}BS)S^{-1} = B,$$

showing that γ is injective.

Also, if $X \in \mathbb{M}_n(\mathbb{C})$, then $Y := SXS^{-1} \in \mathbb{M}_n(\mathbb{C})$, and

$$\gamma(Y) = S^{-1}YS = X,$$

showing that γ is surjective.

Hence γ is an isomorphism.

1.15. Exercise. Let

$$R := \left\{ \begin{bmatrix} a & b \\ -b & a \end{bmatrix} : a, b \in \mathbb{R} \right\} \subseteq \mathbb{M}_2(\mathbb{R}).$$

Verify that R is a subring of $\mathbb{M}_2(\mathbb{R})$ using the usual matrix operations. Define

$$\begin{array}{rccc} \eta: & \mathbb{C} & \to & R \\ & a+bi & \mapsto & \begin{bmatrix} a & b \\ -b & a \end{bmatrix}. \end{array}$$

Verify that η is an isomorphism of rings.

This is an example of what we meant when we suggested that isomorphic rings are just a single ring, presented in more than one way. The isomorphism η allows us

to think of complex numbers as certain 2×2 matrices with real entries, multiplied and added using usual matrix operations.

Any ring theoretic property of $\mathbb{C}$ (the fact that it is a field, for example) will be shared by R and vice-versa. We should be careful to note that $\mathbb{C}$ is isomorphic to $R = \eta(\mathbb{C})$, and *not* to $\mathbb{M}_2(\mathbb{R})$, which is only the co-domain of η. (This comment also gives the reader the opportunity to brush up on the difference between the *range* and the *codomain* of a function, in case "it's been a while".)

1.16. Exercise. Let R be a ring, and suppose that R is isomorphic to a field $\mathbb{F}$. Prove that R must be a field.

The following observations are basic, but very useful.

1.17. Proposition. *Let R and T be rings and suppose that $\varphi : R \to T$ is a homomorphism. Let S be a subring of R.*

(a) *If $r \in R$ and $n \in \mathbb{N}$, then*

$$\varphi(nr) = n\varphi(r) \qquad \text{and} \qquad \varphi(r^n) = \varphi(r)^n.$$

In particular, $\varphi(0) = 0$ and $\varphi(-r) = -\varphi(r)$ for all $r \in R$.

(b)

$$\varphi(S) = \{\varphi(s) : s \in S\} \text{ is a subring of } T.$$

In particular, $\operatorname{ran}\varphi = \varphi(R)$ is a subring of T.

(c) *If S is commutative, then so is $\varphi(S)$.*

(d) *If R is unital, $T \neq \{0\}$ and φ is surjective, then T is unital and $1_T = \varphi(1_R)$.*

Proof. Since this may be the first time we see this kind of argument, we shall be extremely pedantic and we shall use 0_R and 1_R to denote the additive and multiplicative identities of R, and a similar notation will be used for other rings. In the future, however, the reader will be expected to understand which ring one is working with, and therefore which identities one is using. That is, the same notation 0 will be used to denote the additive identity in more than one ring at a time.

(a) Note that $\varphi(0_R) = \varphi(0_R + 0_R) = \varphi(0_R) + \varphi(0_R)$. Thus

$$\begin{aligned} 0_T &= \varphi(0_R) - \varphi(0_R) \\ &= (\varphi(0_R) + \varphi(0_R)) - \varphi(0_R) \\ &= \varphi(0_R) + (\varphi(0_R) - \varphi(0_R)) \\ &= \varphi(0_R) + 0_T \\ &= \varphi(0_R) \end{aligned}$$

Meanwhile, for all $r \in R$,

$$0_T = \varphi(0_R) = \varphi(r + (-r)) = \varphi(r) + \varphi(-r),$$

implying that $\varphi(-r) = -\varphi(r)$. (Note that R and S are abelian groups under addition, so that we have $0_T = \varphi(-r) + \varphi(r)$ for free.)

We shall prove that $\varphi(nr) = n\varphi(r)$ by induction, and leave the (similar) proof that $\varphi(r^n) = \varphi(r)^n$ for all $r \in R$ and $n \in \mathbb{N}$ to the reader.

Let $r \in R$ be arbitrary. First note that if $n = 1$, then $\varphi(1r) = \varphi(r) = 1\varphi(r)$, which provides the initial step in our induction process.

In general, if $\varphi(nr) = n\varphi(r)$, then

$$\begin{aligned}\varphi((n+1)r) &= \varphi(nr + 1r) = \varphi(nr + r)\\ &= \varphi(nr) + \varphi(r) = n\varphi(r) + \varphi(r)\\ &= (n+1)\varphi(r).\end{aligned}$$

This completes the induction step, and the proof that $\varphi(nr) = n\varphi(r)$ for all $n \in \mathbb{N}$, $r \in R$.

(b) Since $0_R = 0_S \in S$, $0_T = \varphi(0_S) \in \varphi(S)$, and thus the latter is non-empty. Let us apply the Subring Test: Let $x, y \in \varphi(S)$. Then there exist $r, s \in S$ such that $x = \varphi(r)$ and $y = \varphi(s)$. Thus

$$x - y = \varphi(r) - \varphi(s) = \varphi(r - s) \in \varphi(S),$$

and

$$xy = \varphi(r)\varphi(s) = \varphi(rs) \in \varphi(S),$$

since $r, s \in S$ and S a subring of R implies that $r - s$ and $rs \in S$.

By the Subring Test, $\varphi(S)$ is a subring of T.

(c) If S is commutative, then for $xy \in \varphi(S)$, there exist $r, s \in S$ such that $x = \varphi(r)$ and $y = \varphi(s)$. Thus

$$xy = \varphi(r)\varphi(s) = \varphi(rs) = \varphi(sr) = \varphi(s)\varphi(r) = yx,$$

proving that $\varphi(S)$ is commutative as well.

(d) Let $t \in T$ be arbitrary. Since φ is surjective, $t = \varphi(r)$ for some $r \in R$. Thus

$$t = \varphi(r) = \varphi(1_R \cdot r) = \varphi(1_R)\varphi(r) = \varphi(1_R) \cdot t,$$

and similarly,

$$t = \varphi(r) = \varphi(r \cdot 1_R) = \varphi(r)\varphi(1_R) = t \cdot \varphi(1_R).$$

Since $t \in T$ was arbitrary, $\varphi(1_R)$ is a multiplicative identity for T. Since this must be unique, $\varphi(1_R) = 1_T$.

□

Here is a wonderful little application of these results.

1.18. Proposition. *Let $n \in \mathbb{N}$ and consider the decimal expansion of n,*

$$n = a_k 10^k + a_{k-1}10^{k-1} + \cdots + a_2 10^2 + a_1 10 + a_0.$$

Then n is divisible by 9 if and only if the sum $a_k + a_{k-1} + \cdots + a_2 + a_1 + a_0$ of its digits is divisible by 9.

Proof. Define the map

$$\begin{array}{rccc}\beta: & \mathbb{N} & \to & \mathbb{Z}_9\\ & m & \mapsto & m \bmod 9.\end{array}$$

We leave it to the reader to verify that β is a ring homomorphism. Thus

$$\begin{aligned}\beta(n) &= \beta(a_k 10^k + a_{k-1}10^{k-1} + \cdots + a_2 10^2 + a_1 10 + a_0)\\ &= \beta(a_k)\beta(10^k) + \beta(a_{k-1})\beta(10^{k-1}) + \cdots + \beta(a_2)\beta(10^2) + \beta(a_1)\beta(10) + \beta(a_0).\end{aligned}$$

But $\beta(10) = 1$, and so by Proposition 1.17 above, $\beta(10^j) = \beta(10)^j = 1^j = 1$ for all $j \geq 1$. Thus

$$\begin{aligned}n \text{ MOD } 9 &= \beta(n)\\ &= \beta(a_k) + \beta(a_{k-1}) + \cdots + \beta(a_2) + \beta(a_1) + \beta(a_0)\\ &= (a_k + a_{k-1} + \cdots + a_2 + a_1 + a_0) \text{ MOD } 9.\end{aligned}$$

In particular, n is divisible by 9 if and only if $\beta(n) = n \text{ MOD } 9 = 0$. But $\beta(n) = 0$ if and only if the sum of the digits of n is divisible by 9.

Putting these two things together, we see that n is divisible by 9 if and only if the sum of its digits is divisible by 9, as claimed.

□

2. Ideals

2.1. We now turn our attention to a special class of subrings of a given ring R. Elements of this class will be called *ideals* of R. They will be of interest because they can always be realised as the kernels of *some* homomorphism from R into *some* ring S, and as mentioned in paragraph 1.4 above, they will play a central role in establishing the three Isomorphism Theorems for rings.

2.2. Definition. *Let R be a ring. A **left ideal** (resp. **right ideal**) of R is a subring K of R with the property that if $k \in K$ and $r \in R$, then $rk \in K$ (resp. $kr \in K$).*

*If K is both a left and a right ideal of R, we say that K is an **ideal** (sometimes also a **two-sided ideal**) of R. We write $K \lhd R$ when K is a (two-sided) ideal of R.*

2.3. Note that the condition that $rk \in K$ for all $r \in R$ and $k \in K$ implies in particular that $jk \in K$ for all $j, k \in K$. As such, in the definition of an ideal, it suffices to ask that K be an additive subgroup of R with the property that rk and $kr \in K$ for all $r \in R$, $k \in K$.

That is, by combining these remarks with the Subring Test 2.4.10, we see that to check that K is an ideal of R, it suffices to prove that

- $k - j \in K$ for all $j, k \in K$; and
- rk and $kr \in K$ for all $k \in K$, $r \in R$.

For ease of reference, we shall call this the **Ideal Test**.

We also point out that if R is commutative, then every left or right ideal is automatically a two-sided ideal.

2.4. Example. Consider the ring $\mathbb{Z}$. Let $m \in \mathbb{N}$ and set $K = m\mathbb{Z} = \{mz : z \in \mathbb{Z}\}$. By a (hopefully by now) routine application of the Subring Test, we see that K is a subring of $\mathbb{Z}$.

If $k \in K$, then $k = mz$ for some $z \in \mathbb{Z}$. Thus for any $r \in \mathbb{Z}$,

$$kr = rk = r(mz) = m(rz),$$

showing that $rk, kr \in K$. Thus K is an ideal of $\mathbb{Z}$.

2.5. Example. Let $R = \mathbb{M}_3(\mathbb{Z})$, and let $K = \left\{ \begin{bmatrix} x & 0 & 0 \\ y & 0 & 0 \\ z & 0 & 0 \end{bmatrix} : x, y, z \in \mathbb{Z} \right\}$.

That K is a subring of R is left as a computation for the reader. If $R = [r_{i,j}] \in \mathbb{M}_3(\mathbb{Z})$ and $K = \begin{bmatrix} x & 0 & 0 \\ y & 0 & 0 \\ z & 0 & 0 \end{bmatrix} \in K$, then

$$\begin{aligned} RK &= \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} x & 0 & 0 \\ y & 0 & 0 \\ z & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} r_{11}x + r_{12}y + r_{13}z & 0 & 0 \\ r_{21}x + r_{22}y + r_{23}z & 0 & 0 \\ r_{31}x + r_{32}y + r_{33}z & 0 & 0 \end{bmatrix} \in K. \end{aligned}$$

Thus K is a left ideal of $\mathbb{M}_3(\mathbb{Z})$. Note that $L := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in K$ and $S := \begin{bmatrix} 1 & 2 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in R$, but

$$LS = S \notin K,$$

proving that K is *not* a right ideal, and *a fortiori* not a two-sided ideal.

The most important example of an ideal stems from the following result. Although its proof is routine, its significance is profound. As we shall eventually see, every ideal of R arises in this manner, for an appropriate choice of S and corresponding homomorphism φ.

2.6. Theorem. *Let R and S be rings and let $\varphi : R \to S$ be a homomorphism. Then*

$$K := \ker \varphi$$

is a two-sided ideal in R.

Moreover, φ is injective if and only if $\ker \varphi = \{0\}$.

Proof. Let $k_1, k_2 \in K$. Then $\varphi(k_1) = 0 = \varphi(k_2)$. Thus

$$\varphi(k_1 - k_2) = \varphi(k_1) - \varphi(k_2) = 0 - 0 = 0,$$

whence $k_1 - k_2 \in K$.

If $r \in R$ is arbitrary, then

$$\varphi(rk) = \varphi(r)\varphi(k) = \varphi(r)0 = 0 = 0\varphi(r) = \varphi(k)\varphi(r) = \varphi(kr),$$

proving that rk and kr both lie in K.

By the Ideal Test, K is a two-sided ideal of R.

By Proposition 1.17 above, we know that $\varphi(0) = 0$ for any homomorphism. If φ is injective, then we conclude that $0 \neq r$ implies that $0 \neq \varphi(r)$, and thus $\ker\varphi = \{0\}$.

Conversely, suppose that $\ker\varphi = \{0\}$. If $\varphi(r_1) = \varphi(r_2)$ for some $r_1, r_2 \in R$, then

$$0 = \varphi(r_1) - \varphi(r_2) = \varphi(r_1 - r_2),$$

implying that $r_1 - r_2 \in \ker\varphi = \{0\}$, whence $r_1 = r_2$. Thus φ is injective.

□

2.7. Because of limitations of time, we shall mostly concentrate on two-sided ideals in our discussions below. Our readers should nevertheless question themselves as to which results hold for one-sided ideals, and for those that do not hold, what can be done to "repair" the results?

2.8. Proposition. *Let R be a ring and J, K be ideals of R. The following are also ideals of R.*

(a) $J + K := \{j + k : j \in J, k \in K\}$.
(b) $J \cap K := \{m : m \in J \text{ and } m \in K\}$.
(c) $JK := \{\sum_{i=1}^{q} j_i k_i : q \geq 1, j_i \in J, k_i \in K, 1 \leq i \leq q\}$.

Proof. In all cases, we shall apply the Ideal Test from paragraph 2.3.

(a) Let $a = j_1 + k_1$ and $b = j_2 + k_2 \in J + K$, where, as the notation suggests, $j_1, j_2 \in J$ and $k_1, k_2 \in K$. Then

$$a - b = (j_1 + k_1) - (j_2 + k_2) = (j_1 - j_2) + (k_1 - k_2).$$

But J and K are rings, by virtue of their being ideals, and thus $j_1 - j_2 \in J$ and $k_1 - k_2 \in K$. It follows that $a - b \in J + K$.

If $r \in R$, then $ra = rj_1 + rk_1$, while $ar = j_1 r + k_1 r$. Since J and K are ideals, rj_1 and $j_1 r \in R$, while rk_1 and $k_1 r \in K$. Hence $ra, ar \in J + K$.

By the Ideal Test, $J + K \lhd R$.

(b) Let $a, b \in J \cap K$. Since each of J and K is a ring, $a - b \in J$ and $a - b \in K$, whence $a - b \in J \cap K$.

Also, given $r \in R$, the fact that both J and K are ideals implies that $ra, ar \in J$ and $ra, ar \in K$, whence $ar, ra \in J \cap K$.

By the Ideal Test, $J \cap K \lhd R$.

(c) Let $a = \sum_{m=1}^{p} j_m k_m$ and $b = \sum_{n=1}^{q} \widehat{j}_n \widehat{k}_n \in JK$, where $j_m, \widehat{j}_n \in J$, $1 \le m \le p$, $1 \le n \le q$ and $k_m, \widehat{k}_n \in K$, $1 \le m \le p$, $1 \le n \le q$. Since J is a ring, $-\widehat{j}_n \in J$ for all $1 \le n \le q$, and so

$$\begin{aligned} a - b &= \left(\sum_{m=1}^{p} j_m k_m \right) - \left(\sum_{n=1}^{q} \widehat{j}_n \widehat{k}_n \right) \\ &= \left(\sum_{m=1}^{p} j_m k_m \right) + \left(\sum_{n=1}^{q} (-\widehat{j}_n) \widehat{k}_n \right) \end{aligned}$$

lies in JK.

If $r \in R$, then $r j_m \in J$ and $k_m r \in K$ for all $1 \le m \le p$, so that

$$ra = r \left(\sum_{m=1}^{p} j_m k_m \right) = \sum_{m=1}^{p} (r j_m) k_m \in JK,$$

and

$$ar = \left(\sum_{m=1}^{p} j_m k_m \right) r = \sum_{m=1}^{p} j_m (k_m r) \in JK.$$

By the Ideal Test, $JK \lhd R$.

□

2.9. Definition. *Let R be a ring and $F \subseteq R$. We denote by $\langle F \rangle$ the smallest ideal of R which contains F, i.e.,*

$$\langle F \rangle := \cap \{K \lhd R : F \subseteq K\}.$$

We refer to this as the ***ideal generated by*** *F.*

If $F = \{k\}$ consists of a single element, we typically write $\langle k \rangle$ instead of $\langle \{k\} \rangle$; also, we say in this case that $\langle F \rangle$ is a ***principal*** *or* ***singly-generated ideal****.*

2.10. Remarks.

(a) It is not hard to show that (and we leave it as an exercise for the reader to show that) if R is a unital ring and $\varnothing \ne F \subseteq R$, then

$$\langle F \rangle = \{ \sum_{i=1}^{n} r_i x_i s_i : n \in \mathbb{N}, r_i, s_i \in R, x_i \in F, 1 \le i \le n \}.$$

When R is non-unital, there is a slight issue that arises with this description. For example, suppose that

$$R = \left\{ \begin{bmatrix} 0 & x \\ 0 & 0 \end{bmatrix} : x \in \mathbb{R} \right\},$$

which forms a non-unital subring of $\mathbb{M}_2(\mathbb{R})$. Let $a = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, and consider $F = \{a\}$. For any $r, s \in R$, it is straightforward to verify that $ras = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$,

and so a would not belong to the set described upon the right-hand side of the above equation, although $a \in \langle a \rangle$ by definition. (The set on the right hand side of the equation would just contain the zero matrix in this case.) As such, this description of the ideal generated by a set F does not hold in general when R is non-unital.

(b) What is true in general is that if R is any ring, $a \in R$, and we define $ma := a + a + \cdots + a$(m times), then $ma \in \langle a \rangle$ for all $m \in \mathbb{N}$, and so $(-m)a := -(ma) \in \langle a \rangle$ as well. Writing $\mathbb{Z}a := \{ma : m \in \mathbb{Z}\}$, we find that in general,

$$\langle a \rangle = \{ma + ra + as + \sum_{k=1}^{n} r_k a s_k : m \in \mathbb{Z}, n \geq 1, r, s, r_k, s_k \in R, 1 \leq k \leq n\}.$$

If we apply this to the example in part (a) above, we find that

$$\langle a \rangle = \left\{ \begin{bmatrix} 0 & m \\ 0 & 0 \end{bmatrix} : m \in \mathbb{Z} \right\}.$$

2.11. Exercise. Find all ideals of $\mathbb{Z}$, proving thereby that every ideal of $\mathbb{Z}$ is a principal ideal. (An integral domain in which every ideal is principal is called a *principal ideal domain*, denoted by PID. Thus $\mathbb{Z}$ is a PID. We shall return to these later.)

2.12. Exercise. Let $R = \mathbb{Z}$, and suppose that $J = \langle 6 \rangle$, $K = \langle 4 \rangle$. Find $J + K$, $J \cap K$ and JK. Generalise your result to arbitrary ideals $J = \langle m \rangle$ and $K = \langle n \rangle$ of $\mathbb{Z}$.

2.13. Example. Let $R = (\mathcal{C}([0,1], \mathbb{R}), +, \cdot)$. We leave it to the reader to verify that $J := \{f \in R : f(x) = 0,\ x \in [0, \frac{1}{3}]\}$ and $K := \{g \in R : g(x) = 0,\ x \in [\frac{1}{4}, 1]\}$ are ideals of R.

- Consider $J \cap K$. If $h \in J \cap K$, then $h \in J$ implies that $h(x) = 0$ for all $x \in [0, \frac{1}{3}]$ and $h(x) = 0$ for all $x \in [\frac{1}{4}, 1]$. But then $h(x) = 0$ for all $x \in [0,1]$, so $h = 0$.

  That is, $J \cap K = \{0\}$.
- Next, consider JK. If $f \in J$ and $g \in K$, then

$$f(x)g(x) = \begin{cases} 0g(x) = 0 & \text{if } x \in [0, \frac{1}{3}] \\ f(x)0 = 0 & \text{if } x \in [\frac{1}{4}, 1] \end{cases}.$$

  Thus $f(x)g(x) = 0$ for all $x \in [0,1]$, so $fg = 0$.

  More generally, if $h \in JK$, then $h = \sum_{i=1}^{n} f_i g_i$ for some $n \in \mathbb{N}$ and some $f_i \in J$ and $g_i \in K$, $1 \leq i \leq n$. Since each $f_i g_i = 0$ from the above computation, we see that $h = \sum_{i=1}^{n} 0 = 0$.

  Thus $JK = \{0\}$.

- As for $J + K$, let $h \in R$ be any continuous function such that $h(x) = 0$, $x \in [\frac{1}{4}, \frac{1}{3}]$. Define

$$f(x) = \begin{cases} 0 & x \in [0, \frac{1}{3}] \\ h(x) & x \in [\frac{1}{3}, 1] \end{cases} \quad \text{and} \quad g(x) = \begin{cases} h(x) & x \in [0, \frac{1}{4}] \\ 0 & x \in [\frac{1}{4}, 1] \end{cases}.$$

  Then f and g are continuous on $[0, 1]$ (you should check this!), $f \in J$ and $g \in K$, so $h = f + g \in J + K$.

  Thus $\{h \in \mathcal{C}([0, 1], \mathbb{R}) : h(x) = 0,\ x \in [\frac{1}{4}, \frac{1}{3}]\} \subseteq J + K$. Conversely, if $f \in J$ and $g \in K$, then $f + g$ is continuous because each of f and g is, and for $x \in [\frac{1}{4}, \frac{1}{3}]$, we see that

$$f + g(x) = f(x) + g(x) = 0 + 0 = 0.$$

  Hence $J + K \subseteq \{h \in \mathcal{C}([0, 1], \mathbb{R}) : h(x) = 0,\ x \in [\frac{1}{4}, \frac{1}{3}]\}$. Combining these two containments,

$$J + K = \{h \in \mathcal{C}([0, 1], \mathbb{R}) : h(x) = 0,\ x \in [\frac{1}{4}, \frac{1}{3}]\}.$$

3. Cosets and quotient rings

3.1. The notion of cosets appears not only in ring theory, but in many areas of algebra, including the theories of linear algebra, groups, rings, fields and algebras over a field. Hopefully, many of you will have already come across this concept in linear algebra.

For example, if we consider the real vector space $\mathcal{V} = \mathbb{R}^3$, then, given any subspace $\mathcal{W}$ of $\mathcal{V}$, we may consider the cosets $x+\mathcal{W} := \{x+w : w \in \mathcal{W}\}$. What are the subspaces of $\mathcal{V}$? The answer depends upon the dimension of $\mathcal{W}$.

- If $\dim \mathcal{W} = 0$, then $\mathcal{W} = \{0\}$, the origin of the vector space. In this case, $x + \mathcal{W} = \{x\}$, a single point in $\mathcal{V}$ - which we may think of as the origin translated by x.
- If $\dim \mathcal{W} = 1$, then $\mathcal{W}$ is a line through the origin. In this case, $x + \mathcal{W}$ is a line parallel to $\mathcal{W}$, passing through $x + 0 = x$.
- If $\dim \mathcal{W} = 2$, then $\mathcal{W}$ is a plane through the origin. In this case, $x + \mathcal{W}$ is a plane parallel to $\mathcal{W}$, passing through x.
- Finally, if $\dim \mathcal{W} = 3$, then $\mathcal{W} = \mathcal{V}$ and $x + \mathcal{W} = \mathcal{V}$ is the entire space.

Note that a line, a plane and the entire vector space each consist of infinitely many points, but they are single quantities unto themselves: we are dealing with a *single* line, a *single* plane or a *single* copy of the entire space. Each coset is a set. The *sets* have many points in them (except in the case where $\mathcal{W} = \{0\}$ above!), but that's neither here nor there.

The concept of a coset is not purely an abstract construction, however. You will recall our earlier reference to your grandmaman's watch, and our earlier threat to resume this thread of conversation. As you shall now see, that threat was anything

but idle. Assuming that your grandmaman was a high-ranking member of the army and had a 24-hour watch with the hours indicated $0000, 0100, 0200, \ldots, 2200, 2300$ but with the date notably absent from the face of her watch (it's been a while since she retired, obviously), we see that 1400 hours refers not just to two in the afternoon today, but in fact two in the afternoon tomorrow, the next day, yesterday and in fact *every day*. If we take a philosophical leap of faith and make the simplifying assumption that time has no beginning and no end (a mild exaggeration, perhaps), then 1400 on your grandmaman's watch would really indicate a member of $\{1400 + 2400n : n \in \mathbb{Z}\}$, which is the set of all two-in-the-afternoons throughout eternity, moving forwards or backwards in time. In other words, *two in the afternoon* on your grandmaman's watch is just the set $1400 + 2400\mathbb{Z}$, a coset of the subring $2400\mathbb{Z}$ of $\mathbb{Z}$.

The collection of all two-in-the-afternoons we think of as a single object, denoted by 1400 on the face of your grandmaman's watch. Observe that if we wanted to know what her watch would say three hours from now, we would add $0300 + 2400\mathbb{Z}$ to $1400 + 2400\mathbb{Z}$ to get $1700 + 2400\mathbb{Z}$, i.e. 5 in the afternoon. Of course, if we wanted to know the time *tomorrow plus three hours from now*, we would add $2700+2400\mathbb{Z}$ to the current time $1400+2400\mathbb{Z}$ to get $4100+2400\mathbb{Z}$. But there is no 4100 on the face of your grandmaman's watch, because this is indicated by 1700, since $4100 = 1700 + 2400$, and thus $4100 + 2400\mathbb{Z}$ is represented by $1700 + 2400\mathbb{Z}$.

The point is, if you understand your grandmaman's watch, then you understand cosets. Of course, the advent of smart-phone technology is quickly making your grandmaman and indeed everyone's grandmaman obsolete, but that unfortunate circumstance is way beyond the scope of these lecture notes.

To summarise: all of this suggests that the concept of a coset is either not as complicated as we are tempted to believe it might be, or that your grandmaman possessed some crazy-mad sophisticated technology. Your humble author prefers the former explanation.

3.2. Definition. *Let R be a ring and $L \subseteq R$ be a subring of R. A **coset** of L is a set of the form*

$$x + L := \{x + m : m \in L\}.$$

*(In other words, it is a translation of L by x.) We say that x is a **representative** of the coset $x + L$.*

The collection of all cosets of L in R is denoted by

$$R/L := \{x + L : x \in R\}.$$

3.3. The astute reader (hopefully you!) will have noticed that we referred to x above as *a* representative of the coset $x + L$, as opposed to *the* representative of $x + L$. There's a very good reason for this. Unless $L = \{0\}$ is the trivial subring, then the representative of $x + L$ is not unique, as the next result proves.

3.4. Proposition. *Let R be a ring and L be a subring of R. Let $x, y \in R$. Then the following statements are equivalent:*

(a) $x + L = y + L$; *that is, both x and y are representatives of the coset $x + L$.*
(b) $x - y \in L$.

Proof.

(a) Suppose that $x + L = y + L$. Since $0 \in L$, $x + 0 = x \in y + L$ and so there exists $m \in L$ such that $x = y + m$. That is, $x - y = m \in L$.
(b) Suppose next that $m := x - y \in L$, so that $x = y + m$. Then for any $n \in L$, $x + n = (y + m) + n = y + (m + n) \in y + L$. Hence $x + L \subseteq y + L$.

But we also have that $y = x - m$, and hence given any $n \in L$, $y + n = (x + (-m)) + n = x + (n - m) \in L$, proving that $y + L \subseteq x + L$.

Taken together, these two statements show that $x + L = y + L$.

□

3.5. Examples.

(a) Let $R = \mathbb{Z}$ and $L = \langle 12 \rangle$. Let us show that $7 + L = 19 + L \neq 4 + L$.

Observe that

$$\begin{aligned} 7 + \langle 12 \rangle &= 7 + \{\ldots, -24, -12, 0, 12, 24, 36, \ldots\} \\ &= \{\ldots, -17, -5, 7, 19, 43, 55, \ldots,\} \\ &= 19 + \langle 12 \rangle, \end{aligned}$$

but that

$$\begin{aligned} 4 + \langle 12 \rangle &= 4 + \{\ldots, -24, -12, 0, 12, 24, 36, \ldots\} \\ &= \{\ldots, -20, -8, 4, 16, 28, 40, \ldots\} \\ &\neq 7 + \langle 12 \rangle. \end{aligned}$$

(For example, $4 \in 4 + \langle 12 \rangle$, but $4 \notin 7 + \langle 12 \rangle$.)

(b) Note that

$$\mathcal{D}_3(\mathbb{R}) = \left\{ \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix} : d_1, d_2, d_3 \in \mathbb{R} \right\}$$

is a subring of $\mathbb{T}_3(\mathbb{R})$. It is not an ideal of $\mathbb{T}_3(\mathbb{R})$ since, for example, the matrix

$$E_{12} := \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in \mathbb{T}_3(\mathbb{R}) \qquad \text{and} \qquad D_1 := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in \mathcal{D}_3(\mathbb{R}),$$

but

$$E_{12} = D_1 \cdot E_{12} \notin \mathcal{D}_3(\mathbb{R}).$$

Fix a matrix, say $T = \begin{bmatrix} 17 & 2 & \pi - e^2 \\ 0 & -12.1 & -\sqrt{2} \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{T}_3(\mathbb{R})$. Then

$$T + \mathcal{D}_3(\mathbb{R}) = \left\{ \begin{bmatrix} x & 2 & \pi - e^2 \\ 0 & y & -\sqrt{2} \\ 0 & 0 & z \end{bmatrix} : x, y, z \in \mathbb{R} \right\}.$$

3.6. In linear algebra, one shows that if $\mathcal{V}$ is a vector space (over a field $\mathbb{F}$) and $\mathcal{W}$ is a subspace of $\mathcal{V}$, then we can view the set $\mathcal{V}/\mathcal{W}$ of cosets of $\mathcal{W}$ in $\mathcal{V}$ as a vector space using the operations

$$(x + \mathcal{W}) + (y + \mathcal{W}) := (x + y) + \mathcal{W}$$

and

$$\kappa(x + \mathcal{W}) := (\kappa x) + \mathcal{W}$$

for all $x, y \in \mathcal{V}$ and $\kappa \in \mathbb{F}$. Unfortunately, in the case of rings, it is not true that the cosets of a subring can always be made into a ring using the canonical operations. This is where the notion of an ideal takes on all of its importance.

3.7. Theorem. *Let R be a ring and L be a subring of R. The set R/L of all cosets of L in R becomes a ring using the operations*

$$(x + L) + (y + L) := (x + y) + L$$

and

$$(x + L)(y + L) := (xy) + L$$

for $x + L, y + L \in R/L$ if and only if L is an ideal of R.

Proof.

- Suppose first that L is an ideal of R. Let us prove that R/L is a ring under the operations defined above. Note that the Subring Test does not apply here – R/L is not contained in anything that we already know to be a ring. As such, we are forced to use the definition of a ring, which involves checking a great many things. Life isn't always easy.

  The first thing that we have to do is to verify that the operations are *well-defined*. (See the Appendix to this Chapter to learn more about when we have to check that operations are well-defined.)

Note that if $x_1 + L = x_2 + L$ and $y_1 + L = y_2 + L$, then $m_1 := x_1 - x_2$ and $m_2 := y_1 - y_2 \in L$ by Proposition 3.4. Thus

$$(x_1 + y_1) - (x_2 + y_2) = (x_1 - x_2) + (y_1 - y_2) = m_1 + m_2 \in L,$$

and so

$$(x_1 + y_1) + L = (x_2 + y_2) + L,$$

again by Proposition 3.4

Similarly,

$$(x_1y_1)-(x_2y_2) = x_1y_1-x_1y_2+x_1y_2-x_2y_2 = x_1(y_1-y_2)+(x_1-x_2)y_2 = x_1m_2+m_1y_2 \in L,$$

since L is an ideal. Thus

$$(x_1y_1) + L = (x_2y_2) + L.$$

That is, the operations of addition and multiplication we are using are well-defined. Now we get to the nitty-gritty of checking that R/L is a ring. Sigh. Nah, "chin up" – let's do it.

For all $r, s, t \in R$, $(r+L)+(s+L) := (r+s)+L \in R/L$, so that R/L is closed under addition. Also,

- $((r+L)+(s+L))+(t+L) = ((r+s)+L)+(t+L) = ((r+s)+t)+L$. Since addition in R is associative we may continue:

$$\begin{aligned}((r+s)+t)+L &= (r+(s+t))+L\\ &= (r+L)+((s+t)+L)\\ &= (r+L)+((s+L)+(t+L)).\end{aligned}$$

  That is,

$$((r+L)+(s+L))+(t+L) = (r+L)+((s+L)+(t+L)).$$

  Hence addition is associative in R/L.
- For all $r+L \in R/L$, $(r+L)+(0+L) = (r+0)+L = r+L = (0+L)+(r+L)$. That is, $0+L$ is an additive identity for R/L.
- For all $r+L \in R/L$,

$$\begin{aligned}(r+L)+((-r)+L) &= (r-r)+L\\ &= 0+L\\ &= ((-r)+r)+L)\\ &= ((-r)+L)+(r+L).\end{aligned}$$

  Hence $(-r)+L = -(r+L)$ is the additive inverse of $r+L$ in R/L.
- $(r+L)+(s+L) = (r+s)+L = (s+r)+L = (s+L)+(r+L)$, and so $(R/L,+)$ is an abelian group under addition.

Next, note that $(r+L)(s+L) := (rs)+L \in R/L$, so that R/L is closed under multiplication. Also,

- $((r+L)(s+L))(t+L) = ((rs)+L)(t+L) = ((rs)t)+L$. Since multiplication in R is associative we may continue:

$$\begin{aligned}((rs)t)+L &= (r(st))+L\\ &= (r+L)((st)+L)\\ &= (r+L)((s+L)(t+L)).\end{aligned}$$

  That is,

$$((r+L)(s+L))(t+L) = (r+L)((s+L)(t+L)).$$

Hence multiplication is associative in R/L.

- $((r+L)+(s+L))(t+L) = ((r+s)+L)(t+L) = ((r+s)t)+L$. Since multiplication distributes over addition in R we find that:

$$\begin{aligned}((r+s)t)+L &= (rt+st)+L \\ &= (rt+L)+(st+L) \\ &= (r+L)(t+L)+(s+L)(t+L).\end{aligned}$$

That is,

$$((r+L)+(s+L))(t+L) = (r+L)(t+L)+(s+L)(t+L).$$

- Finally, $(t+L)((r+L)+(s+L)) = (t+L)((r+s)+L) = (t(r+s))+L$. Since multiplication distributes over addition in R we find that:

$$\begin{aligned}(t(r+s))+L &= (tr+ts)+L \\ &= (tr+L)+(ts+L) \\ &= (t+L)(r+L)+((t+L)(s+L)).\end{aligned}$$

That is,

$$(t+L)((r+L)+(s+L)) = (t+L)(r+L)+(t+L)(s+L).$$

Thus multiplication distributes over addition in R/L.

We have now verified that conditions (A1) - (A5) and (M1) - (M4) from Definition 2.1.2 hold for R/L, and thus it is a ring.

- Conversely, suppose that L is a subring, but *not* an ideal of R. We argue by contradiction. To that end, suppose that R/L is a ring under the operations above. An easy calculation shows that $0+L$ must be the zero element of R/L.

The fact that L is a subring but not an ideal of R implies that there exists $m \in L$ and $r \in R$ such that either $rm \notin L$ or $mr \notin L$. Consider the case where $rm \notin L$. Then $m \in L$ implies that $m+L = 0+L$ and so

$$0+L = (r+L)(0+L) = (r+L)(m+L) = (rm)+L.$$

By Proposition 3.4, we see that $rm \in L$, a contradiction. The case where $mr \notin L$ is similarly handled and is left to the reader.

□

The following result, while easy to prove, is of great importance, because it establishes what ideals of a ring *are*. They are those subrings which arise as the kernel of *some* homomorphism of the ring. Why is this of interest? As stated more than once already in these course notes, the way that we investigate the relation between two comparable mathematical objects (in this course the objects under study are rings) is to use maps between those objects that respect the structure of those objects. This is why we are particularly interested in ring homomorphisms. They help us to understand how those two rings are related.

If R and S are rings, then the kernel of a ring homomorphism $\varphi : R \to S$ gives us a measure of how much information we are losing from the first ring when we

carry it into the second. Anything in the kernel gets sent to 0, and so (for example) if $r \in \ker \varphi$, then so is r^2, r^3, $r + r^2 + r^3$, etc., so the homomorphism by itself can't distinguish between these elements.

We get extra information from the kernel as well; for example, if the kernel of the ring homomorphism φ has exactly 10 elements in it, then not only are these 10 elements sent to 0, but in fact *any* element s in the range of φ is then the image of exactly 10 elements of R. The reason for this is that $\varphi(r_1) = s = \varphi(r_2)$ if and only if $\varphi(r_2 - r_1) = 0$; i.e. if $r_2 = r_1 + k$ for some $k \in \ker \varphi$. In particular, the map φ is injective if and only if its kernel is $\{0\}$, in which case φ loses no information – $\varphi(R)$ is isomorphic to R in this case. So the point is that ring homomorphisms are the object we really want to study.

We originally defined an ideal K of a ring R as a subring with the special property that $k \in K$ and $r \in R$ implies that both rk and kr lie in K. Let us (*very*) temporarily change the name of such subrings to "*multiplication absorbing subrings*". Now the question becomes: why should we care about *multiplication absorbing subrings*? Why did we study them in the first place?

Half of the reason lies in Theorem 2.6, which states that kernels of homomorphisms are in fact *multiplication absorbing subrings*. The other half of the reason is Theorem 3.8 below, which says that these are the only kinds of *multiplication absorbing subrings*. That is, *multiplication absorbing subrings* and kernels of ring homomorphisms are just two ways of looking at the same thing. So why bother with this alternate description? Suppose that we didn't know that these two objects were the same. If one were to present you with a subring L of a ring R and ask you if it is the kernel of some homomorphism (and therefore very important), how would you check if this is so? How would you look for a ring S and a homomorphism $\varphi : R \to S$? You can't say "Hey, Theorem 3.7 is the answer!", because Theorem 3.7 only tells you that a subring L is a *multiplication absorbing subring*. Knowing that L is the kernel of a homomorphism if and only if it is a *multiplication absorbing subring* tells us that instead of looking for *some* ring S and *some* homomorphism $\varphi : R \to S$ whose kernel is L, you only have to apply the *multiplication absorbing subring Test*, which – in our infinite wisdom – we actually refer to as the Ideal Test. Alternatively, it suffices to check that R/L is a ring under the canonical operations.

So why don't we use the expression "multiplication absorbing subring"? My guess is that it is because other mathematicians are jealous of me and my fancy terminology, but this is just speculation.

3.8. Theorem. *Let R be a ring, and let K be an ideal of R. Then there exists a ring S and a homomorphism $\varphi : R \to S$ such that $K = \ker \varphi$.*

Proof. By Theorem 3.7 above, R/K is a ring using the canonical operations. Let

$$\begin{array}{rccc} \varphi : & R & \to & R/K \\ & r & \mapsto & r + K. \end{array}$$

Then

$$\varphi(r_1 + r_2) = (r_1 + r_2) + K = (r_1 + K) + (r_2 + K) = \varphi(r_1) + \varphi(r_2),$$

and

$$\varphi(r_1 r_2) = (r_1 r_2) + K = (r_1 + K)(r_2 + K) = \varphi(r_1)\ \varphi(r_2).$$

Thus φ is a ring homomorphism.

The additive identity element of R/K is $0+K$, since $(r+K)+(0+K) = (r+0)+K = r + K$ for all $r \in R$. By Proposition 3.4,

$$\varphi(r) = r + K = 0 + K$$

if and only if $r = r - 0 \in K$. That is, $\ker \varphi = K$.

□

3.9. Example.
Let $R = \mathbb{R}[x]$ and $r \in \mathbb{R}$. Define $\delta_r(p(x)) = p(r)$.

$$\begin{array}{rccc} \delta_r : & \mathbb{R}[x] & \to & \mathbb{R} \\ & p(x) & \mapsto & p(r). \end{array}$$

Then

$$\delta_r((pq)(x)) = (pq)(r) = p(r)q(r) = \delta_r(p(x))\ \delta_r(q(x)),$$

and

$$\delta_r((p+q)(x)) = (p+q)(r) = p(r) + q(r) = \delta_r(p(x)) + \delta_r(q(x)).$$

Thus δ_r is a homomorphism, often referred to as **evaluation at** r.

Note that $p(x) \in \ker \delta_r$ if and only if $0 = \delta_r(p(x)) = p(r)$. That is,

$$\ker \delta_r = \{p(x) : p(r) = 0\}.$$

Let $f(x) = x - r \in \mathbb{R}[x]$. Then $f(r) = 0$, so $f \in \ker \delta_r$. Note that $\langle f(x) \rangle \subseteq \ker \delta_r$, since if $g(x) \in \langle f(x) \rangle$, then $g(x) = h(x)f(x)$ for some $h(x) \in \mathbb{R}[x]$ (why don't we need to consider finite sums of functions here?), and thus

$$\delta_r(g(x)) = g(r) = h(r)f(r) = h(r)0 = 0.$$

Conversely, if $p(x) \in \ker \delta_r$, then $p(r) = 0$, and so r is a root of $p(x)$. But then $f(x) = x - r$ divides $p(x)$, and so there exists $k(x) \in \mathbb{R}[x]$ such that $p(x) = k(x)f(x) \in \langle f(x) \rangle$.

(Note: the fact that $p(r) = 0$ implies that $x - r$ divides $p(x)$ is something that we know from high school in the setting of polynomials with coefficients in $\mathbb{R}$, but which we shall prove in greater generality in a later chapter (see Corollary 7.1.4)).

3.10. Example. Let $\kappa \in \mathbb{N}$ and consider the map

$$\begin{array}{rccc} \varphi: & \mathbb{Z}/\langle\kappa\rangle & \to & \mathbb{Z}_\kappa \\ & n+\langle\kappa\rangle & \mapsto & n \bmod \kappa. \end{array}$$

Once again, we must check that φ is well-defined, mustn't we? But if $n+\langle\kappa\rangle = m+\langle\kappa\rangle$, then $m-n \in \langle\kappa\rangle$, i.e. $m-n = \kappa z$ for some $z \in \mathbb{Z}$. Thus

$$\varphi(n) = n \bmod \kappa = n + (m-n) \bmod \kappa = m \bmod \kappa = \varphi(m),$$

proving that φ is well-defined.

Then

$$\begin{aligned} \varphi((n+\langle\kappa\rangle)+(m+\langle\kappa\rangle)) &= \varphi((n+m)+\langle\kappa\rangle) \\ &= (n+m) \bmod \kappa \\ &= n \bmod \kappa + m \bmod \kappa \\ &= \varphi(n+\langle\kappa\rangle) + \varphi(m+\langle\kappa\rangle), \end{aligned}$$

and similarly (the verification is left to the reader)

$$\varphi(n+\langle\kappa\rangle)(m+\langle\kappa\rangle)) = \varphi(n+\langle\kappa\rangle)\varphi(m+\langle\kappa\rangle),$$

proving that φ is a homomorphism.

For any $0 \le n < \kappa$, $\varphi(n+\langle\kappa\rangle) = n \bmod \kappa$, and thus φ is surjective.

Furthermore, $n+\langle\kappa\rangle \in \ker\varphi$ if and only if $n \bmod \kappa = 0 \bmod \kappa$, and this happens if and only if κ divides n. That is, it happens if and only if $n \in \langle\kappa\rangle$, or equivalently, when $n+\langle\kappa\rangle = 0+\langle\kappa\rangle$.

Thus the kernel of φ is trivial, and so φ is injective.

By definition, φ is an isomorphism, and so

$$\frac{\mathbb{Z}}{\langle\kappa\rangle} \simeq \mathbb{Z}_\kappa.$$

4. The Isomorphism Theorems

4.1. The importance of the First Isomorphism Theorem we shall prove below cannot be overstated. Such a result holds in most categories (i.e. vector spaces, groups, rings, fields, algebras, etc.), and it is a sublimely useful tool in analysing homomorphisms.

The other two Isomorphism Theorems are also quite useful, though the First Isomorphism Theorem really stands apart.

4.2. Theorem. (***The First Isomorphism Theorem***) *Let R and S be rings and suppose that $\varphi : R \to S$ is a homomorphism. Let $K := \ker \varphi$. The map*

$$\begin{array}{rccc} \tau: & R/K & \to & \varphi(R) \\ & r+K & \mapsto & \varphi(r) \end{array}$$

is an isomorphism. In other notation, $\varphi(R) \simeq R/\ker \varphi$.

Proof. Since τ is defined on cosets in terms of a (generally non-unique) representative of that coset, we shall need to prove that τ is well-defined.

Suppose that $r+K = s+K$ for some $r, s \in R$. Then $r-s \in K$, and so $\varphi(r)-\varphi(s) = \varphi(r-s) = 0$, i.e. $\varphi(r) = \varphi(s)$. Then $\tau(r+K) = \varphi(r) = \varphi(s) = \tau(s+K)$, proving that τ is indeed well-defined.

Note that

$$\tau((r+K)+(s+K)) = \tau((r+s)+K) = \varphi(r+s) = \varphi(r)+\varphi(s) = \tau(r+K)+\tau(s+K),$$

and that

$$\tau((r+K)(s+K)) = \tau((rs)+K) = \varphi(rs) = \varphi(r)\ \varphi(s) = \tau(r+K)\ \tau(s+K),$$

proving that τ is a homomorphism.

By definition, for $\varphi(r) \in \varphi(R)$, $\tau(r+K) = \varphi(r)$, proving that τ is surjective. Also, if $r+K \in \ker \tau$, then $0 = \tau(r+K) = \varphi(r)$, proving that $r \in \ker \varphi$. But $K = \ker \varphi$, and thus $r+K = 0+K$ is the zero element of R/K. That is, by Theorem 2.6, τ is injective.

Thus τ is a bijective homomorphism (i.e. an isomorphism) from R/K onto $\varphi(R)$, completing the proof.

□

4.3. Theorem. (***The Second Isomorphism Theorem***) *Let M and N be ideals of a ring R. Recall that*

$$M+N = \{m+n : m \in M, n \in N\}.$$

Then $M+N$ is an ideal of R, and

$$\frac{M+N}{N} \simeq \frac{M}{M \cap N}.$$

Proof. First we note that by Proposition 2.8, $M+N$ and $M \cap N$ are ideals of R, and as such, they are rings. Clearly N is a subring of $M+N$, while $M \cap N$ is a subring of M. We claim that $M \cap N \lhd M$. In essence, this is the fact that if S is a subring of a ring R and if J is an ideal of R that happens to be contained in S, then J is also an ideal of S. Indeed, clearly J is a subring of S, and if $j \in J, s \in S$, then $s \in R$ and so $sj, js \in J$, proving that $J \lhd S$.

By Theorem 3.7, $\dfrac{M}{M \cap N}$ is a ring under the induced operations.

Consider the map

$$\begin{array}{rccc} \tau: & M+N & \to & M/M\cap N \\ & m+n & \mapsto & m+(M\cap N). \end{array}.$$

Since it is possible that $m_1+n_1=m_2+n_2$ for some $m_1\neq m_2\in M$ and $n_1\neq n_2\in N$, we must again ensure that τ is well-defined.

If $m_1+n_1=m_2+n_2$, then $m_1-m_2=n_2-n_1\in M\cap N$, and so

$$\tau(m_1+n_1)=m_1+(M\cap N)=m_2+(M\cap N)=\tau(m_2+n_2).$$

Hence τ is well-defined.

Next, note that

$$\begin{aligned}\tau((m_1+n_1)+(m_2+n_2)) &= \tau((m_1+m_2)+(n_1+n_2))\\ &= (m_1+m_2)+(M\cap N)\\ &= (m_1+(M\cap N))+(m_2+(M\cap N))\\ &= \tau(m_1+n_1)+\tau(m_2+n_2),\end{aligned}$$

while (keeping in mind that $N\lhd R$ implies that $m_1n_2, n_1m_2, n_1n_2\in N$),

$$\begin{aligned}\tau((m_1+n_1)(m_2+n_2)) &= \tau(m_1m_2+(n_1m_2+m_1n_2+n_1n_2))\\ &= (m_1m_2)+(M\cap N)\\ &= (m_1+(M\cap N))(m_2+(M\cap N))\\ &= \tau(m_1+n_1)\ \tau(m_2+n_2),\end{aligned}$$

proving that τ is a homomorphism.

For any $m\in M$, and hence for any $m+M\cap N$ in $M/(M\cap N)$, we see that

$$\tau(m+0)=m+(M\cap N).$$

Thus τ is surjective.

Finally, suppose that $n\in N$. Then $n=0+n\in M+N$ and $\tau(n)=\tau(0+n)=0+(M\cap N)$ is the zero element of $M/(M\cap N)$. Thus $n\in\ker\tau$. Since $n\in N$ was arbitrary, $N\subseteq\ker\tau$.

Conversely, suppose that $x=(m+n)\in\ker\tau$. Then $\tau(x)=m+(M\cap N)=0+(M\cap N)$, and so $m=m-0\in(M\cap N)\subseteq N$. But $n\in N$ and N is an ideal of R, so $x=m+n\in N$. That is, $\ker\tau\subseteq N$.

Taken together, these two inclusions imply that $\ker\tau=N$.

By the First Isomorphism Theorem,

$$\tau(M+N)=\frac{M}{M\cap N}\simeq\frac{M+N}{\ker\tau}=\frac{M+N}{N}.$$

□

4.4. Theorem. (***The Third Isomorphism Theorem***) *Let M and N be ideals in a ring R, and suppose that $N \subseteq M$. Then*

$$\frac{R}{M} \simeq \frac{R/N}{M/N}.$$

Proof. As we argued in the previous Theorem, since N is an ideal of R, it is a subring of R, and if $n \in N$, $m \in M$, then $m \in R$ and thus $mn, nm \in N$. In other words, N is an ideal of M.

Consider the map

$$\begin{array}{rccc} \tau: & R/N & \to & R/M \\ & r+N & \mapsto & r+M. \end{array}$$

Again - since we are defining τ in terms of a particular representative of a coset, the first thing we must do is to verify that it is well-defined.

Suppose that $r+N = s+N$. Then $r-s \in N \subseteq M$, and thus

$$\tau(r+N) = r+M = s+M = \tau(s+N).$$

That is, τ is indeed well-defined.

For any $r+M \in R/M$, we have that $r+N \in R/N$, and $r+M = \tau(r+N)$, so that τ is surjective.

As for $\ker \tau$, note that $r+N \in \ker \tau$ if and only if $r+M = \tau(r+N) = 0+M$, that is, if and only if $r \in M$. In other words, $r+N \in \ker \tau$ if and only if $r+N \in M/N$.

By the First Isomorphism Theorem,

$$\frac{R}{M} \simeq \frac{R/N}{M/N}.$$

□

4.5. Example. Let us look at the Second Isomorphism Theorem (and its proof) through the prism of an explicit example. It will be enlightening, to say the least.

Suppose that $R = \mathbb{Z}$ (already our example is special - for one thing, it is commutative), $M = \langle 12 \rangle = 12\mathbb{Z}$, and $N = \langle 20 \rangle = 20\mathbb{Z}$. We leave it to the reader as an exercise to prove that $M+N = \langle 4 \rangle = 4\mathbb{Z}$ and $M \cap N = \langle 60 \rangle = 60\mathbb{Z}$.

The Second Isomorphism Theorem says that

$$\frac{4\mathbb{Z}}{20\mathbb{Z}} = \frac{M+N}{N} \simeq \frac{M}{M \cap N} = \frac{12\mathbb{Z}}{60\mathbb{Z}}.$$

So what does an element of $\dfrac{4\mathbb{Z}}{20\mathbb{Z}}$ look like? In general,

$$\frac{M+N}{N} = \{(m+n)+N : m \in M, n \in N\} = \{m+N : m \in M\},$$

since $n+N = N$ for all $n \in N$. In this specific case, we are looking at the cosets of the ideal $20\mathbb{Z}$ of the ring $4\mathbb{Z}$, so the cosets look like

$$\frac{4\mathbb{Z}}{20\mathbb{Z}} = \{0+20\mathbb{Z}, 4+20\mathbb{Z}, 8+20\mathbb{Z}, 12+20\mathbb{Z}, 16+20\mathbb{Z}\}.$$

Similarly,

$$\frac{12\mathbb{Z}}{60\mathbb{Z}} = \{0 + 60\mathbb{Z}, 12 + 60\mathbb{Z}, 24 + 60\mathbb{Z}, 36 + 60\mathbb{Z}, 48 + 60\mathbb{Z}\}.$$

How did our proof proceed? We considered the map

$$\begin{array}{rccc} \tau: & 12\mathbb{Z} + 20\mathbb{Z} & \to & \dfrac{12\mathbb{Z}}{60\mathbb{Z}} \\ & 12r + 20s & \mapsto & 12r + 60\mathbb{Z}. \end{array}$$

Let's think about this for a second. Ok, long enough. We remarked above that $12\mathbb{Z} + 20\mathbb{Z} = 4\mathbb{Z}$. So our map τ is really a map from $4\mathbb{Z}$ to $12\mathbb{Z}/60\mathbb{Z}$. We must specify what τ does to a multiple of 4!

When looking at an element of $4\mathbb{Z}$, we are *first asking ourselves* to express a multiple of four in the form $12r + 20s$. For example, $8 \in 4\mathbb{Z}$, so you might write $8 = 12(-1) + 20(1)$. Then again, I might write $8 = 12(4) + 20(-2)$.

As we have defined τ, you would get $\tau(8) = 12(-1) + 60\mathbb{Z} = -12 + 60\mathbb{Z}$, while I would compute $\tau(8) = 12(4) + 60\mathbb{Z} = 48 + 60\mathbb{Z}$. Fortunately,

$$-12 + 60\mathbb{Z} = \{\ldots, -72, -12, 48, 108, \ldots\} = 48 + 60\mathbb{Z},$$

so the answer you computed agrees with my answer. But what if someone else wrote $8 = 12r + 60s$ for entirely different values of r and s? Would their candidate for $\tau(8)$ agree with ours?

That is what we are doing when we check whether or not τ is well-defined. We are verifying that independent of how we express $8 = 12r + 20s$, all of us agree on what $\tau(8)$ should be. After all, the function τ sees only the number 8. It is you, I and that other person who see the r's and s's. Our computations are different, but if the formula for τ is going to make sense, we had better all agree on the value of $\tau(8)$.

Of course, we can't just check this for 8; we have to check that we all agree on the formula for $\tau(k)$ whenever $k \in 4\mathbb{Z}$. But if two people write $k = 12r_1 + 20s_1$ and $k = 12r_2 + 20s_2$ respectively, then $12r_1 - 12r_2 = 20s_2 - 20s_1$. That means that

$$12r_1 - 12r_2 = 20(s_1 - s_2) \in \langle 20 \rangle = 20\mathbb{Z},$$

and in particular, $12r_1 - 12r_2$ is divisible by 5. Clearly it is divisible by 12 also. But then it is divisible by $\text{LCM}(5, 12) = 60$, so $12r_1 - 12r_2 \in \langle 60 \rangle$, and

$$\tau(k) = 12r_1 + \langle 60 \rangle = 12r_2 + \langle 60 \rangle.$$

In other words, while the two people may have written k completely differently, they agree on what $\tau(k)$ is. Thus τ is well-defined.

The computation that shows that τ is a homomorphism is (this humble author hopes) relatively straight-forward:

$$\begin{aligned} \tau((12r_1 + 20s_1) + (12r_2 + 20s_2)) &= \tau(12(r_1 + r_2) + 20(s_1 + s_2)) \\ &= 12(r_1 + r_2) + \langle 60 \rangle \\ &= (12r_1 + \langle 60 \rangle) + (12r_2 + \langle 60 \rangle) \\ &= \tau(12r_1 + 20s_1) + \tau(12r_2 + 20s_2) \end{aligned}$$

and

$$\begin{aligned}\tau((12r_1+20s_1)(12r_2+20s_2)) &= \tau((12r_1)(12r_2)+(12r_1)(20s_2)\\ &\qquad +(20s_1)(12r_2)+(20s_1)(20s_2))\\ &= \tau((12(12r_1r_2))+20(12r_1s_2+12s_1r_2+20s_1s_2))\\ &= 12(12r_1r_2)+\langle 60\rangle\\ &= ((12r_1)(12r_2))+\langle 60\rangle\\ &= (12r_1+\langle 60\rangle)\ (12r_2+\langle 60\rangle)\\ &= \tau(12r_1+20s_1)\ \tau(12r_2+20s_2).\end{aligned}$$

To show that τ is onto, we consider an arbitrary element y of $\dfrac{12\mathbb{Z}}{60\mathbb{Z}}$, so $y = 12k+60\mathbb{Z}$, where $k\in\{0,1,2,3,4\}$. Then $12k = 12k+0 \in 12\mathbb{Z}+20\mathbb{Z} = 4\mathbb{Z}$, so τ acts on this element and by definition,

$$\tau(12k) = \tau(12k+0) = 12k+\langle 60\rangle = 12k+60\mathbb{Z} = y,$$

and τ is onto.

Finally, we compute $\ker\tau$. Now $x = 12r+20s \in \ker\tau$ if $\tau(x) = 0+60\mathbb{Z}$. But $\tau(x) = 12r+60\mathbb{Z}$, and so we need $12r\in 60\mathbb{Z}$, which means that r must be a multiple of 5, i.e. $r = 5b$ for some $b\in\mathbb{Z}$. But then $x = 12(5b)+20s = 20(3b+s)\in\langle 20\rangle$; i.e. $x\in N$. Conversely, if $x\in\langle 20\rangle = 20\mathbb{Z}$, then we can write $x = 0+20s$, and by definition, $\tau(x) = 0+\langle 60\rangle = 0+60\mathbb{Z}$. Thus $x\in\ker\tau$.

We have proven that $\ker\tau = \langle 20\rangle$. Finally, the First Isomorphism Theorem now assures us that

$$\frac{4\mathbb{Z}}{20\mathbb{Z}} = \frac{M+N}{N} \simeq \operatorname{ran}\tau = \frac{M}{M\cap N} = \frac{12\mathbb{Z}}{60\mathbb{Z}}.$$

And what, pray tell, is the isomorphism given by the First Isomorphism Theorem? Let's call it Φ. Then – keeping in mind that we can always write $4k = (12(2)+20(-1))k = 24k-20k$ – we find that

$$\begin{aligned}\Phi:\qquad \frac{4\mathbb{Z}}{20\mathbb{Z}} &\to \frac{12\mathbb{Z}}{60\mathbb{Z}}\\ 4k+\langle 20\rangle = (24k-20k)+\langle 20\rangle &\mapsto 24k+\langle 60\rangle.\end{aligned}$$

In particular, (choosing $k = 0,1,2,3$ and 4 below), we have

- $\Phi(0+\langle 20\rangle) = 0+\langle 60\rangle$;
- $\Phi(4+\langle 20\rangle) = 24+\langle 60\rangle$;
- $\Phi(8+\langle 20\rangle) = 48+\langle 60\rangle$;
- $\Phi(12+\langle 20\rangle) = 72+\langle 60\rangle = 12+\langle 60\rangle$;
- $\Phi(16+\langle 20\rangle) = 96+\langle 60\rangle = 36+\langle 60\rangle$.

4.6. Examples. The details of the following example are left to the reader.

(a) Let K be an ideal of the ring R. The map

$$\begin{array}{rccc} \pi: & R & \to & R/K \\ & r & \mapsto & r+K \end{array}$$

is a homomorphism, called the **canonical homomorphism** from R onto R/K. Its kernel is $\ker \pi = K$.

(b) For example, if $K = \langle m \rangle$ as an ideal of $\mathbb{Z}$, then the map

$$\begin{array}{rccc} \pi: & \mathbb{Z} & \to & \mathbb{Z}/\langle m \rangle \\ & z & \mapsto & z + m\mathbb{Z} \end{array}$$

is the canonical homomorphism.

4.7. Theorem. *Let R be a unital ring.*

(a) *The map*

$$\begin{array}{rccc} \varphi: & \mathbb{Z} & \to & R \\ & n & \mapsto & n\,1 \end{array}$$

is a homomorphism.

(b) *If* $\text{CHAR}(R) = \kappa > 0$, *then* $\varphi(\mathbb{Z}) \simeq \mathbb{Z}_\kappa$.

(c) *If* $\text{CHAR}(R) = 0$, *then* $\varphi(\mathbb{Z}) \simeq \mathbb{Z}$.

Proof.

(a) Let $n, m \in \mathbb{Z}$ and observe that $\varphi(n+m) = (n+m)1 = n\,1 + m\,1 = \varphi(n) + \varphi(m)$ and that $\varphi(nm) = (nm)\,1 = (n1)\,(m1) = \varphi(n)\varphi(n)$, proving that φ is a homomorphism.

(b) Note that $\ker \varphi =< \kappa >$, and thus by the First Isomorphism Theorem, combined with Example 3.10, we see that

$$\varphi(\mathbb{Z}) \simeq \frac{\mathbb{Z}}{\ker \varphi} = \frac{\mathbb{Z}}{\langle \kappa \rangle} \simeq \mathbb{Z}_\kappa.$$

(c) If $\text{CHAR}(R) = 0$, then for all $n > 0$, $\varphi(n) = n\,1 \neq 0$ and so $\varphi(-n) = -\varphi(n) \neq 0$. In other words, $\ker \varphi = \{0\}$. But then φ is a bijective homomorphism between $\mathbb{Z}$ and $\varphi(\mathbb{Z})$, so

$$\varphi(\mathbb{Z}) \simeq \mathbb{Z}.$$

□

4.8. Example.

(a) Let $n \in \mathbb{N}$ and $R = \mathbb{T}_n(\mathbb{R})$. Then

$$\{kI_n : k \in \mathbb{Z}\} = \left\{ \begin{bmatrix} k & & & \\ & k & & \\ & & \ddots & \\ & & & k \end{bmatrix} : k \in \mathbb{Z} \right\}$$

is isomorphic to $\mathbb{Z}$.

(b) In $\mathbb{Z}_5[i]$, $S := \{0+0i, 1+0i, 2+0i, 3+0i, 4+0i\}$ is a subring which is isomorphic to $\mathbb{Z}_5$.

4.9. Corollary. *Let $p \in \mathbb{N}$ be a prime number.*

(a) *Suppose that $\mathbb{F}$ is a field of characteristic p. Then $\mathbb{F}$ contains a subfield isomorphic to $\mathbb{Z}_p$.*

(b) *Suppose that $\mathbb{F}$ is a field of characteristic 0. Then $\mathbb{F}$ contains a subfield isomorphic to $\mathbb{Q}$.*

Proof.

(a) By Theorem 4.7, since $\mathbb{F}$ is unital and satisfies $\text{CHAR}(\mathbb{F}) = p$, $\mathbb{F}$ contains a subring $\mathbb{G}$ which is isomorphic to $\mathbb{Z}_p$. But p being a prime number implies that $\mathbb{Z}_p$ is a field, so $\mathbb{G}$ is a field.

(b) Again, by Theorem 4.7, $\mathbb{F}$ contains a subring S which is isomorphic to $\mathbb{Z}$. Since $\mathbb{F}$ is a field of characteristic zero, for $0 \neq n \in \mathbb{Z}$, $n1 \neq 0$, and so $(n1)^{-1} \in \mathbb{F}$. Let

$$\mathbb{G} := \{(m1)(n1)^{-1} : m, n \in \mathbb{Z}, n \neq 0\}.$$

We leave it as an exercise for the reader to verify that the map

$$\begin{array}{rccc} \tau: & \mathbb{Q} & \to & \mathbb{G} \\ & \frac{m}{n} & \mapsto & (m1)(n1)^{-1} \end{array}$$

is an isomorphism.

But $\mathbb{Q}$ is a field, and thus so is $\mathbb{G}$. Hence $\mathbb{F}$ contains the subfield $\mathbb{G}$ which is isomorphic to $\mathbb{Q}$.

□

Supplementary Examples.

S4.1. Example. Let R and S be rings. The map $\zeta : R \to S$ defined by $\zeta(r) = 0$ for all $r \in R$ is a homomorphism, called the **zero homomorphism**.

Needless to say, it's not the most interesting homomorphism around. Having said that, one should look at Exercise 4.9 below.

S4.2. Example. *** Let $\mathbb{R}[x]$ denote the commutative ring of polynomials in one variable with real coefficients. Then

$$\frac{\mathbb{R}[x]}{\langle x^2+1\rangle} \simeq \mathbb{C}.$$

S4.3. Example. Let $L := \{a + bi : a, b \in \mathbb{Z} \text{ and } a - b \in 2\mathbb{Z}\} \subseteq \mathbb{Z}[i]$. Then L is an ideal of $\mathbb{Z}[i]$ and

$$\frac{\mathbb{Z}[i]}{L} \simeq \mathbb{Z}_2.$$

S4.4. Example. Let $R := (\mathcal{C}([0,1],\mathbb{R}),+,\cdot)$ be the ring of continuous real-valued functions on $[0,1]$. Let $x_0 \in [0,1]$, and consider the map

$$\begin{array}{rccc} \varepsilon_{x_0}: & \mathcal{C}([0,1],\mathbb{R}) & \to & \mathbb{R} \\ & f & \mapsto & f(x_0). \end{array}$$

For hopefully obvious reasons, we refer to ε_{x_0} as **evaluation at** x_0. Then

$$\begin{aligned} \varepsilon_{x_0}(f+g) &= (f+g)(x_0) = f(x_0) + g(x_0) = \varepsilon_{x_0}(f) + \varepsilon_{x_0}(g), \text{ and} \\ \varepsilon_{x_0}(fg) &= (fg)(x_0) = f(x_0)g(x_0) = \varepsilon_{x_0}(f)\ \varepsilon_{x_0}(g), \end{aligned}$$

proving that ε_{x_0} is a homomorphism of $\mathcal{C}([0,1],\mathbb{R})$ into $\mathbb{R}$. The kernel of this map is the ideal

$$M_{x_0} := \{f \in \mathcal{C}([0,1],\mathbb{R}) : f(x_0) = 0\},$$

and it can be shown that if $N \lhd \mathcal{C}([0,1],\mathbb{R})$ is an ideal and $M_{x_0} \subseteq N \subseteq \mathcal{C}([0,1],\mathbb{R})$, then either $N = M_{x_0}$ or $N = \mathcal{C}([0,1],\mathbb{R})$. In other words, there are no proper ideals which contain M_{x_0}. We say that M_{x_0} is a *maximal ideal* of $\mathcal{C}([0,1],\mathbb{R})$. Maximal ideals will prove important below.

S4.5. Example. Let $\mathbb{F}$ be a field and $\{0\} \neq K \lhd \mathbb{F}$ be a non-trivial ideal of $\mathbb{F}$. Let $0 \neq k \in K$. Then $k^{-1} \in \mathbb{F}$, so $1 = k^{-1}k \in K$. But then for all $b \in \mathbb{F}$, $b = b \cdot 1 \in K$, so $\mathbb{F} \subseteq K$, implying that $K = \mathbb{F}$.

In other words, the only ideals of a field $\mathbb{F}$ are $\{0\}$ and $\mathbb{F}$ itself. We say that $\mathbb{F}$ is **simple**.

S4.6. Example. Let $R = \mathbb{Z}$, $M = 24\mathbb{Z}$ and $N = 15\mathbb{Z}$. Note that M and N are ideals of $\mathbb{Z}$, and

$$M + N = \{m + n : m \in 24\mathbb{Z}, n \in 15\mathbb{Z}\}.$$

Note that $3 = 24 \cdot 2 + 15 \cdot (-3) \in M + N$, and so $3z = 24 \cdot (2z) + 15 \cdot (-3z) \in M + N$ for all $z \in \mathbb{Z}$. Moreover, any element of $M + N$ is divisible by 3, since each of 24 and 15 are. Hence

$$M + N = 3\mathbb{Z}.$$

Also, we leave it to the reader to verify that $M \cap N = 120\mathbb{Z}$. By the Second Isomorphism Theorem,

$$\frac{3\mathbb{Z}}{15\mathbb{Z}} \simeq \frac{24\mathbb{Z}}{120\mathbb{Z}}.$$

Hopefully this will remind the reader of the well-known equation

$$\frac{3}{15} = \frac{24}{120},$$

which our result generalises.

S4.7. Example. Suppose that R is a unital ring and that $s \in R$ is invertible. The map

$$\begin{array}{rccc} \mathrm{Ad}_s : & R & \to & R \\ & x & \mapsto & s^{-1}xs \end{array}$$

is an isomorphism of R onto R. The details are left to the reader. We remind the reader that an isomorphism of a ring R onto *itself* is referred to as an **automorphism** of R.

In some contexts (for example if $R = \mathbb{M}_n(\mathbb{C})$ for some $n \geq 1$), such an automorphism is said to be **inner**, and one is interested in knowing whether all isomorphisms are inner. For example, it can be shown that all *linear* automorphisms of $\mathbb{M}_n(\mathbb{C})$ are inner. (i.e. automorphisms φ that satisfy $\varphi(\kappa X) = \kappa(\varphi(X))$ for all $\kappa \in \mathbb{C}$, $X \in \mathbb{M}_n(\mathbb{C})$.) The reader may wish to try proving this when $n = 2$.

S4.8. Example. Suppose that $n \geq 1$ is an integer and let $u \in \mathbb{M}_n(\mathbb{C})$ be a unitary matrix. The map

$$\begin{array}{rccc} \mathrm{Ad}_u : & \mathbb{M}_n(\mathbb{C}) & \to & \mathbb{M}_n(\mathbb{C}) \\ & x & \mapsto & u^*xu \end{array}$$

is an automorphism of $\mathbb{M}_n(\mathbb{C})$. Unlike the case where $s \in \mathbb{M}_n(\mathbb{C})$ is simply invertible as above, when $s = u$ is unitary, these automorphism preserve adjoints: $\mathrm{Ad}_u(x^*) = (\mathrm{Ad}_u(x))^*$.

If you haven't seen linear algebra yet, then you may feel free to ignore this example. Or better yet, go learn linear algebra and come back to this example!

S4.9. Example. Let $R = \mathbb{Z}$, and let $K = 24\mathbb{Z}$, so that $K \lhd \mathbb{Z}$. The K-radical of $\mathbb{Z}$ as defined in paragraph A4.3 below is

$$\text{RAD}_K := \{r \in \mathbb{Z} : r^n \in 24\mathbb{Z} \text{ for some } n \geq 1\}.$$

Thus, $s \in \text{RAD}_K$ if $24 \mid r^n$ for some $n \geq 1$. Since $24 = 2^3 3$, in order to have $r \in \text{RAD}_K$, we need $2 \mid r$ and $3 \mid r$, and thus $6 \mid r$. Moreover, if $r \in 6\mathbb{Z}$, then $r = 2 \cdot 3 \cdot r_0$ for some $r_0 \in \mathbb{Z}$, so $r^3 \in K$, whence $r \in \text{RAD}_K$.

We have proven that

$$\text{RAD}_{24\mathbb{Z}} := 6\mathbb{Z}.$$

S4.10. Example. in Example 3.3.4 we gave an example of a field $\mathbb{F} = \{0, 1, x, y\}$ with 4 elements. Note that for each element $b \in \mathbb{F}$, $b + b = 0$, and so $\text{CHAR}(\mathbb{F}) = 2$.

The character of a field is not the number of elements in the field.

Appendix

A4.1. The question of when one has to check whether a function is well-defined or not is not as difficult as it is sometimes made out to be. Basically, it comes down to this: one is trying to describe a rule for a set of objects by specifying what that rule does to one particular element of that set. Without knowing in advance that the result is independent of which element of the set you chose, how do you know that your rule even makes sense? To say that a function is "well-defined" is the statement that the rule that you have for describing that function makes sense.

For example: suppose that to each National Hockey League (NHL) team we wish to assign a natural number. Here is the "rule".

> Pick an arbitrary player of that team, and the natural number that we assign to the *team* will be that person's age.

Does this "rule" make sense?

NHL hockey teams have many players. Suppose that the team has two members, one whose age is 23, and the other whose age is 30. Which of these two numbers was the correct one to assign to the team?

The answer is that the "rule" we formulated does not make sense. In other words, the "function" that we described is *not* well-defined.

Here is another possible "rule" - this time we try to assign a colour to each NHL hockey team:

> Pick an arbitrary player on that team. The colour we assign to that team is the colour of that player's right eye.

Does this "rule" make sense?

Again - no! It is entirely possible that one player on the team might have brown eyes, while another member might have green eyes. The function we have described to assign a colour to the country is *not* well-defined.

Undaunted, let's try one more time. Let's assign a number to an NHL team according to the following "rule".

> Pick an arbitrary player of that team. The number we assign to that team is the number of teammates that person has (not counting himself).

This is an interesting one. If the team has 23 players, then – and this is the crucial element – *regardless of which player on the team you pick*, that player will have 22 teammates. While the set of teammates of player A will differ from the set of teammates of player B (for example, our definition states that player B is a teammate of player A, but not of himself), nevertheless, the *number* of teammates does not depend upon the particular player we chose. This way of assigning a number to a

team by choosing an individual player from that team and specifying a rule in terms of something about that particular player makes sense. The function

$$\begin{array}{rccc} f: & \text{NHL teams} & \to & \mathbb{N} \\ & \text{player x on team Y} & \to & \text{number of teammates of player x on team Y} \end{array}$$

is well-defined.

A4.2. Let's see how this works in (mathematical) practice.

Suppose that we wish to determine whether or not the map

$$\begin{array}{rccc} \varphi: & \mathbb{Z}/\langle 12\rangle & \to & \mathbb{Z}/\langle 4\rangle \\ & n+\langle 12\rangle & \mapsto & n+\langle 4\rangle \end{array}$$

is well-defined. Again – in this case, note that $2+\langle 12\rangle = 26+\langle 12\rangle$. The function φ *does not see* "2" *nor* "26". It only sees the coset

$$\{\ldots,-22,-10,2,14,26,38,\ldots\}.$$

So how does φ work? It says - pick *an arbitrary element from this set*, say n and then

$$\varphi(\{\ldots,-22,-10,2,14,26,38,\ldots\}) = \{\ldots,n-4,n,n+4,n+8,n+12,\ldots,\}.$$

Here's the beauty of it: if we had picked $n=2$, then the result would be

$$\begin{aligned} \varphi(\{\ldots,-22,-10,2,14,26,38,\ldots\}) &= \varphi(2+\langle 12\rangle) \\ &= \{\ldots,2-4,2,2+4,2+8,2+12,\ldots\} \\ &= \{\ldots,-2,2,6,10,14,\ldots\}. \end{aligned}$$

If we had picked $n=26$, we would have obtained

$$\begin{aligned} \varphi(\{\ldots,-22,-10,2,14,26,38,\ldots\}) &= \varphi(26+\langle 12\rangle) \\ &= \{\ldots,26-4,26,26+4,26+8,\ldots\} \\ &= \{\ldots,22,26,30,34,\ldots\}. \end{aligned}$$

But

$$\{\ldots,22,26,30,34,\ldots\} = \{\ldots,-2,2,6,10,14,\ldots\}!!!$$

In fact, no matter which two representatives n_1 and n_2 of $n+\langle 12\rangle$ we choose, the fact that $n_1+\langle 12\rangle = n_2+\langle 12\rangle$ implies that $n_1-n_2 \in \langle 12\rangle \subseteq \langle 4\rangle$, and thus $n_1+\langle 4\rangle = n_2+\langle 4\rangle$.

If you understand this example, you should understand "well-definedness", and you are on the path to a better and healthier mathematical lifestyle.

A4.2. The reader might be asking himself/herself: "why is it that if $\mathcal{W}$ is *any* subspace of a vector space $\mathcal{V}$, then $\mathcal{V}/\mathcal{W}$ forms a vector space under the usual operations on cosets, but in the case of a ring R, we must quotient by an *ideal* as opposed to an arbitrary subring S?" Indeed, I would be very happy to learn that the reader has now adopted the strategy I suggested of asking himself/herself (mathematical) questions while reading the notes.

The reason is that in the vector space setting, *all* subspaces occur as the kernels of (vector space) homomorphisms! So it really comes down to the idea that you *always* want to quotient by kernels of morphisms, it's just that the question of which substructures are kernels of morphisms depends upon which mathematical objects you are looking at.

To see that any subspace $\mathcal{W}$ of a vector space $\mathcal{V}$ over a field $\mathbb{F}$ is the kernel of a vector homomorphism (without first appealing to the fact that we can quotient $\mathcal{V}$ by $\mathcal{W}$), note that we can find a basis $\{w_\lambda\}_{\lambda\in\Lambda}$ for $\mathcal{W}$, which we can then extend to a basis $\mathfrak{B} := \{w_\lambda\}_{\lambda\in\Lambda} \cup \{y_\gamma\}_{\beta\in\Gamma}$ for $\mathcal{V}$. (The proof of this depends upon Zorn's Lemma, which appears in Appendix A of these course notes.)

Recall that we can define a linear map (i.e. a vector space homomorphism) $\varphi : \mathcal{V} \to \mathcal{V}$ simply by specifying what it does to a basis (and extending by linearity!). To that end, set

$$\varphi(w_\lambda) = 0 \text{ for all } \lambda \in \Lambda$$
$$\varphi(y_\gamma) = y_\gamma \text{ for all } \gamma \in \Gamma.$$

Clearly $w_\lambda \in \ker\varphi$, and so $\mathcal{W} = \operatorname{span}\{w_\lambda\}_{\lambda\in\Lambda} \subseteq \ker\varphi$.

Conversely, suppose that $v \in \ker\varphi$. Since $v \in \mathcal{V}$ and $\mathfrak{B}$ is a basis for $\mathcal{V}$, we can find $\alpha_1, \alpha_2, \ldots, \alpha_m, \beta_1, \beta_2, \ldots, \beta_n \in \mathbb{F}$ and $\lambda_1, \lambda_2, \ldots, \lambda_m \in \Lambda$, $\gamma_1, \gamma_2, \ldots, \gamma_n \in \Gamma$ such that

$$v = \sum_{i=1}^{m} \alpha_i w_{\lambda_i} + \sum_{j=1}^{n} \beta_j y_{\gamma_j}.$$

Thus

$$\begin{aligned} 0 = \varphi(v) &= \sum_{i=1}^{m} \alpha_i \varphi(w_{\lambda_i}) + \sum_{j=1}^{n} \beta_j \varphi(y_{\gamma_j}) \\ &= \sum_{i=1}^{m} \alpha_i 0 + \sum_{j=1}^{n} \beta_j y_{\gamma_j} \\ &= \sum_{j=1}^{n} \beta_j y_{\gamma_j}. \end{aligned}$$

Since $\{y_\gamma\}_{\beta\in\Gamma}$ is linearly independent, this means that $\beta_j = 0$ for all $1 \le j \le n$, and thus $v = \sum_{i=1}^{m} \alpha_i w_{\lambda_i} \in \mathcal{W}$. That is, $\ker\varphi \subseteq \mathcal{W}$.

Together, these two containments prove that $\mathcal{W} = \ker\varphi$.

So - the issue is that it is "easier" to be the kernel of a vector space homomorphism – *any* subspace of that vector space will do – than it is to be the kernel of a

ring homomorphism – where you must be an *ideal*. But in both cases the key issue is that kernels of homomorphisms are the key! Cray cray.

A4.3. *Let R be a commutative ring, and let $K \lhd R$ be an ideal of R. The K-**radical** of R is the set*

$$\text{RAD}_K := \{r \in R : \text{ there exists } n \in \mathbb{N} \text{ such that } r^n \in K\}.$$

Caveat. Unfortunately, there are several different notions within rings that are referred to as *radicals* of the ring. The one above is but one, and it is not as important as the Jacobson radical, which we shall see in the Assignments.

A4.3. *Let R be a commutative ring and $K \lhd R$ be an ideal of R. Then RAD_K is an ideal of R which contains K.*

Proof. As always, we would like to apply the Ideal Test to verify that this is the case.

That $K \subseteq \text{RAD}_K$ is easy to verify.

Let $r, s \in \text{RAD}_K$ and choose $n, m \in \mathbb{N}$ such that $r^n, s^m \in K$. Consider

$$\begin{aligned}(r-s)^{m+n} = r^{m+n} &- \binom{m+n}{1} r^{m+n-1}s^1 + \binom{m+n}{2} r^{m+n-2}s^2 - \cdots \\ &+ (-1)^{m+n-1}\binom{m+n}{m+n-1} r^1 s^{m+n-1} + (-1)^{m+n} s^{m+n}.\end{aligned}$$

For $0 \leq j \leq m+n$, either $j \leq m$, in which case $r^{m+n-j} \in K$, or $m+1 \leq j \leq m+n$, in which case $s^j \in K$. Since K is an ideal, each term in the above sum lies in K, and thus the sum itself is an element of K.

That is, $r - s \in \text{RAD}_K$.

Let $r \in \text{RAD}_K$ and fix $n \in \mathbb{N}$ such that $r^n \in K$. Let $t \in R$. Then, using the commutativity of R,

$$(tr)^n = t^n r^n \in K,$$

since $r^n \in K$ and $K \lhd R$. Hence $rt = tr \in \text{RAD}_K$.

By the Ideal Test, RAD_K is an ideal of R.

□

Exercises for Chapter 4

Exercise 4.1.

(a) Find all subrings of $\mathbb{Z}$.
(b) Find all ideals of $\mathbb{Z}$.
(c) Let $m, n \in \mathbb{Z}$ be relatively prime. (That is, $\text{GCD}(m, n) = 1$.) Show that $\mathbb{Z}_m \oplus \mathbb{Z}_n$ is isomorphic to $\mathbb{Z}_{mn}$.

Exercise 4.2.

(a) Find a subring of $\mathbb{M}_3(\mathbb{C})$ which is not an ideal.
(b) Find all ideals of $\mathbb{M}_3(\mathbb{C})$.
(c) Can you generalise this result to $\mathbb{M}_n(\mathbb{C})$?

Exercise 4.3.

Let $R = \mathcal{C}([0,1], \mathbb{R})$ and let $K := \{f \in \mathcal{C}([0,1], \mathbb{R}) : f(x) = 0, x \in [\frac{1}{3}, \frac{7}{8}]\}$. Prove or disprove that K is an ideal of R.

For those who enjoy a serious challenge: Let $0 \neq L \neq R$ be a non-trivial, proper ideal of $R = \mathcal{C}([0,1], \mathbb{R})$. Show that there exists $x_0 \in [0,1]$ such that $f(x_0) = 0$ for all $f \in L$.

Exercise 4.4.

A proper ideal M of a ring R (i.e. an ideal that is not R itself) is said to be **maximal** if whenever $N \lhd R$ and M is a *proper* subset of N, we have that $N = R$. That is, M is not properly contained in any *proper* ideal of R.

(a) If the set K of $R = \mathcal{C}([0,1], \mathbb{R})$ in Exercise 3 is in fact an ideal of R, is it maximal?
(b) Find a maximal ideal of R.

Exercise 4.5.

Let R be a ring and

$$\mathcal{D}_n(R) = \left\{ \begin{bmatrix} d_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \cdots & 0 & 0 \\ & & & \ddots & & \\ 0 & 0 & 0 & \cdots & d_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & d_n \end{bmatrix} : d_j \in R, 1 \leq j \leq n \right\}$$

denote the ring of all diagonal matrices in $\mathbb{M}_n(R)$.

Find all ideals of $\mathcal{D}_n(R)$.

Exercise 4.6.

Consider the ideals $J = \langle m \rangle$ and $K = \langle n \rangle$ of $\mathbb{Z}$. Find $J + K$, $J \cap K$, and JK.

Exercise 4.7. Suppose that S is a subring of a ring R.

(a) Suppose that $M \triangleleft R$ is an ideal of R such that $M \subseteq S$. Is M an ideal of S? Prove that it is, or give a counterexample to show that it need not be.
(b) Suppose that $N \triangleleft S$. Is N an ideal of R? Prove that it is, or give a counterexample to show that it need not be.

Exercise 4.8.

Let L be a subring of a ring R. Let $r_1, r_2 \in R$. Prove that either $r_1 + L = r_2 + L$, or $(r_1 + L) \cap (r_2 + L) = \varnothing$.

In other words, the cosets of L **partition** the ring R.

Another way to understand this is the following. Recall that a relation $\sim$ on a set X is said to be an **equivalence relation** if

- $x \sim x$ for all $x \in X$;
- if $x \sim y$, then $y \sim x$; and
- if $x \sim y$ and $y \sim z$, then $x \sim z$.

The **equivalence class** of $x \in X$ under this relation is denoted by $[x] := \{y \in X : x \sim y\}$.

(a) Prove that in any set X equipped with any equivalence relation $\sim$, the equivalence classes partition the set X. That is, $X = \cup_{x \in X}[x]$, and for $x \neq y \in X$, either $[x] = [y]$ or $[x] \cap [y] = \varnothing$.
(b) Define a relation $\sim$ on R via: $r \sim s$ if $r - s \in L$. Prove that $\sim$ is an **equivalence relation on** R.

Exercise 4.9.

Let $m, n \in \mathbb{N}$ with $m < n$. Describe all ring homomorphisms from $\mathbb{M}_n(\mathbb{C})$ to $\mathbb{M}_m(\mathbb{C})$.

Exercise 4.10.

Let $(G, +)$ be an abelian group. A map $\varphi : G \to G$ is said to be **additive** if $\varphi(g+h) = \varphi(g)+\varphi(h)$ for all $g, h \in G$. (In essence, this says that φ is a **group homomorphism** from G *into itself.* A group homomorphism (resp. ring homomorphism) from a group (resp. a ring) into itself is referred to as a **group endomorphism** (resp. a **ring endomorphism**).

We denote by $\textsc{End}\,(G, +)$ the set of all additive maps on G.

(a) Prove that $\textsc{End}\,(G, +)$ is a ring under the operations

$$(\varphi + \psi)(g) := \varphi(g) + \psi(g),$$

and

$$(\varphi * \psi)(g) := \varphi \circ \psi(g) = \varphi(\psi(g))$$

for all $g \in G$.

Given a ring R and an element $r \in R$, define the map $L_r : R \to R$ by $L_r(x) := rx$. In other words, L_r is **left multiplication by** r.

(b) Let $(R, +, \cdot)$ be a ring and recall that $(R, +)$ is an abelian group. Prove that the map

$$\begin{array}{rccc} \lambda: & (R,+,\cdot) & \to & (\text{End}\,(R,+),+,*) \\ & r & \to & L_r \end{array}$$

is an injective ring homomorphism.

In some contexts, this is referred to as the **left regular representation** of the ring R. Where was commutativity of $(R, +)$ used?

Exercise 4.11. In the same way that we worked our way through the Second Isomorphism Theorem in Example 4.5, work your way through the Third Isomorphism Theorem for the special case where $R = \mathbb{Z}$, $M = \langle 10 \rangle$, and $N = \langle 50 \rangle$.

CHAPTER 5

Prime ideals, maximal ideals, and fields of quotients

We owe a lot to Thomas Edison – if it weren't for him, we'd be watching television by candlelight.

Milton Berle

1. Prime and maximal ideals

1.1. One of the most important, if not the most important ring, is the ring of integers $(\mathbb{Z}, +, \cdot)$. In studying more general rings, we are often interested in how much the structure of $\mathbb{Z}$ carries over to a more general setting.

When studying the natural numbers, the prime numbers occupy a special and central role. They have two extremely important properties that are closely related. For one thing, we know that every natural number greater than or equal to 2 can be factored as a product of primes (in a unique way except for the order of the factors). Secondly, we know that if we try to factor a prime number p as a product $p = ab$ where $a, b \in \mathbb{N}$, then either $a = 1$ (and $b = p$), or $b = 1$ (and $a = p$). If we extend our notion of "prime" to integers by agreeing that an integer q is prime if $|q|$ is a prime natural number (so that -7 is now prime, for example), these two results become

- every integer that is neither zero nor invertible (i.e. every integer of absolute value at least equal to 2) can be factored in an essentially unique way as a product of primes; that is, it is unique if we don't care about the order of the terms, and we don't worry about multiplying (an even number of) factors by -1; and
- if $q \in \mathbb{Z}$ is a prime number and $q \mid ab$ for some $a, b \in \mathbb{Z}$, then $q \mid a$ or $q \mid b$.

The first item refers to the fact that we are agreeing that the technically different factorisations $315 = 3 \cdot 3 \cdot 5 \cdot 7 = -7 \cdot -3 \cdot 3 \cdot -5$ should be considered basically the same because the order is irrelevant ($\mathbb{Z}$ is commutative, after all); also $-3 = 3 \cdot -1$ and $-7 = 7 \cdot -1$, and the two invertible elements -1 and -1 are "cancelling each other out anyway". Notice that in $\mathbb{Z}$, 1 and -1 are the only invertible elements. This will play a role later on when looking at factorisation in more general rings.

The "long game" we are chasing is to try to solve polynomial equations $f(x) = 0$ where $f(x) \in \mathbb{F}[x]$ is a non-trivial polynomial with coefficients in a field. If we could factor $f(x) = a_m(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_m)$ for some $\alpha_k \in \mathbb{F}$, $1 \leq k \leq m$,

then we could spot the solutions to the equation immediately as $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$. (Why are these the *only* solutions?) As such, part of our strategy for attacking this problem is to develop a theory of factorisation of polynomials with coefficients in a field. Evolution teaches us that learning from others and from past experience is not the worst way to move forward. We know how to factor integers as a product of primes. Is there some way to extend the notion of *primeness* to rings other than the integers? In particular, would this make sense in the polynomial ring $\mathbb{F}[x]$? Secondly, the first property, when applied to a prime number, says that if $p \in \mathbb{Z}$ is prime and we try to factor it as $p = ab$ for some integers a and b, then one of a and b must have absolute value one, and as such, it must be invertible. We can think of this as saying that prime numbers cannot be *reduced* into two "smaller" components (where I am purposefully avoiding trying to ascribe a precise mathematical meaning to "*reduced*" and "*smaller*" just now). In fact, this characterises prime numbers, as does the second property listed above. That is, if an integer has one of the two properties, it automatically has the second. If we manage to extend both of these properties to more generally rings, will they always describe the same set? Ohhh, the mystery deepens and the tension is palpable!

In the next three Chapters, we shall develop abstract notions of *primeness*, *irreducibility* and of *factorisation*. We shall actually define these notions in more general rings than just polynomial rings over a field, and we shall seek to understand what is so special about the case of $\mathbb{F}[x]$.

Also, in this Chapter, we shall rediscover how to construct the rational numbers from the integers, and we shall see how to extend this construction to integral domains to produce *fields of quotients*.

We remind the reader that a subset B of a set A is said to be a **proper subset** if $B \neq A$. Thus a subset K of a ring R is a **proper ideal** if it is an ideal of R and $K \neq R$. Note that both prime and maximal ideals of a ring R are – by definition – proper ideals (when they exist).

1.2. Definition. *Let R be a ring. A proper ideal K of R is said to be* ***prime*** *if $r, s \in R$ and $rs \in K$ implies that either $r \in K$ or $s \in K$.*

A proper ideal M of R is said to be ***maximal*** *if whenever an ideal $K \lhd R$ satisfies $M \subseteq K \subseteq R$, either $K = M$, or $K = R$.*

1.3. The notion of a *maximal* element in a partially ordered set is not the same as the notion of a *maximum* element. A ring may have many maximal ideals (or none - depending upon the ring). The key observation is that a maximal element need only be bigger than the elements to which it is comparable – a maximum element must first be comparable to everything in the partially ordered set, and must then be bigger than everything. We discuss the definition at greater length in the Appendix to this section.

1.4. Examples.

(a) Recall that $\mathbb{Z}$ is a principal ideal domain, that is, every ideal K of $\mathbb{Z}$ is of the form $K = \langle k \rangle$ for some $k \in \mathbb{Z}$. In order for K to be a proper ideal, we cannot have $k = 1$ nor $k = -1$. Note also that $\langle k \rangle = \langle -k \rangle$, showing that there is no loss of generality in assuming that $k \geq 0$.

- Suppose first that $k = 0$. If $r, s \in \mathbb{Z}$ and $rs \in \langle 0 \rangle = \{0\}$, then either $r = 0 \in \langle 0 \rangle$ or $s = 0 \in \langle 0 \rangle$, and so $\langle 0 \rangle$ is a prime ideal.
- Suppose next that $k = p \in \mathbb{N}$ is a prime number. If $r, s \in R$ and $rs \in \langle p \rangle$, then $rs = pm$ for some integer m. Since p divides the right-hand side of the equation and is prime, it must divide the left-hand side of the equation, and indeed, it must divide either r or s. By relabelling if necessary, we may suppose that p divides r, and thus $r \in \langle p \rangle$. This shows that $\langle p \rangle$ is a prime ideal.
- Suppose that $k = k_1 k_2$ for some $2 \leq k_1, k_2 \in \mathbb{N}$. That is, suppose that k is not prime. Then with $r = k_1$, $s = k_2$, we see that $rs = k \in \langle k \rangle$, but $r, s \notin \langle k \rangle$.

Together, these show that – other than $\langle 0 \rangle$, an ideal $K = \langle k \rangle$ is a prime ideal if and only if k is a prime number (or the negative of a prime number).

(b) Now let us determine the maximal ideals of $\mathbb{Z}$. Again, we have seen that all ideals of $\mathbb{Z}$ are of the form $\langle k \rangle$ for some $0 \leq k \in \mathbb{Z}$, so these are the only ones we need consider, and we can ignore the case where $k = 1$, since $K = \langle 1 \rangle = \langle -1 \rangle = \mathbb{Z}$ is not a proper ideal of $\mathbb{Z}$.

Suppose first that $k = 0$. Then $\langle 0 \rangle$ is contained in any ideal – for example, $\langle 0 \rangle \subseteq \langle 2 \rangle \neq \mathbb{Z}$, showing that $\langle 0 \rangle$ is not maximal.

Similarly, if $k > 0$ is a composite number, say $k = k_1 k_2$ where $2 \leq k_1, k_2 < k$, then

$$\langle k \rangle \subseteq \langle k_1 \rangle,$$

and these two ideals are distinct since $k_1 \in \langle k_1 \rangle$ but $k_1 \notin \langle k \rangle$. Hence $\langle k \rangle$ is not maximal when k is composite.

Finally, suppose that k is prime. If $\langle k \rangle \subseteq \langle m \rangle$ for some $m \in \mathbb{N}$, then $k \in \langle m \rangle$ implies that m divides k. But k prime then implies that either $m = 1$ or $m = k$. If $m = 1$, then $\langle m \rangle = \mathbb{Z}$ is not a proper ideal, while $m = k$ implies that $\langle m \rangle = \langle k \rangle$.

Thus k prime implies that $\langle k \rangle$ is not properly contained in a proper ideal of $\mathbb{Z}$, whence $\langle k \rangle$ is maximal.

We conclude that (except for the zero ideal which is prime but not maximal), the maximal and prime ideals of $\mathbb{Z}$ coincide.

(c) Let $2 \leq n \in \mathbb{N}$. Recall from the Assignments that $\mathbb{M}_n(\mathbb{R})$ is simple. That is, it has no ideals other than the trivial ideals $\{0\}$ and $\mathbb{M}_n(\mathbb{R})$ itself.

The matrix units E_{11} and E_{22} lie in $\mathbb{M}_n(\mathbb{R})$, and $E_{11}E_{22} = 0$, although neither of E_{11} nor E_{22} lies in $\{0\}$. This shows that $\{0\}$ is not a prime ideal in $\mathbb{M}_n(\mathbb{R})$.

It is, however maximal. (Why?)

(d) Let $2 \leq n \in \mathbb{N}$. Let $1 \leq k \leq n$ be an integer, and set

$$M_k := \left\{T = [t_{i,j}] \in \mathbb{T}_n(\mathbb{C}) : t_{k,k} = 0\right\}.$$

We leave it to the reader to show that M_k is in indeed an ideal of $\mathbb{T}_n(\mathbb{C})$. To see that it is maximal, consider the following. If $N \lhd \mathbb{T}_n(\mathbb{C})$ is any ideal which properly contains M_k, then there exists $R \in N$ such that $R \notin M_k$. Thus $r_{k,k} \neq 0$. Let $A = [a_{i,j}] \in \mathbb{T}_n(\mathbb{C})$ be arbitrary, and let $T = [t_{i,j}] \in \mathbb{T}_n(\mathbb{C})$, where $t_{i,j} = a_{i,j}$ if $(i,j) \neq (k,k)$ and $t_{k,k} = 0$. Then $T \in M_k$ and $A = a_{k,k} r_{k,k}^{-1} E_{k,k} R E_{k,k} + T$.

That is, $A \in \langle R, M_k \rangle \subseteq N$. Since $A \in \mathbb{T}_n(\mathbb{C})$ was arbitrary, we conclude that $\mathbb{T}_n(\mathbb{C}) \subseteq N$. That is, there does not exist a proper ideal of $\mathbb{T}_n(\mathbb{C})$ which properly contains M_k, and thus M_k is maximal.

Are there any other maximal ideals in $\mathbb{T}_n(\mathbb{C})$?

(e) Note that each maximal ideal M_k of $\mathbb{T}_n(\mathbb{C})$, $1 \leq k \leq n$ listed in (d) above is also a prime ideal. Suppose that $R = [r_{i,j}]$ and $S = [s_{i,j}] \in \mathbb{T}_n(\mathbb{C})$ and $A = [a_{i,j}] := RS \in M_k$. Then $a_{k,k} = 0$, but note that $a_{k,k} = r_{k,k} s_{k,k}$, so that either $r_{k,k} = 0$, in which case $R \in M_k$, or $s_{k,k} = 0$, in which case $S \in M_k$. That is, M_k is a prime ideal of $\mathbb{T}_n(\mathbb{C})$.

Are there any other prime ideals in $\mathbb{T}_n(\mathbb{C})$?

1.5. Remark. Hopefully we have not become so tired that we have stopped constantly asking ourselves new questions while reading. For example, in Example 1.4(a), what was so special about $\mathbb{Z}$ that made $\langle 0 \rangle$ a prime ideal? Unlike the case of $\mathbb{T}_n(\mathbb{C})$ – where $\langle 0 \rangle$ is *not* a prime ideal – $\mathbb{Z}$ has the property that if $r, s \in \mathbb{Z}$ and $rs = 0$, then $r = 0$ or $s = 0$.

This should look very familiar. It is one of the defining properties of an integral domain. Thus, in every integral domain, $\langle 0 \rangle$ is a prime ideal. Of course, integral domains are commutative (and unital) by definition. Can we think of a non-commutative ring R where $\langle 0 \rangle$ is a prime ideal?

Again - this is the kind of question that should come to mind while reading these notes, and you should always be thinking while reading any mathematical text.

1.6. Example. The ideal $K := \langle x^3 + 1 \rangle \subseteq \mathbb{Z}_3[x]$ is not prime.

To see this, note that $p(x) = x^2 + 2x + 1$ and $q(x) = x + 1$ lie in $\mathbb{Z}_3[x]$ and

$$p(x)q(x) = x^3 + 2x^2 + x + x^2 + 2x + 1 = x^3 + 3x^2 + 3x + 1 = x^3 + 1,$$

though neither $p(x)$ nor $q(x)$ lies in K. (This is Exercise 5.2.)

1.7. Theorem. *Let R be a commutative, unital ring. Suppose that $L \lhd R$. The following statements are equivalent.*

(a) *R/L is an integral domain.*
(b) *L is a prime ideal.*

Proof.

(a) implies (b). First we argue that L is a proper ideal of R. Indeed, since L is an ideal of R, R/L is a ring. The hypothesis that R/L is an integral domain means that it is unital, and it is not hard to see that $1 + L$ must be the multiplicative identity of R/L. Since $1 + L \neq 0 + L$ (in a unital ring, the multiplicative and additive identities must be different), we see that $1 - 0 = 1 \notin L$, so $L \neq R$.

Suppose that $r, s \in R$ and $rs \in L$. Then $r+L, s+L \in R/L$ and $(r+L)(s+L) = rs+L = 0+L$. Since R/L is an integral domain (by hypothesis), either $r + L = 0 + L$, in which case $r \in L$, or $s + L = 0 + L$, in which case $s \in L$. This shows that L is prime.

(b) implies (a). We have seen that R/L is a ring, and $(r + L)(s + L) = (rs) + L = (sr) + L = (s+L)(r+L)$, where the middle equality holds because R is commutative. Moreover, $1+L$ is easily seen to act as the multiplicative identity of R/L, so that the latter is unital. Thus we need only show that $(r + L)(s + L) = 0 + L$ implies that either $r + L = 0$ or $s + L = 0$. But $(r+L)(s+L) = 0+L$ implies that $rs+L = 0+L$, which happens if and only if $rs \in L$. Since L is assumed to be prime, this shows that either $r \in L$ (or equivalently $r + L = 0 + L$) or $s \in L$ (or equivalently $s + L = 0 + L$). Thus R/L is an integral domain.

□

1.8. Theorem. *Let R be a commutative, unital ring, and let $M \triangleleft R$ be an ideal of R. The following statements are equivalent.*

(a) *M is a maximal ideal of R.*
(b) *R/M is a field.*

Proof.

(a) implies (b). Suppose that M is a maximal ideal of R. Arguing as in part (b) of Theorem 1.7 above, we see that R/M is a commutative ring with multiplicative identity $1 + M$. It suffices to prove that if $r_0 \in R$ and $r_0 + M \neq 0 + M$, then $r_0 + M$ is invertible in R/M.

To that end, note that if $r_0 + M \neq 0 + M$, then $r_0 \notin M$. Consider the set $N := \{sr_0 + m : m \in M, s \in R\}$. If $t \in R$, and $sr_0 + m \in N$, then

$$t(sr_0 + m) = (ts)r_0 + tm \in N,$$

since $tm \in M$ because $M \triangleleft R$. Since R is commutative,

$$(sr_0 + m)t \in N.$$

Also,

$$(s_1r_0 + m_1) - (s_2r_0 + m_2) = (s_1 - s_2)r_0 + (m_1 - m_2) \in N,$$

since $s_1 - s_2 \in R$ is clear, while $m_1 - m_2 \in M$ since $M \triangleleft R$ is an ideal. By the Ideal Test, $N \triangleleft R$. That $M \subseteq N$ is clear, and since $1r_0 + 0 = r_0 \in N$ but $r_0 \notin M$, we see that M is a proper subset of N.

The hypothesis that M is a maximal ideal of R and N is an ideal of R, combined with the fact that $M \subsetneq N \subseteq R$ now implies that $N = R$, and thus $1 \in N$. We can therefore write $1 = sr_0 + m$ for some $s \in R$, $m \in M$, and observe that

$$\begin{aligned}(s + M)(r_0 + M) &= (sr_0) + M \\ &= (sr_0 + m) + M \\ &= 1 + M \\ &= (r_0 + M)(s + M),\end{aligned}$$

proving that $r_0 + M$ is invertible, as required.

(b) Suppose next that M is not maximal. Then, either M is not a proper ideal of R, or there exists an ideal $N \lhd R$ satisfying

$$M \subset N \subset R.$$

(Here we remind the reader that $\subset$ means "proper subset", as opposed to $\subseteq$, which includes the possibility of equality.) If $M = R$, then

$$R/M = R/R = \{0 + R\}$$

is just the zero ring. In particular, it is not unital, and thus R/M is certainly not a field. Suppose therefore that there exists $N \lhd R$ as above.

Since $N \neq R$, N does not contain any invertible elements, and in particular, $1 \notin N$.

Let $n \in N \setminus M$. Then $n + M \neq 0 + M$, and for all $r \in R$,

$$(r + M)(n + M) = (rn) + M = (n + M)(r + M).$$

But $rn \in N$, since N is an ideal. If $(rn) + M = 1 + M$, then $rn - 1 \in M$, say $m = rn - 1$, and thus $1 = rn - m \in N$, since $rn \in N$ and $m \in M \subset N$. This contradicts the above paragraph.

This shows that $0 + M \neq n + \mathcal{M}$ is not invertible in R/M, and thus R/M is not a field.

An alternative argument that proves the same thing consists of observing that N/M is a proper ideal of R/M, since $1 + M \notin N/M$. (After all, if $1 + M \in N/M$, then $1 + M = n + M$ and so $1 - n \in M \subseteq N$, i.e. $1 \in N$, contradicting the fact that N is a proper ideal of R.

The presence of multiple ways of proving this result brings to mind certain allusions to taxidermy, but this is neither the time nor the place to discuss this.

□

1.9. Corollary. *Let R be a commutative, unital ring. If $L \lhd R$ is maximal, then L is prime.*

Proof. Suppose that $L \lhd R$ is a maximal ideal. By Theorem 1.8, R/L is a field. But every field is an integral domain, so by Theorem 1.7, L is prime.

□

1.10. Example. The converse to this is false. Recall that $\{0\}$ is a prime ideal in any integral domain, but it is not maximal in the integral domain $\mathbb{Z}$, for example.

2. From integral domains to fields

2.1. Our goal in this section is to use integral domains to construct fields. Our inspiration will be the construction of the field of rational numbers from the set of integers. Let us review (or learn, if we've never seen this before) how this is done. If the student is not familiar with the notion of **equivalence relations** and **equivalence classes** on a non-empty set, now might be the time to look at the Appendix to this Chapter. Perhaps we should rephrase that. Now *is* the time to look at the Appendix to this Chapter.

2.2. We know that the field of rational numbers $\mathbb{Q}$ is obtained from the integral domain $\mathbb{Z}$ by considering quotients of the form

$$\frac{p}{q}, \quad p, q \in \mathbb{Z}, q \neq 0.$$

What we would like to do now is to show that starting with *any* integral domain D, we can always use a similar construction to obtain a field $\mathbb{F}$, which we shall refer to as the **field of quotients** of D.

What we shall have to do is to examine more closely our construction of $\mathbb{Q}$ from $\mathbb{Z}$, and to figure out how to adapt it to our situation.

2.3. One of the issues we face in constructing the rational numbers from the integers as above is the "non-uniqueness" of the representation of a rational number. As we know all too well,

$$\frac{2}{7} = \frac{4}{14} = \frac{-6}{-21},$$

and indeed, we can find infinitely many ways of describing the same rational number.

If we want to say that they all represent the *same* rational number, one approach might be to set up an equivalence relation as follows:

Let $X := \mathbb{Z} \times (\mathbb{Z} \setminus \{0\}) := \{(a, b) : a, b \in \mathbb{Z}, b \neq 0\}$. We define a relation $\sim$ on X via

$$(a, b) \sim (c, d)$$

if $ad = bc$.

Let us verify that this is indeed an equivalence relation.

- $(a,b) \sim (a,b)$ since $ab = ba$.
- If $(a,b) \sim (c,d)$, then $ad = bc$, so $(c,d) \sim (a,b)$.
- If $(a,b) \sim (c,d)$ and $(c,d) \sim (x,y)$, then $ad = bc$ and $cy = dx$. Thus (keeping in mind that we are working in $\mathbb{Z}$!),

$$d(ay) = (ad)y = (bc)y = b(cy) = b(dx) = d(bx).$$

Since $d \neq 0$, we conclude that $ay = bx$, or equivalently that $(a,b) \sim (x,y)$.

Together, these three statements show that $\sim$ is an equivalence relation on X. Let us denote by $[(a,b)]$ the equivalence class of (a,b). That is,

$$[(a,b)] := \{(c,d) \in X : (c,d) \sim (a,b)\}.$$

We denote by $X/\sim$ the set of *all* equivalences classes of elements of X; that is,

$$X/\sim := \{[(a,b)] : (a,b) \in X\}.$$

In this way, we may identify the rational number $\frac{a}{b} \in \mathbb{Q}$ with the equivalence class $[(a,b)] \in X/\sim$.

Notice, incidentally, that the proof that $\sim$ is an equivalence relation on X only used the fact that the cancellation law holds in $\mathbb{Z}$ (see the third item above). But in fact, the cancellation law holds in any integral domain, so that we obtain the following result using exactly the same proof as that above.

2.4. Proposition. *Let D be an integral domain. The relation $\sim$ defined on $X := D \times (D \setminus \{0\})$ by*

$$(a,b) \sim (c,d) \qquad \text{if } ad = bc$$

is an equivalence relation on X.

Let us also remember how we add and multiply rational numbers:

$$\frac{a}{b} + \frac{c}{d} = \frac{ab+bc}{bd} \qquad \text{and} \qquad \frac{a}{b}\frac{c}{d} = \frac{ac}{bd}.$$

If we translate the notation to the ordered pair notation, the next result becomes less surprising.

Proposition 2.4 will be the key to the following theorem.

2.5. Theorem. *Let D be an integral domain. Let $\sim$ be the equivalence relation defined on $X := D \times (D \setminus \{0\})$ in Proposition 2.4 above, and let $\mathbb{F} := X/\sim = \{[(a,b)] : (a,b) \in X\}$. Then $\mathbb{F}$ is a field using the operations:*

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

Proof. Notice that we are dealing with *equivalence classes*. Already, this should trigger a Pavlovian effect, and if nothing else, we should start salivating every time a bell rings.

A fortiori, we are defining the operations of addition and multiplication using *representatives* of these equivalence classes. This is the mathematical equivalence

of someone hitting your leg under the kneecap with a rubber mallet, and your leg should pop forward and trigger the question: *are these operations well-defined*? Let's check!

Suppose that $[(a_1, b_1)] = [(a_2, b_2)]$ and that $[(c_1, d_1)] = [(c_2, d_2)]$ in $\mathbb{F}$. Then $(a_1, b_1) \sim (a, b)$, so $a_1 b = a b_1$, and similarly $c_1 d = c d_1$. Thus

$$\begin{aligned}(ad + bc)(b_1 d_1) &= a b_1 d d_1 + b d_1 b_1 c \\ &= a_1 b d d_1 + b b_1 c_1 d \\ &= (a_1 d_1 + b_1 c_1)(bd).\end{aligned}$$

In other words,

$$[(ad + bc, bd)] = [(a_1 d_1 + b_1 c_1, b_1 d_1)],$$

proving that addition is indeed well-defined.

The proof that the multiplication operation above is well-defined is similar and is left as an exercise for the reader.

There remains to show that $(\mathbb{F}, +, \cdot)$ is a field using these operations.

(A1) That $\mathbb{F}$ is closed under addition is clear, since for all $[(a, b)], [(c, d)] \in \mathbb{F}$, we see that $[(ad + bc), bd] \in \mathbb{F}$.

(A2)

$$\begin{aligned}([(a, b)] + [(c, d)]) + [(x, y)] &= [(ad + bc, bd)] + [(x, y)] \\ &= [((ad + bc)y + (bd)x, bdy)] \\ &= [(ady + bcy + bdx, bdy)] \\ &= [(a, b)] + [(cy + dx), dy)] \\ &= [(a, b)] + ([(c, d)] + [(x, y)]\ ).\end{aligned}$$

Thus addition is associative.

(A3) Consider $\alpha := [(0, 1)] \in \mathbb{F}$. For any $[(a, b)] \in \mathbb{F}$,

$$\begin{aligned}\alpha + [(a, b)] = [(0, 1)] + [(a, b)] &= [(0b + 1a, 1b)] \\ = [(a, b)] &= [(a1 + b0, b1)] \\ = [(a, b)] + [(0, 1)] &= [(a, b)] + \alpha.\end{aligned}$$

Thus $\alpha = [(0, 1)]$ is the neutral element under addition. (We are thinking of $[(0, 1)]$ as "$\frac{0}{1}$". That it is a neutral element under addition should not be surprising!!)

Notice that $(0, 1) \sim (0, d)$ for all $d \in D \setminus \{0\}$, so that $[(0, 1)] = [(0, d)]$

(A4) Let $[(a, b)] \in \mathbb{F}$. Then

$$\begin{aligned}[(a, b)] + [(-a, b)] = [(ab - ba, b^2)] &= [(0, b^2)] \\ = [(0, 1)] &= [(-ab + ba, b^2)] \\ = [(-a, b)] + [(a, b)].\end{aligned}$$

Thus $[(-a, b)] = -[(a, b)]$ is the additive inverse of $[(a, b)]$.

(A5) For all $[(a,b)],[(c,d)] \in \mathbb{F}$,

$$[(a,b)] + [(c,d)] = [(ad+bc,bd)] = [(cb+da,db)] = [(c,d)] + [(a,b)],$$

where the middle equality holds because $(D,+)$ is an abelian group.

We have shown that $(\mathbb{F},+)$ is an abelian group.

The proof that $(\mathbb{F},+,\cdot)$ satisfies conditions (M1), (M2), (M3) and (M4) of Definition 2.1.2 are similar to those above and are left as an exercise for the reader.

We conclude that $(\mathbb{F},+,\cdot)$ is a ring. We also leave it as an exercise for the reader to verify that multiplication is commutative in $\mathbb{F}$, and that $[(1,1)]$ serves as a multiplicative identity for this ring. There remains to show that every non-zero element is invertible.

From above, for all $d \in D$, $[(0,d)] = [(0,1)] := \alpha$ is the neutral element of $\mathbb{F}$ under addition. Suppose that $[(a,b)] \in \mathbb{F}$ with $a \neq 0 \neq b$. Then

$$[(a,b)][(b,a)] = [(ab,ba)] = [(1,1)] = [(ba,ab)] = [(b,a)][(a,b)],$$

and so $[(b,a)] = [(a,b)]^{-1}$ is the multiplicative inverse of $[(a,b)]$.

At long last, we conclude that $(\mathbb{F},+,\cdot)$ is a field.

□

2.6. Example. Let us apply this construction with $D = \mathbb{Z}$. The resulting field $\mathbb{F}$ is then $\mathbb{F} := \{[(p,q)] : p,q \in \mathbb{Z}, q \neq 0\}$. We claim that this is isomorphic to $\mathbb{Q}$. Consider the map

$$\begin{array}{rccc} \Theta: & \mathbb{F} & \to & \mathbb{Q} \\ & [(p,q)] & \mapsto & \frac{p}{q} \end{array}.$$

Once again, we are defining a map on equivalence classes of sets, and since we are defining the map in terms of a particular choice of representative, the first thing that we must do is to check that our map Θ is well-defined.

Suppose that $[(p,q)] = [(r,s)]$. By definition, $q \neq 0 \neq s$ are integers, and $(p,q) \sim (r,s)$, from which we find that $ps = qr$. But then dividing both sides by $qs \neq 0$, we get

$$\Theta([(p,q)]) = \frac{p}{q} = \frac{r}{s} = \Theta([r,s]),$$

proving that Θ is indeed well-defined.

Now

$$\begin{aligned} \Theta([(p,q)] + [(r,s)]) &= \Theta([(ps+rq,qs)] \\ &= \frac{ps+rq}{qs} \\ &= \frac{p}{q} + \frac{r}{s} \\ &= \Theta([(p,q)]) + \Theta([(r,s)]), \end{aligned}$$

and

$$\begin{aligned}\Theta([(p,q)][(r,s)]) &= \Theta([(pr,qs)]\\ &= \frac{pr}{qs}\\ &= \frac{p}{q}\,\frac{r}{s}\\ &= \Theta([(p,q)])\,\Theta([(r,s)]).\end{aligned}$$

Thus Θ is a ring homomorphism.

To see that Θ is injective, it suffices to show that $\ker\Theta = \{[(0,1)]\}$. But $[(p,q)] \in \ker\Theta$ if and only if $\Theta([(p,q)]) = \frac{p}{q} = 0$, which happens if and only if $p = 0$. Hence

$$\ker\Theta = \{[(0,q)] : q \in \mathbb{Z}, 0 \neq q\} = \{[(0,1)]\},$$

since $q \neq 0$ implies that $(0,q) \sim (0,1)$. Hence Θ is injective.

Finally, to see that Θ is surjective, and hence an isomorphism of rings, note that for any $\frac{p}{q} \in \mathbb{Q}$ with $p \in \mathbb{Z}$ and $0 \neq q \in \mathbb{Z}$, we have that

$$\frac{p}{q} = \Theta[(p,q)].$$

Since Θ is an isomorphism, we have just constructed (an isomorphic copy of) $\mathbb{Q}$ from $\mathbb{Z}$.

Of course, the fact that we are able to construct a field using an integral domain is quite interesting, but if that field were to have no other connection to the integral domain, then it would be an idle curiosity at best. In much the same way that the rational numbers contain an isomorphic copy of the integers, we find that the field $\mathbb{F}$ of quotients of an integral domain D always includes an isomorphic of D; in other words, up to a matter of notation – it always includes D.

2.7. Theorem. *Let D be an integral domain and let $\mathbb{F} = \{[(a,b)] : a,b \in D, b \neq 0\}$ be its field of quotients. The map*

$$\begin{array}{rccc}\theta: & D & \to & \mathbb{F}\\ & a & \mapsto & [(a,1)]\end{array}$$

is an injective homomorphism, and thus $\theta(D) \subseteq \mathbb{F}$ is an integral domain which is isomorphic to D.

Proof. That $\theta(D)$ is an integral domain which is isomorphic to D will follow automatically once we show that θ is an injective homomorphism, since it is automatically surjective as a map onto its range $\theta(D)$.

Now

$$\theta(a+b) = [(a+b,1)] = [(a,1)] + [(b,1)] = \theta(a) + \theta(b),$$

and

$$\theta(ab) = [(ab,1)] = [(a,1)][(b,1)] = \theta(a)\,\theta(b),$$

so that θ is a homomorphism. If $\theta(a) = [(0,1)]$, then $[(a,1)] = [(0,1)]$, so $(a,1) \sim (0,1)$. But this happens if and only if $a \cdot 1 = 1 \cdot 0 = 0$, i.e. if $a = 0$. Thus $\ker \theta = \{0\}$, whence θ is injective.

□

2.8. Example. If D is an integral domain, then so is $D[x]$. The corresponding field is

$$\mathbb{F} = \{[(p(x), q(x))] : p(x), q(x) \in D[x], q(x) \neq 0\},$$

with

$$[(p(x), q(x))] + [(r(x), s(x))] := [(p(x)s(x) + q(x)r(x), q(x)s(x))],$$

and

$$[(p(x), q(x))]\ [(r(x), s(x))] := [(p(x)r(x), q(x)s(x))].$$

It is standard to write

$$\frac{p(x)}{q(x)}$$

to denote the equivalence class $[(p(x), q(x))]$. In particular, it is understood that

$$\frac{p(x)}{q(x)} = \frac{r(x)}{s(x)}$$

if and only if $p(x)s(x) = r(x)q(x)$. Using this notation, the above summation and multiplication become the familiar

$$\frac{p(x)}{q(x)} + \frac{r(x)}{s(x)} = \frac{p(x)s(x) + q(x)r(x)}{q(x)s(x)}$$

and

$$\frac{p(x)}{q(x)}\ \frac{r(x)}{s(x)} = \frac{p(x)r(x)}{q(x)s(x)}.$$

Supplementary Examples.

S5.1. Example. Let $R = \mathcal{C}([0,1],\mathbb{R})$, equipped with the usual point-wise addition and multiplication. This ring has an additional structure, namely that it is a vector space over $\mathbb{R}$. (We shall discuss vector spaces over general fields in a later Chapter.)

Given $x_0 \in [0,1]$, the map

$$\begin{array}{rccc} \varepsilon_{x_0}: & \mathcal{C}([0,1],\mathbb{R}) & \to & \mathbb{R} \\ & f & \to & f(x_0) \end{array}$$

is easily seen to be a ring homomorphism. It is also linear and non-zero, since if $g(x) = 1$ for all $x \in [0,1]$, then $g \in \mathcal{C}([0,1],\mathbb{R})$ and $\varepsilon_{x_0}(g) = 1 \neq 0$. By the First Isomorphism Theorem,

$$\mathbb{R} = \operatorname{ran} \varepsilon_{x_0} \simeq \frac{\mathcal{C}([0,1],\mathbb{R})}{\ker \varepsilon_{x_0}}.$$

That $\ker \varepsilon_{x_0}$ is an ideal follows from Chapter 4. Since the co-dimension of $\ker \varepsilon_{x_0}$ is one, this implies that $\ker \varepsilon_{x_0}$ is a maximal ideal of $\mathcal{C}([0,1],\mathbb{R})$.

Less easy to show is the fact that every maximal ideal of $\mathcal{C}([0,1],\mathbb{R})$ is of the form $\ker \varepsilon_{x_0}$ for some $x_0 \in [0,1]$.

S5.2. Example. Let D be an integral domain, and let $R = D \times D = \{(d_1, d_2) : d_i \in D, i = 1,2\}$, equipped with point-wise addition and multiplication.

Let $J \neq D$ be a prime ideal of D. Then for any $a, b \in D$, $ab \in J$ implies that $a \in \mathcal{J}$ or $b \in J$. Let $k \in \mathcal{J}$, and $d \in D \setminus \mathcal{J}$. Then

$$(d,k)\cdot(k,d) = (dk,kd) = (dk,dk) \in J \times J,$$

though neither of (k,d) nor (d,k) belongs to $J \times J$. Thus $J \times J$ is not a prime ideal of R.

The reader would be well-served to generalise this result to ideals of the form $J_1 \times J_2$ of R.

S5.3. Example. If $\mathbb{F}$ is a field, then $\{0\}$ and $\mathbb{F}$ are the only ideals of $\mathbb{F}$, and so $\{0\}$ is a maximal ideal.

S5.4. Example. Let $R = \mathbb{Z}[x]$, the ring of polynomials in x with coefficients in $\mathbb{Z}$. Let $K = \langle x \rangle$, the ideal generated by x.

Suppose that $p(x), q(x) \in \mathbb{Z}[x]$ and that $p(x)q(x) \in K$. Writing $p(x) = p_m x^m + p_{m-1}x^{m-1} + \cdots + p_1 x + p_0$ and $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots q_1 x + q_0$, we see that if $p_0 \neq 0 \neq q_0$, then the constant term of $p(x)q(x)$ is equal to $p_0 q_0$, a contradiction.

In other words, either $p_0 = 0$ (i.e. $p(x) \in K$) or $q_0 = 0$ (i.e. $q(x) \in K$) – or both. Either way, this shows that K is a prime ideal.

S5.5. Example. Let $R = \mathbb{Z}[x]$, the ring of polynomials in x with coefficients in $\mathbb{Z}$. Once again, let $K = \langle x \rangle$, the ideal generated by x. Then $K = \{xq(x) : q(x) \in \mathbb{Z}[x]\}$. Let $N := \{2m + xq(x) : m \in \mathbb{Z}, q(x) \in \mathbb{Z}[x]\}$. We invite the reader to verify that $N \lhd \mathbb{Z}[x]$, and that $K \subsetneq N \subsetneq \mathbb{Z}[x]$.

In other words, K is not maximal. This also follows from the fact that

$$\frac{\mathbb{Z}[x]}{\langle x \rangle} \simeq \mathbb{Z},$$

which is not a field.

S5.6. Example. The ideal $K = \langle x^2 + 1 \rangle$ is a prime ideal in $\mathbb{Z}[x]$. It is left as an exercise for the reader to prove that

$$\frac{\mathbb{Z}[x]}{\langle x^2 + 1 \rangle} \simeq \mathbb{Z}[i],$$

the ring of Gaussian integers. Since the latter is an integral domain but not a field, $\langle x^2 + 1 \rangle$ is a prime ideal which is not a maximal ideal of $\mathbb{Z}[x]$.

S5.7. Example. Note that $\mathbb{Z} \times \{0\}$ is a prime ideal of $\mathbb{Z} \times \mathbb{Z}$, but it is not maximal, since it is contained in the larger ideal $\mathbb{Z} \times 2\mathbb{Z}$.

S5.8. Example. Let $p \in \mathbb{N}$ be a prime number and

$$K := \{q(x) = q_n x^2 + q_{n-1} x^{n-1} + \cdots + q_1 x + pq_0 : n \in \mathbb{N}, q_j \in \mathbb{Z}, 0 \leq j \leq n\} \subseteq \mathbb{Z}[x].$$

That is, K consists of all polynomials whose constant term is divisible by p. Then

$$\frac{\mathbb{Z}[x]}{K} \simeq \mathbb{Z}/\langle p \rangle \simeq \mathbb{Z}_p$$

is a field, so K is a maximal ideal of $\mathbb{Z}[x]$, and therefore it is also a prime ideal of $\mathbb{Z}[x]$.

S5.9. Example. Returning to $\mathcal{C}([0,1], \mathbb{R})$: let $K := \{f \in \mathcal{C}([0,1], \mathbb{R}) : f(x) = 0, x \in [\frac{1}{3}, \frac{2}{3}]\}$. We invite the reader to verify that this is an ideal of $\mathcal{C}([0,1], \mathbb{R})$.

Let $g(x) = \begin{cases} 0 & x \in [0, \frac{1}{2}] \\ x - \frac{1}{2} & x \in [\frac{1}{2}, 1], \end{cases}$ and $h(x) = \begin{cases} \frac{1}{2} - x & x \in [0, \frac{1}{2}] \\ 0 & x \in [\frac{1}{2}, 1] \end{cases}$.

Then $g(x)h(x) = 0 \in K$, but neither $g(x)$ nor $h(x)$ is in K, so K is not prime.

In fact, essentially the same argument shows that if R is not an integral domain, and if K is a non-zero, proper ideal of R, then K is not prime.

S5.10. Example. If we are not dealing with unital rings, there need not exist maximal ideals in general (even if we assume Zorn's Lemma).

Consider $R = (\mathbb{Q}, +, *)$, where $+$ is usual addition of rational numbers, but

$$a * b := 0 \text{ for all } a, b \in \mathbb{Q}.$$

We leave it to the reader to verify that R is indeed a ring.

What should an ideal of R look like? In essence, $K \lhd R$ is $k_1 - k_2 \in K$ for all $k_1, k_2 \in K$, while $a * k = k * a = 0 \in K$ is trivially verified. In other words, K must be a **subgroup** of $\mathbb{Q}$ under addition, and any subgroup of $\mathbb{Q}$ under addition is an ideal of R.

Suppose that $M \subsetneq \mathbb{Q}$ is a proper subgroup under addition.

As per the first Assignment, there exists a prime number $p \in \mathbb{N}$ and an integer $N \in \mathbb{N}$ such that $n \geq N$ implies that $\frac{1}{p^n} \notin M$.

Let $N = \{m + k\frac{1}{p^N} : m \in M, k \in \mathbb{Z}\}$. We claim that N is a subgroup of $(\mathbb{Q}, +)$ and therefore an ideal of $R = (\mathbb{Q}, +, *)$, and that $M \subsetneq N \subsetneq R$, so that M is not a maximal ideal of R.

- Let $a := m_1 + k_1\frac{1}{p^N}$, $b := m_2 + k_2\frac{1}{p^N} \in N$. Then

  $$b - a := (m_2 + k_2\frac{1}{p^N}) - (m_1 + k_1\frac{1}{p^N}) = (m_2 - m_1) + (k_2 - k_1)\frac{1}{p^N} \in N.$$

  Thus N is an additive subgroup of $\mathbb{Q}$, and so it is an ideal of R using $*$ as our multiplication.
- Clearly $M \subset N$ and $M \neq N$ since $\frac{1}{p^N} \in N \setminus M$.

  To see that $N \neq \mathbb{Q}$, note that $\frac{1}{p^{N+1}} \notin N$. Indeed, suppose otherwise. Then

  $$\frac{1}{p^{N+1}} = m + k\frac{1}{p^N} \text{ for some } m \in M, k \in \mathbb{Z},$$

  and so

  $$\frac{1}{p^N} = pm + k\frac{1}{p^{N-1}} \in M,$$

  a contradiction.

Appendix

A5.1. Earlier in this Chapter, and also in the Exercises to the previous Chapter, we defined the notion of a *maximal ideal* M in a ring R, and we also mentioned that there is a difference between the notions of *maximal elements* of partially ordered sets, and of *maximum elements* of partially ordered sets. Let us now examine this more closely. We shall begin, as a wise man used to say, at the "big-inning".

A5.2. Definition. *Let $\varnothing \neq X$ be a set. A* ***relation*** *ρ on a non-empty set X is a subset of the set $X \times X = \{(x, y) : x, y \in X\}$ of ordered pairs of elements of X.*

Often, we write $x \rho y$ to mean the ordered pair $(x, y) \in \rho$. This is especially true when dealing, for example, with the usual relation $\leq$ for real numbers: no one writes $(x, y) \in \leq$; we all write $x \leq y$. Incidentally, the notation $\leq$ is used not only for the relation "less than or equal to" for real numbers; it frequently appears to indicate a specific kind of a relation known as a *partial order* on an arbitrary set, which we now define.

A5.3. Example. Let $X = \mathbb{Z}$. We might define a relation $\sim$ on X by setting $x \sim y$ if $|x - y| \leq 1$. Thus $3 \sim 3$, $3 \sim 4$, $4 \sim 3$, and $4 \sim 5$, but $(3, 5) \notin \sim$ (i.e. $3 \not\sim 5$), since $|3 - 5| = 2 > 1$.

Note that there need not exist an explicit "formula" to describe ρ. *Any* subset of $X \times X$ defines a relation.

A5.4. Definition. *A relation $\leq$ on a non-empty set X is said to be a* ***partial order*** *if, given x, y, and $z \in X$,*

(a) *$x \leq x$;*
(b) *if $x \leq y$ and $y \leq x$, then $x = y$; and*
(c) *if $x \leq y$ and $y \leq z$, then $x \leq z$.*

We refer to the pair $(X, \leq)$ as a ***partially ordered set****, or more succinctly as a* ***poset****.*

A5.5. The prototypical example of a partial order is the partial order of **inclusion** on the power set $P(A)$ of a non-empty set A. That is, we set $P(A) = \{B : B \subseteq A\}$ and for $B, C \in P(A)$, we set $B \subseteq C$ to mean that every member of B is a member of C. It is easy to see that $\subseteq$ is a partial order on $P(A)$.

The word *partial* refers to the fact that given two elements B and C in $P(A)$, they might not be comparable. For example, if $A = \{x, y\}$, $B = \{x\}$ and $C = \{y\}$, then it is not the case that $B \subseteq C$, nor is it the case that $C \subseteq B$. Only *some* subsets of $P(A)$ are comparable to each other.

In dealing with the natural numbers, we observe that they come equipped with a partial order which we typically denote by $\leq$. In this setting, however, *any* two natural numbers are comparable. This leads us to the following notion.

A5.6. Definition. *If (X, ρ) is a partial ordered set, and if $x, y \in X$ implies that either $x \rho y$ or $y \rho x$, then we say that ρ is a* ***total order*** *on X.*

A5.7. Example.

(a) It follows from the above discussion that $(\mathbb{N}, \leq)$ is a *totally ordered set.*
(b) It also follows from the above discussion that if A is a non-empty set with at least two members, then $(P(A), \subseteq)$ is partially ordered, but not totally ordered by inclusion.

A5.8. Definition. *Let $(X, \leq)$ be a partially ordered set. An element $m \in X$ is said to be a* ***maximal element*** *of X if $x \in X$ and $m \leq x$ implies that $x = m$.*

The element $\omega \in X$ is said to be a ***maximum element*** *of X if $x \in X$ implies that $x \leq \omega$.*

A5.9. The key difference between these two notions is that a maximum element is comparable to every other element of X, and is bigger than or equal to it, whereas a *maximal element* need not be comparable to every other element of the set X – it just needs to be bigger than or equal to those elements to which it is actually comparable.

A5.10. Example. Let $\mathcal{A} = \{\varnothing, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}\}$ be the set of *proper* subsets of $X = \{x, y, z\}$, and partially order $\mathcal{A}$ by inclusion $\subseteq$. Then clearly $\{x\} \subseteq \{x, y\}$.

Observe that $\{x, y\}$ is a maximal element of $\mathcal{A}$. *Nothing is bigger than* $\{x, y\}$ in $\mathcal{A}$. But $\{z\} \nsubseteq \{x, y\}$, and so $\{x, y\}$ is not a maximum element of $\mathcal{A}$. Note that $\{x, z\}$ is also a maximal element. Maximal elements need not be unique. (Are maximum elements unique when they exist? Think about it!)

A5.11. Remark. We have not addressed the question of whether or not maximal ideals always exist. As it so happens, the existence of maximal ideals in unital rings is equivalent to *Zorn's Lemma*, or equivalently to the *Axiom of Choice*, both of which are discussed in the Appendix at the end of these course notes.

A5.12. We have seen that a partial order on a non-empty set is a specific kind of relation that mimics the notion of *inclusion* between elements of the power set $P(X)$ of a set X, or the simple concept of *less than or equal to* between two real numbers.

Another important relation tries to mimic the concept of two elements being equivalent. The name chosen for such a relation couldn't be better suited.

A5.13. Definition. *A relation $\equiv$ on a non-empty set X is said to be an* ***equivalence relation*** *if for all $x, y, z \in X$,*

(a) *$x \equiv x$;*
(b) *if $x \equiv y$, then $y \equiv x$; and*
(c) *if $x \equiv y$ and $y \equiv z$, then $x \equiv z$.*

Given $x \in X$, we define the ***equivalence class*** *of x to be the set*

$$[x] := \{y : y \equiv x\},$$

and we denote by $X/\equiv$ the set of all equivalence classes of elements of X. Thus

$$X/\equiv := \{[x] : x \in X\}.$$

We refer to x as a ***representative*** *of its equivalence class.*

A5.14. Thus an equivalence relation differs from a partial order because of item (b) above. In the case of a partial order, the condition $x \leq y$ and $y \leq x$ implies $x = y$ is referred to as **anti-symmetry**. The idea is that a partial order is only symmetric when it has to be. In the case of an equivalence relation, the condition that $x \equiv y$ implies that $y \equiv x$ is referred to as **symmetry**. Equivalence was born to be symmetric.

A5.15. Example. The most obvious notion of equivalence is equality. Let $\varnothing \neq X$ be a set, and define $x \equiv y$ if and only if $x = y$. It is easy to see that this is an equivalence relation. For each $x \in X$, the equivalence class of x is $[x] = \{x\}$.

A5.16. Example. Consider $X = \mathbb{Z}$. For $x, y \in \mathbb{Z}$, set $x \equiv y$ if $x - y \in 2\mathbb{Z}$; i.e. if $x - y = 2m$ for some $m \in \mathbb{Z}$.

Again - we leave it to the reader to verify that this is indeed an equivalence relation.

There are precisely two equivalence classes for $\mathbb{Z}$ under this relation, namely $[0] = 2\mathbb{Z} = \{2m : m \in \mathbb{Z}\}$, and $[1] = \{2m + 1 : m \in \mathbb{Z}\}$.

A5.17. Exercise. Note that in both of the above examples, the underlying set X is a disjoint union of its equivalence classes under the equivalence relation.

Let $\varnothing \neq X$ be a non-empty set and $\equiv$ be an equivalence relation on X.

(a) Prove that if $x, y \in X$, then either $[x] = [y]$, or $[x] \cap [y] = \varnothing$.
(b) Prove that $X = \cup\{[x] : [x] \in X/\equiv\}$.

A5.18. One of the most important thing to remember about equivalence classes is that their representatives are, in general, non-unique. For example, in Example 2.3 above, $[0] = [2] = [246] = [-5484726492]$, so that $0, 2, 246$ and -5484726492 are all representatives of the same equivalence class

$$2\mathbb{Z} = \{\ldots, -4, -2, 0, 2, 4, \ldots\}.$$

Similarly, $[1] = [-2222444466669]$. When dealing with an equivalence class as an *element* of $X/\equiv$, it is important to remember that in general you are looking at a *set* of elements, and that you can't really *see* a particular representative of that set. As a result, any argument or definition that you make about the equivalence class that uses that representative had better not depend upon the representative that you have chosen!

Hey, this sounds familiar: we've seen this kind of thing before, haven't we? (*Yes!* is the correct rhetorical answer to this not-so rhetorical question.) In the Appendix

to Chapter 4, we discussed the notion of well-definedness of functions, especially in relation to cosets of a ring. The point is that if R is a ring, and K is an ideal of that ring, then we may establish an equivalence relation by setting $x \sim y$ if and only if $x - y \in K$. The equivalence class of x is precisely $[x] = \{x + k : k \in K\}$, which we define to be the coset $[x] = x + K$. The point is that when K is an ideal of R, we have discovered a way to turn the collection of *equivalence classes* of elements of R (under the above relation) into a ring. Wicked!

Exercises for Chapter 5

Exercise 5.1.
Let $2 \leq n \in \mathbb{N}$.
(a) Find all maximal ideals of $\mathbb{T}_n(\mathbb{C})$.
(b) Find all prime ideals of $\mathbb{T}_n(\mathbb{C})$.

Exercise 5.2.
Prove that $p(x) = x^2 + 2x + 1$ does not lie in the ideal $\langle x^3 + 1 \rangle$ of $\mathbb{Z}_3[x]$.

Exercise 5.3.
Let D be an integral domain. Let $\sim$ be the equivalence relation defined on $X := D \times (D \setminus \{0\})$ in Proposition 2.4 above, and let $\mathbb{F} := \{[(a,b)] : (a,b) \in X\}$. Prove that the multiplication operation on $\mathbb{F}$ given by

$$[(a,b)][(c,d)] := [(ac,bd)]$$

is well-defined.

Exercise 5.4.
With the hypotheses and notation from Theorem 2.5 above,
(a) Show that conditions (M1), (M2), (M3) and (M4) of Definition 2.1.2 are satisfied by $\mathbb{F}$.
(b) Prove that $[(1,1)]$ serves as a multiplicative identity for $\mathbb{F}$.

Exercise 5.5.
Let R be a ring, and let M and N be prime ideals of R. Either prove that $M \cap N$ is a prime ideal of R, or find a counterexample to this statement if it is false.

Exercise 5.6.
Let R be a ring, and let M and N be maximal ideals of R. Either prove that $M \cap N$ is a maximal ideal of R, or find a counterexample to this statement if it is false.

Exercise 5.7.
Let R and S be rings and $\varphi : R \to S$ be a homomorphism.
(a) Suppose that $P \lhd S$ is a prime ideal. Either prove that $\varphi^{-1}(P) := \{x \in R : \varphi(x) \in P\}$ is a prime ideal of R, or find a counterexample to this statement if it is false.
(b) Suppose that $M \lhd S$ is a maximal ideal. Either prove that $\varphi^{-1}(P) := \{x \in R : \varphi(x) \in P\}$ is a maximal ideal of R, or find a counterexample to this statement if it is false.
(c) Let $K \lhd R$ be an ideal. Prove that $\varphi(K)$ need not be an ideal of S, but that $\varphi(K)$ is an ideal of S if φ is surjective.

Exercise 5.8.

Let R be a unital ring.

(a) Let $K \neq R$ be an ideal of R. Prove that there exists a maximal ideal M such that $K \subseteq M$.

(b) Let $r \in R$ be non-invertible. Prove that there exists a maximal ideal M such that $r \in M$.

(c) Suppose that R has a unique maximal ideal. Prove that the set N of non-invertible elements of R form an ideal in R.

Exercise 5.9.

Let R be a finite, commutative and unital ring. Prove that every prime ideal is maximal.

Exercise 5.10.

Let R be a commutative unital ring. Suppose that ever proper ideal of R is a prime ideal. Prove that R is a field.

CHAPTER 6

Euclidean Domains

Don't ever wrestle with a pig. You'll both get dirty, but the pig will enjoy it.

Cale Yarborough

1. Euclidean Domains

1.1. Algebraists study algebraic structures. That is neither as deep nor as shallow as it may sound. Analysts study algebraic structures that admit some sort of topological structure as well. The rational numbers are a perfect example of what we are trying to get at here.

From the point of view of algebra - the rational numbers are a perfectly legitimate object of study. Heck, they form a field, and fields - from an algebraic standpoint – are *swell*. (That is not a technical term.) Of course, in some (algebraic) ways, the rational numbers are lacking: as a field, $\mathbb{Q}$ is not **algebraically closed**. This can be a severe handicap for an algebra. But all in all, they are still really, really swell.

From the point of view of your average Analyst, the rational numbers are severely lacking not only because they are algebraically incomplete, but more significantly because they are not (analytically) *complete*; that is, it is possible to find Cauchy sequences of rational numbers whose limits are not rational numbers. As a consequence, basic analytical properties such as the Least Upper Bound property and the Intermediate Value Theorem fail abjectly when dealing with the rationals as the base field, and this suggests to the Analyst that the real numbers (or, for the more daring and debonaire – the complex numbers) are the field with which it is preferable to work. *Limits*, one of the many "only true love"'s of the Analyst, tend to be of little or no concern to Algebraists. Since they are interested in alternative, more algebraic properties, the failure of the Intermediate Value Theorem is neither here nor there nor anywhere else to them.

This is not a bad nor a good thing. It is just a thing.

A passing note on Euclidean domains before we embark upon a detailed study of them. The three examples we develop in this Chapter, namely $\mathbb{Z}$, polynomial rings over integral domains (and its generalisation – formal power series), and the Gaussian integers, are the same three examples that appear in most of the (great

many) books that we have consulted. What's more – these are typically *the only three* that appear in those references. See the Appendix for one more example - so-called **Dedekind domains**.[1]

This is not to say that they are not important – and the techniques we shall develop to study them – will be exceedingly important in the study of irreducibility/reducibility of polynomials in the coming chapters.

1.2. Definition. *Let D be an integral domain. A **Euclidean norm** or **valuation** on D is a map*

$$\mu : D \setminus \{0\} \to \mathbb{N}_0 := \mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \ldots\}$$

satisfying

(a) *for all $a, b \in D$ with $b \neq 0$, there exist $q, r \in D$ such that*

$$a = qb + r,$$

where either $r = 0$ or $\mu(r) < \mu(b)$; and

(b) *for all $a, b \in D \setminus \{0\}$, $\mu(a) \leq \mu(ab)$.*

*A **Euclidean domain** is an integral domain equipped with a valuation μ.*

1.3. Example. Suppose that E is a Euclidean domain with valuation μ. Then $\nu(d) = 2^{\mu(d)}$, $d \in E \setminus \{0\}$ defines a valuation on E as well.

1.4. Example. Item (a) in the definition of a valuation should look eerily familiar. In particular, it should look very much like a property enjoyed by the integers referred to as the *division algorithm*. More precisely, **the division algorithm for** $\mathbb{Z}$ says that if a and b are integers, with $b \neq 0$, then there exists an integer q such that

$$a = bq + r,$$

where $0 \leq r < |b|$.

If we set $\mu(d) = |d|$ for all $d \in \mathbb{Z} \setminus \{0\}$, then clearly conditions (a) and (b) of Definition 1.2 hold. Since $\mathbb{Z}$ is an integral domain, we have come to the not-so-surprising conclusion that $(\mathbb{Z}, +, \cdot)$ is a Euclidean domain.

Observe that the conclusion is *not-so-surprising* in large part because the definition of a Euclidean domain is meant to capture exactly this property of $\mathbb{Z}$.

1.5. The second most common and important example of a Euclidean domain is that of a polynomial ring with coefficients in a field. In light of condition (a) above, it behooves us (admit it – how many times have you seen the word "behoove" in a mathematical text, eh?) to prove some version of the division algorithm in this setting.

[1] As it turns out, our plea for more examples (from an earlier version of these notes) was answered by N. Banks. We refer to the reader to the Appendix for a brief discussion of these.

1.6. Theorem. (***The division algorithm for polynomial rings***)
Let $\mathbb{F}$ be a field and $f(x), g(x) \in \mathbb{F}[x]$ with $g(x) \neq 0$. Then there exist unique polynomials $q(x), r(x) \in \mathbb{F}[x]$ satisfying

(I) $f(x) = q(x)g(x) + r(x)$, *and*
(II) *either* $r(x) = 0$ *or* $\deg(r(x)) < \deg(g(x))$.

Proof.

(I) Let us first address the question of the existence of $q(x)$ and $r(x)$.
- If $f(x) = 0$ or if $\deg(f(x)) < \deg(g(x))$, then we set $q(x) = 0$ and $r(x) = f(x)$.
- There remains the case where $f(x) \neq 0$ and $\deg(g(x)) \leq \deg(f(x))$. We argue by induction on the degree of $f(x)$. Since $f(x) \neq 0$, we have that $0 \leq \deg(f(x))$.
  The base case is where $\deg(f(x)) = 0$. Then $f(x) = a_0$ for some $a_0 \in \mathbb{F}$. Since $g(x) \neq 0$, we must have $\deg(g(x)) = 0$, and thus $g(x) = b_0$ for some $0 \neq b_0 \in \mathbb{F}$.
  Setting $q(x) = a_0 b_0^{-1}$ and $r(x) = 0$ does the trick.
  Next, suppose that $N \geq 1$, that $\deg(f(x)) = N$, and that the desired result holds when $f(x)$ is replaced by any polynomial $h(x)$ with $\deg(h(x)) < N$.
  The case where $\deg(f(x)) < \deg(g(x))$ having already been done, we may now suppose that $m := \deg(g(x)) \leq N$. Let us write

$$f(x) = a_N x^N + a_{N-1}x^{N-1} + \cdots + a_1 x + a_0$$
$$g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0.$$

  Let

$$k(x) := f(x) - (a_N b_m^{-1})x^{N-m} g(x)$$

  and observe that $\deg(k(x)) < N$. (Why?)

Since $\deg(k(x)) < N$, our induction hypothesis ensures that there exist $u(x), r(x) \in \mathbb{F}[x]$ such that $r(x) = 0$ or $\deg(r(x)) < g(x)$, and

$$k(x) = u(x)g(x) + r(x).$$

But then

$$\begin{aligned} f(x) &= k(x) + (a_N b_m^{-1})x^{N-m} g(x) \\ &= (u(x) + (a_N b_m^{-1})x^{N-m})g(x) + r(x). \end{aligned}$$

Setting $q(x) = (u(x) + (a_N b_m^{-1})x^{N-m})$ completes the proof of existence.

(II) Next, let us see why $q(x)$ and $r(x)$ satisfying the properties listed above are unique.

To that end, suppose that $u(x)$ and $s(x)$ are also polynomials satisfying conditions (I) and (II) in the statement of the theorem. Thus

$$f(x) = u(x)g(x) + s(x),$$

and either $s(x) = 0$ or $\deg(s(x)) < \deg(g(x))$.

Consider

$$0 = f(x) - f(x) = (u(x) - q(x))g(x) + (s(x) - r(x)),$$

or equivalently,

$$(u(x) - q(x))\ g(x) = r(x) - s(x).$$

It immediately follows that

$$\deg\left((u(x) - q(x))\ g(x)\right) = \deg\left(s(x) - r(x)\right).$$

But from condition (II) applied to each of $r(x)$ and $s(x)$ we find that either $s(x) - r(x) = 0$ or $\deg\left(s(x) - r(x)\right) < \deg\left(g(x)\right)$.

On the other hand, if $u(x) - q(x) \neq 0$, then $\deg\left((u(x) - q(x))\ g(x)\right) \geq \deg\left(g(x)\right)$.

From this it follows that we must have $u(x) - q(x) = 0$, i.e. $u(x) = q(x)$. But then

$$s(x) = f(x) - u(x)g(x) = f(x) - q(x)g(x) = r(x).$$

□

1.7. Example. Let $f(x) = 5x^4 + 3x^3 + 1$ and $g(x) = 3x^2 + 2x + 1 \in \mathbb{Z}_7[x]$.

To divide $f(x)$ by $g(x)$, we proceed with *long division* the way that we would with integers. Thus, to keep trace of the "x^2 place" and the "x^1 place", we write

$$f(x) = 5x^4 + 3x^3 + 0x^2 + 0x^1 + 1.$$

(The reader should not look at us with that expression: the natural number $1004 = 1 \cdot 10^4 + 0 \cdot 10^3 + 0 \cdot 10^1 + 4 \cdot 1$. You've seen this kind of thing before.)

One next has to figure out how many $3x^2$'s "fit" into $5x^4$, and this requires us to divide 5 by 3 in $\mathbb{Z}_7$. Since $5 = 3 \cdot 4$ in $\mathbb{Z}_7$, we shall require $4g(x)$ to begin with. Performing that calculation yield the following.

$$\begin{array}{rr|ccccc} & & & & 4x^2 & & \\ \hline 3x^2 + 2x + 1 & \surd & 5x^4 & +3x^3 & +0x^2 & +0x & +1 \\ & & 5x^4 & +1x^3 & +4x^2 & & \\ \hline & & & 2x^3 & +3x^2 & +0x & \end{array}$$

Next, we must decide how man $3x^2$'s "fit" into $2x^3$, and again – this requires us to divide 2 by 3 in $\mathbb{Z}_7$. Since $2 = 3 \cdot 3$ in $\mathbb{Z}_7$, we continue the above long division with $3x$. Observe that we implicitly use the calculation that $3 - 6 = 0 - 3 = 4$ in $\mathbb{Z}_7$.

$$
\begin{array}{rr|rrrr}
 & & & 4x^2 & +3x & \\
\cline{3-6}
3x^2+2x+1 & \big) & 5x^4 & +3x^3 & +0x^2 & +0x & +1 \\
 & & 5x^4 & +1x^3 & +4x^2 & & \\
\cline{3-5}
 & & & 2x^3 & +3x^2 & +0x & \\
 & & & 2x^3 & +6x^2 & +3x & \\
\cline{3-6}
 & & & & 4x^2 & +4x & +1
\end{array}
$$

Finally, we must decide how man $3x^2$'s "fit" into $4x^2$, and again – this requires us to divide 4 by 3 in $\mathbb{Z}_7$. Since $4 = 3 \cdot 6$ in $\mathbb{Z}_7$, we finish the above long division with 6. Again, observe that we implicitly use the calculation that $4 - 5 = 6$ and $1 - 6 = 2$ in $\mathbb{Z}_7$.

$$
\begin{array}{rr|rrrrr}
 & & & & 4x^2 & +3x & +6 \\
\cline{3-7}
3x^2+2x+1 & \big) & 5x^4 & +3x^3 & +0x^2 & +0x & +1 \\
 & & 5x^4 & +1x^3 & +4x^2 & & \\
\cline{3-5}
 & & & 2x^3 & +3x^2 & +0x & \\
 & & & 2x^3 & +6x^2 & +3x & \\
\cline{3-6}
 & & & & 4x^2 & +4x & +1 \\
 & & & & 4x^2 & +5x & +6 \\
\cline{3-7}
 & & & & & 6x & +2
\end{array}
$$

Our conclusion is that

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

i.e.

$$f(x) = (4x^2 + 3x + 6)g(x) + (6x + 2),$$

and so $q(x) = (4x^2 + 3x + 6)$ and $r(x) = 6x + 2$ in this example.

1.8. A simple reason why we consider the division algorithm for polynomial rings with coefficients in a field as opposed to polynomial rings with coefficients in, say, an integral domain is because we must be able to divide the leading coefficient of the numerator $f(x)$ by the leading coefficient of the divisor $g(x)$. Since $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_0 \neq 0$, and since (without loss of generality) the leading coefficient $b_m \in \mathbb{F}$ is non-zero, given any $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots a_0$, we know that we can divide a_n by b_m. Thus we know that this applies to the remainder term of $f(x)/g(x)$, so long as the degree of that remainder term is at least that of $g(x)$.

The ring $R = \mathbb{Z}$ is an integral domain. If $f(x) = 7x^5 + 1$ and $g(x) = 2x^2$, then for any $q(x) \in \mathbb{Z}[x]$,

$$\deg\left(f(x) - q(x)g(x)\right) \geq 5 = \deg f(x),$$

since no choice of $q(x) \in \mathbb{Z}[x]$ can make the coefficient of x^5 equal to zero.

1.9. Theorem. *Let $\mathbb{F}$ be a field. The map*

$$\begin{array}{rccc} \delta: & \mathbb{F}[x] \setminus \{0\} & \to & \mathbb{N}_0 := \{0, 1, 2, \ldots\} \\ & f(x) & \mapsto & \deg(f(x)) \end{array}$$

defines a valuation on the integral domain $\mathbb{F}[x]$. Thus $\mathbb{F}[x]$ is a Euclidean domain.

Proof.

(a) It is an immediate consequence of the division algorithm for polynomial rings, Theorem 1.6 that there exist polynomials $q(x), r(x) \in \mathbb{F}[x]$ such that

$$f(x) = q(x)g(x) + r(x),$$

where $r(x) = 0$ or $\delta(r(x)) = \deg(r(x)) < \deg(g(x)) = \delta(g(x))$, and thus condition (a) of Definition 1.2 is satisfied.

(b) If $0 \neq f(x), g(x) \in \mathbb{F}[x]$, then

$$\begin{aligned} \delta(f(x)) &:= \deg(f(x)) \\ &\leq \deg(f(x)) + \deg(g(x)) \\ &= \deg(f(x)\, g(x)) \\ &= \delta(f(x)g(x)). \end{aligned}$$

Thus δ is a valuation on $\mathbb{F}[x]$.

□

1.10. We have observed earlier in this manuscript that $\mathbb{Z}$ is a principal ideal domain (i.e. a PID). As it turns out, this is a consequence of the fact that it is a Euclidean domain.

1.11. Theorem. *Let E be a Euclidean domain with valuation μ. Then E is a principal ideal domain.*

Proof. Suppose that $K \lhd E$ is an ideal. If $K = \{0\}$, then $K = \langle 0 \rangle$ is singly-generated. Otherwise, if $K \neq \{0\}$, we can choose $m \in K$ such that

$$\mu(m) = \min\{\mu(k) : 0 \neq k \in K\}.$$

(This uses the fact that any non-empty subset of $\mathbb{N}$ admits a minimum element. We say that $\mathbb{N}$ is well-ordered.) We claim that $K = \langle m \rangle$. If $0 \neq k \in K$, then by condition (a) of Definition 1.2 of a valuation, we can find $q, r \in E$ such that

$$k = qm + r,$$

where $r = 0$ or $\mu(r) < \mu(m)$. Observe first that $m \in K$ implies that $qm \in K$ and so $r = k - qm \in K$ as well. Next, note that $0 \neq r$ implies that $\mu(r) < \mu(m)$ contradicts our choice of m. Thus $r = 0$, and so $k = qm \in \langle m \rangle$. This completes the proof.

□

1.12. Corollary. *Let $\mathbb{F}$ be a field. Then $\mathbb{F}[x]$ is a principal ideal domain.*
Proof. This is an immediate consequence of Theorem 1.9 and Theorem 1.11.

□

2. The Euclidean algorithm

2.1. In this section we shall use valuations on an integral domain to obtain a generalisation of the familiar Euclidean algorithm for integers.

2.2. Definition. *Let R be a commutative ring, and let $a, b \in R$ with $b \neq 0$. We say that b **divides** a if there exists $q \in R$ such that $a = qb$. We write $b|a$ when this holds; otherwise we write $b \nmid a$.*

2.3. Exercise. Let R be a commutative ring and $a, b, c \in R$. The following results are readily verified.

(a) If $a, b \neq 0$, $a|b$ and $b|c$, then $a|c$.
(b) If $a \neq 0$, $a|b$ and $a|c$, then $a|(b+c)$ and $a|(b-c)$.
(c) If $a \neq 0$ and $a \mid b$, then $a|bz$ for all $z \in R$.

2.4. Definition. *Let R be a commutative ring, and $a, b \in R$. An element $d \in R$ is said to be a **greatest common divisor** of a and b*

(I) *$d|a$ and $d|b$ – that is, d is a divisor of both a and b; and*
(II) *If $c \in R$, $c|a$ and $c|b$, then $c|d$.*

We shall write $d = \text{GCD}(a, b)$ to denote the fact that d is a greatest common divisor of a and b (when such a divisor exists).

2.5. Examples.

(a) 7 and -7 are both greatest common divisors of 35 and -49 in $\mathbb{Z}$. This highlights the fact that even if a greatest common divisor exists, it need not be unique.
(b) Let R be a commutative ring, $0 \neq a, b \in R$, and d be a GCD of a and b. Suppose that $x \in R$ is invertible.

Now $d \mid a$ implies that $a = dm$ for some $m \in R$, so $a = (dx)x^{-1}m$, implying that $dx \mid a$. Similarly, $dx \mid b$, showing that dx is a common divisor of a and b.

Moreover, if $c \mid a$ and $c \mid b$, then $c \mid d$, so $d = cn$ for some $n \in R$, implying that $dx = cnx$, and thus $c \mid dx$. Hence dx is a GCD of a and b as well.

Applying this to the example above, we saw that 7 is a GCD of 35 and -49 in $\mathbb{Z}$. Since $-1 \in \mathbb{Z}$ is invertible (with inverse equal to itself), we see that -7 is also a GCD of 35 and -49. We have simply generalised this to other rings which may or may not have many more invertible elements.

(c) Let $\mathbb{F}$ be a field and $a, b \in \mathbb{F}$. Given any $0 \neq d \in \mathbb{F}$, $a = d(d^{-1}a)$ and $b = d(d^{-1}b)$, showing that d is a common divisor of a and b. If $0 \neq c \in \mathbb{F}$, then $d = c(c^{-1}d)$, so $c \mid d$. In other words, d is a GCD of a and b. That's about as far from *unique* as you can get.

(d) Let us apply the argument from (b) to the case $R = \mathbb{Q}[x]$.

Consider $f(x) = (x-1)^2$ and $g(x) = (x-1)(x+2) \in \mathbb{Q}[x]$. It follows from the Euclidean algorithm for polynomial rings that $h(x) = (x-1)$ is a greatest common divisor of $f(x)$ and $g(x)$. (See Question 1 on Assignment 4.)

On the other hand, if $u(x) \in \mathbb{Q}[x]$ is any invertible element, then $u(x)h(x)$ is also a GCD for $f(x)$ and $g(x)$. Of course, in $\mathbb{Q}[x]$, the only invertible elements are the non-zero scalar functions $u(x) = u_0$ with $u_0 \neq 0$, but as we have seen, if we had a general commutative ring, there could be many more.

The greatest common divisor d of two non-zero integers a and b can always be expressed as a $\mathbb{Z}$-linear combination of a and b. This notion carries over to Euclidean domains as well.

2.6. Theorem. *Let E be a Euclidean domain. If $a, b \in E$, then $d := \text{GCD}(a, b)$ exists in E. Moreover, there exist $r, s \in E$ such that*

$$d = ra + sb.$$

Proof. Let $K := \{xa + yb : x, y \in E\}$. We leave it as an exercise for the reader to prove that $K \lhd E$ is an ideal of E. By Theorem 1.11, E is a principal ideal domain, and thus

$$K = \langle d \rangle$$

for some $d \in E$. By definition of K, $d = ra + sb$ for some $r, s \in R$.

There remains to prove that $d = \text{GCD}(a, b)$. Note that $a = 1a + 0b$ and $b = 0a + 1b \in K = \langle d \rangle$, implying that there exist $w_0, z_0 \in E$ such that $a = w_0 d$ and $b = z_0 d$. In other words, d is a common divisor of a and b.

Next, suppose that $k \in K$ divides both a and b. Then, by Exercise 2.3, k divides ra and k divides sb, and thus k divides $ra + sb = d$. This shows that d is a greatest common divisor of a and b.

□

The above proof, while "short and sweet", has the disadvantage that it does not tell us *how* to find a GCD, even when one exists. The process of finding the $\text{GCD}(a, b)$ of two elements a, b of a Euclidean domain E is called the **Euclidean algorithm**, and it will be left as a homework assignment question. See also Exercise 6.7 below.

2.7. Example. Let $E = \mathbb{Q}[x]$, so that E is a Euclidean domain with valuation $\delta = \deg(\cdot)$. Let $f(x) = x^2 - 2x + 1$ and $g(x) = x^2 + x - 2$.

Then $f(x) = (x-1)^2$ and $g(x) = (x-1)(x+2)$, and we recall from Example 2.5 (d) above that $d(x) = (x-1)$ is a greatest common divisor of $f(x)$ and $g(x)$.

Note that with $r(x) = -\frac{1}{3}$ and $s(x) = \frac{1}{3} \in \mathbb{Q}[x]$,

$$\begin{aligned} r(x)f(x) + s(x)g(x) &= \frac{1}{3}(-f(x) + g(x)) \\ &= \frac{1}{3}(-(x-1)(x-1) + (x-1)(x+2)) \\ &= (x-1)\frac{1}{3}(-(x-1) + x + 2) = (x-1) = d(x). \end{aligned}$$

This example was perhaps a little "too easy". We'll examine a more interesting example in the homework assignment questions, once we develop the Euclidean algorithm.

3. Unique Factorisation Domains

3.1. In the next Chapter we shall focus our attention on factorisation in polynomial rings over a field. But first let us get a deeper sense of how factorisation works in integral domains. The two main concepts we now introduce are those of *prime* elements and *irreducible* elements. We shall also study a property of certain rings relating to increasing sequences of ideals which will prove extremely useful – see Definition 3.15 below.

3.2. Definition. *Let D be an integral domain. An element $q \in D$ is said to be:*

(a) *a **unit** in D if q is invertible. We only introduce this terminology because it is the standard one, and now the reader will have seen it should the reader wish to consult other sources. To remove all possibility of confusion between this notion and the notion of a multiplicative unit (i.e. a multiplicative identity), we shall simply refer to these elements as **invertible** in these notes.*

(b) *We say that q is a **prime element** if $q \neq 0$, q is not invertible, and whenever $q|(bc)$ for some $b, c \in D$, then $q|b$ or $q|c$.*

(c) *Let $0 \neq q \in D$ be a non-invertible element. Then q is said to be **irreducible** if, whenever we can write*

$$q = uv$$

*for some $u, v \in D$, then either u or v is invertible. Otherwise, we say that q is **reducible**. (Invertible elements and 0 are not considered to be reducible, nor are they considered to be irreducible.*

(d) *Finally, we say that two elements $a, b \in D$ are **associates** if there exists an invertible element $u \in D$ such that $a = ub$.*

Note that the relation $a \sim b$ if a and b are associates is an equivalence relation on D.

3.3. Examples.

(a) It is not hard to see that the only invertible elements of $\mathbb{Z}$ are 1 and -1; the same is true of $\mathbb{Z}[x]$. Thus the only associates of, say, $q(x) = 12x^4 + 3x^2 - 2x + 1$ are $q(x)$ itself and $-q(x) = -12x^4 - 3x^2 + 2x - 1$.

(b) Consider $D = \mathbb{Z}$, then integers.
- The invertible elements of $\mathbb{Z}$ are exactly $\{-1, 1\}$.
- An element $0 \neq r \in \mathbb{Z}$ is reducible if we can write it as $r = uv$ where neither u nor v is invertible. Since $r \neq 0$, clearly neither u nor v can be equal to zero. Since they do not belong to $\{-1, 1\}$ either, we see that $0 \neq r \in \mathbb{Z}$ is reducible exactly when it is either a composite natural number, or the negative of a composite natural number.
  To be irreducible, therefore, r must either be a prime natural number, or the negative of a prime natural number.
- We leave it to the reader to verify that $0 \neq r \in \mathbb{Z}$ is "ring" prime (i.e. prime in the sense of Definition 3.2 (b)) if and only if it is a prime natural number, or the negative of a prime natural number. Observe, therefore, that in $\mathbb{Z}$, the notions of prime elements and irreducible elements coincide. That prime elements are always irreducible is not unique to $\mathbb{Z}$ – see Proposition 3.4 below.
- Finally, two elements a and b of $\mathbb{Z}$ are associates if and only if $b \in \{a, -a\}$.

(c) The invertible elements of $\mathbb{Q}[x]$ are the non-zero scalar polynomials. Thus if $q(x) \in \mathbb{Q}[x]$, the associates of $q(x)$ are $\{r_0 q(x) : 0 \neq r_0 \in \mathbb{Q}\}$. We shall much more to say about irreducible and prime elements of $\mathbb{Q}[x]$ later.

3.4. Proposition. *Let D be an integral domain. If $p \in D$ is prime, then p is irreducible.*

Proof. Let $p \in D$ be prime. We shall argue by contradiction. To that end, suppose that p is reducible. Then we can find $u, v \in D$, neither of which is invertible, such that

$$p = uv.$$

Since p is prime, either $p \mid u$ or $p \mid v$. By relabelling if necessary (keeping in mind that D is commutative), we may assume without loss of generality that $p \mid u$. Thus there exists $r \in D$ such that $u = pr$, whence

$$p = p\,1 = uv = prv.$$

Since D is an integral domain, it satisfies the cancellation law, and so

$$1 = rv = vr.$$

But then v is invertible, a contradiction.

Thus p is irreducible.

□

To prove that the converse is false requires more work. We begin by introducing the following definition.

3.5. Definition. *Let D be an integral domain. A* ***multiplicative norm*** *on D is a function $\nu : D \to \mathbb{N} \cup \{0\}$ satisfying:*

(I) $\nu(x) = 0$ *if and only if* $x = 0$*;*
(II) $\nu(xy) = \nu(x)\nu(y)$ *for all* $x, y \in D$*.*

3.6. Examples.

(a) Again - we turn to $\mathbb{Z}$ for inspiration. Here we may define $\nu(d) = |d|$ for all $d \in \mathbb{Z}$. Clearly this is a multiplicative norm on $\mathbb{Z}$.
(b) Let $\mathbb{Z}[i] := \{a + bi : a, b \in \mathbb{Z}\}$ denote the Gaussian integers. We may define a multiplicative norm on $\mathbb{Z}[i]$ via

$$|a + bi| := a^2 + b^2.$$

That this is indeed a multiplicative norm is left as an exercise for the reader.
(c) Let $D = \mathbb{Z}[\sqrt{-5}] := \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$. Then

$$\nu(a + b\sqrt{-5}) := a^2 + 5b^2$$

defines a multiplicative norm on D. Again, the verification of this (as well as the verification of the fact that D is an integral domain) is left to the reader.

3.7. Theorem. *Let D be an integral domain and $\nu : D \to \mathbb{N} \cup \{0\}$ be a multiplicative norm on D. Then*

(a) $\nu(u) = 1$ *for every invertible element* $u \in D$*; and*
(b) *if* $\{x \in D : \nu(x) = 1\} = \{x \in D : x \text{ is invertible}\}$*, then* $p := \nu(y) \in \mathbb{N}$ *prime implies that* y *is irreducible in* D*.*

Proof.

(a) First note that $\nu(1) = \nu(1 \cdot 1) = \nu(1)\nu(1)$ implies that $\nu(1) \in \{0, 1\}$. But $1 \neq 0$ implies that $\nu(1) \neq 0$, so $\nu(1) = 1$.
Now suppose that $u \in D$ is invertible. Then

$$1 = \nu(1) = \nu(u \cdot u^{-1}) = \nu(u)\,\nu(u^{-1}).$$

Since $\nu(u), \nu(u^{-1}) \in \mathbb{N}$, it follows that $\nu(u) = \nu(u^{-1}) = 1$.
(b) Suppose next that $\nu(x) = 1$ implies that x is invertible, and let $y \in D$ be an element such that $p := \nu(y) \in \mathbb{N}$ is prime.
If $y = uv$ for some $u, v \in D$, then

$$p = \nu(y) = \nu(uv) = \nu(u)\nu(v),$$

and thus either $p = \nu(u)$, in which case $\nu(v) = 1$ and so v is invertible by hypothesis, or $p = \nu(v)$, in which case $\nu(u) = 1$, and so u is invertible by hypothesis. Either way, we find that y is irreducible in D.

□

3.8. Example. Although an element of $\mathbb{Z}$ is prime if and only if it is irreducible, and although prime elements of an integral domain are always irreducible, in general the two concepts are different, as we shall now see.

Let $D = \mathbb{Z}[\sqrt{-5}]$, equipped with the multiplicative norm

$$\nu(a + b\sqrt{-5}) = a^2 + 5b^2.$$

Let $r := 1+2\sqrt{-5}$. Then $\nu(r) = 1+5(2^2) = 21$. We shall prove that r is irreducible, but not prime in D.

$r = 1 + 2\sqrt{-5}$ IS IRREDUCIBLE:

Suppose that $r = uv$ for some $u, v \in D$. Then $21 = \nu(r) = \nu(uv) = \nu(u)\nu(v)$, and thus (by relabelling u and v if necessary), we may assume that either $\nu(u) = 1$ and $\nu(v) = 21$, or that $\nu(u) = 3$ and $\nu(v) = 7$.

Writing $u = u_1 + u_2\sqrt{-5}$, we see that $\nu(u) \in \{1, 3\}$ implies that $u_1^2 + 5u_2^2 \in \{1, 3\}$, whence $u_2 = 0$ and $u = u_1 \in \{-1, 1\}$. Thus u is invertible, proving that r is irreducible.

$r = 1 + 2\sqrt{-5}$ IS NOT PRIME:

Note that

$$r(1 - 2\sqrt{-5}) = (1 + 2\sqrt{-5})(1 - 2\sqrt{-5}) = 1 - 4(-5) = 21 = 3 \cdot 7.$$

Clearly r divides the left-hand side of this equation. If it were prime, then it would have to divide 3 or 7.

Suppose $3 = rx$ for some $x \in D$. Then

$$9 = \nu(3) = \nu(r)\nu(x) = 21\nu(x),$$

which is impossible with $\nu(x) \in \mathbb{N}$. Similarly, if $7 = rx$ for some $x \in D$, then

$$49 = \nu(7) = \nu(r)\nu(x) = 21\nu(x),$$

which is impossible with $\nu(x) \in \mathbb{N}$. Thus $r = 1 + 2\sqrt{-5}$ is not prime.

3.9. Remarks. Let D be an integral domain.

(a) If $0 \neq a, b \in D$ and $\langle a \rangle = \langle b \rangle$, then a and b must be associates.

Indeed, since $a \in \langle a \rangle = \langle b \rangle$, there exists $u \in D$ such that $a = bu$. Similarly, since $b \in \langle b \rangle \subseteq \langle a \rangle$, there exists $v \in D$ such that $b = av$.

Thus $a1 = a = bu = avu$. Since D is an integral domain, it satisfies the cancellation law, and so $vu = uv = 1$, proving that both u and v are invertible. From this it follows that a and b are associates.

Of course, the contrapositive of this is that if a and b are *not* associates, then $\langle a\rangle \neq \langle b\rangle$.

(b) This fails if we are dealing with a general (non-unital) ring. Consider

$$R = \left\{ \begin{bmatrix} 0 & x & y \\ 0 & 0 & z \\ 0 & 0 & 0 \end{bmatrix} : x, y, z \in \mathbb{R} \right\},$$

and let $a = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, and $b = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. Then a and b cannot be associates, since R has no identity element, and as such, no invertible elements.

Now $a \in \langle a\rangle$ by definition, and so $a^2 \in \langle a\rangle$, since the latter is an ideal of R. But then $b = a + a^2 \in \langle a\rangle$, whence $\langle b\rangle \subseteq \langle a\rangle$.

Conversely, $b \in \langle b\rangle$ by definition, and arguing as above, $a = b - b^2 \in \langle b\rangle$, whence $\langle a\rangle \subseteq \langle b\rangle$.

Putting these two containments together,

$$\langle a\rangle = \langle b\rangle$$

even though a and b are not associates.

(c) If $0 \neq d \in D$ is reducible and $d = ab$ for non-invertible $a, b \in D$, then $\langle d\rangle \subsetneq \langle a\rangle$ and $\langle d\rangle \subsetneq \langle b\rangle$.

We shall prove that $\langle d\rangle \subsetneq \langle a\rangle$; the other case is similar.

Suppose otherwise. Then $a \in \langle d\rangle$ implies that $a = dx$ for some $x \in D$. Thus

$$d = ab = dxb.$$

Again, since D is an integral domain, the cancellation law implies that $xb = bx = 1$. Hence b is invertible, a contradiction.

Thus $\langle d\rangle \subsetneq \langle a\rangle$.

(d) If $p \in D$ is irreducible and q is an associate of p, then q is also irreducible.

Suppose that $q = pu$ for some invertible element $u \in D$, and write $q = ab$ for some $a, b \in D$. We must prove that a or b is invertible.

Now $q = ab$ implies that $pu = ab$, or equivalently that $p = u^{-1}ab = (u^{-1}a)b$. Since p is irreducible, either $u^{-1}a$ is invertible, or b is invertible. Of course, if b is invertible, we are done. On the other hand, if $u^{-1}a$ is invertible, then $a^{-1} = (u^{-1}a)^{-1}u^{-1}$, showing that a is invertible as well.

(It is not hard to show that the invertible elements of a unital ring form a group.)

When dealing with principal ideal domains, the distinction between irreducibility and primeness disappears, and we obtain one more equivalent property.

3.10. Theorem. *Let D be a* PID*, and $p \in D$. The following statements are equivalent.*

(a) *p is prime.*
(b) *p is irreducible.*
(c) *$\langle p \rangle$ is maximal.*

Proof.

(a) implies (b): Since every principal ideal domain is an integral domain, this follows immediately from Proposition 3.4.

(b) implies (a): Suppose now that $p \in D$ is irreducible. If $a, b \in D$ and $p \mid (ab)$, then there exists $u \in D$ such that $ab = pu = up$.

Consider $K := \langle a, p \rangle \lhd D$. Since D is a principal ideal domain, $K = \langle k \rangle$ for some $k \in K$. Thus $a, p \in \langle k \rangle$, implying that $p = kx$ and $a = ky$ for some $x, y \in D$. But p is irreducible, so by definition, either k is invertible, or x is invertible.

- If x is invertible, then $k = px^{-1}$, and thus $a = ky = px^{-1}y$ is divisible by p.
- If k is invertible, then $K = D$, and so $1 \in K = \langle a, p \rangle$, which implies that there exist $r, s \in D$ such that

$$1 = ra + sp.$$

Hence

$$b = rab + spb = rup + sbp = (ru + sb)p,$$

so that b is divisible by p.

We have shown that if $p \mid (ab)$, then $p \mid a$ or $p \mid b$. That is, p is prime.

(b) implies (c). Suppose that p is irreducible. Let $K \lhd D$, and suppose that $\langle p \rangle \subseteq K \subseteq D$. Since D is a PID, $K = \langle k \rangle$ for some $k \in D$. Thus $p \in \langle k \rangle$, so $k \mid p$; i.e. there exists $y \in D$ such that $p = ky$. Since p is irreducible, either k or y is invertible. If k is invertible, then $K = D$, whereas if y is invertible, then p and k are associates, and so by Remark 3.9, $K = \langle p \rangle$.

Thus $\langle p \rangle$ is maximal.

(c) implies (b). Suppose that $\langle p \rangle$ is maximal. If p were reducible, then we could write $p = ab$, where neither a nor b is invertible. By Remark 3.9,

$$\langle p \rangle \subsetneq \langle a \rangle.$$

By the maximality of $\langle p \rangle$, it follows that $\langle a \rangle = D$, and thus $1 \in \langle a \rangle$. But since D is commutative, $\langle a \rangle = \{xa : x \in D\}$, and thus there exists $b \in D$ such that $ba = 1 = ab$, proving that a is invertible, a contradiction of our hypothesis. Thus p must be irreducible.

□

A particularly useful application of the above theorem is the following:

3.11. Corollary. *Let $\mathbb{F}$ be a field. A polynomial $p(x) \in \mathbb{F}[x]$ is prime if and only if $p(x)$ is irreducible.*

Proof. This is an immediate corollary of the above theorem combined with the fact that if $\mathbb{F}$ is a field, then $\mathbb{F}[x]$ is a PID (Corollary 1.12).

□

3.12. Definition. *An integral domain D is said to be a* ***unique factorisation domain*** *– denoted* UFD *– if*

(a) *every non-invertible element of D other than zero can be factored as a product of finitely many irreducible elements of D, and*
(b) *if $p_1p_2\cdots p_r = q_1q_2\cdots q_s$ are two factorisations of the same (non-zero, non-invertible) element of D, then $r = s$ and there exists a permutation σ of $\{1, 2, \ldots, r\}$ such that p_j and $q_{\sigma(j)}$ are associates, $1 \leq j \leq r$.*

3.13. Example. Both of these properties are familiar in $\mathbb{Z}$, where – as we have seen (or by applying Theorem 3.10 above) – an integer is irreducible if and only if it is prime. Thus $\mathbb{Z}$ is a UFD.

3.14. Example. Note that every field is a UFD, since $\mathbb{F}$ admits no non-zero, non-invertible elements. This is cheap, but true.

Our next goal is to prove that every PID is a UFD. We shall use the very important concept of rings alluded to at the beginning of this section to do this.

3.15. Definition. *A ring R is said to be* ***Noetherian*** *– or alternatively, R satisfies the* ***ascending chain condition*** *– if whenever*

$$J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots \subseteq J_n \subseteq \cdots$$

is an increasing sequence of ideals in R, there exists $N \geq 1$ such that $m \geq N$ implies that $J_m = J_N$.

3.16. Examples.

(a) Let $n \in \mathbb{N}$ and $R = \mathbb{T}_n(\mathbb{R})$. If $K \lhd \mathbb{T}_n(\mathbb{R})$ and $K := [k_{i,j}] \in K$, then for each $1 \leq i \leq j \leq n$, $L_{i,j} = E_{i,i}KE_{j,j} \in K$. Thus we see that K is spanned by the (non-zero) matrix units it contains. Since we only have finitely many matrix units to choose from, $\mathbb{T}_n(\mathbb{R})$ can have at most finitely many ideals. In particular, any increasing chain of ideals must stabilise and so $\mathbb{T}_n(\mathbb{R})$ is Noetherian.

We shall determine all of the ideals of $\mathbb{T}_n(\mathbb{R})$ in the Assignments.

(b) Let $n \in \mathbb{N}$ and $R = \mathbb{M}_n(\mathbb{R})$. Then, as seen in the Assignments, R is **simple**; that is, the only ideals of R are $\{0\}$ and R itself. Clearly any ascending chain must stabilise.

(c) A more interesting example is $R = \mathbb{Z}$. Since $\mathbb{Z}$ is a principal ideal domain, given any ideal $J \lhd \mathbb{Z}$, there exists $m \in \mathbb{Z}$ such that $J = \langle m \rangle$.

Suppose that

$$J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots \subseteq J_n \subseteq \cdots$$

is an increasing sequence of ideals in $\mathbb{Z}$, and choose $m_k \in \mathbb{Z}$ such that $J_k = \langle m_k \rangle$ for each $k \geq 1$. Note that

$$\langle m_1 \rangle \subseteq \langle m_2 \rangle \subseteq \langle m_3 \rangle \subseteq \cdots$$

implies that $m_k \in \langle m_{k+1} \rangle$, or equivalently that $m_{k+1} | m_k$ for each $k \geq 1$.

Since m_1 has at most finitely many non-trivial (i.e. non-invertible) factors, there must exist $N \in \mathbb{N}$ such that $n \geq N$ implies that $|m_n| = |m_N|$, and thus

$$J_n = \langle m_n \rangle = \langle m_N \rangle = J_N \text{ for all } n \geq N.$$

Thus $\mathbb{Z}$ is Noetherian.

(d) Let $R = \prod_{n=1}^{\infty} \mathbb{R} := \{(x_n)_{n=1}^{\infty} : x_n \in \mathbb{R} \text{ for all } n \geq 1\}$. For each $k \geq 1$, set

$$J_k = \{(x_n)_n \in R : x_m = 0 \text{ if } m \geq k\}.$$

We leave it to the reader to verify that J_k is an ideal of R for each $k \geq 1$, and clearly

$$J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots \subseteq J_n \subseteq \cdots.$$

That $J_k \neq J_{k+1}$ for all $k \geq 1$ is also a (very routine) exercise. Thus R is *not* Noetherian.

(e) Let $R = \mathcal{C}([0,1], \mathbb{R})$. For each $n \geq 1$, let

$$J_n := \{f \in \mathcal{C}([0,1], \mathbb{R}) : f(x) = 0 \text{ for all } x \in [0, \frac{1}{n}]\}.$$

Again, we leave it to the reader to verify that J_n is an ideal of R for each $n \geq 1$, and that

$$J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots \subseteq J_n \subseteq \cdots.$$

We also leave it to the reader to verify that $J_n \neq J_{n+1}$ for any $n \geq 1$.

Thus $\mathcal{C}([0,1], \mathbb{R})$ is *not* Noetherian.

3.17. Theorem. *Every* PID *is Noetherian.*

Proof. Suppose that D is a PID and that

$$J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots \subseteq J_k \subseteq \cdots$$

is an ascending chain of ideals of D. Let $J = \bigcup_{k=1}^{\infty} J_k$.

We claim that J is an ideal of D. That $J \neq \varnothing$ is clear, since $J_1 \neq \varnothing$ and $J_1 \subseteq J$. Next, given $a, b \in J$, there exist k_1 and $k_2 \in \mathbb{N}$ such that $a \in J_{k_1}$ and $b \in J_{k_2}$. It follows that $a, b \in J_{k_0}$, where $k_0 = \max(k_1, k_2)$. Since J_{k_0} is an ideal of D, $a + b \in J_{k_0} \subseteq J$. Thus J is closed under addition.

If $a \in J$ and $r \in D$, then $a \in J_k$ for some $k \geq 1$, and $J_k \lhd D$, so $ra = ar \in J_k \subseteq J$. By the Ideal Test, J is an ideal of D.

Since D is a PID, there exists an element $m \in D$ such that $J = \langle m \rangle$.

Now $m \in J$ implies that there exists $N \geq 1$ such that $m \in J_N$. But then

$$J = \langle m \rangle \subseteq J_N \subseteq J,$$

implying that $J_N = J$. If $k \geq N$, then $J = J_N \subseteq J_k \subseteq J$, further implying that $J_k = J$ for all $k \geq N$.

By definition, D is Noetherian.

□

3.18. Corollary. *Suppose that D is a* PID. *If $0 \neq d \in D$ and d is not invertible, then d is a product of finitely many irreducible elements of D.*

Proof. The proof will proceed in two steps. First we shall show that d admits and irreducible divisor. Then we shall prove that it is a product of irreducible divisors.

STEP 1. Suppose that every divisor of d is reducible. Since $d \mid d$ is obvious, we can then write $d = a_1 b_1$, where a_1, b_1 are not invertible. Let $J_1 := \langle b_1 \rangle$.

Clearly $b_1 \mid b_1$, and thus $b_1 \mid d$. By hypothesis, b_1 is reducible, so we may write $b_1 = a_2 b_2$ for non-invertible b_2. Set $J_2 = \langle b_2 \rangle$.

In general, given $k \geq 2$ and having written $d = a_1 a_2 \cdots a_k b_k$, where each $a_1, a_2, \ldots, a_k$ and b_k are non-invertible, we note that $b_k \mid b_k$ and hence $b_k \mid d$. By hypothesis, b_k is reducible, and so we may write $b_k = a_{k+1} b_{k+1}$ for non-invertible $a_{k+1}, b_{k+1} \in D$. We set $J_{k+1} = \langle b_{k+1} \rangle$.

By Remark 3.9, $J_k = \langle b_k \rangle \subsetneq \langle b_{k+1} \rangle = J_{k+1}$, $k \geq 1$. But then

$$J_1 \subseteq J_2 \subseteq J_3 \subseteq \cdots \subseteq J_n \subseteq \cdots$$

is an increasing sequence of ideals of D which does not stabilise, contradicting the fact that every PID is Noetherian.

This contradiction proves that d admits an irreducible divisor.

STEP 2. To prove that it d is a product of irreducible divisors, we shall once again argue by contradiction. Suppose, to the contrary, that d can not be written as a finite product of irreducible elements of D. Then d itself is reducible, otherwise it is a product of (one) irreducible element, namely itself.

From STEP 1, there exist an irreducible element q_1 and an element $y_1 \in D$ such that $d = q_1 y_1$. We claim that y_1 is not invertible. Indeed, if y_1 were invertible, then d and q_1 would be associates. By Remark 3.9 (d), this would mean that d is irreducible, contradicting our current hypothesis.

If y_1 were irreducible, then $d = q_1 y_1$ would be a product of irreducible elements, contradicting our current hypothesis. Thus, applying STEP 1 to y_1, we may write $y_1 = q_2 y_2$, where q_2 is irreducible and $y_2 \in D$. Note that $d = q_1 y_1 = q_1 q_2 y_2$.

If y_2 were irreducible, then d would be a product of irreducible elements, a contradiction. But y_2 is not invertible - otherwise y_1 and q_2 are associates, contradicting the fact that y_1 is reducible.

In general, given $k \geq 2$ and $q_1, q_2, \ldots, q_k \in D$ irreducible, $y_1, y_2, \ldots y_{k-1} \in D$ reducible and satisfying $y_j = q_{j+1} y_{j+1}$ for all $1 \leq j \leq k-1$, we get

$$d = q_1 q_2 \cdots q_k y_k.$$

If y_k were irreducible, then clearly d would be a product of irreducible elements of D, a contradiction.

But as always, y_k is not invertible, otherwise y_{k-1} and q_k would be associates, and y_{k-1} would have been irreducible, a contradiction.

Observe that by Remark 3.9, $\langle y_k \rangle \neq \langle y_{k+1} \rangle$ whenever y_k and y_{k+1} are not associates. Combining this with the fact that $y_{k+1} \mid y_k$ implies that

$$\langle y_1 \rangle \subsetneq \langle y_2 \rangle \subsetneq \langle y_2 \rangle \subsetneq \cdots,$$

contradicting the fact that every PID is Noetherian (Theorem 3.17).

We conclude that d can indeed be written as a finite product of irreducible elements of D, as claimed.

□

3.19. Theorem. *Every* PID *is a* UFD.

Proof. Let D be a PID, and let $0 \neq d \in D$ be a non-invertible element. By Corollary 3.18, we can write

$$d = q_1 \, q_2 \cdots q_m$$

for some integer $m \geq 1$, where each q_j is an irreducible element of D.

Suppose that $d = p_1 \, p_2 \cdots p_n$, where $n \in \mathbb{N}$ and each p_j is irreducible in D. Without loss of generality, we may assume that $m \leq n$. (Otherwise, we interchange the sequence of p_j's and q_i's.)

Recall that an element of a PID is irreducible if and only if it is prime. Thus each q_j is prime, $1 \leq j \leq m$. Since $q_1 \mid p_1 \, p_2 \cdots p_n$, there exists $k \geq 1$ such that $q_1 \mid p_k$. By reordering the terms $p_1, p_2, \ldots, p_n$ (principal ideal domains are commutative, after all!), we may assume that $k = 1$. Thus $q_1 \mid p_1$, and so there exists $u_1 \in D$ such that $p_1 = q_1 u_1$. But p_1 is irreducible, and thus either q_1 is invertible, or u_1 is. But q_1 is irreducible, so it is non-invertible. Hence u_1 is invertible.

That is,

$$d = p_1 \, p_2 \cdots p_n = q_1 \, u_1 \, p_2 \cdots p_n = q_1 \, q_2 \cdots q_m.$$

Since every PID is an integral domain, the cancellation law holds, and so

$$(u_1^{-1} q_2) \, q_3 \cdots q_m = p_2 \, p_3 \, p_4, \ldots, p_n.$$

Note that $u_1^{-1} q_2$ is irreducible (hence prime), being the associate of an irreducible element in D.

Since $(u_1^{-1} q_2) \mid p_2 \, p_3 \, p_4, \ldots, p_n$, there exists $2 \leq k \leq n$ such that $(u_1^{-1} q_2) \mid p_k$. By reordering the p_j's if necessary, we may assume that $j = 2$, i.e. that $(u_1^{-1} q_2) \mid p_2$. Write $p_2 = (u_1^{-1} q_2) u_2$ for some $u_2 \in D$. Since p_2 is irreducible, and thus either $(u_1^{-1} q_2)$ is invertible, or u_2 is. But $(u_1^{-1} q_2)$ is irreducible, so it is non-invertible. Hence u_2 is invertible.

Thus

$$(u_1^{-1}q_2)\,q_3\cdots q_m = p_2\,p_3\cdots p_n = (u_1^{-1}q_2)\,u_2\,p_3\,p_4\cdots p_n.$$

Again, we may apply the cancellation law to conclude that

$$q_3\,q_4\cdots q_m = u_2\,p_3\,p_4\cdots p_n,$$

or equivalently, that

$$(u_2^{-1}q_3)\,q_4\cdots q_m = p_3\,p_4\cdots p_n.$$

Proceeding in this manner, we see that after m steps (and potentially m reorderings of the p_j terms), we may assume that $p_1 = u_1q_1$ and $p_j = (u_{j-1}^{-1}u_j)q_j$, $2 \le j \le m$. At that stage, we have

$$\begin{aligned} q_1\,q_2\cdots q_m &= (u_1q_1)(u_1^{-1}u_2)q_2)(u_2^{-1}u_3q_3)\cdots(u_{m-1}^{-1}u_mq_m)\,p_{m+1}\,p_{m+2}\cdots p_n \\ &= u_m(q_1\,q_2\cdots q_m)\,p_{m+1}\,p_{m+2}\cdots p_n. \end{aligned}$$

Applying the cancellation law yet again yields

$$1 = u_m\,p_{m+1}\,p_{m+2}\cdots p_n.$$

If $n > m$, then this contradicts the fact that each p_j, $m \le j \le n$ is irreducible, hence non-invertible. Thus $n = m$, and each pair (q_j, p_j) consists of associates, as required.

□

The next result is familiar - but it's certainly nice to have a formal proof of this fact.

3.20. Corollary. *$\mathbb{Z}$ is a unique factorisation domain.*

Proof. As we have already noted, $\mathbb{Z}$ is a PID, from which the claim follows.

□

More importantly for us – and this will certainly be useful in the next Chapter – we have:

3.21. Corollary. *Let $\mathbb{F}$ be a field. Then $\mathbb{F}[x]$ is a unique factorisation domain.*

Proof. By Corollary 1.12, $\mathbb{F}[x]$ is a principal ideal domain. The result now follows from Theorem 3.19.

□

Supplementary Examples.

S6.1. Example. Most sources list at most three, and often just two, examples of Euclidean domains. The first two are invariably the ring $\mathbb{Z}$ of integers and polynomial rings with coefficients in a field, which we have dutifully mentioned above. Let us now consider the Gaussian integers, which is typically mentioned third, if at all.

Recall that the Gaussian integers are the set

$$\mathbb{Z}[i] := \{a + bi : a, b \in \mathbb{Z}\},$$

where $i^2 = -1$. We think of these as a unital (and necessarily commutative) subring of $\mathbb{C}$. Note that $\mathbb{C}$ is a field, and thus it is an integral domain. If $a + bi \neq 0 \neq c + di$ are non-zero Gaussian integers, then they are also non-zero complex numbers, so

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i \neq 0.$$

As with $\mathbb{Z}$, we may define a Euclidean valuation on $\mathbb{Z}[i] \setminus \{0\}$ through the map $\mu(a + bi) := |a + bi|^2 = a^2 + b^2$.

That μ satisfies condition (b) of a Euclidean valuation is left as an exercise for the reader to verify. Let us verify condition (a).

Let $a = w + xi$ and $b = y + zi \in \mathbb{Z}[i]$. We must find $q, r \in D$ such that $a = qb + r$, where either $r = 0$ or $\mu(r) < \mu(b)$. We consider two cases.

- Suppose first that $b = k$ for some $0 < k \in \mathbb{N}$. It is not hard to see that we can write

  $$w = u_1 k + v_1, \quad x = u_2 k + v_2,$$

  where $\max(|v_1|, |v_2|) \leq \dfrac{k}{2}$. Thus

  $$a = w + xi = (u_1 k + v_1) + (u_2 k + v_2)i = (u_1 + u_2 i)k + (v_1 + v_2 i) = qk + r,$$

  where $q = (u_1 + u_2 i)$ and $r = (v_1 + v_2 i)$. Note that if $r \neq 0$, then

  $$\mu(r) = \mu(v_1 + v_2 i) = v_1^2 + v_2^2 \leq \left(\frac{k}{2}\right)^2 + \left(\frac{k}{2}\right)^2 < k^2 = \mu(k) = \mu(b).$$

- Now suppose that $b = y + zi \neq 0$. Let $k = \overline{b}b = (y - zi)(y + zi)$, and observe that $0 < y^2 + z^2 = k \in \mathbb{N}$. Let us now apply the above argument to $a\overline{b}$ and k to see that we may write

  $$a\overline{b} = qk + r_0,$$

  where $r_0 = 0$ or $\mu(r_0) < \mu(k)$.

  We pause to observe that for all $s + ti \in \mathbb{Z}[i]$, $\mu(s + ti) = s^2 + t^2 = |s + ti|^2$, the square of the absolute value of the complex number $s + ti$, and thus

  $$\begin{aligned} \mu((m + ni)(s + ti)) &= |(m + ni)(s + ti)|^2 \\ &= |m + ni|^2 \ |s + ti|^2 \\ &= \mu(m + ni) \ \mu(s + ti). \end{aligned}$$

Thanks to this observation, we see that

$$\begin{aligned}\mu(a-qb)\mu(\bar{b}) &= \mu(a\bar{b}-qb\bar{b})\\ &= \mu(a\bar{b}-qk)\\ &= \mu(r_0)\\ &< \mu(k) = \mu(b\bar{b})\\ &= \mu(b)\mu(\bar{b}).\end{aligned}$$

Since $\mu(\bar{b}) = \mu(b) = y^2 + z^2 \neq 0$ (recall that $b \neq 0$), we may divide both sides of this inequality by the positive integer $\mu(\bar{b})$ to obtain

$$\mu(a-qb) < \mu(b).$$

Let $r = a - qb$. Then $\mu(r) < \mu(b)$ and

$$a = qb + r,$$

as required to prove that μ is a valuation on $\mathbb{Z}[i]$, and completing the proof that $\mathbb{Z}[i]$ is a Euclidean domain.

S6.2. Example. In Exercise 6.6, you are asked to prove that if E is a Euclidean domain with Euclidean valuation μ, then

(a) $\mu(1) = \min\{\mu(d) : 0 \neq d \in D\}$, and
(b) $x \in E$ is invertible if and only if $\mu(x) = \mu(1)$.

Note that if $\mathbb{F}$ is a field and $\delta = \deg(\cdot)$ is the usual valuation on $\mathbb{F}[x]$, then

$$\delta(1) = \deg(1) = 0$$

is the minimum possible degree for any non-zero element of $\mathbb{F}[x]$. It is achieved at all $p(x) = p_0 \neq 0$.

In $\mathbb{Z}$ with valuation $\mu(n) = |n|$, we see that $\mu(1) = 1$ is the minimum possible value of μ (why?), achieved at $n = 1$ and $n = -1$.

S6.3. Example. Consider the Euclidean domain $E := \mathbb{Q}[x]$, equipped with the valuation $\delta(q(x)) := \deg(q(x))$, $q(x) \in \mathbb{Q}[x]$.

Let's see how to use the division algorithm to find the greatest common divisor of $f(x) = x^4 + x^3 - x - 1$ and $g(x) = x^3 + 1$.

By definition, we may find $q^{(1)}(x), r^{(1)}(x) \in E$ such that

$$f(x) = q^{(1)}(x)g(x) + r^{(1)}(x),$$

and either $r^{(1)}(x) = 0$ or $\deg(r^{(1)}(x)) < \deg(g(x))$. This is precisely the division algorithm: the computation

$$
\begin{array}{r|llll}
 & & & x & +1 \\
\hline
x^3+1 & x^4 & +x^3 & +0x^2 \; -x & -1 \\
 & x^4 & & +x & \\
\hline
 & & x^3 & +0x^2 \; -2x & -1 \\
 & & x^3 & & +1 \\
\hline
 & & & -2x & -2
\end{array}
$$

shows that $f(x) = (x+1)(x^3+1) + (-2x-2)$, so we set $q^{(1)}(x) = x+1$ and $r^{(1)}(x) = (-2x-2)$. Anything that divides both $f(x)$ and $g(x)$ must divide $f(x) - q^{(1)}(x)g(x) = r^{(1)}(x)$, so we look for common divisors of $g(x)$ and $r^{(1)}(x)$.

The advantage of doing this is that we have reduced the problem to the case where the maximum degree of the two polynomials whose greatest common divisor we are hunting has been reduced from $\deg(f(x)) = 4$ to $\deg(g(x)) = 3$. If we continue doing this process, it must eventually stop, since we cannot have a degree that is negative!

Consider

$$
\begin{array}{r|llll}
 & & -\frac{1}{2}x^2 & +\frac{1}{2}x & -\frac{1}{2} \\
\hline
-2x-2 & x^3 & +0x^2 & +0x & +1 \\
 & x^3 & +x^2 & & \\
\hline
 & & -x^2 & & +1 \\
 & & -x^2 & -x & \\
\hline
 & & & x & +1 \\
 & & & x & +1 \\
\hline
 & & & & 0
\end{array}
$$

Thus $g(x) = x^3+1 = (-\frac{1}{2}x^2 + \frac{1}{2}x - \frac{1}{2})r^{(1)}(x)$, and so the greatest common divisor of $g(x) = x^3+1$ and $r^{(1)}(x) = -2x-2$ must be $r^{(1)}(x) = -2x-2$.

But then – as we argued above – this must also divide $f(x)$, and so $r^{(1)}(x) = \text{GCD}(f(x), g(x))$. Note that $r^{(1)}(x)$ and $h(x) := x+1$ are associates, since $-\frac{1}{2} \in \mathbb{Q}[x]$ is invertible. So we could just as well say that

$$h(x) = x+1 = \text{GCD}(f(x), g(x)).$$

Indeed,

$$f(x) = x^4 + x^3 - x - 1 = (x^3-1)(x+1) = (x-1)(x^2+x+1)(x+1),$$

while

$$g(x) = x^3 + 1 = (x+1)(x^2-x+1).$$

We leave it to the reader to verify that (x^2+x+1) and (x^2-x+1) have no common factors.

S6.4. Example. It is interesting to look for the **least common multiple** $\text{LCM}(f(x), g(x))$, where $f(x), g(x) \in \mathbb{Q}[x]$ are the polynomials from Example S6.3 above.

The definition of the least common multiple is drawn from the definition used for positive integers. We shall say that $h(x) = \text{LCM}(f(x), g(x))$ if

- $f(x) \mid h(x)$ and $g(x) \mid h(x)$, so that $h(x)$ is a multiple of both $f(x)$ and $g(x)$, and
- if $k(x) \in \mathbb{Q}[x]$ is any multiple of both $f(x)$ and $g(x)$, then $h(x) \mid k(x)$.

The decomposition

$$f(x) = (x-1)(x^2+x+1)(x+1)$$
$$g(x) = (x+1)(x^2-x+1)$$

obtained there shows that $(x+1), (x-1), (x^2+x+1)$ and (x^2-x+1) must all divide the least common multiple. Thus

$$h(x) := \text{LCM}(f(x), g(x)) = (x+1)(x-1)(x^2+x+1)(x^2-x+1) = x^6 - 1.$$

S6.5. Example. We know that $\mathbb{Z}$ is a PID, and that $\langle 17 \rangle$ is a prime ideal of $\mathbb{Z}$. Then

$$\frac{\mathbb{Z}}{\langle 17 \rangle} \simeq \mathbb{Z}_{17}$$

is a field, as we have seen earlier.

S6.6. Example. Consider the ring $\mathbb{R}[x]$ of polynomials with coefficients in $\mathbb{R}$. Since $\mathbb{R}$ is a field and therefore a PID, we see that $\mathbb{R}[x]$ is also a PID, by Corollary 1.12. By Theorem 3.17, $\mathbb{R}[x]$ is Noetherian.

For each $n \in \mathbb{N}$, set $K_n := \langle x^n \rangle = \{x^n q(x) : q(x) \in \mathbb{R}[x]\}$. It is not hard to verify that

$$\cdots \subsetneq K_3 \subsetneq K_2 \subsetneq K_1.$$

In other words, we can find a decreasing sequence of distinct ideals of $\mathbb{R}[x]$.

A ring R is said to be **Artinian** – or alternatively, R satisfies the **descending chain condition** – if whenever

$$J_1 \supseteq J_2 \supseteq J_3 \supseteq \cdots \supseteq J_n \supseteq \cdots$$

is a *decreasing* sequence of ideals in R, there exists $N \geq 1$ such that $m \geq N$ implies that $J_m = J_N$.

We conclude that $\mathbb{R}[x]$ is Noetherian, but not Artinian.

S6.7. Example. Suppose that R is a ring and that $K \triangleleft R$ is an ideal. We leave it as an Assignment Exercise for the reader to prove that the ideals of R/K are of the form J/K, where $J \triangleleft R$ is an ideal which contains K.

From this it readily follows that if R is Noetherian, then so is R/K. (Check!)

S6.8. Example. Suppose that R and $\mathcal{S}$ are rings and that R is Noetherian. If $\varphi : R \to \mathcal{S}$ is a homomorphism, then $\varphi(R)$ is Noetherian as well.

Indeed, by the First Isomorphism Theorem ,

$$\varphi(R) \simeq \frac{R}{\ker \varphi},$$

and we saw in the previous example that R/K is Noetherian for any ideal of R.

S6.9. Example. Let D be a PID, and let $K = \langle b \rangle \lhd D$ be an ideal of D. Then D/K is a PID as well.

Again, we remark that any ideal of D/K is of the form J/K, where $J \lhd D$ is an ideal which contains K. Since $J = \langle d \rangle$ for some $d \in D$, we see that $J/K = \langle d + K \rangle \lhd D/K$.

S6.10. Example. For those of you with a good background in Complex Analysis, you may be interested to learn that the ring

$$H(\mathbb{D}) := \{f : \mathbb{D} \to \mathbb{C} \mid f \text{ is holomorphic}\}$$

is not Artinian, since the ideals $K_n = \{f \in H(\mathbb{D}) : f \text{ has a zero of degree at least } n \text{ at } 0\}$, $n \geq 1$ are descending and all distinct.

Appendix

A6.1. Although we won't prove it, it can be shown that if D is a UFD, then so is $D[x]$. In particular, $\mathbb{Z}[x]$ is a UFD. Having said this, we *have proven* that if D is a PID, then so is $D[x]$, and thus $D[x]$ is a UFD. This gives an explicit proof of the fact that $\mathbb{Z}[x]$ is a UFD.

A6.2. As we have already alluded to, in a Euclidean domain we may apply a version of division algorithm finitely often to find the greatest common divisor of two elements. In this context, it is referred to as the **Euclidean algorithm**. It's proof will appear as an Assignment question.

A6.3. Recall that we defined a **multiplicative norm** ν on an integral domain D be a function $\nu : D \to \mathbb{N} \cup \{0\}$ satisfying

(I) $\nu(x) = 0$ if and only if $x = 0$; and
(II) $\nu(xy) = \nu(x)\nu(y)$ for all $x, y \in D$.

Let's find out what these might look like when $D = \mathbb{Z}$, and use that inspiration to figure out what's happening in more general PID's.

In the case where $D = \mathbb{Z}$ and $\nu : \mathbb{Z} \to \mathbb{N} \cup \{0\}$ is a multiplicative norm, by definition we know that $\nu(0) = 0$. Furthermore, by Theorem 3.7, we also know that $\nu(1) = \nu(-1) = 1$, since 1 and -1 are invertible in $\mathbb{Z}$.

Now, for each $2 \leq n \in \mathbb{Z}$, we can factor n as a product $n = p_1 p_2 \cdots p_k$ of prime numbers (each positive). By definition of a multiplicative norm (and with the aide of a routine induction argument), we find that

$$\nu(n) = \nu(p_1)\nu(p_2)\cdots\nu(p_k).$$

In other words, if we know the value of $\nu(p)$ for each prime number $p \in \mathbb{N}$, then we now know the value of $\nu(n)$ for each $n \geq 2$. We also know that

$$\nu(-n) = \nu(-1)\nu(n) = 1 \cdot \nu(n),$$

which means that we know the value of $\nu(m)$ for all $m := (-n) \leq -2$. Thus knowing the value of $\nu(p)$ for every prime is equivalent to understanding the value of $\nu(n)$ for all $n \in \mathbb{Z}$.

What restrictions do we have on the possible values of $\nu(p)$, p prime? None, other than $\nu(p) \geq 1$ (since $p \neq 0$ implies that $\nu(p) \neq 0$). Thus for each $p \geq 2$ prime, we can choose $m_p \in \mathbb{N}$, set $\nu(p) := m_p$, and set

$$\nu(n) = m_1 m_2 \cdots m_k$$

whenever $n = p_1 p_2 \cdots p_k$. What makes this well-defined is that $\mathbb{Z}$ is a PID, which allows us to

- write every non-zero, non-invertible $n \in \mathbb{Z}$ as a product of prime elements of $\mathbb{Z}$ (i.e. $|p| \in \mathbb{N}$ is a prime number in the usual sense),
- observe that prime elements and irreducible elements coincide, and

- use the fact that every PID is a UFD to note that the decomposition of non-zero, non-invertible elements of $\mathbb{Z}$ into products of primes is essentially unique (up to the order of the terms and replacing primes by their associates). Since $\nu(a) = \nu(b)$ when a and b are associates, and since $\nu(a)\nu(b) = \nu(b)\nu(a)$ because multiplication in $\mathbb{N} \cup \{0\}$ is commutative, this means that $\nu(n)$ is well-defined by the above formula.

This allows us to show that there exist an uncountable number of multiplicative norms on $\mathbb{Z}$; indeed, if we set $\mathcal{P} := \{2, 3, 5, 7, 11, \ldots\}$ to denote the set of prime natural numbers, then each function $f : \mathcal{P} \to \mathbb{N}$ determines a multiplicative norm

$$\nu_f(p) := f(p) \text{ for all } p \in \mathcal{P},$$

and the correspondence is bijective. How many such functions are there? Precisely $|\mathbb{N}|^{|\mathcal{P}|}$, where $|X|$ denotes the cardinality of a set X. In our case,

$$|\mathbb{N}|^{|\mathcal{P}|} = (\aleph_0)^{\aleph_0} = 2^{\aleph_0} = c,$$

the **cardinality of the continuum**. Of course, $c = |\mathbb{R}|$, so there are as many multiplicative norms on $\mathbb{Z}$ as there are points in $\mathbb{R}$, in the sense that there exists a bijection between these two sets. Finding that bijection is another matter, however. (We point out that these cardinality arguments are just CULTURE. Don't worry if you haven't seen them before.)

A6.4. As mentioned above, we can extend this analysis to more general PID's. If D is a PID, then D is a UFD. To define a multiplicative norm ν on D, it suffices to choose a positive integer m_p for every irreducible (equivalently every prime) element of D, and then to set

- $\nu(0) = 0$;
- $\nu(u) = 1$ whenever $u \in D$ is invertible, and
- $\nu(p) := m_p$.

The value of $\nu(d)$ for an arbitrary non-zero, non-invertible element of D is then entirely determined by its essentially unique factorisation as a product of primes.

The moral of the story is that PIDS are good.

A6.5. Don't look at these notes that way. You've seen this phenomenon before. Recall that if $\mathcal{V}$ and $\mathcal{W}$ are vector spaces over $\mathbb{R}$, and if $\mathfrak{B} := \{b_\lambda\}_{\lambda \in \Lambda}$ is a basis for $\mathcal{V}$, then to define a linear map $T : \mathcal{V} \to \mathcal{W}$ it suffices to choose any $w_\lambda \in \mathcal{W}$, $\lambda \in \Lambda$ and to set each $Tb_\lambda := w_\lambda$.

If $v \in \mathcal{V}$ is arbitrary, then there exist $m \geq 1$, real numbers $r_1, r_2, \ldots, r_m$ and indices $\lambda_1, \lambda_2, \ldots, \lambda_m$ such that

$$v = r_1 b_{\lambda_1} + r_2 b_{\lambda_2} + \cdots r_m b_{\lambda_m}.$$

If T is to be linear, then we must have

$$\begin{aligned} Tv &= T(r_1 b_{\lambda_1} + r_2 b_{\lambda_2} + \cdots r_m b_{\lambda_m}) \\ &= r_1 T b_{\lambda_1} + r_2 T b_{\lambda_2} + \cdots r_m T b_{\lambda_m} \\ &= r_1 w_{\lambda_1} + r_2 w_{\lambda_2} + \cdots r_m w_{\lambda_m}. \end{aligned}$$

(We leave it to the reader to check that this is well-defined.)

In other words – in vector space theory, it suffices to know what a linear map does to a basis to completely understand that linear map. In the setting of a PID, it suffices to know know a multiplicative norm does to the prime elements (i.e. the irreducible elements) of that PID to completely understand the multiplicative norm.

It is not that these situations are identical: but there are definite parallels, and understanding one allows us to better understand the other. This is what makes mathematics interesting – not just the attention that we get from rich and attractive people at political pyjama parties.

A6.6. In an earlier version of this text, I put out a plea for more examples of Euclidean domains. After consulting a number of elementary abstract algebra textbooks, I was struck by the fact that virtually every such textbook always referred to the three examples we mention in Section 6.1, and to no other examples.

Mr. Nic Banks answered my plea for more examples. A **discrete valuation ring** (a.k.a. DVR) D may be defined as a PID with precisely one non-zero prime ideal. By Exercise 6.11 below, this is the same as asking that it has precisely one non-zero maximal ideal. (The expression **local pid** is also used.) It can be shown that every DVR is a Euclidean domain. As explained to me by Mr. Banks, the motivating examples for DVR's come from number theory and algebraic geometry. In particular, "DVR's have very simple ideal factorisations ... and ideal factorisations are essential in problems like Fermat's Last Theorem".

A concrete example of a DVR is the set of p-adic integers, where $p \in \mathbb{N}$ is a prime number. Unfortunately, this is a bit beyond the scope of these notes, but not much. We include the definition for those willing to challenge themselves.

Definition. *An **inverse system of rings** is a sequence $(R_n)_n$ together with a sequence $(\varphi_n)_n$ of homomorphisms*

$$\cdots \longrightarrow R_{n+1} \xrightarrow{\varphi_n} R_n \longrightarrow \cdots \longrightarrow R_2 \xrightarrow{\varphi_1} R_1.$$

*The **inverse limit***

$$R := \varprojlim R_n$$

*is the subset of the direct product $\prod_n R_n := \{(r_n)_n : r_n \in R_n \text{ for all } n \geq 1\}$ for which $\varphi_n(r_{n+1}) = r_n$ for all $n \geq 1$. For each $m \geq 1$, this gives rise to a **projection map** $\pi_m : R \to R_n$ such that $\pi_m((r_n)_n) = r_m$.*

(For those who are interested in such things – know that inverse limits may be defined in any category.)

If we fix a prime number $p \in \mathbb{N}$, the **ring of p-adic integers** $\mathbb{Z}_p$ is the inverse limit

$$\mathbb{Z}_p = \varprojlim \mathbb{Z}/p^n\mathbb{Z}$$

of the inverse system of rings $(\mathbb{Z}/p^n\mathbb{Z})$ using the homomorphisms

$$\begin{array}{rccc} \varphi_n: & \mathbb{Z}/p^{n+1}\mathbb{Z} & \to & \mathbb{Z}/p^n\mathbb{Z} \\ & z+\langle p^{n+1}\rangle & \to & z+\langle p^n\rangle. \end{array}$$

(Note: Here is another example of the lack of imagination of certain mathematicians insofar as choosing good notation is concerned: this version of $\mathbb{Z}_p$ – i..e. the inverse limit – has **nothing to do** with the field $\mathbb{Z}_p := \mathbb{Z}/\langle p\rangle$. It is a mystery as deep and almost as important as the mystery behind how they get the caramel into the Caramilk bar to determine why the same notation would have been chosen for both.)

The identity of this ring is $e := (1+\langle p\rangle, 1+\langle p^2\rangle, 1+\langle p^3\rangle, \ldots)$, and the map

$$\begin{array}{rccc} \tau: & \mathbb{Z} & \to & \lim_{\leftarrow}\mathbb{Z}/p^n\mathbb{Z} \\ & z & \mapsto & (1+\langle p\rangle, 1+\langle p^2\rangle, 1+\langle p^3\rangle, \ldots) \end{array}$$

is an injective ring homomorphism. Thus $\mathbb{Z}_p$ has characteristic 0, and contains (an isomorphic copy of) $\mathbb{Z}$ as a subring.

Aren't you glad I asked?

A6.7. Let us recall our definition of a **valuation** on an integral domain D: a **valuation** on D is a map

$$\mu: D\smallsetminus\{0\} \to \mathbb{N}_0 := \mathbb{N}\cup\{0\} = \{0,1,2,3,\ldots,\}$$

satisfying

(a) for all $a, b \in D$ with $b \neq 0$, there exist $q, r \in D$ such that

$$a = qb + r,$$

where either $r = 0$ or $\mu(r) < \mu(b)$; and

(b) for all $a, b \in D\smallsetminus\{0\}$, $\mu(a) \leq \mu(ab)$.

A **Euclidean domain** is an integral domain equipped with a valuation μ.

A6.8. Suppose that we are given a map

$$\nu: D\smallsetminus\{0\} \to \{0,1,2,3,\ldots\}$$

satisfying only the first condition (a), that is: for all $a, b \in D$ with $b \neq 0$, there exist $q, r \in D$ such that

$$a = qb + r,$$

where either $r = 0$ or $\nu(r) < \nu(b)$.

For $0 \neq x \in D$, define $\mu(x) := \min\{\nu(xy) : 0 \neq y \in D\}$. (Note that $\mu(x)$ exists because $\{0,1,2,\ldots\}$ is well-ordered.)

We claim that μ is a valuation on D.

- Clearly $\mu(a) \in \{0,1,2,\ldots\}$ for all $0 \neq a \in D$.

- If $0 \neq a, b \in D$, then
$$\begin{aligned}\mu(a) &:= \min\{\nu(ay) : 0 \neq y \in D\} \\ &\leq \min\{\nu(aby) : 0 \neq y \in D\} \\ &= \mu(ab).\end{aligned}$$
- Now let $a, b \in D$ with $b \neq 0$. Choose $0 \neq y_0 \in D$ such that $\mu(b) = \min\{\nu(by) : 0 \neq y \in D\}$. Since b, y_0 are both non-zero and D is an integral domain, $by_0 \neq 0$.

  Now, choose $q, r \in D$ such that
$$a = q(by_0) + r,$$
  where either $r = 0$ or $\nu(r) < \nu(by_0)$.

  Suppose that $r \neq 0$. Then $\nu(r) < \nu(by_0) = \mu(b)$. Moreover,
$$\mu(r) := \min\{\nu(ry) : 0 \neq y \in D\} \leq \nu(r \cdot 1) = \nu(r),$$
  and so
$$\mu(r) \leq nu(r) < \mu(b).$$
  Thus μ satisfies both conditions from A6.7, and so μ is a valuation on D, as claimed.

Thus if ν satisfies only condition (a) of a valuation, then it induces a valuation μ on D by setting $\mu(x) = \min\{\nu(xy) : 0 \neq y \in D\}$ for each $0 \neq x \in D$. Fascinating.

Exercises for Chapter 6

Exercise 6.1.
Divide $f(x) = 3x^4 + 2x^3 + x + 2$ by $g(x) = x^2 + 4$ in $\mathbb{Z}_5[x]$.

Exercise 6.2.
Prove that $p(x) = x^2 + 2x + 1$ does not lie in the ideal $\langle x^3 + 1 \rangle$ of $\mathbb{Z}_3[x]$.

Exercise 6.3.
Let D be a Euclidean domain, $a, b, c \in D \setminus \{0\}$, and suppose that d is a GCD of a and b. Show that if $a \mid (bc)$, then

$$ad^{-1} \mid c.$$

Exercise 6.4.

(a) Prove **Fermat's Theorem on sums of squares**: Let $p \in \mathbb{N}$ be an odd prime. Then $p = a^2 + b^2$ for some $a, b \in \mathbb{N}$ if and only if $p \equiv 1 \bmod 4$.
(b) Conclude that if $p \in \mathbb{N}$ is prime and $p \equiv 3 \bmod 4$, then p is a prime element in the ring $\mathbb{Z}[i]$ of Gaussian integers.

Exercise 6.5.
Verify that the map $\mu : \mathbb{Z}[i] \setminus \{0\} \to \mathbb{N}$ defined by $\mu(a + bi) = a^2 + b^2$ satisfies

$$\mu((a + bi)(c + di)) = \mu(a + bi)\,\mu(c + di),$$

which was left as an exercise when establishing the fact that μ is a Euclidean valuation on the set $\mathbb{Z}[i]$ of Gaussian integers.

Exercise 6.6.
Let E be a Euclidean domain with Euclidean valuation μ. Prove that

(a) $\mu(1) = \min\{\mu(d) : 0 \neq d \in D\}$, and
(b) $x \in E$ is invertible if and only if $\mu(x) = \mu(1)$.

Exercise 6.7. The Euclidean algorithm
Let E be a Euclidean domain with Euclidean valuation μ. Let $0 \neq a, b \in E$. Choose $q_1, r_1 \in E$ such that

$$a = q_1 b + r_1,$$

where either $r_1 = 0$ or $0 \leq \mu(r_1) < \mu(b)$.

(a) Prove that if $r_1 \neq 0$, then the set of common divisors of a and b is the same as the set of common divisors of b and r_1.

Let $k \geq 1$, and suppose that we have chosen $r_1, r_2, \ldots, r_k$ as above.

If $r_k = 0$, we stop.

If $r_k \neq 0$, choose $q_{k+1}, r_{k+1} \in E$ such that

$$b = q_{k+1} r_k + r_{k+1},$$

where either $r_{k+1} = 0$ or $0 \leq \mu(r_{k+1}) < \mu(r_k)$.

(b) Prove that either $r_1 = 0$, or there exists $m \geq 2$ such that $r_m = 0 \neq r_{m-1}$.

(c) Prove that either

- $r_1 = 0$, in which case b is a greatest common divisor of a and b, or that
- r_{m-1} is a greatest common divisor of a and b, where m is chosen as in part (b).

This process to find the greatest common divisor of two elements of a Euclidean domain is called the **Euclidean algorithm**.

Exercise 6.8.

Let E be a Euclidean domain with Euclidean valuation μ. Prove that if $x \in E$ and x is not invertible, then for any $a \in D$,

$$\mu(a) < \mu(ab).$$

Exercise 6.9.

Let

$$f(x) = x^{10} - 3x^9 + 3x^8 - 11x^7 - 11x^5 + 19x^4 - 13x^3 + 8x^2 - 9x + 3$$

and

$$g(x) = x^6 - 3x^5 + 3x^4 - 9x^3 + 5x^2 - 5x + 2$$

be elements of $E := \mathbb{Q}[x]$. Find the greatest common divisor $d(x)$ of $f(x)$ and $g(x)$.

Exercise 6.10.

Let E be a Euclidean domain with Euclidean valuation μ. In Example 1.3, we saw that if $\nu(d) := 2^{\mu(d)}$, $d \in E$, then ν is again a Euclidean valuation on E.

Observe that $\nu = f \circ \mu$, where $f(x) = 2^x$, $x \geq 0$.

Do there exist other functions $g : [0, \infty) \to [0, \infty)$ such that $\varphi_g := g \circ \mu$ is a Euclidean valuation on E whenever μ is? Can we characterise such functions g? (We do not yet have an answer to this ourselves, although it is entirely possible that such an answer is known.)

Exercise 6.11.

Let D be a PID. Prove that the following conditions are equivalent.

(a) D admits a unique non-zero prime ideal.

(b) D admits a unique non-zero maximal ideal.

CHAPTER 7

Factorisation in polynomial rings

A computer once beat me at chess, but it was no match for me at kick boxing.

Emo Philips

1. Divisibility in polynomial rings over a field

From our work in the previous Chapter, we know that if $\mathbb{F}$ is a field, then $\mathbb{F}[x]$ is a unique factorisation domain. As such, given any non-zero, non-invertible element $0 \neq f(x) \in \mathbb{F}[x]$, we can factor $f(x)$ as a product of irreducible elements in an *essentially unique way*, meaning up to the order of the terms and the replacement of some factors by their associates.

In this Chapter we further investigate what it means for an element of $\mathbb{F}[x]$ to be irreducible. Along the way, we shall more closely investigate the notion of factorisation over more general integral domains, specifically in the setting of $\mathbb{Z}[x] \subseteq \mathbb{Q}[x]$.

1.1. Definition. *Let $\mathbb{F}$ be a field and $0 \neq f(x) \in \mathbb{F}[x]$. We say that $\alpha \in \mathbb{F}$ is a **zero** or a **root** of $f(x)$ if $f(\alpha) = 0$.*

*We define the **multiplicity** of the root α to be*

$$\text{MULT}(\alpha) := \max\{k \in \mathbb{N} : (x-\alpha)^k \mid f(x) \text{ but } (x-\alpha)^{k+1} \nmid f(x)\}.$$

1.2. Example. In $\mathbb{R}[x]$, $\alpha = 3$ is a root of multiplicity 2 for the polynomial $f(x) = x^3 - 7x^2 + 15x - 9$. (Check!)

1.3. Proposition. *Let $\mathbb{F}$ be a field, $0 \neq f(x) \in \mathbb{F}[x]$ and $\alpha \in \mathbb{F}$. Then $f(\alpha) = r_0$, where $r(x) = r_0$ is the remainder of "$f(x)/(x-\alpha)$". More precisely, there exists $q(x) \in \mathbb{F}[x]$ such that $f(x) = q(x)(x-\alpha) + f(\alpha)$.*

Proof. Using the Division Algorithm, we may write $f(x) = q(x)(x-\alpha) + r(x)$, where either $0 \leq \deg r(x) < \deg(x-\alpha) = 1$, or $r(x) = 0$.

In the first case, we deduce that $r(x) = r_0$ is a constant polynomial with $r_0 \neq 0$. Note that in this case,

$$f(\alpha) = q(\alpha)(\alpha - \alpha) + r(\alpha) = q(\alpha)0 + r_0 = r_0.$$

In the second case, $r(x) = r_0 = 0$, and

$$f(\alpha) = q(\alpha)(\alpha - \alpha) = 0 = r_0.$$

□

The following result is an immediate consequence of Proposition 1.3 above.

1.4. Corollary. *Let $\mathbb{F}$ be a field, $0 \neq f(x) \in \mathbb{F}[x]$ and $\alpha \in \mathbb{F}$. The following statements are equivalent:*

(a) *$f(\alpha) = 0$, i.e. α is a root of $f(x)$; and*
(b) *$x - \alpha$ divides $f(x)$.*

1.5. Lemma. *Let $\mathbb{F}$ be a field, $0 \neq f(x) \in \mathbb{F}[x]$, and $\alpha \in \mathbb{F}$. If $k = \text{MULT}(\alpha)$ is the multiplicity of α, then we may write*

$$f(x) = (x - \alpha)^k g(x),$$

where $g(\alpha) \neq 0$ and $\deg(g(x)) = \deg(f(x)) - k$.

Proof. By definition of multiplicity, $(x - \alpha)^k \mid f(x)$, and so we may write $f(x) = (x - \alpha)^k g(x)$ for some $g(x) \in \mathbb{F}[x]$.

Suppose that $g(\alpha) = 0$. Then, by Corollary 1.4, $(x - \alpha) \mid g(x)$, and so we may write $g(x) = (x - \alpha)h(x)$ for some $h(x) \in \mathbb{F}[x]$. But then

$$f(x) = (x - \alpha)^k g(x) = (x - \alpha)^{k+1} h(x),$$

contradicting the fact that $\text{MULT}(\alpha) = k$.

Thus $g(\alpha) \neq 0$.

Finally, by Remark 3.1.8, since every field is an integral domain,

$$\deg(f(x)) = \deg((x - \alpha)^k g(x)) = \deg((x - \alpha)^k) + \deg(g(x)) = k + \deg(g(x)),$$

completing the proof.

□

1.6. Theorem. *Let $\mathbb{F}$ be a field and $0 \neq f(x) \in \mathbb{F}[x]$. The number of roots of $f(x)$, counted according to multiplicity, is at most $\deg f(x)$.*

Proof. We shall argue by induction on the degree of $f(x)$. To that end, let $n := \deg(f(x))$.

If $n = 0$, then $f(x) = a_0 \neq 0$, and so $f(x)$ has no roots; i.e. it has zero roots. This establishes the first step of the induction argument.

Now suppose that $N = \deg(f(x)) \in \mathbb{N}$ and that the result holds for all polynomials $h(x)$ of degree less than N. That is, suppose that if $\deg(h(x)) < N$, then the number of roots of $h(x)$, counted according to multiplicity, is at most $\deg(h(x))$.

If $f(x)$ has no roots, then clearly we are done. Otherwise, let $\alpha \in \mathbb{F}$ be a root of $f(x)$ of multiplicity $k \geq 1$. By Lemma 1.5, we may write

$$f(x) = (x-\alpha)^k g(x)$$

for some $0 \neq g(x) \in \mathbb{F}[x]$, and $g(\alpha) \neq 0$. In particular, this shows that $N = \deg(f(x)) \geq k$, and from above,

$$\deg(g(x)) = N - k < N.$$

By our induction hypothesis, the number of roots of $g(x)$ counted according to multiplicity is at most $\deg(g(x))$. But $\beta \in \mathbb{F}$ is a root of $f(x)$ if and only if $\beta = \alpha$ or β is a root of $g(x)$. (In fact, since $g(\alpha) \neq 0$, these form disjoint sets, but we don't even need this.)

Hence the number of roots of $f(x)$, counted according to multiplicity, is at most

$$k + \deg(g(x)) = k + (N-k) = N = \deg(f(x)),$$

completing the induction step and the proof.

□

1.7. Example. The above theorem fails in general if we consider polynomials with coefficients in commutative rings which are not fields. For example, consider $R = \mathcal{C}([0,1],\mathbb{R})$. We define three functions in R, namely

- $r_0(y) = 0$;
- $r_1(y) = \begin{cases} 0 & 0 \leq y \leq \frac{1}{2} \\ (y-\frac{1}{2})^2 & \frac{1}{2} \leq y \leq 1 \end{cases}$; and
- $r_2(y) = \begin{cases} 0 & 0 \leq y \leq \frac{1}{2} \\ \sin(y-\frac{1}{2}) & \frac{1}{2} \leq y \leq 1 \end{cases}$.

We leave it to the reader to verify that each of these three functions is indeed continuous on $[0,1]$. Consider

$$f(x) = r_2(y)x^2 + r_1(y)x + r_0(y).$$

(Observe that the coefficients just so happen to be functions of $y \in [0,1]$ but this is a polynomial in x with strange coefficients!)

For each $k \in \mathbb{N}$, consider the continuous function

$$\alpha_k(y) := \begin{cases} (y-\frac{1}{2})^k & 0 \leq y \leq \frac{1}{2} \\ 0 & \frac{1}{2} \leq y \leq 1 \end{cases}.$$

(Again - that each of these functions lies in $\mathcal{C}([0,1],\mathbb{R})$ is left as an exercise.)

Note that for each such $k \geq 1$,

$$f(\alpha_k) = r_2(y)\alpha_k^2(y) + r_1(y)\alpha_k(y) + r_0(y) = 0.$$

Thus $f(x)$ has infinitely many roots in $\mathcal{C}([0,1],\mathbb{R})$, despite being a polynomial of degree 2 with coefficients in that ring.

1.8. In the example above, we see that R is a commutative, unital ring, but it is not an integral domain. Indeed, although we shall not require it here, it is not hard to see that Theorem 1.6 extends to polynomials over an integral domain.

Suppose that D is an integral domain, and that $0 \neq f(x) \in D[x]$. As seen in Chapter 5 (more specifically, see Theorem 5.2.7), we can think of D as being a subring of its field $\mathbb{F}$ of quotients, and thus we may also think of $f(x)$ as an element of $\mathbb{F}[x]$. By Theorem 1.6 above, $f(x)$ has at most $\deg(f(x))$ roots in $\mathbb{F}$. But any root of $f(x)$ in D (i.e. any element $\beta \in D$ such that $f(\beta) = 0$) is automatically a root of $f(x)$ in $\mathbb{F}$, and so $f(x)$ has at most $\deg(f(x))$ roots in D.

1.9. Example.

- Let $n \in \mathbb{N}$ and let $f(x) = x^n - 1 \in \mathbb{C}[x]$. If $\omega := e^{2\pi i/n}$, then $\omega, \omega^2, \ldots, \omega^{n-1}, \omega^n = 1$ are n distinct roots of $f(x)$. By Theorem 1.6, these are the *only* roots of $f(x)$ in $\mathbb{C}$.
- Observe that $f(x) = x^4 - 1$ only has two roots in $\mathbb{R}$. Indeed,

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

  Now $g(x) = x^2 - 1$ has two real roots, namely $a_1 = 1$ and $a_2 = -1$, but $h(x) = x^2 + 1$ has no real roots.

Let us recall Corollary 6.3.21.

1.10. Theorem. *Let $\mathbb{F}$ be a field. Then $\mathbb{F}[x]$ is a* UFD. *As such, if $f(x) \in \mathbb{F}[x]$ is a polynomial of degree at least one, then*

(a) *$f(x)$ can be factored in $\mathbb{F}[x]$ as a product of irreducible polynomials, and*
(b) *except for the order of the terms and for non-zero scalar terms (which correspond to the invertible elements of $\mathbb{F}[x]$), the factorisation is unique.*

1.11. Remark. Recall also from Theorem 6.3.10 that an element of $\mathbb{F}[x]$ is prime if and only if it is irreducible. Thus every non-invertible element of $\mathbb{F}[x]$ can be factored as a product of finitely many *prime* elements of $\mathbb{F}[x]$, in the same (essentially) unique way.

1.12. Proposition. *Let $\mathbb{F}$ be a field, and let $q(x), u(x), v(x) \in \mathbb{F}[x]$. Suppose that $q(x)$ is irreducible and that*

$$q(x) \mid u(x)v(x).$$

Then, either $q(x) \mid u(x)$ or $q(x) \mid v(x)$.

Proof. By Remark 1.11, $q(x)$ is prime in $\mathbb{F}[x]$. The result now follows from this.

□

1.13. Example.

(a) The polynomial $f(x) = x^4 - 1 \in \mathbb{R}[x]$ can be factored as a product irreducible polynomials over $\mathbb{R}$ as

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

The significance of item (b) in Theorem 1.10 is that we can also factor $f(x)$ as

$$f(x) = (3(x^2 + 1))(\pi(x + 1))(\frac{1}{3\pi}(x - 1)),$$

and each of these terms is also irreducible.

On the other hand, the three terms in the second factorisation are merely associates of the three terms in the first factorisation, written in a different order.

(b) Note that although $g(x) = x^2 + 1$ is irreducible in $\mathbb{R}[x]$, it is reducible in $\mathbb{C}[x]$. Indeed, there we can factor

$$f(x) = x^4 - 1 = (x - 1)(x + 1)(x - i)(x + i).$$

Those last two factors do not lie in $\mathbb{R}[x]$.

1.14. Culture. Although we shall not prove it here, the field $\mathbb{C}$ has a wonderfully useful property. *Every polynomial $f(x) \in \mathbb{C}[x]$ can be factored into linear terms.* Stated another way, every complex polynomial of degree $n \in \mathbb{N}$ has *exactly* n roots, counted according to multiplicity.

A field that enjoys this property is said to be **algebraically closed**. Thus $\mathbb{C}$ is algebraically closed. Since $f(x) = x^2 + 1$ can not be factored into linear terms in $\mathbb{R}[x]$, $\mathbb{R}$ is *not* algebraically closed.

While it might not seem like much, this is actually the reason why people who study operator theory (a version of linear algebra where the linear maps act on normed vector spaces) tend to use $\mathbb{C}$ as opposed to $\mathbb{R}$ for their base field. The eigenvalues of a matrix $T \in \mathbb{M}_n(\mathbb{R})$ can be found by considering the characteristic polynomial of T. If this polynomial has no real roots, then T has no eigenvalues whatsoever. But for any $Y \in \mathbb{M}_n(\mathbb{C})$, the minimal polynomial always factors into linear terms, and so Y has exactly n eigenvalues, counted according to multiplicity. The usefulness of the eigenvalues of a matrix (and of their generalisation to operators, namely *spectrum*) cannot be overstated.

Hold on, that's a bit hyperbolic. What we mean by that is that eigenvalues are *really important*. That is, they are really important to the study of matrices, and to those disciplines (I'm talking to you, Physics) that rely heavily upon matrix theory. If you are alone in a cage with a starving lion, a new hair-do and your wits, you will generally be excused for not having the eigenvalues of a matrix be one the first things that spring to your mind, and your estimation of their general usefulness will undoubtedly not be of the *impossible to overstate* variety. But if you are invited to a pyjama party at Angela Merkel's house, well, the sky's the limit.

2. Reducibility in polynomial rings over a field

2.1. We wish to factor polynomials in $\mathbb{F}[x]$. The question of whether a polynomial $q(x)$ can be factored is as much a question of the field as of the polynomial, as we shall soon see. In other words – one can not speak of an *irreducible polynomial* – instead, one has to speak of a polynomial being *irreducible over a given field*. Indeed, it is precisely this notion that is at the heart of extension theory for fields, which we shall examine in Chapter 9.

We mention in passing that we shall adopt the popular convention whereby we use the expression $q(x)$ *is irreducible over* $\mathbb{F}$ to mean the same thing as $q(x)$ *is irreducible in* $\mathbb{F}[x]$.

2.2. Recall that if D is an integral domain, then $0 \neq q \in D$ is reducible if and only if we can find two non-invertible elements u and $v \in D$ such that

$$q = uv.$$

Since $D[x]$ is an integral domain whenever D is, it follows that $0 \neq q(x) \in D[x]$ is reducible if and only if there exist non-invertible elements $u(x)$ and $v(x) \in D[x]$ such that

$$q(x) = u(x)v(x).$$

2.3. Remarks.

(a) Every field $\mathbb{F}$ is clearly an integral domain. Moreover, it is a routine exercise (left to the reader) to show that $u(x) \in \mathbb{F}[x]$ is invertible if and only if $u(x) = u_0$ for some $0 \neq u_0 \in \mathbb{F}$. It follows that in this case, $q(x) \in \mathbb{F}[x]$ is reducible if and only if there exist $u(x)$ and $v(x)$ of degree at least one such that

$$q(x) = u(x)v(x).$$

That is, $q(x)$ is reducible if and only if we can factor $q(x) = u(x)v(x)$ with $1 \leq \deg(u(x)), \deg(v(x)) < \deg(q(x))$.

Thus $0 \neq r(x) \in \mathbb{F}[x]$ is irreducible if and only if it can *not be* expressed as a product of two elements of $\mathbb{F}[x]$, each having degree at least one and therefore strictly lower than that of $r(x)$.

(b) This is not the case for polynomials over general integral domains. The polynomial $f(x) = 2x + 6 \in \mathbb{Z}[x]$ is *reducible*, because we may write $f(x) = g(x)h(x)$, where $g(x) = 2$ and $h(x) = x + 3$, neither of which is invertible in $\mathbb{Z}[x]$. Clearly $\deg(g(x)) = 0$ while $\deg(h(x)) = 1$.

2.4. Examples.

(a) For example, the polynomial $f(x) = x^2 - 2$ factors as

$$f(x) = x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2}) \text{ over } \mathbb{R},$$

but it is irreducible over $\mathbb{Q}$. If it were reducible over $\mathbb{Q}$, we could write

$$f(x) = g(x)h(x),$$

where $g(x), h(x) \in \mathbb{Q}[x]$ would necessarily each have degree one. Thus $g(x) = a_1x + a_0$ and $h(x) = b_1x + b_0$, with $a_i, b_i \in \mathbb{Q}$ and $a_1 \neq 0 \neq b_1$. But then

$$f(-a_0/a_1) = g(-a_0/a_1)h(-a_0/a_1) = 0,$$

contradicting the fact that $f(x)$ does not have any rational roots.

Similarly, the polynomial $g(x) = x^2 + 1$ is irreducible as a polynomial in $\mathbb{R}[x]$, but it is reducible over $\mathbb{C}$, since $g(x) = (x - i)(x + i)$ in $\mathbb{C}[x]$.

(b) Using Proposition 2.5 below, it is not hard to see that the polynomial $f(x) = 9x^2 + 3$ is irreducible over $\mathbb{Q}$. On the other hand, the fact that $g(x) = 3$ and $h(x) = 3x^2 + 1$ are both non-invertible in $\mathbb{Z}[x]$ implies that

$$f(x) = g(x)h(x)$$

is in fact *reducible* over $\mathbb{Z}$.

2.5. Proposition. *Let $\mathbb{F}$ be a field and $f(x) \in \mathbb{F}[x]$. Suppose that $\deg(f(x)) \in \{2, 3\}$. The following statements are equivalent.*

(a) *$f(x)$ is reducible in $\mathbb{F}[x]$.*
(b) *$f(x)$ has a root in $\mathbb{F}[x]$.*

Proof.

(a) implies (b). Suppose that $f(x)$ is reducible over $\mathbb{F}$. Then we may write $f(x) = g(x)h(x)$ for some $g(x), h(x) \in \mathbb{F}[x]$. Since $\deg(f(x)) = \deg(g(x)) + \deg(h(x))$ (why?), it follows that one of $g(x)$ and $h(x)$ must have degree equal to 1. By relabelling if necessary, we may assume that it is $g(x)$. But then

$$g(x) = a_1x + a_0 \in \mathbb{F}[x]$$

with $a_1 \neq 0$ has a root in $\mathbb{F}$, namely $\alpha = -a_0/a_1$. Hence α is a root of $f(x)$ in $\mathbb{F}$ as well.

(b) implies (a). Suppose that $\alpha \in \mathbb{F}$ is a root of multiplicity $1 \leq k \leq \deg(f(x))$ for $f(x)$. By Corollary 1.4, $x - \alpha$ divides $f(x)$, and so

$$f(x) = (x - \alpha)g(x)$$

for some $g(x) \in \mathbb{F}[x]$ of degree $\deg(g(x)) = \deg(f(x)) - 1 \in \{1, 2\}$. By Remark 2.3, $f(x)$ is reducible over $\mathbb{F}$.

□

2.6. Examples. This is an especially useful device, especially when working over small fields, where one can try to locate roots by trial and error!

(a) Consider $f(x) = x^3 + 3x + 2 \in \mathbb{Z}_5[x]$. If it were to be reducible over $\mathbb{Z}_5$, it would have to have a root $\alpha \in \mathbb{Z}_5$.

Now

$$f(0) = 2 \quad f(1) = 1 \quad f(2) = 1$$
$$f(3) = 3 \quad f(4) = 3$$

shows that $f(x)$ has no roots in $\mathbb{Z}_5$, whence – by Proposition 2.5 – $f(x)$ is irreducible over $\mathbb{Z}_5$. (Compare this with trying to factor $f(x)$ over $\mathbb{Z}_5$!)

(b) Consider $f(x) = x^3 + 3x + 2 \in \mathbb{Z}_3[x]$. Note that $f(0) = 2$, but $f(1) = 0$ in $\mathbb{Z}_3$. By Proposition 2.5, $f(x)$ is reducible over $\mathbb{Z}_3$, and $(x-1)$ is one of its factors. Indeed,

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

is a factorisation of $f(x)$ into irreducible factors over $\mathbb{Z}_3$. (Why is $h(x) = x-1$ irreducible over $\mathbb{Z}_3$?)

We have already done all of the work necessary to prove the next result. Now we just need to reap the rewards of our efforts.

2.7. Theorem. *Let $\mathbb{F}$ be a field and $p(x) \in \mathbb{F}[x]$. The following statements are equivalent.*

(a) *$p(x)$ is prime.*
(b) *$p(x)$ is irreducible over $\mathbb{F}$.*
(c) *$\langle p(x)\rangle$ is a maximal ideal of $\mathbb{F}[x]$.*
(d) *$\mathbb{F}[x]/\langle p(x)\rangle$ is a field.*

Proof. Since $\mathbb{F}$ is a field, $\mathbb{F}[x]$ is a PID, by Corollary 6.1.12. The result is now a simple application of Theorem 6.3.10 and Theorem 5.1.8 to this setting.

□

2.8. Proposition. *Let $p \in \mathbb{N}$ be a prime number and suppose that $q(x) \in \mathbb{Z}_p[x]$ is an irreducible polynomial of degree n over $\mathbb{Z}_p$. Then*

$$\mathbb{G} := \frac{\mathbb{Z}_p[x]}{\langle q(x)\rangle}$$

is a field with p^n elements.

Proof. By Theorem 2.7, it is clear that $\mathbb{G}$ is a field. The elements of $\mathbb{G}$ are of the form

$$f(x) + \langle q(x)\rangle,$$

where $f(x) \in \mathbb{Z}_p[x]$ is any polynomial. By the division algorithm, we can always write $f(x) = h(x)q(x) + r(x)$, where $r(x) = 0$ or $0 \le \deg(r(x)) < \deg(q(x)) = n$. That is,

$$f(x) + \langle q(x)\rangle = r(x) + \langle q(x)\rangle.$$

Note that

$$r(x) = r_{n-1}x^{n-1} + r_{n-2}x^{n-2} + \cdots + r_1x + r_0$$

for some $r_j \in \mathbb{Z}_p$, $0 \le j \le p$, which means that we have *at most* p^n distinct cosets in $\mathbb{G}$. on the other hand, given any two *distinct* choices of $r(x)$ and $s(x) = s_{n-1}x^{n-1} +$

$s_{n-2}x^{n-2} + \cdots + s_1x + s_0$, the fact that $\deg(s(x) - r(x)) < n$ and $s(x) - r(x) \neq 0$ implies that the cosets are different.

This shows that we have *exactly* p^n cosets in $\mathbb{G}$. In other words, $\mathbb{G}$ is a field with exactly p^n elements, as claimed.

□

2.9. Example. Consider $q(x) = x^3 + 2x + 2 \in \mathbb{Z}_3[x]$. Now $q(0) = q(1) = q(2) = 2 \neq 0$ in $\mathbb{Z}_3$. Since $q(x)$ is a polynomial of degree 3 with no roots in $\mathbb{Z}_3$, we see from Proposition 2.5 above that $q(x)$ is irreducible.

By Proposition 2.8, we conclude that $\mathbb{G} := \mathbb{Z}_3[x]/\langle q(x)\rangle$ is a field with $3^3 = 27$ elements.

The operations on $\mathbb{G}$ are the usual operations in a quotient ring:

$$(r(x) + \langle q(x)\rangle) + (s(x) + \langle q(x)\rangle) = (r(x) + s(x)) + \langle q(x)\rangle,$$

and

$$(r(x) + \langle q(x)\rangle)(s(x) + \langle q(x)\rangle) = (r(x)s(x)) + \langle q(x)\rangle.$$

Of course, if we start with $\deg(r(x)), \deg(s(x)) < 3$, then $\deg(r(x) + s(x)) < 3$, and so $r(x) + s(x)$ is the "canonical" choice of representative for the sum of the two cosets, if we agree that we always want the representatives to have degree less than 3. (This is "nice", but not strictly necessary, mind you.)

As for multiplication, it is not unlikely that $\deg(r(x)s(x)) > 3$, in which case we use the Division Algorithm to write

$$(r(x)s(x)) = t(x)q(x) + w(x),$$

where $w(x) = 0$ or $\deg(w(x)) < \deg(q(x)) = 3$. As before, $w(x) + \langle q(x)\rangle = (r(x)s(x)) + \langle q(x)\rangle$, and so

$$(r(x) + \langle q(x)\rangle)(s(x) + \langle q(x)\rangle) = w(x) + \langle q(x)\rangle.$$

For example,

$$\begin{aligned}((x^2 + 1) + \langle q(x)\rangle) \cdot ((x^2 + x + 2) + \langle q(x)\rangle) &= (x^4 + x^3 + 3x^2 + x + 2) + \langle q(x)\rangle \\ &= (x^4 + x^3 + x + 2) + \langle q(x)\rangle,\end{aligned}$$

since $3 = 0$ in $\mathbb{Z}_3$.

Now, by using the Division Algorithm, we obtain that

$$x^4 + x^3 + x + 2 = (x + 1)(x^3 + 2x + 2) + (-2x^2) = (x)(x^3 + 2x + 2) + (x^2),$$

and so

$$\begin{aligned}((x^2 + 1) + \langle q(x)\rangle) \cdot ((x^2 + x + 2) + \langle q(x)\rangle) &= (x^4 + x^3 + x + 2) + \langle q(x)\rangle \\ &= (x^2) + \langle q(x)\rangle.\end{aligned}$$

3. Factorisation of elements of $\mathbb{Z}[x]$ over $\mathbb{Q}$

3.1. As previously mentioned, the question of whether or not a given polynomial is reducible or irreducible depends to a large extent over which domain we are trying to factor it.

For example, in $\mathbb{Z}[x]$, the polynomial $p(x) = 6x$ is reducible, since we may write $p(x) = q(x)r(x)$, where $q(x) = 2x$ and $r(x) = 3$, neither of which is invertible. The same polynomial, when viewed as an element of $\mathbb{Q}[x]$, is irreducible.

In light of this, it may be surprising to learn that the reducibility of a polynomial $q(x) \in \mathbb{Z}[x]$ over $\mathbb{Q}$ implies its reducibility over $\mathbb{Z}$. We say it *may* be... let's learn it and find out just how surprised we are!

3.2. Definition. *Let $0 \neq q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0 \in \mathbb{Z}[x]$ be a polynomial of degree n. We define the* ***content*** *of $q(x)$ to be*

$$\text{CONTENT}(q(x)) := |\text{GCD}(q_n, q_{n-1}, \ldots, q_1, q_0)|.$$

If $\text{CONTENT}(q(x)) = 1$, we say that $q(x)$ is ***primitive****.*

3.3. Examples.

(a) Let $q(x) = 10x^2 + 4x + 18$. Then $\text{CONTENT}(q(x)) = 2$.
(b) Let $r(x) = 10x^2 + 7x + 18$. Then $\text{CONTENT}(r(x)) = 1$, and $r(x)$ is primitive.
(c) Let $s(x) = -5$. Then $\text{CONTENT}(s(x)) = 5$.

3.4. Note. We remark that may authors define the content of a non-zero polynomial $q(x) \in \mathbb{Z}[x]$ without the absolute value sign. The only difference is that this allows both 2 and -2 to be the content of $q(x)$ in the example above; and we would have to modify our notion of primitive to include the case where

$$\text{GCD}(q_n, q_{n-1}, \ldots, q_1, q_0) = -1.$$

We are simply trying to be pragmatic.

The notion of content can be extended to polynomials over a general integral domain D. In that setting, the notion of an absolute value makes no sense, and we have to agree that the content of a non-zero polynomial over D should be the set of all greatest common divisors of its coefficients.

3.5. The proof of Gauss' Lemma below is extremely clever – almost *too* clever, but it has the advantage that it is relatively easy to follow. It relies on the following simple but effective observation.

Given a polynomial $r(x) = r_n x^n + r_{n-1}x^{n-1} + \cdots + r_1 x + r_0 \in \mathbb{Z}[x]$, and given a prime number $p \in \mathbb{N}$, we shall denote by $\overline{r}(x)$ the polynomial

$$\overline{r}(x) := \overline{r}_n x^n + \overline{r}_{n-1}x^{n-1} + \cdots + \overline{r}_1 x + \overline{r}_0 \in \mathbb{Z}_p[x],$$

where $\overline{r}_j := r_j \text{ MOD } p \in \mathbb{Z}_p$, $0 \leq j \leq n$.

We leave it as an exercise for the reader to prove that the map

$$\begin{array}{rccc} \Delta: & \mathbb{Z}[x] & \to & \mathbb{Z}_p[x] \\ & r(x) & \mapsto & \overline{r}(x) \end{array}$$

is a homomorphism.

3.6. Gauss' Lemma. *Suppose that $0 \neq q(x), r(x) \in \mathbb{Z}[x]$ are primitive. Then $q(x)r(x)$ is also primitive.*

Proof. Suppose that $\gamma := \text{CONTENT}(q(x)r(x))$. Then $\gamma \in \mathbb{N}$, and our goal is to prove that $\gamma = 1$. The "trick" is to realise that if $\gamma \neq 1$, then there exists a prime number p which divides γ, and thus p divides each coefficient of $q(x)r(x)$.

Now the magic happens. Write $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0$ and $r(x) = r_m x^m + r_{m-1}x^{m-1} + \cdots + r_1 x + r_0$, where $n = \deg(q(x))$ and $m = \deg(r(x))$. Then

$$\overline{q}(x) := \overline{q}_n x^n + \overline{q}_{n-1}x^{n-1} + \cdots + \overline{q}_1 x + \overline{q}_0$$

and

$$\overline{r}(x) := \overline{r}_m x^m + \overline{r}_{m-1}x^{m-1} + \cdots + \overline{r}_1 x + \overline{r}_0$$

lie in $\mathbb{Z}_p[x]$, the ring of polynomials over the *field* $\mathbb{Z}_p$, where each $\overline{q}_j := q_j \bmod p \in \mathbb{Z}_p$, $0 \leq j \leq n$ (and a similar statement holds for each $\overline{r}_j$).

The fact that $q(x)$ is primitive, i.e. that $\text{CONTENT}(q(x)) = 1$, implies that not every coefficient of $q(x)$ is divisible by p, and hence $\overline{q}(x) \neq 0$ in $\mathbb{Z}_p[x]$. Similarly, the fact that $r(x)$ is primitive implies that $\overline{r}(x) \neq 0$ in $\mathbb{Z}_p[x]$. Now $\mathbb{Z}_p$ is a field and hence an integral domain, and therefore $\mathbb{Z}_p[x]$ is also an integral domain. But then

$$\overline{qr}(x) = \overline{q}(x)\overline{r}(x) \neq 0$$

in $\mathbb{Z}_p[x]$.

If p had divided γ, then we would have $\overline{qr}(x) = 0$, since each coefficient of $\overline{qr}(x)$ would be zero!

Thus γ is not divisible by any prime number, and so $\gamma = 1$, i.e. $q(x)r(x)$ is primitive.

□

3.7. Remark. Let $\mathbb{F}$ and $\mathbb{E}$ be fields, and suppose that $\mathbb{F} \subseteq \mathbb{E}$. Suppose furthermore that $f(x) \in \mathbb{F}[x]$ is reducible over $\mathbb{F}$. It is then easy to see that $f(x)$ is also reducible over $\mathbb{E}$. For if we can write

$$f(x) = g(x)h(x)$$

where $g(x), h(x) \in \mathbb{F}[x]$ each have degree at least one, then it is clear that $g(x), h(x) \in \mathbb{E}[x]$ as well, proving that $f(x)$ is reducible over $\mathbb{E}$.

If $f(x)$ is *irreducible* over $\mathbb{F}$, then unfortunately no universal conclusions are possible. For example, if we set $p(x) = x^2 + 1$, then $p(x) \in \mathbb{Q}[x]$, $\mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}$, and $p(x)$ is irreducible over both $\mathbb{Q}$ and $\mathbb{R}$, but $p(x) = (x + i)(x - i)$ is reducible over $\mathbb{C}$.

The "problem", so to speak, with $D[x]$ where D is an integral domain (as opposed to a field) is that any non-invertible element $d_0 \in D$ gives rise to a non-invertible polynomial $g(x) = d_0$, and "skews" the issue of whether a polynomial in $D[x]$ is reducible in $\mathbb{F}[x]$ when $\mathbb{F}$ is a field and $D \subseteq \mathbb{F}$.

For example, we recall that $q(x) = 9x^2 + 3 = 3(3x^2 + 1)$ is reducible in $\mathbb{Z}[x]$, but irreducible in $\mathbb{Q}[x]$.

In light of these examples, the next result becomes interesting.

3.8. Theorem. *Let $q(x) \in \mathbb{Z}[x]$. If $q(x)$ is reducible over $\mathbb{Q}$, then $q(x)$ is reducible over $\mathbb{Z}$. Moreover, if*

$$q(x) = g(x)h(x)$$

in $\mathbb{Q}[x]$ with $\deg(g(x)), \deg(h(x)) < \deg(q(x))$, then it is possible to factor

$$q(x) = g_1(x)h_1(x)$$

in $\mathbb{Z}[x]$ with $\deg(g_1(x)) = \deg(g(x))$ and $\deg(h_1(x)) = \deg(h(x))$.

Proof.

Suppose that $q(x)$ is reducible over $\mathbb{Q}$, and choose $g(x), h(x) \in \mathbb{Q}[x]$ such that $q(x) = g(x)h(x)$ with $\deg g(x),\ \deg h(x) \geq 1$. We shall prove that $q(x)$ is reducible over $\mathbb{Z}$ by proving that there exist polynomials $g_0(x), h_0(x) \in \mathbb{Z}[x]$ such that

- $\deg(g_0(x)) = \deg(g(x))$,
- $\deg(h_0(x)) = \deg(h(x))$, and
- $q(x) = g_0(x)h_0(x)$.

Case One. Suppose that $q(x)$ is primitive.

Write $g(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x^1 + a_0$, $h(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0$, where $a_n \neq 0 \neq b_m$.

Set $\kappa = \min\{k \in \mathbb{N} : k\,g(x) \in \mathbb{Z}[x]\}$ and $\lambda := \min\{k \in \mathbb{N} : k\,h(x) \in \mathbb{Z}[x]\}$. (Note that if we write the coefficients of $g(x)$ as rational numbers each of whose numerator and denominator are relatively prime, then κ is the just the absolute value of least common multiple of those denominators. A similar remark holds for $h(x)$.)

Now

$$\kappa\lambda\ q(x) = (\kappa g(x))(\lambda h(x)).$$

Write $\kappa g(x) = \gamma_{\kappa g} g_0(x)$ and $\lambda h(x) = \gamma_{\lambda h} h_0(x)$ where $\gamma_{\kappa g} := \text{CONTENT}(\kappa g(x))$ and $\gamma_{\lambda h} := \text{CONTENT}(\lambda h(x))$. Thus g_0 and h_0 are primitive elements of $\mathbb{Z}[x]$, and

$$\kappa\lambda\ q(x) = \gamma_{\kappa g}\gamma_{\lambda h}(g_0(x)h_0(x)).$$

Since $g_0(x)$ and $h_0(x)$ are primitive, it follows from Gauss's Lemma 3.6 that $g_0(x)h_0(x)$ is also primitive. Hence

$$\begin{aligned}\kappa\lambda &= \text{CONTENT}(\kappa\lambda q(x)) \\ &= \text{CONTENT}(\gamma_{\kappa g}\gamma_{\lambda h}(g_0(x)h_0(x))) \\ &= \gamma_{\kappa g}\gamma_{\lambda h}.\end{aligned}$$

But then (keeping in mind that $\mathbb{Z}$ is an integral domain and so we have cancellation there!)

$$q(x) = g_0(x)h_0(x)$$

factors in $\mathbb{Z}[x]$ as well, since

$$\deg(g_0(x)) = \deg(g(x)) \geq 1 \qquad \text{and} \qquad \deg(h_0(x)) = \deg(h(x)) \geq 1.$$

Case Two. If $q(x)$ is not primitive, then we can write $q(x) = \gamma_q\ q_0(x)$, where $q_0(x) \in \mathbb{Z}[x]$ *is* primitive, and note that

$$q_0(x) = (g_1(x))(h(x)),$$

where $g_1(x) = \gamma_q^{-1} g(x) \in \mathbb{Q}[x]$ has the same degree as $g(x)$. By Case One above, $q_0(x)$ factors over $\mathbb{Z}$, say $q_0(x) = r(x)s(x)$ with $r(x), s(x) \in \mathbb{Z}[x]$ having degree at least one, and thus $q(x) = (\gamma_q r(x))s(x)$ factors over $\mathbb{Z}$ as well.

□

3.9. Corollary. *Suppose that $q(x) = x^n + q_{n-1}x^{n-1} + q_{n-2}x^{n-2} + \cdots + q_1 x + q_0$ is a monic polynomial in $\mathbb{Z}[x]$, and that $q_0 \neq 0$. If $q(x)$ has a root in $\mathbb{Q}$, then it has a root – say α – in $\mathbb{Z}$, and $\alpha \mid q_0$.*

Proof. By Corollary 1.4, the fact that $\beta \in \mathbb{Q}$ is a root of $q(x)$ implies that $x - \beta$ divides $q(x)$ over $\mathbb{Q}$.

Thus we may write $q(x) = (x - \beta)h(x)$ for some $h(x) \in \mathbb{Q}[x]$. As per the first paragraph of the proof of Theorem 3.8 above, we see that we can then find two polynomials $g_0(x), h_0(x)$ in $\mathbb{Z}[x]$ such that

- $\deg(g_0(x)) = \deg(x - \beta) = 1$,
- $\deg(h_0(x)) = \deg(h(x)) = n - 1$, and
- $q(x) = g_0(x)h_0(x)$.

Write $g_0(x) = (a_1 x - a_0)$ and $h_0(x) = b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_1 x + b_0 \in \mathbb{Z}[x]$. Since $q(x)$ is monic, we see that $a_1 b_{n-1} = 1$, and so there is no loss of generality in assuming that $a_1 = 1 = b_{n-1}$. Let $\alpha = a_0$. Then $q(\alpha) = g_0(\alpha)h_0(\alpha) = 0h_0(\alpha) = 0$, and $q_0 = \alpha b_0$ is divisible by α.

□

3.10. Example. Consider $q(x) = x^4 - 2x^2 + 8x + 1 \in \mathbb{Z}[x]$. We wish to determine whether or not $q(x)$ is reducible over $\mathbb{Q}$.

Suppose that $q(x)$ is reducible over $\mathbb{Q}$ (and hence over $\mathbb{Z}$). A moment's thought shows that we can find polynomials $g(x)$ and $h(x) \in \mathbb{Z}[x]$ such that $q(x) = g(x)h(x)$ and either:

- $\deg(g(x)) = 1$, $\deg(h(x)) = 3$, or
- $\deg(g(x)) = 2 = \deg(h(x))$.

CASE ONE. $\deg(g(x)) = 1$, $\deg(h(x)) = 3$.

In this case, by Corollary 3.9 above, $q(x)$ has a root $\alpha \in \mathbb{Z}$ which divides $q_0 = 1$. That is, $\alpha \in \{-1, 1\}$. But

$$q(-1) = 1 - 2 - 8 + 1 = -8 \neq 0 \neq 8 = 1 - 2 + 8 + 1 = q(1).$$

Since $q(x)$ has no roots in $\mathbb{Z}$, it has no roots in $\mathbb{Q}$, and so this decomposition is not possible.

CASE TWO. $\deg(g(x)) = 2 = \deg(h(x))$.

In this case, we can write $g(x) = a_2x^2 + a_1x + a_0$ and $h(x) = b_2x^2 + b_1x + b_0 \in \mathbb{Z}[x]$. Thus

$$\begin{aligned} q(x) &= x^4 - 2x^2 + 8x + 1 \\ &= (a_2x^2 + a_1x + a_0)(b_2x^2 + b_1x + b_0) \\ &= (a_2b_2)x^4 + (a_2b_1 + a_1b_2)x^3 + (a_2b_0 + a_1b_1 + a_0b_2)x^2 + (a_1b_0 + a_0b_1)x + a_0b_0. \end{aligned}$$

Since $a_2b_2 = 1$, we either have $a_2 = 1 = b_2$ or $a_2 = -1 = b_2$. In the second case, we shall simply multiply both $g(x)$ and $h(x)$ by -1; in other words, without loss of generality, we may assume that $a_2 = 1 = b_2$. Thus

$$q(x) = x^4 + (a_1 + b_1)x^3 + (b_0 + a_1b_1 + a_0)x^2 + (a_1b_0 + a_0b_1)x + a_0b_0.$$

Let us now try to solve for a_i, b_i, $0 \leq i \leq 2$. From $a_0b_0 = 1$, we see that either $a_0 = 1 = b_0$, or $a_0 = -1 = b_0$.

- If $a_0 = 1 = b_0$, then by comparing the coefficients of x and x^3 we find that

  $$8 = a_1b_0 + a_0b_1 = a_1 + b_1 = 0,$$

  a contradiction, while
- if $a_0 = -1 = b_0$, then by comparing the coefficients of x and x^3 we find that

  $$8 = a_1b_0 + a_0b_1 = -a_1 - b_1 = 0,$$

  another contradiction.

We conclude that this decomposition is also impossible.

Hence $q(x)$ is irreducible over $\mathbb{Z}$, and thus $q(x)$ is irreducible over $\mathbb{Q}$ as well.

3.11. We next introduce the first of two irreducibility tests. Recall that the map

$$\begin{array}{rccc} \Delta: & \mathbb{Z}[x] & \to & \mathbb{Z}_p[x] \\ & r(x) & \mapsto & \overline{r}(x) \end{array}$$

defined in paragraph 3.5 is a homomorphism.

3.12. The Mod-p Test. *Let $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0 \in \mathbb{Z}[x]$ be a polynomial with $\deg(q(x)) = n \geq 1$. Let $p \in \mathbb{N}$ be a prime number.*

If $\deg(\overline{q}(x)) = \deg(q(x))$ and $\overline{q}(x)$ is irreducible over $\mathbb{Z}_p$, then $q(x)$ is irreducible over $\mathbb{Q}$.

Proof. We argue the contrapositive of our desired statement. Suppose that $q(x)$ is reducible over $\mathbb{Q}$. Since $\mathbb{Q}$ is a field, this means that we can write

$$q(x) = g(x)h(x),$$

where $\max(\deg(g(x)), \deg(h(x))) < n$.

From the comment preceding the statement of this result,

$$\overline{q}(x) = \overline{g}(x)\ \overline{h}(x).$$

Note also that $\deg(\overline{g}(x)) \leq \deg(g(x)) < n$ and $\deg(\overline{h}(x)) \leq \deg(h(x)) < n$.

Thus $\overline{q}(x)$ is reducible in $\mathbb{Z}_p[x]$.

□

3.13. Remarks.

(a) The converse of this statement is false! Consider $q(x) = x^4 + 1$. It can be shown that $q(x)$ is irreducible over $\mathbb{Q}$, but $\overline{q}(x) = x^4 + 1$ is reducible in $\mathbb{Z}_p[x]$ for all primes p.

The proof of this relies on group theory and will be omitted.

(b) It is tempting to ask, and we shall allow ourselves to give into the temptation, whether the irreducibility of $\overline{q}(x)$ over $\mathbb{Z}_p$ in the Mod-p Test implies that $q(x)$ is irreducible not only over $\mathbb{Q}$, but also over $\mathbb{Z}$.

Alas, this fails. Consider the polynomial $p(x) = 2x^2 + 4 \in \mathbb{Z}[x]$. Then $\overline{p}(x) = 2x^2 + 4 \in \mathbb{Z}_5[x]$ is a polynomial of degree 2, and thus is reducible over $\mathbb{Z}_5$ if and only if it has a root in $\mathbb{Z}_5$. But

$$\begin{array}{ll} \overline{p}(0) = 4 & \overline{p}(3) = 2 \\ \overline{p}(1) = 1 & \overline{p}(4) = 1, \\ \overline{p}(2) = 2 & \end{array}$$

showing that $\overline{p}(x)$ has no roots in $\mathbb{Z}_5$, and is therefore irreducible over $\mathbb{Z}_5$. By the Mod-p Test, $p(x)$ is irreducible over $\mathbb{Q}$. On the other hand, we can factor $p(x) = g(x)h(x)$ as a product of non-invertible elements by setting $g(x) = 2$ and $h(x) = x + 2 \in \mathbb{Z}[x]$, proving that $p(x)$ is reducible over $\mathbb{Z}$.

3.14. Eisenstein's Criterion. *Let $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0 \in \mathbb{Z}[x]$ be a polynomial of degree n. Suppose that there exists a prime number $p \in \mathbb{N}$ such that*

(I) $p \nmid q_n$.
(I) $p \mid q_j$ *for* $0 \leq j \leq n-1$, *but*
(I) $p^2 \nmid q_0$.

Then $q(x)$ is irreducible over $\mathbb{Q}$.

Proof.

We shall argue by contradiction. Suppose that $q(x)$ is reducible over $\mathbb{Q}$. By Theorem 3.8, it follows that it is reducible over $\mathbb{Z}$ with factors of degree at least one.

Thus we can find an integer $1 \leq m < n$, and polynomials $g(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots a_1 x + a_0$ and $h(x) = b_{n-m}x^{n-m} + b_{n-m-1}x^{n-m-1} + \cdots b_1 x^1 + b_0 \in \mathbb{Z}[x]$ such that

$$q(x) = g(x)h(x).$$

Then $q_0 = a_0 b_0$. Since $p \mid q_0$ but $p^2 \nmid q_0$, it follows that either

- $p \mid a_0$, $p^2 \nmid a_0$ and $p \nmid b_0$, or
- $p \mid b_0$, $p^2 \nmid b_0$ and $p \nmid a_0$.

By interchanging $g(x)$ and $h(x)$ if necessary, we may assume without loss of generality that the first case holds.

Note that $p \nmid q_n$ implies that $p \nmid a_m$ and $p \nmid b_{n-m}$.

Let $\kappa_0 := \min\{1 \leq k \leq m : p \mid a_{k-1} \text{ but } p \nmid a_k\}$. Observe that there exists $0 \leq j \leq \kappa_0$ such that

$$q_{\kappa_0} = b_0 a_{\kappa_0} + b_1 a_{\kappa_0 - 1} + \cdots + b_j a_{\kappa_0 - j}.$$

By definition of κ_0, we have that $p \mid b_i a_{\kappa_0 - i}$, for $1 \leq i \leq \kappa_0 - j$, but $p \nmid b_0 a_{\kappa_0}$, whence $p \nmid q_{\kappa_0}$. This forces $\kappa_0 = n$, as per the hypotheses of the theorem.

But the definition of κ_0 meant that $\kappa_0 \leq m < n$, a contradiction. This completes the proof.

□

3.15. Examples.

(a) Consider $q(x) = 7x^3 - 6x^2 + 3x + 9 \in \mathbb{Z}[x]$.

Let us set $p = 2$ and apply the Mod-2 Test.

Then $\overline{q}(x) = x^3 - 0x^2 + x + 1 = x^3 + x + 1 \in \mathbb{Z}_2[x]$. This has degree three. If it were to be reducible over $\mathbb{Z}_2$, it would need to have roots in $\mathbb{Z}_2$, by Proposition 2.5.

But $\overline{q}(0) = 0 + 0 + 1 = 1 \neq 0$ and $\overline{q}(1) = 1 + 1 + 1 = 1 \neq 0$ in $\mathbb{Z}_2$, so that $\overline{q}(x)$ is irreducible over $\mathbb{Z}_2$, and therefore $q(x)$ is irreducible over $\mathbb{Q}$ by the Mod-2 Test.

(b) Consider $q(x) = 25x^5 - 9x^4 + 3x^2 - 12 \in \mathbb{Z}[x]$.

Set $p = 3$. Then $p \mid -12$, $p \mid 3$, $p \mid -9$, while $p \nmid 25$ and $p^2 \nmid 12$. By Eisenstein's Criterion, $q(x)$ is irreducible over $\mathbb{Q}$.

(c) Consider $q(x) = 3x^4 + 5x + 1 \in \mathbb{Z}[x]$.

Since the constant term in $q(x)$ is 1, Eisenstein's Criterion will not help us here. Let us try the Mod-p Test. Note that we can't take $p = 3$, because that would reduce the degree of $q(x)$, which we are not allowed to do. Let us try $p = 2$.

In this case $\overline{q}(x) = x^4 + x + 1 \in \mathbb{Z}_2[x]$. If this is reducible over $\mathbb{Z}_2[x]$, then we can find $\overline{r}(x), \overline{s}(x) \in \mathbb{Z}_2[x]$ such that

$$\overline{q}(x) = \overline{r}(x)\overline{s}(x)$$

and $\max(\deg(\overline{r}(x)), \deg(\overline{s}(x))) < 4$. Hence, after relabelling $\overline{r}(x)$ and $\overline{s}(x)$ if necessary, we are led to two possibilities:

(i) $\deg(\overline{r}(x)) = 1$ and $\deg(\overline{s}(x)) = 3$, or
(ii) $\deg(\overline{r}(x)) = 2 = \deg(\overline{s}(x))$.

(The case where $\deg(\overline{s}(x)) = 1$ and $\deg(\overline{r}(x)) = 3$ is the case where we simply relabel $\overline{r}(x)$ and $\overline{s}(x)$.)

If $\deg(\overline{r}(x)) = 1$, then $\overline{r}(x) = \overline{1}x + \overline{r_0}$. But then $\overline{r_0} \in \mathbb{Z}_2$ is a root of $\overline{r}(x)$, meaning that it is a root of $\overline{q}(x)$. Since $\overline{q}(0) = 1 = \overline{q}(1)$, we see that this is impossible.

Thus we must consider the case where $\overline{r}(x) = x^2 + \overline{r_1}x + \overline{r_0}$ and $\overline{s}(x) = x^2 + \overline{s_1} + \overline{s_0}$. (The fact that these both have degree two means that the coefficients of x^2 are both $\overline{1}$, which we have suppressed in the notation above.)

Thus

$$\begin{aligned}
\overline{q}(x) &= x^4 + x + 1 \\
&= \overline{r}(x)\overline{s}(x) \\
&= (x^2 + \overline{r_1}x + \overline{r_0})(x^2 + \overline{s_1}x + \overline{s_0}) \\
&= x^4 + (\overline{r_1} + \overline{s_1})x^3 + (\overline{r_0} + \overline{r_1 s_1} + \overline{s_0})x^2 + (\overline{r_1 s_0} + \overline{r_0 s_1})x + \overline{r_0 s_0}.
\end{aligned}$$

Now we try to solve for the r_i's and s_i's. Since $\overline{r_0 s_0} = 1$, we must have $\overline{r_0} = 1 = \overline{s_0}$. Hence the coefficient of x is $1 = \overline{r_1} + \overline{s_1}$, while the coefficient of x^3 is $0 = \overline{r_1} + \overline{s_1}$, a contradiction.

Thus $\overline{q}(x)$ is irreducible over $\mathbb{Z}_2$, and so by the Mod-2 Test, $q(x)$ is irreducible over $\mathbb{Q}$.

Supplementary Examples.

S7.1. Example. Let $p \in \mathbb{N}$ be a prime number. The p^{th} **cyclotomic polynomial**

$$\Phi_p(x) := \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible over $\mathbb{Q}$.

Proof. As we have seen in Theorem 3.8, if $\Phi_p(x)$ is reducible over $\mathbb{Q}$, then $\Phi_p(x)$ is reducible over $\mathbb{Z}$.

Now,

$$\begin{aligned} g(x) := \Phi_p(x+1) &= \frac{(x+1)^p - 1}{(x+1) - 1} \\ &= \frac{x^p + \binom{p}{1}x^{p-1} + \cdots + \binom{p}{p-1}x^1 + 1^p - 1}{x} \\ &= x^{p-1} + \binom{p}{1}x^{p-2} + \cdots + \binom{p}{p-2}x + \binom{p}{p-1} \\ &= x^{p-1} + \binom{p}{1}x^{p-2} + \cdots + \binom{p}{p-2}x + p. \end{aligned}$$

Astoundingly, this polynomial $g(x)$ satisfies Eisenstein's Criterion for p, and thus $g(x)$ is irreducible over $\mathbb{Z}$.

Suppose now that

$$\Phi_p(x) = g(x)h(x) \in \mathbb{Z}[x]$$

is a non-trivial factorisation (i.e. $\max(\deg(g(x)), \deg(h(x))) < p - 1$). Then

$$g(x) = \Phi_p(x+1) = g(x+1)h(x+1)$$

is also a non-trivial factorisation (*check!*), contradicting the irreducibility of $g(x)$.

□

S7.2. Example. Let $p(x) = x^2 + 1 \in \mathbb{R}[x]$. Note that $p(x)$ is irreducible since $\deg p(x) = 2$ and $p(x)$ has no roots in $\mathbb{R}$; i.e. $p(x) \geq 1 > 0$ for all $x \in \mathbb{R}$.

Note, however that the polynomial $q(x) = x^4 + 2x^2 + 1 = (x^2 + 1)(x^2 + 1)$ is reducible, despite the fact that it has no roots over $\mathbb{R}$.

S7.3. Example. One of the most beautiful results concerning the field of complex numbers is that it is **algebraically closed**. This is the statement that if $p(x) \in \mathbb{C}[x]$ is a non-constant polynomial, then $p(x)$ admits a root in $\mathbb{C}$, and is usually referred to as the **Fundamental theorem of algebra**.

The standard, and the most accessible proof of this derives from the theory of complex variables, and is beyond the scope of this course. It is remarkable, to say the least, that such a "purely algebraic" concept can be proven using very "analytic techniques". This is why mathematics is better than chocolate, and almost as good as potato chips. Of course, here we are not being so impudent as to compare

mathematics to bacon and hickory-flavoured potato chips, which are in a league of their own.

As an immediate consequence – of the Fundamental Theorem of Algebra, and not of the sublime and incomparable exquisiteness of bacon and hickory-flavoured potato chips – if $p(x) \in \mathbb{C}[x]$ is a polynomial of degree $n \geq 1$, there exist $\alpha_1, \alpha_2, \ldots, \alpha_n$ (not necessarily distinct) and $\beta \in \mathbb{C}$ such that

$$p(x) = \beta(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n).$$

S7.4. Example. This has the following remarkable consequence. Suppose that $p(x) \in \mathbb{R}[x]$ is a monic polynomial of degree $n \geq 1$. From the Fundamental Theorem of algebra, thinking of $p(x)$ as a polynomial in $\mathbb{C}[x]$, we can find n elements $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{C}$ such that

$$p(x) = (x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n).$$

Since the coefficients of $p(x)$ are all real numbers, we see that $\overline{p}(x) = p(x)$. Thus

$$p(x) = \overline{p}(x) = (x - \overline{\alpha_1})(x - \overline{\alpha_2})\cdots(x - \overline{\alpha_n}).$$

In other words, $\alpha \in \mathbb{C}$ is a root of $p(x)$ if and only if $\overline{\alpha}$ is a root of $p(x)$.

Suppose that $p(x)$ is irreducible over $\mathbb{R}$. If $\deg(p(x)) \neq 1$, then from above, we may find a complex root $\alpha \in \mathbb{C}$ for $p(x)$ and its complex conjugate $\overline{\alpha}$ is also a root for $p(x)$. Thus

$$p(x) = (x - \alpha)(x - \overline{\alpha})q(x) = (x^2 + 2\mathrm{Re}(\alpha) + |\alpha|^2)q(x).$$

Note that $r(x) := x^2 + \mathrm{Re}(\alpha) + |\alpha|^2 \in \mathbb{R}[x]$. Thus

$$\overline{p}(x) = p(x) = r(x)q(x) = \overline{r}(x)\overline{q}(x) = r(x)\overline{q}(x).$$

Since $\mathbb{R}[x]$ is an integral domain (by virtue of the fact that $\mathbb{R}$ is a field), we see that $q(x) = \overline{q}(x)$, implying that $q(x) \in \mathbb{R}[x]$ as well.

Next, since $r(x)$ is a polynomial of degree 2, the irreducibility of $p(x)$ over $\mathbb{R}$ implies that $q(x)$ must be invertible in $\mathbb{R}[x]$, and thus must have degree zero. This shows that the only irreducible polynomials over $\mathbb{R}$ have degree 1 or 2.

The contrapositive of this statement says that if $\deg(p(x)) \geq 3$, then $p(x) \in \mathbb{R}[x]$ is reducible.

S7.5. Example. Consider $p(x) = \frac{3}{11}x^3 + \frac{5}{11}x + \frac{1}{11} \in \mathbb{Q}[x]$. Then $q(x) := 11p(x) = 3x^3 + 5x + 1 \in \mathbb{Z}[x]$ is primitive. Let's apply the Mod-2 Test.

$$\overline{q}(x) = x^3 + x + 1 \in \mathbb{Z}_2[x]$$

is a polynomial of degree 3 with no roots in $\mathbb{Z}_2$, and thus is irreducible there. By the Mod-2 Test, $q(x)$ is irreducible over $\mathbb{Q}$, and thus $p(x)$ is irreducible over $\mathbb{Q}$ as well, being an associate of $q(x)$.

S7.6. Example. Consider the polynomial $p(x) = x^3 + 5 \in \mathbb{Q}[x]$. The Mod-2 Test (resp. the Mod-3 Test) fails, since $\overline{p}(x)$ has a root in $\mathbb{Z}_2$ (resp. in $\mathbb{Z}_3$). That it fails does not mean that $p(x)$ is irreducible - it simply means that the Test doesn't tell us one way or another.

Note that the Mod-5 Test also fails, since $\overline{p}(x) = x^3 \in \mathbb{Z}_5[x]$ has $x = 0$ as a root of multiplicity 3.

On the other hand, with $p = 7$, we find that $\overline{p}(x) = x^3 + 5$ has *no* roots in $\mathbb{Z}_7$ (check!), and so $\overline{p}(x)$ is irreducible in $\mathbb{Z}_7[x]$ (it has degree 3 and no roots). By the Mod-7 Test, $p(x) \in \mathbb{Q}[x]$ is irreducible.

S7.7. Example. Let's look at the polynomial $p(x) = x^3 + 5 \in \mathbb{Q}[x]$ once again. Since the coefficients are integers, we know that it is reducible over $\mathbb{Q}$ only if it is reducible over $\mathbb{Z}$, and furthermore, we can choose the factors of $p(x)$ in $\mathbb{Z}[x]$ to have the same degrees as those in $\mathbb{Q}[x]$. Since $\deg(p(x)) = 3$, this can only happen if one (or more) of the irreducible factors has degree 1. Furthermore, the fact that the leading coefficient of $p(x)$ is equal to 1 means that the leading coefficients of the factors must either both be 1 or -1. Without loss of generality, we may assume that they are both one, for otherwise we simply multiply each factor by -1.

Thus there exists $\alpha \in \mathbb{Z}$ such that $q(x) = (x - \alpha) \mid p(x)$. Equivalently,

$$p(x) = (x - \alpha)r(x)$$

for some $r(x) = x^2 + r_1 x + r_0 \in \mathbb{Z}[x]$. In particular, $p(x)$ has a root in $\mathbb{Z}$! Thus $5 = \alpha r_0$, where $\alpha, r_0 \in \mathbb{Z}$. This forces $\alpha, r_0 \in \{-5, -1, 1, 5\}$. But $p(-5) = -120$, $p(-1) = 4$, $p(1) = 6$ and $p(5) = 130$. Since none of these are roots of $p(x)$, $p(x)$ must be irreducible over $\mathbb{Z}$, and hence irreducible over $\mathbb{Q}$.

Fortunately, we have arrived to the same conclusion as before. Some might say that that was to be expected, but perhaps we can view this as an opportunity for us to show a bit more humility and learn to feel less entitled.

S7.8. Example. Let $q(x) = x^3 + 9x^2 + 12x + 6 \in \mathbb{Z}[x]$. Let $p = 3 \in \mathbb{N}$, so that p is prime. Clearly $p \nmid 1$, while $p \mid 9$, $p \mid 12$, $p \mid 6$ but $p^2 = 9 \nmid 6$.

By Eisenstein's Criterion, $q(x)$ is irreducible over $\mathbb{Z}[x]$.

S7.9. Example. Let $p(x) = x^3 + 5 \in \mathbb{Q}[x]$ yet again. If $r = 5$, then $r \nmid 1$, $5 \mid 0$, $5 \mid 0$ and $5 \mid 5$ but $5^2 \nmid 5$, so again, by Eisenstein's Criterion, $p(x)$ is irreducible over $\mathbb{Q}$.

The reader may or may not agree that this evokes certain memories of cats and taxidermists.

S7.10. Example. Suppose that $\mathbb{F}$ is a field and that $p(x), q(x) \in \mathbb{F}[x]$ satisfy $\text{GCD}(p(x), q(x)) = 1$. (We say that $p(x)$ and $g(x)$ are **relatively prime**.)

If $r(x) \in \mathbb{F}[x]$ and $\deg(r(x)) < \deg(p(x)) + \deg(q(x))$, then there exist polynomials $f(x)$ and $g(x) \in \mathbb{F}[x]$

$$\frac{r(x)}{p(x)q(x)} = \frac{f(x)}{p(x)} + \frac{g(x)}{q(x)}.$$

Indeed, by Euclid's algorithm, we know that we may write

$$1 = \text{GCD}(p(x), q(x)) = t(x)p(x) + s(x)q(x)$$

for some $t(x), s(x) \in \mathbb{F}[x]$. Thus

$$\frac{r(x)}{p(x)q(x)} = \frac{r(x)t(x)p(x)}{p(x)q(x)} + \frac{r(x)s(x)q(x)}{p(x)q(x)} = \frac{r(x)t(x)}{q(x)} + \frac{r(x)s(x)}{p(x)}.$$

What extra conditions would we need to impose upon $f(x)$ and $g(x)$ to guarantee that they are unique?

Appendix

A7.1. As we have seen, every Euclidean domain is a Principle Ideal Domain PID, and every Principle Ideal Domain is a Unique Factorisation Domain UFD. I've got this friend (whose identity there is absolutely no good reason to reveal) who remembers this using the mnemonic ED PIDUFFED. This friend tells me that he's remembered this result for over 40 years using this mnemonic. Well, you don't get to choose your friends, I suppose.

A7.2. The proof of the Fundamental Theorem of Algebra using Complex Analysis relies on a result known as **Liouville's Theorem** which states that if $f : \mathbb{C} \to \mathbb{C}$ is a holomorphic, then f cannot be bounded. That is, there does not exist $0 < M \in \mathbb{R}$ such that

$$|f(z)| \leq M \text{ for all } z \in \mathbb{C}.$$

Suppose that $p(z) = p_0 + p_1 z + \cdots + p_n z^n \in \mathbb{C}[x]$ is a polynomial of degree $n \geq 2$ which has no roots in $\mathbb{C}$. Then $p(z)$ is holomorphic on $\mathbb{C}$ and so $h(z) := \frac{1}{p(z)}$ is also holomorphic on $\mathbb{C}$ (because $p(z) \neq 0$ for all $z \in \mathbb{C}$).

The trick is to show that $h(z)$ is bounded. To do this, one first finds $0 < R \in \mathbb{C}$ such that

$$M_1 := \sup\{|h(z)| : |z| > R\} < \infty,$$

and then uses the so-called **compactness** of $\{z \in \mathbb{C} : |z| \leq R\}$ to argue that

$$M_2 := \sup\{|h(z)| : |z| \leq R\} < \infty.$$

Finally, setting $M := \max(M_1, M_2)$, we find that

$$|h(z)| \leq M \text{ for all } z \in \mathbb{C},$$

a contradiction of Liouville's Theorem.

This is helpful to know only insofar as one is more likely to believe Liouville's result without proof more than the Fundamental Theorem of Algebra, or one actually knows Liouville's Theorem.

Exercises for Chapter 7

Exercise 7.1.
Prove that $p(x) = x^2 - 2$ is irreducible over $\mathbb{Q}$.

Exercise 7.2.
Prove that $p(x) = x^3 + x^2 + 2$ is irreducible over $\mathbb{Q}$.

Exercise 7.3.
Let $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0 \in \mathbb{Q}[x]$ be a polynomial of degree $n \geq 1$. Prove that there exist integers, $a, b \in \mathbb{Z}$ and $p_0, p_1, \ldots, p_n \in \mathbb{Z}$ with $\text{GCD}(p_0, p_1, \ldots, p_n) = 1$ such that

$$q(x) = \frac{q}{b}(p_n x^n + p_{n-1}x^{n-1} + \cdots + p_1 x + p_0).$$

Exercise 7.4.
Let $p(x) = x^4 - 2x^3 + x + 1$. Prove that $p(x)$ is irreducible over $\mathbb{Q}$.

Exercise 7.5.
Prove that $p(x) = 16x^5 - 9x^4 + 3x^2 + 6x - 21 \in \mathbb{Q}[x]$. Prove that $p(x)$ is irreducible over $\mathbb{Q}$.

Exercise 7.6.
Let $\mathbb{F}$ be a field. Prove that $\mathbb{F}[x, y]$ is not a PID.

Exercise 7.7.
Apply the division algorithm to $f(x) = 6x^4 - 2x^3 + x^2 - 3x + 1$ and $g(x) = x^2 + x - 2 \in \mathbb{Z}_7[x]$ to find $f(x)/g(x)$.

Exercise 7.8.
Find all of the roots of $p(x) = 5x^3 + 4x^2 - x + 9$ in $\mathbb{Z}_{12}$.

Exercise 7.9.
Find an invertible element $p(x)$ of $\mathbb{Z}_4[x]$ such that $\deg(p(x)) > 1$.

Exercise 7.10.
Give two different factorisations of $p(x) = x^2 + x + 8$ in $\mathbb{Z}_{10}[x]$.

CHAPTER 8

Vector spaces

There are so many different kinds of stupidity, and cleverness is one of the worst.

Thomas Mann

1. Definition and basic properties

1.1. Most of us will have seen vector spaces (over $\mathbb{R}$) in a first course in linear algebra. Vector spaces – at least finite-dimensional vector spaces over a general field – are in many ways similar to the familiar n-dimensional $\mathbb{R}^n$ as a vector space over $\mathbb{R}$.

Our goal in this Chapter is not to fully develop the theory of vector spaces. Rather, we concentrate only on those results we shall need to discuss field extensions in the next chapter. As such, we shall demonstrate that every vector space admits a basis, and in the case where the vector space $\mathcal{V}$ admits a *finite* basis over a field $\mathbb{F}$, we shall then prove that every basis for $\mathcal{V}$ over $\mathbb{F}$ has the same number of elements, allowing us to define the *dimension* of that space. (A more general result asserts that for *any* vector space $\mathcal{V}$ over $\mathbb{F}$, given any two bases $\mathfrak{B}$ and $\mathfrak{C}$ for $\mathcal{V}$, there exists a bijection from $\mathfrak{B}$ to $\mathfrak{C}$. We shall not require that here.)

1.2. Definition. *Let $\mathbb{F}$ be a field. A **vector space** $\mathcal{V}$ over $\mathbb{F}$ is a non-empty set whose elements are called **vectors**. The set $\mathcal{V}$ is equipped with two operations, namely **addition** (denoted by $+$) and **scalar multiplication** (denoted by $\cdot$ or simple juxtaposition) which satisfy the following:*

(a) $(\mathcal{V}, +)$ *is an abelian group.*
- *Given $x, y \in \mathcal{V}$, $x + y \in \mathcal{V}$.*
- *Given $x, y, z \in \mathcal{V}$, $(x + y) + z = x + (y + z)$.*
- *There exists a neutral element $0 \in \mathcal{V}$ such that $0 + x = x = x + 0$ for all $x \in \mathcal{V}$.*
- *Given $x \in \mathcal{V}$, there exists $y \in \mathcal{V}$ such that $x + y = 0 = y + x$. (As it is unique, we write $-x$ for y.)*
- *$x + y = y + x$ for all $x, y \in \mathcal{V}$.*

(b) *Scalar multiplication satisfies the following.*
- *Given $\kappa \in \mathbb{F}$, $x \in \mathcal{V}$, we have that $\kappa x \in \mathcal{V}$.*

- $1x = x$ *for all* $x \in \mathcal{V}$.
- $\kappa(x+y) = \kappa x + \kappa y$ *for all* $\kappa \in \mathbb{F}$, $x, y \in \mathcal{V}$.
- $(\kappa+\lambda)x = \kappa x + \lambda x$ *for all* $\kappa, \lambda \in \mathbb{F}$, $x \in \mathcal{V}$.
- $\kappa(\lambda x) = (\kappa\lambda)x$ *for all* $\kappa, \lambda \in \mathbb{F}$, $x \in \mathcal{V}$.

1.3. Examples.

(a) Let $n \in \mathbb{N}$ be an integer. If $p \in \mathbb{N}$ is a prime, then $\mathbb{Z}_p^n := \{(x_1, x_2, \ldots, x_n) : x_j \in \mathbb{Z}_p, 1 \le j \le n\}$ forms a vector space over $\mathbb{Z}_p$ using the familiar operations

$$(x_1, x_2, \ldots, x_n) + (y_1, y_2, \ldots, y_n) := (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$$

and

$$\kappa(x_1, x_2, \ldots, x_n) := (\kappa x_1, \kappa x_2, \ldots, \kappa x_n)$$

for all $(x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n) \in \mathbb{Z}_p^n$ and $\kappa \in \mathbb{Z}_p$.

(b) Let $\mathcal{V} = \mathcal{C}([0,1], \mathbb{C}) := \{f : [0,1] \to \mathbb{C} : f \text{ is continuous}\}$. Then $\mathcal{C}([0,1], \mathbb{C})$ is a vector space over $\mathbb{C}$ using the operations

$$(f+g)(x) := f(x) + g(x),$$

and

$$(\kappa f)(x) = \kappa(f(x))$$

for all $f, g \in \mathcal{C}([0,1], \mathbb{C})$ and $\kappa \in \mathbb{C}$.

(c) If $\mathcal{V}$ is a vector space over $\mathbb{C}$, then $\mathcal{V}$ is a vector space over $\mathbb{R}$ and over $\mathbb{Q}$, using the same operations, only restricting the scalar multiplication to scalars in $\mathbb{R}$ or to scalars in $\mathbb{Q}$ respectively.

1.4. Definition. *A* ***subspace*** $\mathcal{W}$ *of a vector space* $\mathcal{V}$ *over a field* $\mathbb{F}$ *is a non-empty subset of* $\mathcal{V}$ *which is also a vector space over* $\mathbb{F}$.

We write $\mathcal{W} \le \mathcal{V}$ *to indicate that* $\mathcal{W}$ *is a subspace of* $\mathcal{V}$.

The familiar **Subspace Test** for subspaces of $\mathbb{R}^n$ from linear algebra carries over to a completely general setting, and we leave the proof to the reader.

1.5. The Subspace Test. A *non-empty* subset $\mathcal{W}$ of a vector space $\mathcal{V}$ over a field $\mathbb{F}$ is a subspace of $\mathcal{V}$ if and only if

(a) $x, y \in \mathcal{W}$ implies that $x + y \in \mathcal{W}$, and
(b) $x \in \mathcal{W}$ and $\kappa \in \mathbb{F}$ implies that $\kappa x \in \mathcal{W}$.

Alternatively, we may replace these two conditions by the single condition that $\kappa x + y \in \mathcal{W}$ for all $\kappa \in \mathbb{F}$ and $x, y \in \mathcal{W}$.

1.6. Notation. Given a subset X of a vector space $\mathcal{V}$ over a field $\mathbb{F}$, we denote by

$$\langle X \rangle := \cap\{\mathcal{W} : \mathcal{W} \le \mathcal{V} \text{ and } X \subseteq \mathcal{W}\}.$$

This defines the smallest subspace (relative to the partial order of inclusion) of $\mathcal{V}$ that contains X.

1.7. Definition. *Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$. A subset $X \subseteq \mathcal{V}$ is said to be* ***linearly dependent*** *if there exist $x_1, x_2, \ldots, x_m \in X$ and $\kappa_1, \kappa_2, \ldots, \kappa_m \in \mathbb{F} \setminus \{0\}$ such that*

$$\kappa_1 x_1 + \kappa_2 x_2 + \cdots + \kappa_m x_m = 0.$$

Otherwise, we say that X is ***linearly independent****.*

1.8. Example. The set $\{\sin x, \cos x, \sin(2x), \cos(2x)\}$ is linearly independent in the vector space $\mathcal{C}((-1,1), \mathbb{R})$ over $\mathbb{R}$.

Consider

$$0 = \kappa_1 \sin x + \kappa_2 \cos x + \kappa_3 \sin(2x) + \kappa_4 \cos(2x).$$

By differentiating three times we obtain the three new equations

$$0 = \kappa_1 \sin x + \kappa_2 \cos x + \kappa_3 \sin(2x) + \kappa_4 \cos(2x) \tag{1}$$

$$0 = \kappa_1 \cos x - \kappa_2 \sin x + 2\kappa_3 \cos(2x) - 2\kappa_4 \sin(2x) \tag{2}$$

$$0 = -\kappa_1 \sin x - \kappa_2 \cos x - 4\kappa_3 \sin(2x) - 4\kappa_4 \cos(2x) \tag{3}$$

$$0 = -\kappa_1 \cos x + \kappa_2 \sin x - 8\kappa_3 \cos(2x) + 8\kappa_4 \sin(2x). \tag{4}$$

These equations must hold for all $x \in (-1,1)$, and so setting $x = 0$, we obtain that

$$0 = \kappa_2 + \kappa_4 \tag{5}$$

$$0 = \kappa_1 + 2\kappa_3 \tag{6}$$

$$0 = -\kappa_2 - 4\kappa_4 \tag{7}$$

$$0 = -\kappa_1 - 8\kappa_3. \tag{8}$$

An easy calculation now shows that $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4 = 0$, and thus that the set $\{\sin x, \cos x, \sin(2x), \cos(2x)\}$ is indeed linearly independent.

The student who is not already familiar with partially and totally ordered sets should consult the Appendix to Chapter 5.

1.9. Definition. *Let $(X, \leq)$ be a partially ordered set. A* ***chain*** *in X is a non-empty subset $\mathcal{C} \subseteq X$ which is totally ordered. Thus, given $c_1, c_2 \in \mathcal{C}$, either $c_1 \leq c_2$ or $c_2 \leq c_1$.*

Let $\varnothing \neq Y \subseteq X$. An ***upper bound in X for Y*** *is an element $u \in X$ such that $y \leq u$ for all $y \in Y$.*

1.10. Example. Let $X := \{a, b, c, d\}$, and partially order the power set $\mathcal{P}(X) := \{Y : Y \subseteq X\}$ by inclusion: i.e. $Y_1 \leq Y_2$ if and only if $Y_1 \subseteq Y_2$.

Then

$$\mathcal{C} := \{\varnothing, \{a\}, \{a, c\}, \{a, c, d\}\}$$

is a chain in $\mathcal{P}(X)$, while

$$\mathcal{D} := \{\varnothing, \{a\}, \{b\}, \{a, b\}\}$$

is not a chain in $\mathcal{P}(X)$, since $\{a\}$ and $\{b\}$ are not comparable. That is, $\{a\} \nsubseteq \{b\}$ and $\{b\} \nsubseteq \{a\}$.

1.11. Definition. *Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$. Let*

$$\mathfrak{L} := \{L \subseteq \mathcal{V} : L \text{ is linearly independent}\}.$$

We partially order $\mathfrak{L}$ by inclusion, so that $L_1 \leq L_2$ if and only if $L_1 \subseteq L_2$.

*A **basis** for $\mathcal{V}$ is a maximal element $\mathcal{B}$ of $\mathfrak{L}$.*

The proof that every vector space over a field $\mathbb{F}$ admits a basis is much deeper than it may sound. It is known to be equivalent to the **Axiom of Choice**, and to **Zorn's Lemma**, which we now state. We shall go into further detail on the Axiom of Choice in Appendix A. In fact, we shall now prove "half" of the equivalence by using Zorn's Lemma to prove that every vector space admits a basis.

1.12. Zorn's Lemma. *Let $(X, \leq)$ be a non-empty poset. Suppose that every chain $\mathcal{C}$ in X has an upper bound in X. Then X admits a maximal element.*

1.13. Theorem. *Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$. Then $\mathcal{V}$ admits a basis.*

Proof. Since a basis is by definition a *maximal* (in terms of inclusion) linearly independent set, this should point one in the direction of how to apply Zorn's Lemma. It suggests that we should first partially order all linearly independent subsets of $\mathcal{V}$. We shall also need to show that the collection of such sets is non-empty, and then we shall have to consider chains of linearly independent sets.

Notice that we have required very little creativity to come up with this plan of attack. The only *inspiration* required was to notice that Zorn's Lemma is a tool used to produce maximal elements in certain partially ordered sets. Having formulated our strategy, let's begin.

Let

$$\mathfrak{L} := \{L \subseteq \mathcal{V} : L \text{ is linearly independent}\}.$$

Given $L_1, L_2 \in \mathfrak{L}$, set $L_1 \leq L_2$ if and only if $L_1 \subseteq L_2$. In other words, we partially order $\mathfrak{L}$ by inclusion.

First we need to know that $\mathfrak{L} \neq \varnothing$. We are inspired by the case where $\mathcal{W} := \{0\}$. We know that $\{0\}$ is always a linearly *dependent* set (why?), and yet the above statement assures us that $\mathcal{W}$ admits a basis. The only possibility is to take that basis to be the empty set.

Now in general, is $\varnothing \in \mathfrak{L}$? That is, is $\varnothing \subseteq \mathcal{V}$ linearly independent? To answer this, one should perhaps ask whether or not $\varnothing$ is linearly *dependent*!

Since we can not find $x_1, x_2, \ldots, x_n \in \varnothing$ in the first place, $\varnothing$ can *not* be linearly dependent, and thus it is linearly independent. But then $\varnothing \in \mathfrak{L}$, and so $\mathfrak{L} \neq \varnothing$!

The rest of the proof is more straightforward. Let

$$\mathcal{C} = \{L_\lambda\}_{\lambda \in \Lambda}$$

be a chain in $\mathfrak{L}$. Thus each $L_\lambda \in \mathfrak{L}$. Define $L := \cup_{\lambda\in\Lambda} L_\lambda$.

We claim that L is linearly independent. Suppose that $x_1, x_2, \ldots, x_n \in L$. Then there exist $L_{\lambda_1}, L_{\lambda_2}, \ldots, L_{\lambda_n}$ such that $x_j \in L_{\lambda_j}$, $1 \le j \le n$. But $\mathcal{C}$ is a chain, so one of these L_{λ_j}'s must include all of the others, say $L_{\lambda_{j_0}}$. Hence

$$x_j \in L_{\lambda_j} \subseteq L_{\lambda_{j_0}}, 1 \le j \le n.$$

Since $L_{\lambda_{j_0}} \in \mathfrak{L}$, it is linearly independent, whence $\{x_1, x_2, \ldots, x_n\}$ is also linearly independent (why?).

This shows that L is linearly independent, and so $L \in \mathfrak{L}$ and clearly L is an upper bound for $\mathcal{C}$, since $L_\lambda \subseteq L$ for all $\lambda \in \Lambda$.

By Zorn's Lemma, $\mathfrak{L}$ has a maximal element, and this is precisely the definition of a basis for $\mathcal{V}$.

□

1.14. Exercise. Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$, and let $L \subseteq \mathcal{V}$ be a linearly independent set. Then there exists a basis $\mathcal{B}$ for $\mathcal{V}$ such that $L \subseteq \mathcal{B}$.

1.15. Theorem. *Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$, and let $\mathcal{B}$ be a basis for $\mathcal{V}$. Then*

(a) *each element $x \in \mathcal{V}$ can be expressed as a finite linear combination of elements of $\mathcal{B}$; and*
(b) *except for the presence of terms with a zero coefficient, this can be done in a unique way.*

Proof.

(a) Define

$$\mathcal{W} := \operatorname{span}\mathcal{B} = \{\sum_{j=1}^{n} \kappa_j x_j : n \in \mathbb{N}, \kappa_j \in \mathbb{F}, x_j \in \mathcal{B}, 1 \le j \le n\}.$$

We leave it as an exercise for the reader to prove that $\mathcal{W} \le \mathcal{V}$; that is, that $\mathcal{W}$ is a subspace of $\mathcal{V}$. Suppose that $\mathcal{W} \ne \mathcal{V}$, i.e. that there exists $y \in \mathcal{V} \setminus \mathcal{W}$.

We claim that $\mathcal{B} \cup \{y\}$ is linearly independent in $\mathcal{V}$, contradicting the maximality of $\mathcal{B}$ as a linearly independent set in $\mathcal{V}$.

Indeed, otherwise we can find $x_1, x_2, \ldots, x_n \in \mathcal{V}$ and $\kappa_1, \kappa_2, \ldots, \kappa_{n+1} \in \mathbb{F}$ such that

$$\kappa_1 x_1 + \kappa_2 x_2 + \cdots + \kappa_n x_n + \kappa_{n+1} y = 0.$$

If $\kappa_{n+1} \ne 0$, then

$$y = \frac{-1}{\kappa_{n+1}}(\kappa_1 x_1 + \kappa_2 x_2 + \cdots + \kappa_n x_n) \in \mathcal{W},$$

a contradiction.

Thus $\kappa_{n+1} = 0$, and so

$$\kappa_1 x_1 + \kappa_2 x_2 + \cdots + \kappa_n x_n = 0,$$

implying that $\kappa_j = 0$ for all j since $\mathcal{B}$ is linearly independent.

This contradiction shows that $\mathcal{W} = \mathcal{V}$, i.e. that $\mathcal{V}$ consists of all finite linear combinations of its basis elements.

(b) This is left as an exercise for the reader.

□

1.16. Definition. *A vector space $\mathcal{V}$ over a field $\mathbb{F}$ is said to be* ***finite-dimensional*** *if it admits a finite basis.*

1.17. Theorem. *Let $\mathcal{V}$ be a finite-dimensional vector space over a field $\mathbb{F}$. Any two bases $\mathcal{B}_1$ and $\mathcal{B}_2$ for $\mathcal{V}$ have the same number of elements.*

Proof. We shall argue by contradiction. Suppose that $\mathcal{B} := \{x_1, x_2, \dots x_m\}$ is a basis for $\mathcal{V}$, and that $Y := \{y_1, y_2, \dots, y_{m+1}\}$ is linearly independent in $\mathcal{V}$. By Exercise 1.14 above, $Y \subseteq \mathcal{B}_2$ for some basis $\mathcal{B}_2$ of $\mathcal{V}$. We shall show that this leads to a contradiction.

By Theorem 1.15, $\mathcal{V} = \operatorname{span}\mathcal{B}$, and thus

$$y_1 \in \operatorname{span}\mathcal{B},$$

i.e.,

$$y_1 = \kappa_1 x_1 + \kappa_2 x_2 + \cdots + \kappa_m x_m$$

for some choice of $\kappa_j \in \mathbb{F}$, $1 \le j \le m$. Since $y_1 \neq 0$, not all κ_j's are equal to zero. By reindexing the x_j's if necessary, we may assume that $\kappa_1 \neq 0$.

By Exercise 1 below, $\operatorname{span}\{y_1, x_2, x_3, \dots, x_m\} = \mathcal{V}$.

Thus $y_2 \in \operatorname{span}\{y_1, x_2, x_3, \dots, x_m\}$, say

$$y_2 = \alpha_1 y_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m.$$

Arguing as before, not all of $\alpha_2, \alpha_3, \dots, \alpha_m$ can be zero, otherwise $\{y_1, y_2\}$ are not linearly independent. Say $\alpha_2 \neq 0$. Then (exercise)

$$\operatorname{span}\{y_1, y_2, x_3, \dots, x_m\} = \mathcal{V}.$$

Continuing in this manner, we get to

$$\operatorname{span}\{y_1, y_2, y_3, \dots, y_m\} = \mathcal{V}.$$

But then $y_{m+1} \in \operatorname{span}\{y_1, y_2, y_3, \dots, y_m\}$, contradicting the fact that our original set $\{y_1, y_2, y_3, \dots, y_m, y_{m+1}\}$ was linearly independent.

Hence any linearly independent subset Y of $\mathcal{V}$ has at most m elements.

If $\mathfrak{C} \subseteq \mathcal{V}$ were a basis with fewer than m elements, then the above argument would show that $\mathcal{B}$ is not linearly independent, a contradiction.

Thus *all* bases for $\mathcal{V}$ have exactly m elements.

□

Theorem 1.17 is what allows us to make the following definition.

1.18. Definition. *Let $\mathcal{V}$ be a finite-dimensional vector space over a field $\mathbb{F}$. The* ***dimension*** *of $\mathcal{V}$ over $\mathbb{F}$ is the number of elements in a basis for $\mathcal{V}$, denoted by* $\dim_{\mathbb{F}} \mathcal{V}$.

1.19. Examples.

(a) If $n \in \mathbb{N}$ and $p \in \mathbb{N}$ is prime, then $\dim_{\mathbb{Z}_p} \mathbb{Z}_p^n = n$.

(b) The base field is extremely important when determining the dimension of a vector space. For example,

$$\dim_{\mathbb{C}} \mathbb{C} = 1, \qquad \text{and } \dim_{\mathbb{R}} \mathbb{C} = 2.$$

Indeed, $\mathcal{B} := \{1, i\}$ is easily seen to be a basis for $\mathbb{C}$ over $\mathbb{R}$, while $\mathcal{D} := \{1\}$ is a basis for $\mathbb{C}$ over itself.

(c) What should $\dim_{\mathbb{Q}} \mathbb{C}$ be?

Supplementary Examples.

S8.1. Example. Let $\mathbb{F}$ be a field. For each $1 \le n \in \mathbb{N}$, $\mathbb{F}^n := \{(x_1, x_2, \dots, x_n) : x_j \in \mathbb{F}, 1 \le j \le n\}$ is a vector space over $\mathbb{F}$ using the operations

$$(x_1, x_2, \dots, x_n) + (y_1, y_2, \dots, y_n) := (x_1 + y_1, x_2 + y_2, \dots, x_n + y_n)$$

and

$$\kappa(x_1, x_2, \dots, x_n) := (\kappa x_1, \kappa x_2, \dots, \kappa x_n).$$

In particular, if $n = 1$, this shows that $\mathbb{F}$ is a vector space over $\mathbb{F}$. The verification is left to the reader.

S8.2. Example. If $\varnothing \ne X$ is a set, $\mathbb{F}$ is a field and $f : X \to \mathbb{F}$ is a function, we define the **support** of f to be

$$\text{SUPP}(f) := \{x \in X : f(x) \ne 0\}.$$

The set $\mathbb{F}^X := \{f : X \to \mathbb{F} : f \text{ a function}\}$ is a vector space over $\mathbb{F}$, using the operations

$$(f + g)(x) := f(x) + g(x) \qquad \text{and} \qquad (\kappa f)(x) := \kappa(f(x)), \quad x \in X.$$

The set

$$\text{C}_{00}(X, \mathbb{F}) := \{f \in \mathbb{F}^X : \text{SUPP}(f) \text{ is finite}\}$$

is a subspace of $\mathbb{F}^X$.

A basis for $\text{C}_{00}(X, \mathbb{F})$ is the set $\mathfrak{B} := \{\delta_x : x \in X\}$, where

$$\begin{array}{rccl} \delta_x : & X & \to & \mathbb{F} \\ & y & \to & \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \ne x. \end{cases} \end{array}$$

If X is finite, then $\mathbb{F}^X = \text{C}_{00}(X, \mathbb{F})$.

S8.3. Example. Let $\mathbb{F}$ be a field. Then $\mathbb{F}[x]$ is a vector space over $\mathbb{F}$, where

$$\kappa(q_n x^n + q_{n-1} x^{n-1} + \cdots + q_1 x + q_0) := (\kappa q_n x^n + \kappa q_{n-1} x^{n-1} + \cdots + \kappa q_1 x + \kappa q_0)$$

for all $q(x) = q_n x^n + q_{n-1} x^{n-1} + \cdots + q_1 x + q_0 \in \mathbb{F}[x]$. (We are using the usual addition in $\mathbb{F}[x]$.)

Again, the verification of this is left to the reader.

S8.4. Example. The ring $\mathcal{C}([0,1], \mathbb{R}) := \{f : [0,1] \to \mathbb{R} : f \text{ is continuous}\}$ is a vector space over $\mathbb{R}$, using point-wise addition and setting

$$(\kappa f)(x) := \kappa(f(x)), \quad x \in [0,1], \quad \kappa \in \mathbb{R}.$$

S8.5. Example. Let $\mathbb{F}$ be a field, and $n \in \mathbb{N}$. Then $\mathbb{M}_n(\mathbb{F})$ is a vector space over $\mathbb{F}$, where

$$[x_{i,j}] + [y_{i,j}] := [x_{i,j} + y_{i,j}]$$

and

$$\kappa[x_{i,j}] := [\kappa x_{i,j}].$$

Note that $\dim_{\mathbb{F}} \mathbb{M}_n(\mathbb{F}) = n^2$. The reader is left to discover a basis for $\mathbb{M}_n(\mathbb{F})$ over $\mathbb{F}$.

S8.6. Example. Let $\mathbb{F}$ be a field and $\mathcal{V}$ be a vector space over $\mathbb{F}$. If $\mathcal{W}$ is a subspace of $\mathcal{V}$, and if $\mathcal{Y}$ is a subspace of $\mathcal{W}$, then $\mathcal{Y}$ is a subspace of $\mathcal{V}$.

S8.7. Example. Let $\varnothing \neq \Lambda$ be a set, and let $\mathbb{F}$ be a field. If, for each $\lambda \in \Lambda$, we have that $\mathcal{V}_\lambda$ is a vector space over $\mathbb{F}$, then

$$\prod_{\lambda \in \Lambda} \mathcal{V}_\lambda$$

is a vector space over $\mathbb{F}$, where

$$(x_\lambda)_\lambda + (y_\lambda)_\lambda := (x_\lambda + y_\lambda)_\lambda,$$

and

$$\kappa(x_\lambda)_\lambda := (\kappa x_\lambda)_\lambda,$$

for all $\kappa \in \mathbb{F}$, $(x_\lambda)_\lambda, (y_\lambda)_\lambda \in \prod_{\lambda \in \Lambda} \mathcal{V}_\lambda$.

S8.8. Example. Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Let $\mathcal{P}_n(\mathbb{F}) := \{q_n x^n + q_{n-1} x^{n-1} + \cdots + q_1 x + q_0 : q_j \in \mathbb{F}, 0 \leq j \leq n\}$. Then $\mathcal{P}_n(\mathbb{F})$ is a finite-dimensional subspace of the infinite-dimensional vector space $\mathbb{F}[x]$ over $\mathbb{F}$, and

$$\dim_{\mathbb{F}} \mathcal{P}_n(\mathbb{F}) = n + 1.$$

A basis for $\mathcal{P}_n(\mathbb{F})$ is $\mathfrak{B} := \{1, x, x^2, \ldots, x^n\}$.

S8.9. Example. Consider the polynomial $q(x) = x^2 + x + 1$ over $\mathbb{Z}_2$, and observe that it is irreducible over $\mathbb{Z}_2$, by virtue of the fact that its degree is 2 and it has no roots in $\mathbb{Z}_2$.

As we have seen, $\mathbb{F} := \mathbb{Z}_2[x]/\langle x^2 + x + 1\rangle$ is a field, and a general element of $\mathbb{F}$ looks like

$$b_1 x + b_0 + \langle q(x)\rangle : b_1, b_0 \in \mathbb{Z}_2.$$

Then $\mathbb{F}$ is a vector space over $\mathbb{Z}_2$, where

$$(b_1 x + b_0 + \langle q(x)) + (d_1 x + d_0 + \langle q(x)) := (b_1 + d_1)x + (b_0 + d_0) + \langle q(x),$$

and

$$\kappa(b_1 x + b_0 + \langle q(x)) := (\kappa b_1 x + \kappa b_0) + \langle q(x)\rangle.$$

A basis $\mathfrak{B}$ for $\mathbb{F}$ over $\mathbb{Z}_2$ is $\{1 + \langle q(x)\rangle, x + \langle q(x)\rangle\}$, and thus

$$\dim_{\mathbb{Z}_2}(\mathbb{F}) = 2.$$

S8.10. Example. Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$ and $\mathcal{W}, \mathcal{Y}$ be subspaces of $\mathcal{V}$. We say that $\mathcal{V}$ is an **(internal) direct sum** of $\mathcal{W}$ and $\mathcal{Y}$ if

- $\mathcal{W} \cap \mathcal{Y} = \{0\}$, and
- $\mathcal{V} = \mathcal{W} + \mathcal{Y} := \{w + y : w \in \mathcal{W}, y \in \mathcal{Y}\}$.

In this case we write $\mathcal{V} = \mathcal{W} \oplus \mathcal{Y}$.

Note that if $\text{CHAR}(\mathbb{F}) \neq 2$, then $\mathbb{F}^4 = \mathcal{W} \oplus \mathcal{Y}$, where $\mathcal{W} = \{(w, -w, 0, z) : w, z \in \mathbb{F}\}$ and $\mathcal{Y} = \{(x, x, y, 0) : x, y \in \mathbb{F}\}$, whereas if $\text{CHAR}(\mathbb{F}) = 2$, then $\mathcal{Y} \cap \mathcal{W} \neq \{(0, 0, 0, 0)\}$.

Appendix

A8.1. Vector spaces over general fields behave for the most part like vector spaces over $\mathbb{R}$ or over $\mathbb{C}$, although we must always be aware of those few cases where a property of the underlying field (such as finiteness, or the finiteness of its character) plays a role (e.g. Example 1.12).

Since we assume that the reader has already seen vector spaces over $\mathbb{R}$ in a linear algebra course, this chapter should be quite familiar, with the possible exception of the proof that every vector space admits a basis. We should point out that when a vector space is infinite-dimensional, it may be very, very difficult to actually describe a basis, and Theorem 1.13 merely asserts that one exists.

A8.2. Students who have seen Calculus or Fourier Analysis may have seen us write, for example,

$$\exp(x) := e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

This is NOT a linear combination of the functions $g_n(x) := \frac{x^n}{n!}$, $n \geq 0$. This is a *series expansion*. Be aware that we can NEVER add up infinitely many things in mathematics. The most general thing we can do is to take *finite* sums and then take *limits* of these.

The above expression really means exactly that. We are claiming that

$$\exp(\cdot) = \lim_{N \to \infty} \sum_{n=1}^{N} g_n.$$

What we mean by the limit depends in general on the circumstance. In this instance, we might be referring to point-wise limits, or to uniform convergence on closed and bounded (i.e. so-called *compact* subsets) of $\mathbb{R}$, or some other form of convergence altogether.

This study of such series expansions falls outside the scope of a course on rings and fields. We only bring it up to emphasise the fact that in vector spaces, we are only ever allowed to consider *finite* sums of vectors, and in any setting – linear algebra or otherwise – linear combinations only involve finitely many vectors. Be warned.

Exercises for Chapter 8

Exercise 8.1.

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over a field $\mathbb{F}$. A map $\varphi : \mathcal{V} \to \mathcal{W}$ is said to be $\mathbb{F}$-**linear** if $\varphi(\kappa x + y) = \kappa\varphi(x) + \varphi(y)$ for all $x, y \in \mathcal{V}$ and $\kappa \in \mathbb{F}$.

Let $\mathcal{L}(\mathcal{V}, \mathcal{W}) := \{\varphi : \varphi : \mathcal{V} \to \mathcal{W} \text{ is } \mathbb{F} - \textbf{linear}\}$. If $\mathcal{V} = \mathcal{W}$, we abbreviate the notation to $\mathcal{L}(\mathcal{V})$.

Prove that $\mathcal{L}(\mathcal{V}, \mathcal{W})$ is a vector space over $\mathbb{F}$ using the operations

$$(\varphi + \psi)(x) := \varphi(x) + \psi(x), \quad x \in \mathcal{V}$$

and

$$(\kappa\varphi)(x) := \kappa(\varphi(x)), \quad x \in \mathcal{V}, \kappa \in \mathbb{F}.$$

Note. Both $\mathcal{V}$ and $\mathcal{W}$ must be vector spaces over the *same* field $\mathbb{F}$. Also, if we are only dealing with one field $\mathbb{F}$ at a time, and there is no risk of confusion, we sometimes relax a bit and simply refer to these maps as *linear* (as opposed to $\mathbb{F}$-linear). Let's face it – life is stressful enough and sometimes we just need to let our hair down.

Exercise 8.2.

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over a field $\mathbb{F}$, and let $\varphi \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. Prove that

$$\ker \varphi := \{x \in \mathcal{V} : \varphi(x) = 0\}$$

is a subspace of $\mathcal{V}$, while $\operatorname{ran} \varphi := \{\varphi(x) : x \in \mathcal{V}\}$ is a subspace of $\mathcal{W}$.

Exercise 8.3.

Let $\mathcal{V}$ be a vector space over the field $\mathbb{F}$. Prove that $(\mathcal{L}(\mathcal{V}), +, \circ)$ is a ring with the operations

$$(\varphi + \psi)(x) := \varphi(x) + \psi(x), \quad x \in \mathcal{V}$$

and

$$(\varphi \circ \psi)(x) := \varphi(\psi(x)), \quad x \in \mathcal{V}.$$

Exercise 8.4.

Let $\mathcal{V}$ be a vector space over the field $\mathbb{F}$, and let $\mathfrak{L} \subseteq \mathcal{V}$ be a linearly independent set. Show that $\mathfrak{L}$ can be extended to a basis $\mathfrak{B}$ for $\mathcal{V}$. That is, show that there exists a basis $\mathfrak{B} \subseteq \mathcal{V}$ such that $\mathfrak{L} \subseteq \mathfrak{B}$.

Exercise 8.5.

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces over a field $\mathbb{F}$, and let $\varphi \in \mathcal{L}(\mathcal{V}, \mathcal{W})$. We say that φ is a **(vector space) isomorphism** if φ is bijective.

Prove that the following statements are equivalent.

(a) $\varphi \in \mathcal{L}(\mathcal{V}, \mathcal{W})$ is an isomorphism.
(b) Given a basis $\mathfrak{B} = \{b_\lambda : \lambda \in \Lambda\}$ for $\mathcal{V}$, the set $\mathfrak{D} := \{\varphi(b_\lambda) : \lambda \in \Lambda\}$ is a basis for $\mathcal{W}$.

Exercise 8.6.

Let $\mathcal{V}$ be a vector space over a field $\mathbb{F}$, and let $0 \neq \varphi \in \mathcal{L}(\mathcal{V}, \mathbb{F})$. Show that

$$\ker \varphi$$

is a maximal proper subspace of $\mathcal{V}$. That is, if $\mathcal{W} \leq \mathcal{V}$ is a subspace of $\mathcal{V}$ and

$$\ker \varphi \leq \mathcal{W} \leq \mathcal{V},$$

then either $\mathcal{W} = \ker \varphi$ or $\mathcal{W} = \mathcal{V}$.

Exercise 8.7.

Let $\mathcal{V}$ be a vector space over the field $\mathbb{F}$, and let $\mathcal{W}$ be a subspace of $\mathcal{V}$. We define a relation $\sim$ on $\mathcal{V}$ as follows: given $x, y \in \mathcal{V}$, we set

$$x \sim y \text{ if and only if } x - y \in \mathcal{W}.$$

(a) Prove that $\sim$ is an equivalence relation on $\mathcal{V}$.

(b) As we did in the case of rings, we refer to the equivalence class $x + \mathcal{W} := [x]$ of $x \in \mathcal{V}$ as a **coset of** $\mathcal{W}$. Prove that the set $\mathcal{V}/\mathcal{W} := \{[x] : x \in cV\}$ of cosets of $\mathcal{W}$ form a vector space using the operations

$$(x + \mathcal{W}) + (y + \mathcal{W}) := (x + y) + \mathcal{W},$$

and

$$\kappa(x + \mathcal{W}) := (\kappa x) + \mathcal{W}.$$

Of course, the first thing you will have to do is to verify that these operations are well-defined.

Exercise 8.8.

Prove the **First Isomorphism Theorem for vector spaces**:

If $\mathcal{V}$ and $\mathcal{W}$ are vector spaces over a field $\mathbb{F}$ and $\varphi : \mathcal{V} \to \mathcal{W}$ is an $\mathbb{F}$-linear map, then

$$\operatorname{ran} \varphi = \varphi(\mathcal{V}) \simeq \frac{\mathcal{V}}{\ker \varphi}.$$

Exercise 8.9.

Prove the **Second Isomorphism Theorem for vector spaces**:

Let $\mathcal{W}$ and $\mathcal{Y}$ are be subspaces of a vector space $\mathcal{V}$ over a field $\mathbb{F}$, and set

$$\mathcal{W} + \mathcal{Y} := \{w + y : w \in \mathcal{W}, y \in \mathcal{Y}\}.$$

Prove that $\mathcal{W} + \mathcal{Y}$ is a subspace of $\mathcal{V}$, and

$$\frac{\mathcal{W} + \mathcal{Y}}{\mathcal{Y}} \simeq \frac{\mathcal{W}}{\mathcal{W} \cap \mathcal{Y}}.$$

Exercise 8.10.

Prove the **Third Isomorphism Theorem for vector spaces**:

Let $\mathcal{W}$ and $\mathcal{Y}$ are be subspaces of a vector space $\mathcal{V}$ over a field $\mathbb{F}$, and suppose that $\mathcal{Y} \leq \mathcal{W}$. Prove that

$$\frac{\mathcal{V}}{\mathcal{W}} \simeq \frac{\mathcal{V}/\mathcal{Y}}{\mathcal{W}/\mathcal{Y}}.$$

CHAPTER 9

Extension fields

We're intellectual opposites. Well, I'm intellectual and you're opposite.

Mae West

1. A return to our roots (of polynomials)

1.1. Many of the problems which have arisen in the history of mathematics have had at their heart the problem of finding solutions to polynomial equations. For example, given a square whose sides each have length 1, the Pythagorean Theorem assures us that the length x of a diagonal of the square is given by the solution to $x^2 = 1^2 + 1^2 = 2$, or equivalently, to the equation $x^2 - 2 = 0$. In this way, one sees the desire to study very basic geometrical problems immediately creates the need to extend the rational numbers to a larger field where this equation admits a solution. Unsurprisingly, we turn to the field of real numbers. But even there, we are met with disappointment. The very simple quadratic polynomial equation $q(x) = x^2+1 \in \mathbb{R}[x]$ is irreducible. Seeing how it has degree 2, we conclude that it has no roots in $\mathbb{R}$. Again, we remedy this by introducing a larger field, namely $\mathbb{C}$, where $i = \sqrt{-1}$ exists and is a solution to that equation.

The goal in this Chapter is to generalise this phenomenon to more general fields. Given $\mathbb{F}$ and a polynomial $q(x) \in \mathbb{F}[x]$, we wish to show that it is always possible to construct a larger field $\mathbb{E}$ that contains $\mathbb{F}$ and also contains all of the roots of $q(x)$. The construction is somewhat abstract, and may not be to everyone's taste, even though it is perfectly acceptable from a mathematical viewpoint.

1.2. Definition. *Let $\mathbb{F}$ be a field. An **extension field** of $\mathbb{F}$ is a pair $(\mathbb{E}, \iota)$, where $\mathbb{E}$ is a field and $\iota : \mathbb{F} \to \mathbb{E}$ is an **embedding**, that is, an injective homomorphism. In other words, $\mathbb{E}$ is an extension of $\mathbb{F}$ if $\mathbb{F}^\iota := \iota(\mathbb{F})$ is a **subfield** of $\mathbb{E}$ for some isomorphism ι from $\mathbb{F}$ to $\mathbb{F}^\iota$.*

1.3. Remarks.

- We note that this is not the most common definition of an extension field of a field $\mathbb{F}$. Many authors define an extension field $\mathbb{E}$ of $\mathbb{F}$ simply as a field $\mathbb{E}$ which contains $\mathbb{F}$ as a subfield, and we may do this if we identify $\mathbb{F}$

with its image $\mathbb{F}^\iota := \iota(\mathbb{F}) \subseteq \mathbb{E}$. There is no harm in doing this – indeed, this is how we should interpret the notion of an extension field, provided that we are careful about keeping track of this identification. In writing these course notes however, we have taken the viewpoint that when first learning a subject, it is good to be overly cautious, meticulous and probably a bit pedantic, and to do the bookkeeping in plain sight for all to see.

- At times we shall have to simultaneously consider multiple extension fields of a field $\mathbb{F}$. When we are only dealing with one extension field $(\mathbb{E}, \iota)$ however, we shall use the notation $b^\iota := \iota(b)$ in the hope that it will make the arguments more readable. *This will also apply to polynomials in the following way:* suppose that $q(x) = q_n x^n + q_{n-1}x^{n-1} + \cdots + q_1 x + q_0 \in \mathbb{F}[x]$ and that $(\mathbb{E}, \iota)$ is an extension of $\mathbb{F}$. We shall denote by
$$q^\iota(x) := q_n^\iota x^n + q_{n-1}^\iota x^{n-1} + \cdots + q_1^\iota x + q_0^\iota$$
the corresponding polynomial in $\mathbb{E}[x]$ with coefficients $q_j^\iota = \iota(q_j)$, $0 \le j \le n$.
- A brief comment on terminology: we shall also say that $\mathbb{E}$ *is an extension field of* $\mathbb{F}$ *via* ι to mean that $(\mathbb{E}, \iota)$ is an extension field of $\mathbb{F}$.
- Finally, let $(\mathbb{E}, \iota)$ be an extension field of $\mathbb{F}$ and $q(x) \in \mathbb{F}[x]$. Then, in the same way that we are (mentally) identifying $\mathbb{F}$ and $\mathbb{F}^\iota$, we should be identifying $q(x)$ and $q^\iota(x)$. Thus "finding a root for $q(x)$" in $\mathbb{E}$ *really means* finding a root for $q^\iota(x)$ in $\mathbb{E}$! Despite this, we shall continue to use the common **abuse of terminology** and refer to a root of $q^\iota(x)$ in $\mathbb{E}$ as a "root of $q(x)$".

1.4. Examples.

(a) $(\mathbb{R}, \iota)$ is an extension of $\mathbb{Q}$, where $\iota(q) = q$ for all $q \in \mathbb{Q}$.

Note that in this example, if $q(x) = \frac{3}{2}x^2 - \frac{17}{18}x + \frac{222}{1033} \in \mathbb{Q}[x]$, then
$$q^\iota(x) = \frac{3}{2}^\iota x^2 - \frac{17}{18}^\iota x + \frac{222}{1033}^\iota = \frac{3}{2}x^2 - \frac{17}{18}x + \frac{222}{1033} \in \mathbb{R}[x].$$
Hopefully the reader can see why we would want to abuse the terminology in cases such as these, and refer to a root of $q^\iota(x)$ in $\mathbb{R}$ as a "root of $q(x)$".

(b) $(\mathbb{C}, \iota)$ is an extension of $\mathbb{R}$, where $\iota(x) = x \cdot 1 + 0 \cdot \sqrt{-1} = x + i0$ for all $x \in \mathbb{R}$, and therefore $(\mathbb{C}, \iota|_{\mathbb{Q}})$ is also an extension of $\mathbb{Q}$. [Note that ι and i are different! The first is an injective homomorphism, while the second is the square root of -1 in $\mathbb{C}$!]

(c) There does not exist an embedding $\varphi : \mathbb{Z}_7 \to \mathbb{Z}_{19}$ for which $(\mathbb{Z}_{19}, \varphi)$ is an extension of $\mathbb{Z}_7$. Indeed, if $1 \in \mathbb{Z}_7$ denotes the multiplicative identity in $\mathbb{Z}_7$, then the fact that $\text{CHAR}(\mathbb{Z}_7) = 7$ implies that
$$0 = \varphi(7) = \varphi(1) + \varphi(1) + \varphi(1) + \varphi(1) + \varphi(1) + \varphi(1) + \varphi(1).$$
But the only element $a \in \mathbb{Z}_{19}$ that satisfies $7a = (a + a + a + a + a + a + a) = 0$ is $a = 0$. Thus $\varphi(1) = 0 = \varphi(0)$, contradicting the fact that φ is injective.

In constructing extension fields which contain a root to a given polynomial, the next result is the key to our construction. As previously mentioned – the construction is perhaps more abstract than some of us might have preferred.

1.5. Kronecker's Theorem – The fundamental theorem of field theory. *Let $\mathbb{F}$ be field and $0 \neq q(x)$ be a non-constant polynomial in $\mathbb{F}[x]$. Then there exists an extension field $(\mathbb{E}, \iota)$ of $\mathbb{F}$ such that $q^\iota(x)$ has a root in $\mathbb{E}$.*

Proof. If $q(x)$ already has a root in $\mathbb{F}$, then we simply set $\mathbb{E} := \mathbb{F}$ and $\iota(a) = a$ for all $a \in \mathbb{F}$. Thus we suppose that $q(x)$ does not have any roots in $\mathbb{F}$.

Since $\mathbb{F}$ is a field, $\mathbb{F}[x]$ is a UFD. Thus we can factor $q(x)$ as a product of irreducible polynomials over $\mathbb{F}$. Let $p(x)$ denote one of these irreducible factors of $q(x)$. That is, $q(x) = p(x)s(x)$ for some $s(x) \in \mathbb{F}[x]$, and $p(x)$ is irreducible. Write $p(x) = p_n x^n + p_{n-1}x^{n-1} + \cdots + p_1 x + p_0$.

Now consider $\mathbb{E} := \mathbb{F}[x]/\langle p(x)\rangle$. By Theorem 7.2.7, $\mathbb{E}$ is a field. The map

$$\begin{array}{rccc} \iota: & \mathbb{F} & \to & \mathbb{E} \\ & a & \mapsto & a + \langle p(x)\rangle \end{array}$$

is easily seen to be an injective homomorphism. That is, for all $a, b \in \mathbb{F}$,

$$\iota(ab) = ab + \langle p(x)\rangle = (a + \langle p(x)\rangle)(b + \langle p(x)\rangle) = \iota(a)\iota(b),$$

and

$$\iota(a+b) = (a+b) + \langle p(x)\rangle = (a + \langle p(x)\rangle) + (b + \langle p(x)\rangle) = \iota(a) + \iota(b).$$

Moreover, $\ker \iota = \{0\}$, since $p(x)$ irreducible implies that $\deg p(x) \geq 1$, and thus $0 \neq a \in \mathbb{F}$ implies that $0 \neq \iota(a) = a + \langle p(x)\rangle \in \mathbb{E}$.

Thus $(\mathbb{E}, \iota)$ is an extension field of $\mathbb{F}$.

We now leave it to the reader to verify that $q^\iota(x) = p^\iota(x)s^\iota(x)$, and that $p^\iota(x)$ is irreducible in $\mathbb{F}^\iota[x]$. (This should be very easy if one understands what an isomorphism is.)

It is clear that any root of $p^\iota(x)$ is also a root of $q^\iota(x)$. As such, we have reduced the problem to finding roots of irreducible polynomials.

Here's the *kicker*. Consider $\alpha := x + \langle p(x)\rangle \in \mathbb{E}$. Then $\alpha^j = x^j + \langle p(x)\rangle$ for all $j \geq 1$ and so

$$\begin{aligned} p^\iota(\alpha) &= p_n^\iota \alpha^n + p_{n-1}^\iota \alpha^{n-1} + \cdots + p_1^\iota \alpha^1 + p_0^\iota \\ &= p_n^\iota(x^n + \langle p(x)\rangle) + p_{n-1}^\iota(x^{n-1} + \langle p(x)\rangle) + \cdots + p_1^\iota(x + \langle p(x)\rangle) + p_0^\iota \\ &= (p_n + \langle p(x)\rangle)(x^n + \langle p(x)\rangle) + (p_{n-1} + \langle p(x)\rangle)(x^{n-1} + \langle p(x)\rangle) + \cdots \\ &\qquad \cdots + (p_1 + \langle p(x)\rangle)(x + \langle p(x)\rangle) + (p_0 + \langle p(x)\rangle) \\ &= (p_n x^n + p_{n-1}x^{n-1} + \cdots + p_1 x + p_0) + \langle p(x)\rangle \\ &= p(x) + \langle p(x)\rangle \\ &= 0 + \langle p(x)\rangle \in \mathbb{E}. \end{aligned}$$

In other words, α is a root of $p^\iota(x)$, and hence of $q^\iota(x)$ in $\mathbb{E}$!

Are you kidding me? Nay! Am I kidding you? Nay nay!

□

Remark. We shall refer to the embedding ι of $\mathbb{F}$ into $\mathbb{E} = \mathbb{F}[x]/\langle p(x)\rangle$ defined in the proof of Kronecker's Theorem by $\iota(a) = a + \langle p(x)\rangle$ for all $a \in \mathbb{F}$ as the *canonical embedding* of $\mathbb{F}$ into $\mathbb{E}$. Thus

$$\mathbb{F}^\iota := \iota(\mathbb{F}) = \{a^\iota := a + \langle p(x)\rangle : a \in \mathbb{F}\}$$

is a subfield of $\mathbb{E}$ which is isomorphic to $\mathbb{F}$.

1.6. Examples.

(a) Let $\mathbb{F} = \mathbb{Q}$ and $q(x) = x^2 + 1 \in \mathbb{Q}[x]$. Note that in this example, $q(x)$ is already irreducible over $\mathbb{Q}$, so $q(x)$ will play the role of $p(x)$ from Kronecker's Theorem. By the argument in the proof of Kronecker's Theorem 1.5,

$$\mathbb{E} := \mathbb{Q}[x]/\langle x^2 + 1\rangle$$

is an extension of $\mathbb{Q}$ via the canonical embedding $a^\iota := \iota(a) = a + \langle x^2 + 1\rangle,\ a \in \mathbb{Q}$.

Moreover, $q(x) = q_2x^2 + q_0$, where $q_2 = q_0 = 1$ and thus

$$q^\iota(x) = q_2^\iota x^2 + q_0^\iota = (1 + \langle x^2 + 1\rangle)x^2 + (1 + \langle x^2 + 1\rangle).$$

This might seem a bit confusing at first, but notice that if we write $K := \langle q(x)\rangle = \langle x^2 + 1\rangle$, then we are saying that

$$q^\iota(x) = (q_2 + K)x^2 + (q_0 + K).$$

It is important to keep in mind that an *element* of $\mathbb{E}$ is a coset of K, so $q_2^\iota := q_2 + K$ and $q_0^\iota := q_0 + K$ are elements of $\mathbb{E}$. That is, $q^\iota(x) = q_2^\iota x^2 + q_0^\iota \in \mathbb{E}[x]$, as required.

Let's check that α is indeed a root for $q^\iota(x)$:
Setting $\alpha = x + K = x + \langle x^2 + 1\rangle \in \mathbb{E}$,

$$\begin{aligned} q^\iota(\alpha) &= (q_2 + K)\alpha^2 + (q_0 + K) \\ &= (q_2 + K)(x + K)^2 + (q_0 + K) \\ &= (q_2 + K)(x^2 + K) + (q_0 + K) \\ &= (q_2x^2 + q_0) + K \\ &= q(x) + \langle q(x)\rangle \\ &= (x^2 + 1) + \langle x^2 + 1\rangle \\ &= 0 + \langle x^2 + 1\rangle. \end{aligned}$$

The reader will observe that we never needed to invoke $\mathbb{C}$ or "$\sqrt{-1}$" to get a root for this polynomial! Indeed, this is one way to *define* $\sqrt{-1}$! (How sweet is that?)

Observe also that

$$\begin{aligned}(-\alpha)^2 + 1^\iota &= (-x + \langle x^2+1\rangle)^2 + (1 + \langle x^2+1\rangle)\\ &= ((-x)^2 + \langle x^2+1\rangle) + (1 + \langle x^2+1\rangle)\\ &= (x^2+1) + \langle x^2+1\rangle\\ &= 0 + \langle x^2+1\rangle.\end{aligned}$$

Thus $-\alpha$ is also a root of $q^\iota(x)$. (Of course, we could have simply noted that $(-\alpha)^2 + 1^\iota = (\alpha)^2 + 1^\iota$. We are simply trying to emphasise what α "*really looks like*".)

Thus $q^\iota(x) = (x-\alpha)(x+\alpha)$ factors into linear terms in $\mathbb{E}$.

(b) Let

$$q(x) = (x^2+2)(x^3+x^2+1) \in \mathbb{Z}_5[x].$$

Note that $g(x) := x^2+2$ and $h(x) := x^3+x^2+1$ are both irreducible polynomials over $\mathbb{Z}_5$, since neither has a root in $\mathbb{Z}_5$ and they each have degree in $\{2,3\}$.

By Kronecker's Theorem,

$$\mathbb{E}_1 := \mathbb{Z}_5[x]/\langle x^2+2\rangle$$

and

$$\mathbb{E}_2 := \mathbb{Z}_5[x]/\langle x^3+x^2+1\rangle$$

are two extension fields of $\mathbb{Z}_5$ via the maps $\iota_1(a) = a + \langle x^2+2\rangle$ and $\iota_2(a) = a + \langle x^3+x^2+1\rangle$ respectively. Note that each of $\mathbb{E}_j$ contains a root α_j of $q^{\iota_j}(x), j = 1,2$. But

$$\mathbb{E}_1 = \{a_1x + a_0 + \langle x^2+2\rangle : a_0, a_1 \in \mathbb{Z}_5\}$$

has 25 elements, whereas

$$\mathbb{E}_2 = \{b_2x^2 + b_1x + b_0 + \langle x^3+x^2+1\rangle : b_0, b_1, b_2 \in \mathbb{Z}_5\}$$

has 125 elements.

This shows that not all extensions of a given field $\mathbb{F}$ need to be isomorphic to each other. That is, there can exist many fundamentally different (i.e. non-isomorphic) extensions of a given field $\mathbb{F}$.

1.7. Definition. *Let $(\mathbb{E},\iota)$ be an extension field of a field $\mathbb{F}$, and let $q(x) \in \mathbb{F}[x]$. We say that $q(x)$ **splits over** $\mathbb{E}$ if $q^\iota(x)$ can be written as a product of linear terms with coefficients in $\mathbb{E}$.*

*The field $\mathbb{E}$ is said to be a **splitting field** for $q(x) \in \mathbb{F}[x]$ if*

(a) *$q(x)$ splits over $\mathbb{E}$, and*
(b) *If $\mathbb{F}^\iota \subseteq \mathbb{K} \subseteq \mathbb{E}$ for some **proper** subfield $\mathbb{K}$ of $\mathbb{E}$, then $q(x)$ does not split over $\mathbb{K}$.*

1.8. Examples.

(a) Consider $q(x) = x^2 - 2 \in \mathbb{Q}[x]$. Then $q(x)$ does *not* split over $\mathbb{Q}$, but it does split over $\mathbb{R}$. Indeed,

$$q^\iota(x) = x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2}) \in \mathbb{R}[x].$$

Note also that $q(x)$ splits over $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$, and thus $\mathbb{R}$ is not the splitting field for $q(x)$. As it turns out, $\mathbb{Q}(\sqrt{2})$ is the splitting field for $q(x) \in \mathbb{Q}[x]$.

(b) The same polynomial $q(x) = x^2 - 2$ can be thought of as an element of $\mathbb{R}[x]$. In this case, $\mathbb{Q}(\sqrt{2})$ would not be considered a splitting field for $q(x)$ over $\mathbb{R}$, since any splitting field for $q(x)$ over $\mathbb{R}$ would have to contain the domain for the coefficients, namely $\mathbb{R}$. In other words, if we are thinking of $q(x)$ as a polynomial with real coefficients, then $q(x)$ already splits in $\mathbb{R}$, so $\mathbb{R}$ is the splitting field of $q(x)$.

1.9. Notation. Let $\mathbb{F}$ be a field and $(\mathbb{E}, \iota)$ be an extension field of $\mathbb{F}$. If $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{E}$, we denote by $\mathbb{F}(\alpha_1, \alpha_2, \ldots \alpha_n)$ the smallest subfield of $\mathbb{E}$ containing $\mathbb{F}^\iota$ and $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$.

If

$$q^\iota(x) = q_n^\iota(x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n) \in \mathbb{E}[x],$$

then $\mathbb{F}(\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a splitting field for $q(x)$ over $\mathbb{F}$. The proof of this is left as an exercise for the reader.

1.10. Theorem.

Let $\mathbb{F}$ be a field and $0 \neq q(x) \in \mathbb{F}[x]$. Then there exists a splitting field $\mathbb{K}$ for $q(x)$ over $\mathbb{F}$.

Proof. We shall prove this by induction on the degree, say n, of $q(x)$. Of course, if $n = 0$, then $q(x) = q_0$ has no roots (since we assumed that $q(x) \neq 0$), and so $\mathbb{F}$ is the splitting field for $q(x)$ over $\mathbb{F}$.

If $n = 1$, then $q(x) = q_1 x + q_0$ with $q_1 \neq 0$, and thus $\alpha := -q_1^{-1} q_0 \in \mathbb{F}$ is the unique root of $q(x)$. In other words, $\mathbb{F}$ is again a splitting field for $q(x)$.

Let $N > 1$ be fixed, and suppose that the result holds for all polynomials of degree less than N. Suppose next that $q(x) \in \mathbb{F}[x]$ is a polynomial of degree N. By Kronecker's Theorem 1.5, there exists an extension field $(\mathbb{E}, \iota)$ of $\mathbb{F}$ in which $q^\iota(x)$ has a root, say $\alpha_1 \in \mathbb{E}$.

Thus

$$q^\iota(x) = (x - \alpha_1)g(x)$$

for some $g(x) \in \mathbb{E}[x]$ with $\deg(g(x)) = N - 1 < N$. By our induction hypothesis, there exists an extension field $(\mathbb{H}, \tau)$ of $\mathbb{E}$ such that $g(x)$ splits in $\mathbb{K}$; that is, there exist $\alpha_2, \alpha_3, \ldots, \alpha_N \in \mathbb{H}$ such that

$$g(x) = g_{N-1}^\tau(x - \alpha_2)(x - \alpha_3)\cdots(x - \alpha_N) \in \mathbb{K}[x].$$

Let $\mathbb{K} := \mathbb{F}(\alpha_1, \alpha_2, \ldots, \alpha_N) \subseteq \mathbb{H}$. Then $(\mathbb{K}, \tau \circ \iota)$ is a splitting field for $q(x)$ over $\mathbb{F}$.

□

1.11. Example.
Let $q(x) = x^4 + 4x^2 - 45 = (x^2 - 5)(x^2 + 9) \in \mathbb{Q}[x]$, and let $\iota : \mathbb{Q} \to \mathbb{C}$ denote the inclusion map $\iota(q) = q$ for all $q \in \mathbb{Q}$. The roots of $q^\iota(x)$ in $\mathbb{C}$ are $\{\sqrt{5}, -\sqrt{5}, 3i, -3i\}$. Thus, a splitting field for $q(x)$ over $\mathbb{Q}$ is $\mathbb{K} := \mathbb{Q}(\sqrt{5}, 3i) = \mathbb{Q}(\sqrt{5}, i)$.

In particular,

$$\mathbb{K} = \{(a + b\sqrt{5}) + (c + d\sqrt{5})i : a, b, c, d \in \mathbb{Q}\}.$$

It is worth noting that $\mathbb{K}$ is a vector space over $\mathbb{Q}$ with basis $\{1, \sqrt{5}, i, \sqrt{5}i\}$, and thus

$$\dim_{\mathbb{Q}} \mathbb{K} = 4.$$

That $\mathbb{K}$ is a vector space over $\mathbb{Q}$ is not a coincidence. We shall study this phenomenon presently.

1.12. Exercise. What is the dimension of $\mathbb{R}$ as a vector space over $\mathbb{Q}$?

1.13. Theorem. *Let $\mathbb{F}$ be a field and $0 \neq q(x) \in \mathbb{F}[x]$ be an irreducible polynomial over $\mathbb{F}$. Suppose that $(\mathbb{E}, \iota)$ is an extension field of $\mathbb{F}$ containing a root α of $q(x)$. Then*

(a) $\mathbb{F}(\alpha) \simeq \mathbb{F}[x]/\langle q(x)\rangle$; *and*
(b) *if* $\deg(q(x)) = n \geq 1$, *then*

$$\mathbb{F}(\alpha) = \{b_{n-1}^\iota \alpha^{n-1} + b_{n-2}^\iota \alpha^{n-2} + \cdots + b_1^\iota \alpha + b_0^\iota : b_j \in \mathbb{F}, 0 \leq j \leq n-1\}.$$

In particular $\mathbb{F}(\alpha)$ is an n-dimensional vector space over $\mathbb{F}$.

Proof. Since $q(x)$ is irreducible over $\mathbb{F}$, by Theorem 7.2.7, $\langle q(x)\rangle$ is a maximal ideal and $\mathbb{F}[x]/\langle q(x)\rangle$ is a field.

(a) Consider the so-called **evaluation map**

$$\begin{array}{rccc} \Theta: & \mathbb{F}[x] & \to & \mathbb{F}(\alpha) \\ & f(x) & \mapsto & f^\iota(\alpha). \end{array}$$

It is routine to verify (and the reader should check) that Θ is a surjective homomorphism. Now $q^\iota(\alpha) = 0$, so $q(x) \in \ker \Theta$. Since $\ker \Theta$ is an ideal, $\langle q(x)\rangle \subseteq \ker \Theta$. By maximality of $\langle q(x)\rangle$ combined with the fact that $\Theta(1) = 1^\iota \neq 0$ (and hence $\ker \Theta \neq \mathbb{F}[x]$), it follows that

$$\langle q(x)\rangle = \ker \Theta.$$

By the First Isomorphism Theorem 4.4.2,

$$\mathbb{F}(\alpha) \simeq \mathbb{F}[x]/\langle q(x)\rangle,$$

via the map

$$\begin{array}{rccc} \widehat{\Theta}: & \dfrac{\mathbb{F}[x]}{\langle q(x)\rangle} & \to & \mathbb{F}(\alpha) \\ & f(x) + \langle q(x)\rangle & \mapsto & f^\iota(\alpha). \end{array}$$

(b) Since

$$\mathbb{F}[x]/\langle q(x)\rangle = \{b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_1 x + b_0 + \langle q(x)\rangle : b_j \in \mathbb{F}, 0 \le j \le n-1\},$$

we see that

$$\begin{aligned}\mathbb{F}(\alpha) &= \widehat{\Theta}(\mathbb{F}[x]/\langle q(x)\rangle) \\ &= \{b^\iota_{n-1}\alpha^{n-1} + b^\iota_{n-2}\alpha^{n-2} + \cdots + b^\iota_1\alpha + b^\iota_0 : b_j \in \mathbb{F}, 0 \le j \le n-1\}.\end{aligned}$$

□

1.14. Remark. It is worth noting that the isomorphism between $\mathbb{F}(\alpha)$ and $\mathbb{F}[x]/\langle q(x)\rangle$ in Theorem 1.13 above is given by the map

$$\widehat{\Theta}(f(x) + \langle q(x)\rangle) = f^\iota(\alpha).$$

This follows immediately from the First Isomorphism Theorem and its proof.

Also, let us keep in mind that for $b \in \mathbb{F}$, the map $\delta : b \mapsto b + \langle q(x)\rangle$ is an embedding of $\mathbb{F}$ to $\mathbb{F}[x]/\langle q(x)\rangle$, and

$$\widehat{\Theta}(b + \langle q(x)\rangle) = b^\iota \in \mathbb{F}(\alpha).$$

Thus we may think of the map $\kappa := \widehat{\Theta} \circ \delta : \mathbb{F} \to \mathbb{F}(\alpha)$ satisfying $\kappa(b) = b^\iota$ as a canonical embedding of $\mathbb{F}$ into $\mathbb{F}(\alpha)$ induced by this construction.

1.15. Corollary. *Let $\mathbb{F}$ be a field and $0 \neq q(x) \in \mathbb{F}[x]$ be irreducible. Suppose that $(\mathbb{E}_1, \iota_1)$ and $(\mathbb{E}_2, \iota_2)$ are extension fields of $\mathbb{F}$ such that*

- *$\mathbb{E}_1$ contains a root α of $q^{\iota_1}(x)$, and*
- *$\mathbb{E}_2$ contains a root β of $q^{\iota_2}(x)$.*

Then there exists an isomorphism $\Gamma : \mathbb{F}(\alpha) \to \mathbb{F}(\beta)$ such that $\Gamma(\iota_1(b)) = \iota_2(b)$ for all $b \in \mathbb{F}$ and $\Gamma(\alpha) = \beta$.

Proof. This follows immediately from Theorem 1.13 and Remark 1.14. As we saw there, there exist isomorphisms $\widehat{\Theta}_1 : \mathbb{F}[x]/\langle q(x)\rangle \to \mathbb{F}(\alpha)$ and $\widehat{\Theta}_2 : \mathbb{F}[x]/\langle q(x)\rangle \to \mathbb{F}(\beta)$ such that

- $\widehat{\Theta}_j(b + \langle q(x)\rangle) = \iota_j(b) = b^{\iota_j}$ for all $b \in \mathbb{F}$, $j = 1, 2$;
- $\widehat{\Theta}_1(x + \langle q(x)\rangle) = \alpha$; and
- $\widehat{\Theta}_2(x + \langle q(x)\rangle) = \beta$.

It now suffices to set

$$\Gamma := \widehat{\Theta}_2 \circ \widehat{\Theta}_1^{-1}.$$

□

1.16. Example. Let $q(x) = x^6 - 2 \in \mathbb{Q}[x]$. Observe that $q(x)$ is irreducible over $\mathbb{Q}$ (why?) and that $\alpha := \sqrt[6]{2}$ is a root of $q(x)$ in $\mathbb{R}$. Let $\iota : \mathbb{Q} \to \mathbb{R}$ be the inclusion map $\iota(b) = b$, so that $b^\iota = b$ for all $b \in \mathbb{Q}$.

By Theorem 1.13,

$$\begin{aligned}\mathbb{Q}(\alpha) &= \{b_5^\iota \alpha^5 + b_4^\iota \alpha^4 + \cdots + b_1^\iota \alpha + b_0^\iota : b_j \in \mathbb{Q}, 0 \le j \le 5\} \\ &= \{b_5 \sqrt[6]{2}^5 + b_4 \sqrt[6]{2}^4 + \cdots + b_1 \sqrt[6]{2} + b_0 : b_j \in \mathbb{Q}, 0 \le j \le 5\}\end{aligned}$$

is a vector space of dimension 6 over $\mathbb{Q}$ with basis

$$\mathcal{B} := \{1, \sqrt[6]{2}, \sqrt[6]{2}^2, \sqrt[6]{2}^3, \sqrt[6]{2}^4, \sqrt[6]{2}^5\}.$$

Note, however, that $\mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt[6]{2})$ is *not* a splitting field for $q(x)$ over $\mathbb{Q}$. If $\omega \in \mathbb{C}$ is a primitive sixth root of unity (i.e. $\omega = e^{\frac{2\pi i}{6}}$ and thus $\{1, \omega, \omega^2, \ldots, \omega^5\}$ is the set of *all* roots of unity in $\mathbb{C}$), then $\{\sqrt[6]{2}, \omega\sqrt[6]{2}, \omega^2\sqrt[6]{2}, \omega^3\sqrt[6]{2}, \omega^4\sqrt[6]{2}, \omega^5\sqrt[6]{2}\}$ are all roots of $q(x)$. Let's see why.

First note that

$$\alpha_j^6 = (\omega^j \sqrt[6]{2})^6 = \omega^{6j}(\sqrt[6]{2})^6 = (\omega^6)^j 2 = 1^j \cdot 2 = 2,$$

whence $q(\alpha_j) = \alpha_j^6 - 2 = 2 - 2 = 0$. Thus each α_j is a root of $q(x)$. But $\deg(q(x)) = 6$, and so $q^\iota(x) \in \mathbb{K}[x]$ is a polynomial of degree 6, it can have at most six roots. Thus

$$q^\iota(x) = (x - \alpha_0)(x - \alpha_1)\cdots(x - \alpha_4)(x - \alpha_5)$$

splits over $\mathbb{K}$, and $(\mathbb{K}, \iota)$ is a splitting field for $q(x)$.

Note that $\mathbb{K}$ is the smallest field in $\mathbb{C}$ containing all of the roots of $q(x)$. Thus it must contain $\alpha_1 = \sqrt[6]{2}$ and $\alpha_2 = \omega\sqrt[6]{2}$. But then it also must contain $\alpha_1^{-1}\alpha_2 = \omega$. Hence $\mathbb{K} \supseteq \mathbb{Q}(\omega, \sqrt[6]{2})$. Since $\mathbb{Q}(\omega, \sqrt[6]{2})$ contains each α_j, $0 \le j \le 5$, we conclude that $\mathbb{K} = \mathbb{Q}(\omega, \sqrt[6]{2})$.

What is the dimension of $\mathbb{K}$ over $\mathbb{Q}$?

The astute reader (which may or may not be the same person as the perspicacious reader mentioned earlier) will have noticed that we have been referring to *a* splitting field for a polynomial over a field, as opposed to *the* splitting field. The following theorem will set things right.

But first, note that if $\theta : \mathbb{F}_1 \to \mathbb{F}_2$ is an isomorphism of fields, and if $f(x) = b_N x^N + b_{N-1}x^{N-1} + \cdots + b_1 x + b_0 \in \mathbb{F}_1[x]$, then we extend our previous notation by setting

$$f^\theta(x) := b_N^\theta x^N + b_{N-1}^\theta x^{N-1} + \cdots + b_1^\theta x + b_0^\theta \in \mathbb{F}_2[x],$$

where $b_j^\theta := \theta(b_j)$, $0 \le j \le N$.

1.17. Theorem. *Let $\mathbb{F}$ be a field and suppose that $f(x) \in \mathbb{F}[x]$. If $(\mathbb{K}_1, \iota_1)$ and $(\mathbb{K}_2, \iota_2)$ are splitting fields for $f(x)$ over $\mathbb{F}$, then there exists an isomorphism*

$$\Phi : \mathbb{K}_1 \to \mathbb{K}_2$$

which satisfies $\Phi(\iota_1(b)) = \iota_2(b)$ for all $b \in \mathbb{F}$.

Proof. Let $1 \leq N := \deg(f(x))$.

If $N = 1$, then $\mathbb{K}_1 = \mathbb{F} = \mathbb{K}_2$, so we may take $\Phi := \iota_2 \circ \iota_1^{-1}$, and we are done.

Now suppose that $N \geq 2$, and that the result holds for all polynomials of degree less than N. Since $\mathbb{F}[x]$ is a UFD by Theorem 7.1.10, we can factor $f(x)$ as a product of irreducible polynomials. Let $q(x)$ be one of the irreducible factors of $f(x)$.

Let $\alpha_1 \in \mathbb{K}_1$ be a root of $q^{\iota_1}(x)$. (Note that all roots of $q^{\iota_1}(x)$ are roots of $f^{\iota_1}(x)$, and hence must lie in $\mathbb{K}_1$.) Similarly, we can find a root $\beta_1 \in \mathbb{K}_2$ of $q^{\iota_2}(x)$.

By Corollary 1.15, there exists an isomorphism $\theta : \mathbb{F}(\alpha_1) \to \mathbb{F}(\beta_1)$ such that $\theta(\iota_1(b)) = \iota_2(b)$ for all $b \in \mathbb{F}$ and $\theta(\alpha_1) = \beta_1$.

It follows that

$$f^{\iota_2}(x) = (f^{\iota_1})^{\theta}(x).$$

If we then write $f^{\iota_1}(x) = (x - \alpha_1)h(x)$ for some $h(x) = h_{N-1}x^{N-1} + h_{N-2}x^{N-2} + \cdots + h_1 x + h_0 \in \mathbb{F}(\alpha_1)[x]$, then

$$\begin{aligned} f^{\iota_2}(x) &= (f^{\iota_1})^{\theta}(x) \\ &= (x - (\alpha_1)^{\theta})h^{\theta}(x) \\ &= (x - \beta_1)(h^{\theta}_{N-1}x^{N-1} + h^{\theta}_{N-2}x^{N-2} + \cdots + h^{\theta}_1 x + h^{\theta}_0) \qquad \in \mathbb{F}(\beta_1)[x]. \end{aligned}$$

Moreover, $(\mathbb{K}_1, \eta_1)$ is a splitting field for $h(x)$ over $\mathbb{F}(\alpha_1)$, where η_1 is the inclusion map of $\mathbb{F}(\alpha_1)$ into $\mathbb{K}_1$, that is, $\eta_1(z) = z$ for all $z \in \mathbb{F}(\alpha_1)$. In particular, $\eta_1(\iota_1(b)) = \iota_1(b)$ for all $b \in \mathbb{F}$ and $\eta_1(\alpha_1) = \alpha_1$.

Also, $(\mathbb{K}_2, \eta_2)$ is another splitting field for $h(x)$ over $\mathbb{F}(\alpha_1)$, where $\eta_2(z) = z^{\theta} := \theta(z)$ for all $z \in \mathbb{F}(\alpha_1)$. That is, η_2 first implements the isomorphism between $\mathbb{F}(\alpha_1)$ and $\mathbb{F}(\beta_1)$, and then embeds $\mathbb{F}(\beta_1)$ into $\mathbb{K}_2$ via the inclusion map.

By our induction hypothesis – noting that $\deg(h(x)) = N - 1 < N$ for $j = 1, 2$ – we may find an isomorphism $\Phi : \mathbb{K}_1 \to \mathbb{K}_2$ with $\Phi(\eta_1(z)) = \Phi(z) = \eta_2(z)$ for all $z \in \mathbb{F}(\alpha_1)$; that is,

$$\Phi(\eta_1(\iota_1(b))) = \Phi(\iota_1(b)) = \eta_2(\iota(b)) = (\iota_1(b))^{\theta} = \theta(\iota_1(b)) = \iota_2(b) \text{ for all } b \in \mathbb{F},$$

(and $\Phi(\eta_1(\alpha_1)) = \Phi(\alpha_1) = \eta_2(\alpha_1) = \theta(\alpha_1) = \beta_1$).

□

We agree with the reader; this is an awful lot of book-keeping. On the bright side, it is mostly book-keeping and not theory. The main lines of the proof (after the base case of the induction) were as follows:

- find one root α_1 of $f(x)$ in $\mathbb{K}_1$, another root β_1 of $f(x)$ in $\mathbb{K}_2$;
- factor $f^{\iota_1}(x) = (x - \alpha_1)h(x)$ over $\mathbb{F}(\alpha_1)$, and use the isomorphism $\theta : \mathbb{F}(\alpha_1) \to \mathbb{F}(\beta_1)$ to obtain the factorisation $f^{\iota_2}(x) = (x - \beta_1)h^{\theta}(x)$;
- observe that $\mathbb{K}_1$ is a splitting field for $h(x)$ over $\mathbb{F}(\alpha_1)$ and that $\mathbb{K}_2$ is a splitting field of $h(x)$ over $\mathbb{F}(\beta_1)$ (this is where the book-keeping of the embeddings gets most hairy); and

- then use the fact that the remaining factor $h(x)$ has degree less than N, which allows you to apply the induction hypothesis to conclude that $\mathbb{K}_1$ and $\mathbb{K}_2$ are isomorphic.

Almost all of the work above went into keeping track of where the coefficients of the different factors were.

1.18. Remark. The previous result shows that any two splitting fields for a polynomial $f(x) \in \mathbb{F}[x]$ are isomorphic, and thus – *up to isomorphism* – we can speak about *the* splitting field of a polynomial $f(x) \in \mathbb{F}[x]$.

Given the construction of Φ from θ in the above result, it is not hard to see that with a bit more work, we can also ensure that

$$\Phi(\alpha_j) = \beta_j, \quad 1 \leq j \leq N.$$

That is, without loss of generality, we have that $\mathbb{K}_1 := \mathbb{F}(\alpha_1, \alpha_2, \ldots, \alpha_N)$ and $\mathbb{K}_2 := \mathbb{F}(\beta_1, \beta_2, \ldots, \beta_N)$ are isomorphic via an isomorphism that "fixes" $\mathbb{F}$ (in other words, sends $\iota_1(b)$ to $\iota_2(b)$ for all $b \in \mathbb{F}$ and which satsifies $\Phi(\alpha_j) = \beta_j$ for all j.

It is important to note that *when we do this* for any fixed $1 \leq j \leq N$, α_j and β_j *must be roots of the same irreducible factor* of $q(x)$. The reader may wish to review Example 1.6 (b) to see why this is the case.

1.19. Definition. *Let $\mathbb{F}$ be a field and let $(\mathbb{E}, \iota)$ be an extension field of $\mathbb{F}$. An element $\alpha \in \mathbb{E}$ is said to be* ***algebraic over*** *$\mathbb{F}$ if there exists a (non-zero) polynomial $q(x) \in \mathbb{F}[x]$ such that $q^\iota(\alpha) = 0$. Otherwise, we say that α is* ***transcendental*** *over $\mathbb{F}$.*

We say that $\mathbb{E}$ itself is ***algebraic over*** *$\mathbb{F}$ if every element of $\mathbb{E}$ is algebraic over $\mathbb{F}$. Otherwise we say that $\mathbb{E}$ is* ***transcendental*** *over $\mathbb{F}$.*

1.20. Examples.

(a) $\alpha := (\sqrt{2} + \sqrt{3})$ is algebraic over $\mathbb{Q}$. To see this, consider the extension field $(\mathbb{R}, \iota)$ of $\mathbb{Q}$ where $\iota(b) = b$ for all $b \in \mathbb{Q}$. Note that

$$(\sqrt{2} + \sqrt{3})^2 = 5 + 2\sqrt{6},$$

and thus

$$((\sqrt{2} + \sqrt{3})^2 - 5)^2 - 24 = 0.$$

In other words, α is a root of $q^\iota(x)$, where

$$q(x) = (x^2 - 5)^2 - 24 = x^4 - 10x^2 + 1 \in \mathbb{Q}[x].$$

We tend not to bother writing ι when it consists of the inclusion map; in this case we think of $\mathbb{F}$ as a subfield of $\mathbb{E}$, and an element $q(x) \in \mathbb{F}[x]$ as an element of $\mathbb{E}[x]$. This is especially true when $\mathbb{F} = \mathbb{Q}$ and $\mathbb{E} = \mathbb{R}$, or even a subfield of $\mathbb{R}$ that contains $\mathbb{Q}$.

This leads us to say that a real number α is algebraic if there exists $q(x) \in \mathbb{Q}[x]$ such that $q(\alpha) = 0$. Recall that this is equivalent to the condition that there exists $p(x) \in \mathbb{Z}[x]$ such that $p(\alpha) = 0$.

(b) Although it is *much, much* harder to show, both π^k and e^j are transcendental over $\mathbb{Q}$ for all $j, k \in \mathbb{N}$.

(c) $i := \sqrt{-1}$ is algebraic over $\mathbb{R}$; it is a root of $q(x) = x^2 + 1$. Indeed, we leave it as an exercise for the reader to show that $\mathbb{C}$ is an algebraic extension of $\mathbb{R}$.

(d) Are $\beta := \pi + e$ and $\gamma := e\pi$ algebraic or transcendental over $\mathbb{Q}$? An answer to this would certainly get you a Ph.D., since both questions remains open to this day. By Exercise 10, at least one of them is transcendental.

Recall that $\mathbb{F}(x)$ denotes the field of quotients of the integral domain $\mathbb{F}[x]$.

1.21. Theorem. *Let $(\mathbb{E}, \iota)$ be an extension field of a field $\mathbb{F}$.*

(a) *If $\alpha \in \mathbb{E}$ is algebraic over $\mathbb{F}$, then*

$$\mathbb{F}(\alpha) \simeq \mathbb{F}[x]/\langle q(x)\rangle,$$

where α is a root of $q^\iota(x)$ for some $q(x) \in \mathbb{F}[x]$ which satisfies

$$\deg(q(x)) = \min\{\deg(r(x)) : r(x) \in \mathbb{F}[x], r^\iota(\alpha) = 0\}.$$

(b) *If $\beta \in \mathbb{E}$ is transcendental over $\mathbb{F}$, then*

$$\mathbb{F}(\beta) \simeq \mathbb{F}(x).$$

Proof.

(a) Let us prove that if $q(x)$ is chosen as above, then $q(x)$ is irreducible. If we can do that, then the result follows immediately from Theorem 1.13 (a).

Indeed, if $q(x) = g(x)h(x)$ for some $g(x), h(x) \in \mathbb{F}[x]$ with

$$\max\{\deg(g(x)), \deg(h(x))\} < \deg(q(x)),$$

then either $g(\alpha) = 0$ or $h(\alpha) = 0$, otherwise $q(\alpha) \neq 0$.

But then this contradicts our choice of $q(x)$. Hence $q(x)$ is irreducible.

(b) We leave it as an assignment exercise for the reader to verify that the evaluation map

$$\begin{array}{rccc} \theta: & \mathbb{F}(x) & \to & \mathbb{F}(\beta) \\ & [(f(x), g(x))] & \mapsto & f(\alpha)(g(\alpha))^{-1} \end{array}$$

is a (well-defined!!) isomorphism.

□

1.22. Corollary. *Let $\mathbb{F}$ be a field and $(\mathbb{E}, \iota)$ be an extension field of $\mathbb{F}$. If $\alpha \in \mathbb{E}$ is algebraic over $\mathbb{F}$, then there exists a unique **monic** irreducible polynomial $q(x) \in \mathbb{F}[x]$ such that*

(a) *$q^\iota(\alpha) = 0$, and*
(b) *if $r(x) \in \mathbb{F}[x]$ and $r^\iota(\alpha) = 0$, then $q(x) \mid r(x)$ in $\mathbb{F}[x]$.*

We refer to the polynomial $q(x)$ above as the **minimal polynomial** for α over $\mathbb{F}$.

Proof. Let $\mathcal{R}_\alpha := \{0 \neq r(x) \in \mathbb{F}[x] \text{ and } r^\iota(\alpha) = 0\}$. Since $\alpha \in \mathbb{F}$ is algebraic, $\mathcal{R}_\alpha \neq \varnothing$, and thus

$$n := \min\{\deg(r(x)) : r(x) \in \mathcal{R}_\alpha\} < \infty.$$

Let $p(x) \in \mathcal{R}_\alpha$ be a polynomial of degree n. Multiplying $p(x)$ by the inverse p_n^{-1} of its leading coefficient yields a monic polynomial $q(x)$, also of degree n, such that $q^\iota(\alpha) = 0$.

If $r(x) \in \mathcal{R}_\alpha$, then $\deg(r(x)) \geq \deg(q(x))$. We can use the division algorithm to write

$$r(x) = f(x)q(x) + g(x)$$

for some $f(x), g(x) \in \mathbb{F}[x]$ with $g(x) = 0$ or $\deg(g(x)) < n$. Since $r^\iota(\alpha) = q^\iota(\alpha) = 0$, we conclude that $g^\iota(\alpha) = 0$. But then $g(x) = 0$, for otherwise we contradict the minimality of n.

Thus $q(x) \mid r(x)$ in $\mathbb{F}[x]$, as claimed.

□

1.23. Definition. *Let $\mathbb{F}$ be a field and $(\mathbb{E}, \iota)$ be an extension field of $\mathbb{F}$. Then $\mathbb{E}$ is a vector space over $\iota(\mathbb{F})$, and we define the **degree of** $\mathbb{E}$ over $\mathbb{F}$ to be the dimension of the vector space $\mathbb{E}$ over $\iota(\mathbb{F})$. We write*

$$[\mathbb{E} : \mathbb{F}] := \dim_{\iota(\mathbb{F})} \mathbb{E}.$$

1.24. Examples.

(a) $[\mathbb{C} : \mathbb{R}] = 2$, and $\{1, i\}$ is a basis for $\mathbb{C}$ over $\mathbb{R}$.
(b) $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$, and $\{1, \sqrt{2}\}$ is a basis for $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$.
(c) $[\mathbb{Q}(\sqrt[5]{2}) : \mathbb{Q}] = 5$, and $\{1, \sqrt[5]{2}, \sqrt[5]{4}, \sqrt[5]{8}, \sqrt[5]{16}\}$ is a basis for $\mathbb{Q}(\sqrt[5]{2})$ over $\mathbb{Q}$.
(d) $[\mathbb{R} : \mathbb{Q}] = \infty$. The proof of this is left as an exercise for the reader.

1.25. Theorem. *Let $(\mathbb{E}, \iota)$ be an extension field for $\mathbb{F}$ and suppose that*

$$[\mathbb{E} : \mathbb{F}] < \infty.$$

Then $\mathbb{E}$ is an algebraic extension of $\mathbb{F}$.

Proof. Set $n := [\mathbb{E} : \mathbb{F}] < \infty$, and let $\alpha \in \mathbb{E}$. Then $\{1, \alpha, \alpha^2, \ldots, \alpha^n\}$ has $n+1$ elements, so it must be linearly dependent over $\mathbb{F}^\iota = \iota(\mathbb{F})$. In other words, there exist $q_0, q_1, q_2, \ldots, q_{n+1} \in \mathbb{F}$ (not all equal to zero) such that

$$q_0^\iota + q_1^\iota \alpha + q_2^\iota \alpha^2 + \cdots + q_{n+1}^\iota \alpha^{n+1} = 0.$$

This is the statement that α is a root of the non-zero polynomial $q^\iota(x)$, where $q(x) = q_{n+1}x^{n+1} + q_n x^2 + \cdots + q_1 x + q_0 \in \mathbb{F}[x]$. Hence α is algebraic over $\mathbb{F}$.

Since $\alpha \in \mathbb{E}$ was arbitrary, $\mathbb{E}$ is an algebraic extension of $\mathbb{F}$.

□

1.26. The converse is false. Note that

$$\mathbb{K} := \mathbb{Q}(\sqrt{2}, \sqrt[3]{2}, \sqrt[4]{2}, \sqrt[5]{2}, \ldots)$$

is an algebraic extension of $\mathbb{Q}$, yet $[\mathbb{K} : \mathbb{Q}] = \infty$.

1.27. Theorem. *Let $\mathbb{F}$ be a field and $(\mathbb{E}, \iota)$ be an extension field of $\mathbb{F}$. Let $(\mathbb{K}, \eta)$ be an extension field of $\mathbb{E}$. Suppose that $[\mathbb{E} : \mathbb{F}] < \infty$ and that $[\mathbb{K} : \mathbb{E}] < \infty$. Then*

$$[\mathbb{K} : \mathbb{F}] = [\mathbb{K} : \mathbb{E}]\ [\mathbb{E} : \mathbb{F}].$$

Proof. Let $\mathfrak{K} := \{\kappa_1, \kappa_2, \ldots, \kappa_n\}$ be a basis for $\mathbb{K}$ over $\eta(\mathbb{E})$, and $\mathfrak{E} := \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ be a basis for $\mathbb{E}$ over $\iota(\mathbb{F})$. Observe that $\eta \circ \iota : \mathbb{F} \to \mathbb{K}$ is an embedding.

Next, set $\mathfrak{L} := \{\alpha_i^\eta \kappa_j : 1 \le i \le m, 1 \le j \le n\}$. We claim that $\mathfrak{L}$ is a basis for $\mathbb{K}$ over $\eta \circ \iota(\mathbb{F})$. If this is the case, then $|\mathfrak{L}| = mn$ shows that $[\mathbb{K} : \mathbb{F}] = mn$.

(I) Let $\beta \in \mathbb{K}$. Since $\mathfrak{K}$ is a basis for $\mathbb{K}$ over $\eta(\mathbb{E})$, there exist $\gamma_1, \gamma_2, \ldots, \gamma_n \in \mathbb{E}$ such that

$$\beta = \gamma_1^\eta\ \kappa_1 + \gamma_2^\eta\ \kappa_2 + \cdots + \gamma_n^\eta\ \kappa_n.$$

Temporarily fix $1 \le j \le n$. Since $\gamma_j \in \mathbb{E}$ and since $\mathfrak{E}$ is a basis for $\mathbb{E}$ over $\mathbb{F}^\iota$, there exist $y_{1,j}, y_{2,j}, \ldots, y_{m,j} \in \mathbb{F}$ such that

$$\gamma_j = y_{1,j}^\iota\, \alpha_1 + y_{2,j}^\iota\, \alpha_2 + \cdots + y_{m,j}^\iota\, \alpha_m.$$

Hence

$$\begin{aligned}\beta &= \sum_{j=1}^{n} \gamma_j^{\eta}\ \kappa_j \\ &= \sum_{j=1}^{n} \eta(\gamma_j)\ \kappa_j \\ &= \sum_{j=1}^{n} \eta\left(\sum_{i=1}^{m} \iota(y_{i,j})\alpha_i\right)\kappa_j \\ &= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} \eta\circ\iota(y_{i,j})\alpha_i^{\eta}\right)\kappa_j \\ &= \sum_{j=1}^{n}\sum_{i=1}^{m} y_{i,j}^{\eta\circ\iota}(\alpha_i^{\eta}\kappa_j).\end{aligned}$$

We conclude that $\operatorname{span}_{\eta\circ\iota(\mathbb{F})} \mathfrak{L} = \mathbb{K}$, and so $[\mathbb{K}:\mathbb{F}] \le mn = [\mathbb{K}:\mathbb{E}]\,[\mathbb{E}:\mathbb{F}]$.

(II) Next suppose that $y_{i,j} \in \mathbb{F}$, $1 \le i \le m,\ 1 \le j \le n$ satisfy

$$\sum_{j=1}^{n}\sum_{i=1}^{m} y_{i,j}^{\eta\circ\iota}(\alpha_i^{\eta}\kappa_j) = 0.$$

For each $1 \le j \le n$, define $\gamma_j := \sum_{i=1}^{m} y_{i,j}^{\iota}\alpha_i \in \mathbb{E}$. Then

$$0 = \sum_{j=1}^{n}\sum_{i=1}^{m} y_{i,j}^{\eta\circ\iota}(\alpha_i^{\eta}\kappa_j) = \left(\sum_{j=1}^{n} \gamma_j^{\eta}\kappa_j\right).$$

Since $\mathfrak{K}$ is a basis for $\mathbb{K}$ over $\eta(\mathbb{E})$, we may deduce that for each $1 \le j \le n$,

$$0 = \gamma_j^{\eta} = \eta(\gamma_j).$$

But η is injective and so for each $1 \le j \le n$, we have that

$$0 = \sum_{i=1}^{m} y_{i,j}^{\iota}\alpha_i.$$

Now $\mathfrak{E}$ is a basis for $\mathbb{E}$ over $\iota(\mathbb{F})$, and so $y_{i,j}^{\iota} = 0$ for all $1 \le i \le m$, and this for all $1 \le j \le n$. Hence $\mathfrak{L}$ is linearly independent over $\mathbb{F}$, and so it is a basis for $\mathbb{K}$ over $\mathbb{F}$, completing the proof of the Theorem.

□

1.28. Examples.

(a) $[\mathbb{Q}(\sqrt{2},\sqrt{3}):\mathbb{Q}] = 4 = [\mathbb{Q}(\sqrt{2},\sqrt{3}):\mathbb{Q}(\sqrt{2})]\ [\mathbb{Q}(\sqrt{2}):\mathbb{Q}]$.
A basis for this extension is $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$.

Supplementary Examples.

S9.1. Example. Suppose that $\mathbb{E}$ is an extension field for the field $\mathbb{F}$ with $\mathbb{F} \subseteq \mathbb{E}$, and that $[\mathbb{E} : \mathbb{F}] = 2$. Note that $\{1\}$ is linearly independent in $\mathbb{E}$, since $1 \neq 0$, and so $\{1\}$ can be extended to a basis $\mathfrak{B} := \{1, \alpha\}$ for $\mathbb{E}$ over $\mathbb{F}$.

It follows that if $\gamma \in \mathbb{E}$, then there exists $a, b \in \mathbb{F}$ such that $\gamma = a \cdot 1 + b \cdot \alpha$. In particular, $\alpha^2 = a_0 \cdot 1 + b_0 \cdot \alpha$ for some $a_0, b_0 \in \mathbb{F}$.

Thus

$$\begin{aligned}\gamma^2 &= a^2 \cdot 1 + 2ab\alpha + b^2\alpha^2 \\ &= a^2 \cdot 1 + 2ab\alpha + b^2(a_0 \cdot 1 + b_0 \cdot \alpha) \\ &= (a^2 + a_0 b^2) \cdot 1 + (2ab + b_0 b^2)\alpha.\end{aligned}$$

It follows that $\{1, \gamma, \gamma^2\}$ are linearly dependent over $\mathbb{F}$, which implies that γ satisfies a polynomial of degree at most 2 over $\mathbb{F}$.

S9.2. Example. Let $0 \neq p(x) = p_2x^2 + p_1x + p_0 \in \mathbb{Z}_2[x]$ be an irreducible polynomial of degree 2. Then $p_2 = 1$, otherwise $\deg(p(x)) < 2$. Also, in order to be irreducible, $p(x)$ can not have roots in $\mathbb{Z}_2$, so

$$p(0) = p_1 \cdot 0 + p_0 = p_0 \neq 0.$$

Thus $p_0 = 1$. Also,

$$p(1) = p_2 + p_1 + p_0 = 1 + p_1 + 1 = p_1 \neq 0,$$

so $p_1 = 1$.

Hence $p(x) = x^2 + x + 1$. That is, there exists a unique irreducible polynomial of degree 2 over $\mathbb{Z}_2$.

What does this say about extension fields over $\mathbb{Z}_2$ of degree 2?

S9.3. Example. The polynomial $q(x) = x^4 + x + 1 \in \mathbb{Z}_2[x]$ is irreducible. Indeed, it has no roots in $\mathbb{Z}_2$, and so the only possible factorisation into polynomials of lower degree would have to involve two polynomials $r(x)$ and $s(x) \in \mathbb{Z}_2[x]$, each of degree 2.

Writing $r(x) = x^2 + r_1x + r_0$ and $s(x) = x^2 + s_1x + s_0$, we see that $q(x) = r(x)s(x)$ implies that $r_0s_0 = 1$ implies that $r_0 = s_0 = 1$.

Thus

$$\begin{aligned}q(x) = r(x)s(x) &= (x^2 + r_1x + 1)(x^2 + s_1x + 1) \\ &= x^4 + (r_1 + s_1)x^3 + (1 + r_1s_1 + 1)x^2 + (r_1 + s_1)x + 1,\end{aligned}$$

from which we deduce that $r_1 + s_1 = 0$, $r_1s_1 = 1$, and $r_1 + s_1 = 1$. The first and third equation clearly contradict each other, so no such factorisation is possible.

It follows that $\mathbb{F} := \mathbb{Z}_2[x]/\langle x^4 + x + 1\rangle$ is a field, and that a general element of $\mathbb{F}$ is of the form

$$b_3x^3 + b_2x^2 + b_1x + b_0 + \langle x^4 + x + 1\rangle,$$

for some $b_0, b_1, b_2, b_3 \in \mathbb{Z}_2$. There are 16 such possibilities, and so $|\mathbb{F}| = 16 = 2^4$.

S9.4. Example. Let $p \in \mathbb{N}$ be prime. Given any integer $n \geq 1$, it can be shown that there exists an irreducible polynomial of degree n over $\mathbb{Z}_p$, and thus – by generalising the argument presented in Example 1.26, there exists a field with p^n elements.

S9.5. Example. It can be shown that the polynomial $q(x) = x^4+1$ is irreducible over $\mathbb{Z}$, but reducible over $\mathbb{Z}_p$ for all primes p. In other words – this is the worst case scenario in terms of trying to use the Mod-p Test, since that Test will fail for all primes p.

Here's a proof that $q(x)$ is irreducible over $\mathbb{Z}$. Indeed, $q(m) \geq 1$ for all $m \in \mathbb{Z}$ implies that q has no roots in $\mathbb{Z}$, and hence the only possible factorisation is as a product of quadratic terms. Suppose such a factorisation exists. Since the product of the leading terms has a coefficient of 1, they must both be 1 or both be -1. By multiplying both terms by -1 in the second case, we may assume without loss of generality that each quadratic term is a monic polynomial.

Let $r(x) = x^2 + r_1x + r_0$ and $s(x) = x^2 + s_1x + s_0$. By considering the constant term of $q(x) = r(x)s(x)$, we see that either $r_0 = s_0 = 1$ or $r_0 = s_0 = -1$. Now

$$q(x) = r(x)s(x) = x^4 + (r_1 + s_1)x^3 + (r_0 + s_0 + r_1s_1)x^2 + (r_0s_1 + r_1s_0)x + r_0s_0.$$

Thus we must solve

- $r_1 + s_1 = 0$;
- $r_0 + s_0 + r_1s_1 = 0$;
- $r_0s_1 + r_1s_0 = 0$;
- $r_0 = s_0 \in \{-1, 1\}$.

From the first and second equations, we see that $r_1s_1 = -r_1^2 \in \{-2, 2\}$, which is unsolvable with $r_1 \in \mathbb{Z}$.

S9.6. Example. Here's a more sophisticated proof of the fact that $q(x) = x^4+1$ is irreducible over $\mathbb{Z}$. Suppose that

$$q(x) = r(x)s(x)$$

for some $r(x), s(x) \in \mathbb{Z}[x]$.

Then

$$(x+1)^4 + 1 = q(x+1) = r(x+1)s(x+1),$$

showing that $q(x+1)$ can be factored in $\mathbb{Z}[x]$. But

$$q(x+1) = (x+1)^4 + 1 = x^4 + 4x^3 + 6x^2 + 4x + 2$$

cannot be factored in $\mathbb{Z}[x]$ by Eisenstein's Criterion, applied with $p = 2$.

S9.7. Example. Let $\mathbb{F}$ be a field and $q(x) \in \mathbb{F}[x]$ be a monic irreducible polynomial of degree d over $\mathbb{F}$. As we have seen, if $\mathbb{E} = \mathbb{F}[x]/\langle q(x)\rangle$, then $\mathbb{E}$ is a field, and $\varphi : \mathbb{F} \to \mathbb{E}$ defined by $\varphi(b) = b + \langle q(x)\rangle$ is an embedding of $\mathbb{F}$ into $\mathbb{E}$, so that $\mathbb{E}$ is an extension field of $\mathbb{F}$.

Let $\alpha := x + \langle q(x)\rangle$. Since

$$\mathbb{E} = \{r_{d-1}x^{d-1} + r_{d-2}x^{d-2} + \cdot + r_1x + r_0 + \langle q(x)\rangle : r_j \in \mathbb{F}, 0 \leq j \leq d-1\},$$

we see that $\mathfrak{B} := \{1, \beta, \beta^2, \ldots, b^{d-1}\}$ is a basis for $\mathbb{E}$ over $\mathbb{F}$.

S9.8. Example. Note that π^2 is transcendental over $\mathbb{Q}$, but that π^2 is algebraic over $\mathbb{Q}(\pi)$. Thus *transcendental* by itself is "not a thing"; you must be transcendental (or algebraic, as the case may be) over a particular field.

S9.9. Example. Let $\mathbb{F}$ be a field of characteristic p, where $p \in \mathbb{N}$ is a prime number. If $\mathbb{E}$ is an extension field of $\mathbb{F}$, then $\text{CHAR}(\mathbb{E}) = p$.

To see this, note that $\mathbb{E}$ is a vector space over $\mathbb{F}$. Let $\mathfrak{B} = \{e_\lambda\}_{\lambda\in\Lambda}$ be a basis for $\mathbb{E}$ over $\mathbb{F}$, and $y \in \mathbb{E}$. Then there exist $b_1, b_2, \ldots, b_k \in \mathbb{F}$, $\lambda_1, \lambda_2, \ldots, \lambda_k \in \Lambda$ such that

$$y = b_1 e_{\lambda_1} + b_2 e_{\lambda_2} + \cdots + b_k e_{\lambda_k}.$$

Hence

$$\begin{aligned} py &= (pb_1)e_{\lambda_1} + (pb_2)e_{\lambda_2} + \cdots + (pb_k)e_{\lambda_k} \\ &= 0e_{\lambda_1} + 0e_{\lambda_2} + \cdots + 0e_{\lambda_k} \\ &= 0. \end{aligned}$$

It follows that the characteristic of $\mathbb{E}$ divides p, hence equals p, as the latter is prime.

A similar statement holds if $\mathbb{F}$ has characteristic zero. The proof is left to the reader.

S9.10. Example. Suppose that $\mathbb{F}$ is a field and that $g(x) \in \mathbb{F}[x]$. Recall from Example 3.3.1 that if $g(x)$ has a root r of multiplicity at least two in $\mathbb{F}$, then $(x - r) \mid g'(x)$, the *derivative* of $g(x)$.

Stated another way, if $\text{GCD}(g(x), g'(x)) = 1$, then $g(x)$ has distinct roots in $\mathbb{F}[x]$.

Suppose that $\text{CHAR}(\mathbb{F}) = p$, a prime, $n \in \mathbb{N}$, and that $q(x) = x^{p^n} - x$. Then $q'(x) = p^n x^{p^n-1} - 1 = 0 - 1 = -1 \in \mathbb{F}[x]$, and so $\text{GCD}(q(x), q'(x)) = 1$. Let $\mathbb{K}$ be a splitting field for $q(x)$ over $\mathbb{F}$. In $\mathbb{K}$, $q(x)$ will have p^n *distinct* roots, say $\mathbb{E} := \{\alpha_1, \alpha_2, \ldots, \alpha_{p^n}\}$.

Curious, isn't it, that we should use $\mathbb{E}$ to denote a set of roots, instead of a field? But in fact, if $a, b \in \mathbb{E}$, then $a^{p^n} = a$ and $b^{p_n} = b$, so $(ab)^{p^n} = a^{p^n} b^{p^n} = ab \in \mathbb{E}$ and (by the binomial expansion, noting that each coefficient in the expansion other than the first and the last is divisible by p^n, and therefore by p,

$$(b - a)^{p^n} = b^{p^n} - a^{p^n} = b - a,$$

so that $b - a \in \mathbb{E}$. In other words, $\mathbb{E}$ is a subring of $\mathbb{K}$. If $0 \neq a \in \mathbb{E}$, then $a^{p^n} = a$ implies that $a^{p^{n-1}} = 1$ in $\mathbb{K}$, so that $a^{-1} = a^{p^{n-2}} \in \mathbb{E}$. Thus $\mathbb{E}$ is in fact a subfield of $\mathbb{K}$, and $\mathbb{E}$ has exactly p^n elements.

That is, for any prime p and any $n \in \mathbb{N}$, there exists a field with p^n elements.

Appendix

A9.1. In many texts, an extension field $\mathbb{E}$ of a field $\mathbb{F}$ is defined as a field which *contains* $\mathbb{F}$. Technically, this is false. What the authors really mean is that $\mathbb{E}$ contains an *isomorphic copy*, say $\mathbb{K}$ of $\mathbb{F}$, and that they are identifying $\mathbb{K}$ with $\mathbb{F}$.

Suppose we consider the set $\mathbb{E} := \{\begin{bmatrix} x & y \\ -y & x \end{bmatrix} : x, y \in \mathbb{R}\} \subseteq \mathbb{M}_2(\mathbb{R})$. Recall that $\mathbb{E}$ is isomorphic to $\mathbb{C}$, the field of complex numbers. It is readily seen that the map

$$\begin{array}{rccc} \varphi : & \mathbb{R} & \to & \mathbb{E} \\ & x & \to & \begin{bmatrix} x & 0 \\ 0 & x \end{bmatrix} \end{array}$$

is an injective homomorphism of $\mathbb{R}$ into $\mathbb{E}$. Now, $\mathbb{K} := \varphi(\mathbb{R})$ is a field which, *in terms of all of its field properties*, behaves exactly the same as $\mathbb{R}$. Surely we can agree that elements of $\mathbb{E}$ are not actually complex numbers, even though they behave exactly the same way (again, in terms of their field properties) as complex numbers. If we were to agree amongst friends to *identify* $\mathbb{K}$ and $\mathbb{R}$ – in other words, if we were to agree not to quibble and just say that we are going to treat $\mathbb{K}$ and $\mathbb{R}$ as the same thing – then we could say that $\mathbb{E}$ contains $\mathbb{R}$.

We would be lying, of course, but if we had sufficient mathematical sophistication to understand that we are lying, to understand the nature of our lie, and to understand how to eliminate the lie by reintroducing the map φ and painstakingly keeping track of the embedding, then we could live fruitful and productive lives and never talk about this again.

This is precisely what these sinful authors to which I have earlier alluded are doing. When they define an extension field $\mathbb{E}$ of a field $\mathbb{F}$ as a field which contains $\mathbb{F}$, they are agreeing to conflate the copy $\mathbb{K} \subseteq \mathbb{E}$ of $\mathbb{F}$ and $\mathbb{F}$ itself.

So what does this have to do with Definition 1.7? If we were to adopt the strategy of those naughty authors referenced above, then we could simply say that $\mathbb{E}$ is a splitting field for a polynomial $q(x) \in \mathbb{F}[x]$ if, whenever $\mathbb{K}$ is an extension of $\mathbb{F}$ and $q(x)$ splits in $\mathbb{K}$, then $\mathbb{K} \supseteq \mathbb{E}$.

A9.2. Although we shall not prove it here, it is known that the number of irreducible polynomials of degree n over a field $\mathbb{F}_q$, where $q = p^m$ for some prime p and positive integer m is given by

$$N(q, n) := \frac{1}{n} \sum_{d|n} \mu(d) q^{n/d},$$

where μ is the so-called **Möbius function**. More specifically, $\mu(d)$ is the sum of the *primitive* d^{th} *roots of unity* in $\mathbb{C}$, and takes on values in $\{-1, 0, 1\}$. In particular, $N(q, n) \geq 1$ for all q, n.

A9.3. There is still some time left to continue asking questions while reading the text. For instance, how many algebraic elements over $\mathbb{Q}$ are there in $\mathbb{R}$? Stated more precisely, what is the cardinality of the set

$$\Lambda := \{\alpha \in \mathbb{R} : \alpha \text{ is algebraic over } \mathbb{Q}\}?$$

Is Λ a field?

Exercises for Chapter 9

Exercise 9.1.
Prove that $\mathbb{Q}[x]/\langle x^2+1\rangle$ is isomorphic to $\mathbb{Q}[i]$.

Exercise 9.2.
Show that $\alpha = \sqrt{\frac{1}{3}+\sqrt{7}}$ is algebraic over $\mathbb{Q}$, and find its minimal polynomial.

Exercise 9.3.
Show that $\beta = \sqrt{\sqrt[3]{2}-i}$ is algebraic over $\mathbb{Q}$, and find its minimal polynomial.

Exercise 9.4.
Find a basis for $\mathbb{Q}(\sqrt{3},\sqrt{6})$ over $\mathbb{Q}$. What is the degree of this extension?

Exercise 9.5.
Find a basis for $\mathbb{Q}(\sqrt{2},\sqrt{6}+\sqrt{10})$ over $\mathbb{Q}(\sqrt{3}+\sqrt{5})$. What is the degree of this extension?

Exercise 9.6.
Prove that $\mathbb{Z}_2[x]/\langle x^3+x^2+1\rangle$ is a field with 8 elements, and construct a multiplication table for this field.

Exercise 9.7.
Is e algebraic over $\mathbb{Q}(e^7)$?

Exercise 9.8.
Prove or disprove that $\mathbb{Q}(\sqrt{7}) \simeq \mathbb{Q}(\sqrt{19})$.

Exercise 9.9.
Let $\mathbb{F}$ be a field of characteristic 5, and let $b \in \mathbb{F}$. Prove that $p(x) = x^5 - b$ is either irreducible over $\mathbb{F}$, or splits in $\mathbb{F}$.

Exercise 9.10.
Let $\alpha, \beta \in \mathbb{R}$ be transcendental over $\mathbb{Q}$. Prove that either $\alpha\beta$ or $\alpha+\beta$ is transcendental. (Recall from Example 1.20 (d) that we do not know whether $\pi+e$ and πe are algebraic or transcendental over $\mathbb{Q}$. This shows that at least one of the two is transcendental, but does not tell us which.)

Exercise 9.11.
Recall that a set X is said to be **denumerable** if there exists a bijection $\varphi : \mathbb{N} \to X$. (We say that X is **countable** if X is either finite or denumerable.)

Let $(\mathbb{R}, \iota)$ be the standard extension of $\mathbb{Q}$ (i.e. $\iota(q) = q$ for all $q \in \mathbb{Q}$). Prove that the set

$$\Omega := \{\alpha \in \mathbb{R} : \alpha \text{ is algebraic over } \mathbb{Q}\}$$

is denumerable.

Exercise 9.12.
Prove that the set Ω of algebraic real numbers defined in Exercise 9.11 is a field.

Exercise 9.13.
Let $\mathbb{K} := \mathbb{Q}(\sqrt{2}, \sqrt[3]{2}, \sqrt[4]{2}, \sqrt[5]{2}, \ldots)$ be the extension field of $\mathbb{Q}$ from Paragraph 9.1.26. Prove that $\mathbb{K}$ is algebraic over $\mathbb{Q}$.

CHAPTER 10

Straight-edge and Compasses constructions

We are all here on earth to help others; what on earth the others are here for I don't know.

W.H. Auden

1. An ode to Wantzel

1.1. In this Chapter we shall see how **Pierre Laurent Wantzel** (I don't know why, but I like him already) managed to use field theory to prove lots of stuff in geometry. So yes, that's a thing now.

1.2. The story begins, as they say, with the ancient Greeks. By that we do not mean that the Greeks themselves were ancient, but rather that we are referring to Greeks who lived in ancient times. More specifically, we are interested in **Plato**, born circa 425 BC. An Athenian philosopher, he held that the only "perfect" geometrical forms were straight lines and circles, and as such, he was only interested in geometrical constructions that could be accomplished with a straight-edge and a pair of compasses. [1]

Although he was not a mathematician, he nevertheless taught mathematics, and one of his students, **Eudoxos of Cnidus**, is considered to have been one of the greatest mathematicians of those days. In fact, much of what is to be found in *Elements*, **Euclid's** famous[2] treatise on mathematics - written circa 300 BC – is actually the work of Eudoxos.

[1] The Cambridge Dictionary asserts that *a compass* is "a device for finding direction with a needle that can move easily and that always points to magnetic north", whereas (a pair of) *compasses* refers to "a V-shaped device that is used for drawing circles or measuring distances on maps". You could see where a pair of compasses might have been more useful than a compass when performing geometrical constructions.

[2] The word "famous" is a curious beast when applied in a mathematical context. In this case, we feel justified in using this expression because Euclid's *Elements* is well-known *outside* of the mathematical world. In fact, if one were to walk up to some random individual on the street and ask them to name a famous mathematical text, then it is hard to think of another text the random person might cite. Of course, it is far more likely that the person to whom one is posing this question would ignore one, or – in a worst case scenario – would visit some extreme violence upon one's person, but that is an altogether different matter. It is worth bearing in mind that being

We being with the first of the postulates stated in Euclid's *Elements*.

1.3. Euclid's postulates. Let $\mathcal{P} \subseteq \mathbb{R}^2$ be a set containing at least two distinct points. We shall consider the following two geometric operations, hereafter referred to as **admissible operations**:

(I) Given two distinct points A and B in $\mathcal{P}$, there is a unique line which contains them, and hence a unique line segment connecting A to B.

(II) Given points $\mathrm{A}, \mathrm{B}, \mathrm{C}$ in $\mathcal{P}$ with $\mathrm{A} \neq \mathrm{B}$, we can construct a circle centred at C with radius equal to the distance $d(\mathrm{A}, \mathrm{B})$ from A to B. (Here we are using the usual **Euclidean distance** between two points $\mathrm{A} = (a_1, a_2)$ and $\mathrm{B} = (b_1, b_2)$ in $\mathbb{R}^2$, namely $d(\mathrm{A}, \mathrm{B}) := \sqrt{|a_1 - b_1|^2 + |a_2 - b_2|^2}$.)

The Greeks of Plato's day referred to the construction of a geometric figure using only a straight-edge and compasses (and the two admissible operations above) as the **plane method**. We should think of these operations as telling us which lines and circles we are allowed to "draw".

1.4. Euclid elucidated a number of these straight-edge-and-compasses constructions in *Elements*. More specifically, he demonstrated (amongst a great many other things) that with only the plane method, it is possible to find the midpoint of a given line segment, or to bisect a given angle, or to construct – given a line L and a point X not on that line – a second line, parallel to L, and passing through the point X. We shall demonstrate some of these constructions in the Appendix to this chapter.

Our present goal is to try to describe which geometric figures may be constructed using the plane method, starting from a minimal set $\mathcal{P}$ consisting of two distinct points. By rotating, translating and scaling if necessary, we may assume without loss of generality that those two points are $\mathrm{P}_0 = (0, 0)$ and $\mathrm{P}_1 = (1, 0)$. Thus, throughout the remainder of this chapter

$$\textit{we shall always assume that } \{\mathrm{P}_0 = (0, 0), \mathrm{P}_1 = (1, 0)\} \subseteq \mathcal{P} \subseteq \mathbb{R}^2.$$

In order to describe these geometric figures, we must first properly define what it means to "construct" such a geometric object.

It is clear, for example, that since we can construct line segments which connect pairs of distinct points using the plane method, we can construct a square S_1 whose area is 1 provided that the points $\mathrm{Q} = (1, 0)$ and $\mathrm{R} = (1, 1)$ can be "constructed" from the set $\mathcal{P}$ using the plane method – the points $\mathrm{P}_0 = (0, 0)$ and $\mathrm{P}_1 = (1, 0)$ already lying in $\mathcal{P}$ by hypothesis. To do this, we simply connect P_0 to P_1, P_1 to R, R to Q, and Q to P_0 via line segments.

We can similarly "construct" a square S_2 whose area is twice that of S_1 provided that we can somehow "construct" the points $\mathrm{X} = (\sqrt{2}, 0)$, $\mathrm{Y} = (0, \sqrt{2})$ and Z =

"famous" in mathematics today means being known to literally hundreds of other mathematicians studying the same branch of mathematics, and to absolutely no one who actually matters.

$(\sqrt{2}, \sqrt{2})$ from $\mathcal{P}$ using the plane method, and "constructing" an arbitrary polygon can be done provided that we can somehow "construct" the vertices of that polygon, after which we may the appropriate vertices with line segments, which the plane method allows us to draw.

Thus by "constructing" a more general geometric figure using the plane method, we shall mean constructing the points which define the object, and connecting them using the lines, line segments and circles prescribed by the plane method. Of course, being able to do so relies upon our having a rigorous definition of which *points* we can "construct" from $\mathcal{P}$ using the plane method.

1.5. Definition. *Let $\mathcal{P} \subseteq \mathbb{R}^2$ be a set containing the points $\mathrm{P}_0 = (0,0)$ and $\mathrm{P}_1 = (1,0)$. A point $\mathrm{Q} \in \mathbb{R}^2$ is said to be **constructible in one step** from $\mathcal{P}$ if it is a point of intersection of two non-parallel lines, two distinct circles, or a line and a circle obtained from admissible operations from $\mathcal{P}$.*

*We say that $\mathrm{R} \in \mathbb{R}^2$ is **constructible from** $\mathcal{P}$ if there exist finitely many points $\mathrm{Q}_0, \mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{R} = \mathrm{Q}_n \in \mathbb{R}^2$ such that each for each $1 \leq k \leq n$, Q_k is constructible in one step from $\mathcal{P} \cup \{\mathrm{Q}_0, \mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_{k-1}\}$.*

*A real number $a \in \mathbb{R}$ is said to be **constructible from** $\mathcal{P}$ if the point $\mathrm{A} := (a,0)$ is constructible from $\mathcal{P}$, and in the case where $\mathcal{P} = \{\mathrm{P}_0, \mathrm{P}_1\}$, we abbreviate this to the expression $a \in \mathbb{R}$ **is a constructible real number**. We shall denote by $\Gamma_{\mathcal{P}}$ the set of all real numbers which are constructible from $\mathcal{P}$, and by Γ° the set of all constructible real numbers, so that $\Gamma^\circ = \Gamma_{\{\mathrm{P}_0, \mathrm{P}_1\}}$.*

1.6. Remark. Suppose that $\mathcal{P} \subseteq \mathbb{R}^2$ is a set containing the points $\mathrm{P}_0 = (0,0)$ and $\mathrm{P}_1 = (1,0)$, and that $\mathrm{Q} = (q_1, q_2)$ and $\mathrm{R} = (r_1, r_2)$ are constructible from $\mathcal{P}$. Then the x-axis, i.e. the set $\{(x,0) : x \in \mathbb{R}\}$, corresponds to the line through P_0 and P_1, which we may draw using the first admissible operation from Paragraph 1.3 above. (Note: although the line itself may be drawn, this is not the same as saying that every point $(x,0)$ on that line may be constructed. Constructible points come from *intersections* of lines and circles with lines and circles.)

Moreover, the second admissible operation from Euclid's postulates allows us to construct a circle of radius $\varrho := d(\mathrm{Q}, \mathrm{R})$ centred at $(0,0)$, which clearly must intersect the x-axis at the point $(\varrho, 0)$. It follows that $\varrho \in \mathbb{R}$ is constructible from $\mathcal{P}$, i.e. $\varrho \in \Gamma_{\mathcal{P}}$. Conversely, if $\varrho \in \Gamma_{\mathcal{P}}$, then the point $(\varrho, 0) \in \mathbb{R}^2$ is constructible from $\mathcal{P}$, and $\varrho = d((0,0), (\varrho, 0))$.

Thus we may also view those real numbers constructible from $\mathcal{P}$ as the set of all possible distances between points in $\mathbb{R}^2$ which are themselves constructible from $\mathcal{P}$.

This is not the only interesting way to interpret the set $\Gamma_{\mathcal{P}}$. Indeed, suppose that the point $\mathrm{R} = (r_1, r_2)$ is constructible from $\mathcal{P}$.

- as we have just seen, we may construct the x-axis from $\mathcal{P}$ by the first admissible operation.
- By Paragraph A10.1, we may also construct the y-axis.

- By Paragraph A10.2, we may draw a line parallel to the x-axis passing through R. This line will intersect the y-axis at $(0, r_2)$. The circle centred at P_0 with radius $r_2 = d((0,0),(0,r_2))$ intersects the x-axis at $(r_2, 0)$. Thus $r_2 \in \Gamma_{\mathcal{P}}$.
- By Paragraph A10.2, we may also draw a line parallel to the y-axis passing through R. This line will intersect the x-axis at $(r_1, 0)$, and thus $r_1 \in \Gamma_{\mathcal{P}}$ as well.

In other words, given a point $R = (r_1, r_2) \in \mathbb{R}^2$ which is constructible from $\mathcal{P}$, the points $(r_1, 0)$ and $(0, r_2)$ are also constructible from $\mathcal{P}$, and therefore $r_1, r_2 \in \Gamma_{\mathcal{P}}$.

Conversely, if r_1 and $r_2 \in \Gamma_{\mathcal{P}}$ are real numbers constructible from $\mathcal{P}$, then

- by definition, $(r_1, 0) \in \mathcal{P}$.
- By Paragraph A10.1, we may draw a line L perpendicular to the x-axis passing through $(r_1, 0)$. Since the point $(r_2, 0)$ is constructible, we may draw a circle C centred at $(r_1, 0)$ with radius $\varrho := |r_2|$ which intersects the line L at the points $(r_1, -|r_2|)$ and $(r_1, |r_2|)$. Clearly R is one of those two points.

Thus the point $R := (r_1, r_2)$ is constructible from $\mathcal{P}$.

From this we see that the

$$\Gamma_{\mathcal{P}} = \{\alpha, \beta \in \mathbb{R} : Q := (\alpha, \beta) \in \mathbb{R}^2 \text{ is constructible from } \mathcal{P}\}.$$

2. Enter fields

2.1. This is where things get interesting for those of us who have spent the past few weeks learning about rings and fields. Our immediate goal is to prove that if $\{P_0, P_1\} \subseteq \mathcal{P}$, then $\Gamma_{\mathcal{P}}$ is a field. First we observe that if $\alpha \in \Gamma_{\mathcal{P}}$, then the circle of radius $|\alpha|$ centred at P_0 intersects the x-axis at $(-|\alpha|, 0)$ and at $(|\alpha|, 0)$, In other words, $\alpha \in \Gamma_{\mathcal{P}}$ implies that $-\alpha \in \Gamma_{\mathcal{P}}$ as well.

2.2. Proposition. *Let $\alpha, \beta \in \Gamma_{\mathcal{P}}$. Then $\alpha + \beta$ and $\alpha - \beta$ lie in $\Gamma_{\mathcal{P}}$.*

Proof. First note that $|\alpha|, |\beta| \in \Gamma_{\mathcal{P}}$ by the comments of Paragraph 2.1.

Set $A := (|\alpha|, 0)$. The second admissible operation allows us to draw a circle of radius $|\beta|$ centred at A, which intersects the x-axis at the points $(|\alpha| - |\beta|, 0)$ and $(|\alpha| + |\beta|, 0)$. Hence $|\alpha| - |\beta|, |\alpha| + |\beta| \in \Gamma_{\mathcal{P}}$.

But then $-|\alpha| + |\beta|, -|\alpha| - |\beta| \in \Gamma_{\mathcal{P}}$, again, by Paragraph 2.1. Since

$$\{\alpha + \beta, \alpha - \beta\} \subseteq \{|\alpha| - |\beta|, |\alpha| + |\beta|, -|\alpha| + |\beta|, -|\alpha| - |\beta|\}$$

for all real numbers α, β, we conclude that $\alpha + \beta, \alpha - \beta \in \Gamma_{\mathcal{P}}$.

□

2.3. Proposition. *Let $\alpha, \beta \in \Gamma_{\mathcal{P}}$ with $\alpha \neq 0$. Then*

$$\alpha\beta \in \Gamma_{\mathcal{P}} \text{ and } \frac{\beta}{\alpha} \in \Gamma_{\mathcal{P}}.$$

Proof. First note that by the argument in Paragraph 2.1, it suffices to consider the case where $\alpha, \beta > 0$. If $\alpha = 1$, there is nothing to prove, and so we henceforth assume that $\alpha \neq 1$.

(a) Let $\mathrm{A} = (\alpha, 0)$ and $\mathrm{B} = (0, \beta)$, which are constructible from $\mathcal{P}$.
(b) Using Paragraph A.10.3, we can draw a line L through the point $\mathrm{P}_1 = (1, 0)$ which is parallel to the line through A and B. That line will intersect the y-axis at the point $\mathrm{C} := (0, \gamma)$, and thus $\gamma \in \mathbb{R} \in \Gamma_{\mathcal{P}}$.

The triangles $(\mathrm{P}_0\,\mathrm{A}\,\mathrm{B})$ and $(\mathrm{P}_0\,\mathrm{P}_1\,\mathrm{C})$ are similar, and thus

$$\frac{\alpha}{1} = \frac{d(\mathrm{P}_0, \mathrm{A})}{d(\mathrm{P}_0, \mathrm{P}_1)} = \frac{d(\mathrm{P}_0, \mathrm{B})}{d(\mathrm{P}_0, \mathrm{C})} = \frac{\beta}{\gamma}.$$

That is, $\dfrac{\beta}{\alpha} = \gamma \in \Gamma_{\mathcal{P}}$.

Since $1 \in \Gamma_{\mathcal{P}}$, the above argument with $\beta = 1$ implies that $\gamma_1 := \dfrac{1}{\alpha} \in \Gamma_{\mathcal{P}}$ and therefore that

$$\alpha\beta = \frac{\beta}{\gamma_1} \in \Gamma_{\mathcal{P}}.$$

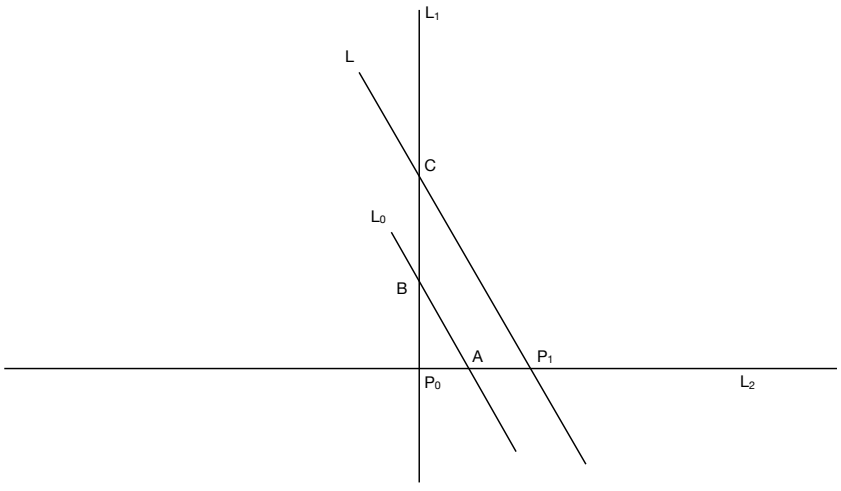

Figure 10.0

□

2.4. Theorem. *Suppose that* $\{\mathrm{P}_0 = (0,0), \mathrm{P}_1 = (1,0)\} \subseteq \mathcal{P} \subseteq \mathbb{R}^2$. *Then*

$$\mathbb{Q} \subseteq \Gamma_{\mathcal{P}} \subseteq \mathbb{R}$$

is a field. Thus

$$\Gamma_{\mathcal{P}} = \mathbb{Q}(\{\alpha, \beta : (\alpha, \beta) \in \mathcal{P}\}).$$

Proof. That $\Gamma_{\mathcal{P}} \subseteq \mathbb{R}$ is clear from the definition.

Note that $\mathcal{P}$ has at least two distinct points, and that $\mathrm{P}_1 = (1,0) \in \mathcal{P}$ implies that $1 \in \Gamma_{\mathcal{P}} \neq \varnothing$. By Proposition 2.2 and Proposition 2.3, we see that α, $\beta \in \Gamma_{\mathcal{P}}$ implies that $\alpha - \beta$ and $\alpha\beta$ lie in $\Gamma_{\mathcal{P}}$. By the Subring Test, $\Gamma_{\mathcal{P}}$ is a unital subring. Since any unital subring of $\mathbb{R}$ contains $\mathbb{Z}$, $\Gamma_{\mathcal{P}}$ contains $\mathbb{Z}$.

Since $1 \in \Gamma_{\mathcal{P}}$ and $\beta \neq 0$ implies that $\dfrac{1}{\beta} \in \Gamma_{\mathcal{P}}$ by Proposition 2.3, we see that in fact, $\Gamma_{\mathcal{P}}$ is a field. But $\mathbb{Z} \subseteq \Gamma_{\mathcal{P}}$, and thus $\mathbb{Q} \subseteq \Gamma_{\mathcal{P}}$ as well.

Finally, by the comments in Remark 1.6, we know that $\Gamma_{\mathcal{P}} = \{\alpha, \beta : (\alpha, \beta) \in \mathcal{P}\}$. Since we now know that $\Gamma_{\mathcal{P}}$ is a field which also contains $\mathbb{Q}$, it must be the smallest field which contains $\mathbb{Q}$ and $\{\alpha, \beta : (\alpha, \beta) \in \mathcal{P}\}$. In other words,

$$\Gamma_{\mathcal{P}} = \mathbb{Q}(\{\alpha, \beta : (\alpha, \beta) \in \mathcal{P}\}).$$

□

2.5. Corollary. *The set* Γ° *of constructible real numbers is a subfield of* $\mathbb{R}$ *that contains* $\mathbb{Q}$.

2.6. Remark. Although we have determined that the set $\Gamma^\circ = \Gamma_{\{\mathrm{P}_0, \mathrm{P}_1\}}$ of constructible real numbers is a subfield of $\mathbb{R}$ that contains $\mathbb{Q}$, we still don't know much about it. The next result establishes a relation between straight-edge and compasses constructions and field extensions, which will be the key to understanding Γ°, and to establishing the impossibility of certain classical geometric constructions in the next section of this Chapter.

2.7. Theorem. *Suppose that* $\{\mathrm{P}_0, \mathrm{P}_1\} \subseteq \mathcal{P} \subseteq \mathbb{R}^2$, *and suppose that a point* $\mathrm{R} = (r_1, r_2) \in \mathbb{R}^2$ *is constructible in one step from* $\mathcal{P}$. *Then*

$$[\Gamma_{\mathcal{P} \cup \{\mathrm{R}\}} : \Gamma_{\mathcal{P}}] \in \{1, 2\}.$$

Proof. It is clear from the definition of $\Gamma_{\mathcal{P} \cup \{\mathrm{R}\}}$ that $\Gamma_{\mathcal{P} \cup \{\mathrm{R}\}} = \Gamma_{\mathcal{P}}(r_1, r_2)$. There remains to find the degree of the extension. This leads us to consider three possibilities for R.

(a) R is the point of intersection of two non-parallel lines L_1 determined by $\mathrm{A} = (a_1, a_2)$ and $\mathrm{B} = (b_1, b_2) \in \mathcal{P}$, and L_2 determined by $\mathrm{C} = (c_1, c_2)$ and $\mathrm{D} = (d_1, d_2) \in \mathcal{P}$.

We shall argue the case where the slopes of L_1 and L_2 are non-zero real numbers. The cases where one line is horizontal and/or one line is vertical are left as (routine) exercises for the reader.

Let us write L_1 in the form $L_1 = \{(x,y) \in \mathbb{R}^2 : y = m_1x + n_1\}$. For $(x,y) \in L_1$ and $(x,y) \neq \{A, B\}$, we have that

$$\frac{x-a_1}{x-b_1} = \frac{y-a_2}{y-b_2},$$

and therefore

$$y = \frac{a_2-b_2}{a_1-b_1}x + \frac{a_1b_2-a_2b_1}{a_1-b_1}.$$

Set $m_1 = \dfrac{a_2-b_2}{a_1-b_1}$ and $n_1 = \dfrac{a_1b_2-a_2b_1}{a_1-b_1}$, and observe that $m_1, n_1 \in \Gamma_{\mathcal{P}}$, since $a_1, a_2, b_1, b_2 \in \Gamma_{\mathcal{P}}$.

A similar argument shows that we may write

$$L_2 = \{(x,y) \in \mathbb{R}^2 : y = m_2x + n_2\},$$

where $m_2, n_2 \in \Gamma_{\mathcal{P}}$.

Thus $R \in L_1 \cap L_2$ implies that

$$r_2 = m_1r_1 + n_1 = m_2r_1 + n_2,$$

whence

$$r_1 = \frac{n_2-n_1}{m_1-m_2}, \qquad r_2 = \frac{m_1(n_2-n_1)}{m_1-m_2} + n_1.$$

Observe that $r_1, r_2 \in \Gamma_{\mathcal{P}}$, and in this case,

$$\Gamma_{\mathcal{P}}(r_1,r_2) = \Gamma_{\mathcal{P}\cup\{r_1,r_2\}} = \Gamma_{\mathcal{P}}.$$

Thus $[\Gamma_{\mathcal{P}}(r_1,r_2) : \Gamma_{\mathcal{P}}] = 1$ in this case.

(b) R is the point of intersection of a line L determined by $A = (a_1,a_2)$ and $B = (b_1,b_2) \in \mathcal{P}$, and a circle T centred at a point $c = (c_1,c_2) \in \mathcal{P}$ and containing point $d = (d_1,d_2) \in \mathcal{P}$. Observe that the radius of the circle is $\rho := \sqrt{(d_1-c_1)^2 + (d_2-c_2)^2}$, and thus $\rho^2 \in \Gamma_{\mathcal{P}}$.

Again, we shall argue the case where L is parallel to neither axis, and leave the other cases as an exercise for the reader. From part (a) above, we see that we may express L as

$$L = \{(x,y) \in \mathbb{R}^2 : y = mx + n\},$$

where $m, n \in \Gamma_{\mathcal{P}}$ are fixed constants and $m \neq 0$.

Now $R = (r_1,r_2) \in L \cap T$ implies that

$$r_2 = mr_1 + n, \qquad \text{and} \qquad (r_1-c_1)^2 + (r_2-c_2)^2 = \rho^2.$$

Note that $r_2 \in \Gamma_{\mathcal{P}}(r_1)$, and thus $\Gamma_{\mathcal{P}}(r_1,r_2) = \Gamma_{\mathcal{P}}(r_1)$ in this case. To solve for r_1, we must solve the equation

$$(r_1-c_1)^2 + ((mr_1+n)-c_2)^2 = \rho^2,$$

or equivalently,

$$(1+m^2)r_1^2 + (-2c_1 + 2mn - 2mc_2)r_1 + (c_1^2 + n^2 - 2nc_2 + c_2^2 - \rho^2) = 0.$$

But this is a quadratic equation, and thus $[\Gamma_{\mathcal{P}}(r_1) : \Gamma_{\mathcal{P}}] \in \{1, 2\}$ (depending upon whether or not this equation has roots in $\Gamma_{\mathcal{P}}$.)

(c) The third and last possibility is that R is the point of intersection of two circles T_1 (centred at $\mathrm{A} = (a_1, a_2) \in \mathcal{P}$ and containing $\mathrm{B} = (b_1, b_2) \in \mathcal{P}$) and T_2 (centred at $\mathrm{C} = (c_1, c_2) \in \mathcal{P}$ and containing $\mathrm{D} = (d_1, d_2) \in \mathcal{P}$). As in part (b), we note that if ρ_k is the radius of the circle T_k, then $\rho_k \in \Gamma_{\mathcal{P}}$, $k = 1, 2$.

Thus $\mathrm{R} = (r_1, r_2) \in \mathrm{T}_1 \cap \mathrm{T}_2$ implies that

$$(r_1 - a_1)^2 + (r_2 - a_2)^2 = \rho_1^2,$$

and

$$(r_1 - c_1)^2 + (r_2 - c_2)^2 = \rho_2^2.$$

Taking the difference of these two equations, we obtain

$$r_2 = \alpha r_1 + \beta,$$

where $\alpha, \beta \in \Gamma_{\mathcal{P}}$. In particular, $r_2 \in \Gamma_{\mathcal{P}}(r_1)$, and so $\Gamma_{\mathcal{P}}(r_1, r_2) = \Gamma_{\mathcal{P}}(r_1)$ again! Although we don't need the exact values of α and β, for those who may be checking:

$$\alpha = \frac{a_1 - c_1}{a_2 - c_2}, \qquad \text{and} \qquad \beta = \frac{(a_1^2 - c_1^2) + (a_2^2 - c_2^2) - (\rho_1^2 - \rho_2^2)}{2(a_2 - c_2)}.$$

Substituting this into the equation for L_1, we see that

$$(r_1 - a_1)^2 + (\alpha r_1 + \beta - a_2)^2 = \rho_1^2,$$

so that r_1 satisfies a quadratic equation over $\Gamma_{\mathcal{P}}$, and thus

$$[\Gamma_{\mathcal{P}}(r_1, r_2) : \Gamma_{\mathcal{P}}] = [\Gamma_{\mathcal{P}}(r_1) : \Gamma_{\mathcal{P}}] \in \{1, 2\}.$$

□

2.8. Remark. In fact, the above proof shows that if $\mathrm{R} = (r_1, r_2) \in \mathbb{R}^2$ can be constructed in one step from $\mathcal{P}$, then $[\Gamma_{\mathcal{P}}(r_1) : \Gamma_{\mathcal{P}}] \in \{1, 2\}$ and $r_2 \in \mathbb{F}_{\mathcal{P}}(r_1)$, so that $[\Gamma_{\mathcal{P}}(r_1, r_2) : \Gamma_{\mathcal{P}}] \in \{1, 2\}$.

2.9. Corollary. *Suppose that* $\mathrm{R} = (r_1, r_2) \in \Gamma_{\mathcal{P}}$. *Then*

$$[\Gamma_{\mathcal{P}}(r_1, r_2) : \Gamma_{\mathcal{P}}] \in \{2^m : 0 \leq m \in \mathbb{Z}\}.$$

Proof. Choose $\mathrm{Q}_0 = \mathrm{P}_1 = (1, 0) \in \mathcal{P}$ and $\mathrm{Q}_k = (a_k, b_k) \in \mathbb{R}^2$ such that each Q_k is constructible in one step from $\mathcal{P} \cup \{\mathrm{Q}_0, \mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_{k-1}\}$, $1 \leq k \leq n$, and $\mathrm{Q}_n = (a_n, b_n) = \mathrm{R}$.

To simplify the notation, set $\mathcal{P}_k = \mathcal{P} \cup \{\mathrm{Q}_0, \mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_k\}$, $1 \leq k \leq n$, and observe that q_k is constructible in one step from $\mathcal{P}_{k-1}$. By the above Remark 2.8,

$$[\Gamma_{\mathcal{P}_{k-1}}(a_k, b_k) : \Gamma_{\mathcal{P}_{k-1}}] \in \{1, 2\}.$$

But $\Gamma_{\mathcal{P}_{k-1}}(a_k, b_k) = \Gamma_{\mathcal{P}_k}$, and thus

$$[\Gamma_{\mathcal{P}_k} : \Gamma_{\mathcal{P}_{k-1}}] \in \{1, 2\}.$$

Finally, by Theorem 1.27,

$$[\Gamma_{\mathcal{P}_n} : \Gamma_{\mathcal{P}}] = [\Gamma_{\mathcal{P}_n} : \Gamma_{\mathcal{P}_{n-1}}][\Gamma_{\mathcal{P}_{n-1}} : \Gamma_{\mathcal{P}_{n-2}}] \cdots [\Gamma_{\mathcal{P}_1} : \Gamma_{\mathcal{P}}] \in \{2^m : 0 \leq m \in \mathbb{Z}\},$$

say

$$[\Gamma_{\mathcal{P}_n} : \Gamma_{\mathcal{P}}] = 2^{m_0}$$

for some $0 \leq m_0 \in \mathbb{Z}$.

Now $\Gamma_{\mathcal{P}} \subseteq \Gamma_{\mathcal{P}}(r_1, r_2)$ is a subfield of $\Gamma_{\mathcal{P}_n}$, and thus

$$2^{m_0} = [\Gamma_{\mathcal{P}_n} : \Gamma_{\mathcal{P}}] = [\Gamma_{\mathcal{P}_n} : \Gamma_{\mathcal{P}(r_1, r_2)}]\ [\Gamma_{\mathcal{P}(r_1, r_2)} : \Gamma_{\mathcal{P}}],$$

which implies that $[\Gamma_{\mathcal{P}(r_1, r_2)} : \Gamma_{\mathcal{P}}]$ divides 2^{m_0}; i.e.

$$\Gamma_{\mathcal{P}}(r_1, r_2) = 2^{n_0} \text{ for some } 0 \leq n_0 \leq m_0.$$

□

2.10. Corollary. *Let* $\varrho \in \mathbb{R}$ *be a constructible real number. Then*

$$[\mathbb{Q}(\varrho) : \mathbb{Q}] \in \{2^m : 0 \leq m \in \mathbb{Z}\}.$$

Proof. Observe that $\varrho \in \Gamma^\circ$ if and only if $\mathrm{R} := (\varrho, 0)$ is constructible from $\mathcal{P}^\circ := \{\mathrm{P}_0, \mathrm{P}_1\}$. By Corollary 2.9 above, this implies that

$$[\Gamma_{\mathcal{P}^\circ}(\varrho, 0) : \Gamma_{\mathcal{P}^\circ}] \in \{2^m : 0 \leq m \in \mathbb{Z}\}.$$

To complete the proof, it suffices to note that $\Gamma_{\mathcal{P}^\circ} = \mathbb{Q}(0, 1) = \mathbb{Q}$.

□

3. Back to geometry

3.1. We started this journey into the realm of constructible numbers because of our deep admiration of Euclid, and our appreciation of his weakness for lines and circles. While the Greeks were masters of the plane method - there were three constructions that fell beyond their reach, and which would subsequently become famous,[3]namely: DOUBLING THE CUBE, SQUARING THE CIRCLE, and TRISECTING AN ANGLE. The question of whether or not these constructions were possible using the plane method remained open for approximately 2000 years, until the French mathematician Pierre **Laurent** Wantzel (1814-1848) proved that the problems of doubling the cube and trisecting the angle lay not with the Greeks, but with the nature of field extensions related to constructible real numbers. The same Pierre **Laurent** Wantzel also proved that a regular polygon is constructible if and only if the number of its sides is a product of a power of 2 and (any number – including zero) of prime natural numbers. As for the problem of squaring the circle – it was resolved when Lindemann proved that π is transcendental over $\mathbb{Q}$, as we shall discover below.

3.2. Doubling the cube. Since the set Γ° of constructible real numbers is a field that contains $\mathbb{Q}$, it is clear that we can construct the points $(0,0)$, $(1,0)$, $(1,1)$ and $(0,1)$ from the set $\mathcal{P}^\circ := \{\mathrm{P}_0, \mathrm{P}_1\}$. By connecting these points with the appropriate line segments, it follows that we may construct a square of area equal to 1. By thinking "outside the box", so to speak, we may view this as the face of a cube in $\mathbb{R}^3$ whose volume is equal to 1.

Of course - using a straight-edge and compasses construction, we can "double the area of the square", because if we set

$$\varrho := d((0,0),(1,1)),$$

then $\varrho = \sqrt{2}$ is constructible, and so we may also construct the points $(0,0), (\sqrt{2},0)$, $(\sqrt{2},\sqrt{2})$ and $(0,\sqrt{2})$, which are the vertices of a square of area equal to 2.

It occurred to the Greeks to ask:

> starting with $\mathcal{P}^\circ := \{\mathrm{P}_0, \mathrm{P}_1\}$, can we construct the face of a cube whose volume is 2?

Despite the efforts of a great many and many great mathematicians, this problem remained unsolved for approximately 2000 years. Thanks to Pierre **Laurent** Wantzel, we are now in a position to resolve this problem. Take that, Euclid.

3.3. Theorem. (**Wantzel**) *The cube cannot be doubled in volume using a straight-edge-and-compasses construction.*

Proof.

If it were possible to construct the face of this cube, then we could construct a line segment of length $\sqrt[3]{2}$, and in particular, we would conclude that $\alpha := \sqrt[3]{2} \in \Gamma^\circ$.

[3]See our previous note on fame in mathematics.

Note that α is a root of the polynomial equation

$$q(x) = x^3 - 2 \in \mathbb{Z}[x].$$

By Eisenstein's Criterion (with $p = 2$) (Theorem 7.3.14), we see that $q(x)$ is *irreducible over* $\mathbb{Q}$. By Theorem 9.1.13, we see that $\mathbb{Q}(\sqrt[3]{2}) \simeq \mathbb{Q}[x]/\langle q(x)\rangle$ is a field extension of degree 3 over $\mathbb{Q}$. By Corollary 2.10 above, $\sqrt[3]{2} \notin \Gamma^\circ$.

Thus we cannot construct the face of a cube of volume 2 in $\mathbb{R}^2$ using a straight-edge and compasses construction and the points P_0 and P_1.

□

3.4. Squaring the circle. Starting with $\mathcal{P}^\circ = \{\mathrm{P}_0, \mathrm{P}_1\}$, we can construct a circle C whose radius is $d(\mathrm{P}_0, \mathrm{P}_1) = 1$. The area it inscribes is then $\textsc{Area}(\mathrm{C}) = \pi(1^2) = \pi$.

A second question of geometry which troubled the Greeks was:

> *starting with $\mathcal{P}^\circ := \{\mathrm{P}_0, \mathrm{P}_1\}$, can we construct a square whose area is that inscribed by the circle* C *of radius* 1*?*

Of course, we now recognise that the length of a side of such a square would have to be $\alpha := \sqrt{\pi}$.

At this stage, we shall unfortunately have to appeal to a result a bit outside the scope of these notes and appeal to the following wonderful theorem of **Carl Louis Ferdinand von Lindemann** in 1882 (see [**vL82**] for a proof of this fact).

3.5. Theorem. (**Lindemann**) *The number π is transcendental. That is, π is not the root of any non-zero polynomial with rational coefficients.*

If $\sqrt{\pi}$ were constructible (i.e. if we were to have $\sqrt{\pi} \in \Gamma^\circ$), then by Corollary 2.10, we would know that $[\mathbb{Q}(\sqrt{\pi}) : \mathbb{Q}] < \infty$. Since $\pi \in \mathbb{Q}(\sqrt{\pi})$, this would imply that π is algebraic, a contradiction.

Thus we cannot construct a square whose area is that of the unit circle C.

3.6. Trisecting an angle. Paragraph A.10.4 shows that it is always possible to *bisect* an constructible angle.

A third question of geometry which troubled and eluded the ancient Greeks was that of "trisecting the angle". Properly phrased, the question is whether or not one could *always* trisect an angle *which one could already construct using a straight-edge and compasses.*

For example, if one begins with the angle $\theta = \pi$, then trisecting the angle θ is certainly possible using the plane method. Similarly, one can "trisect" the angle $\theta' := \frac{\pi}{2}$, because one can construct an angle of $\frac{\pi}{6}$. (We leave these two problems as exercises for the interested reader.)

Let us consider the problem of trisecting other angles. We first construct two circles C_1 and C_2 of radius $1 \in \Gamma^\circ$ centred at P_0 and $\mathrm{P}_1 = (1, 0)$. They will intersect at the point $\mathrm{Z} = (\frac{1}{2}, \frac{\sqrt{3}}{2})$, and the triangle $(\mathrm{P}_0\,\mathrm{P}_1\,\mathrm{Z})$ is equilateral. Thus $\angle(\mathrm{Z}\mathrm{P}_0\,\mathrm{P}_1)$ is $\frac{\pi}{3}$ radians (or 60°) is a "constructible angle". Is it possible to trisect this angle?

That is, is it possible to construct a line L which intersects the unit circle C at the point W such that the angle $\angle(\mathrm{WP}_0\,\mathrm{P}_1)$ measures $\frac{\pi}{9}$ radians?

3.7. Theorem. (**Wantzel**) *The angle $\frac{\pi}{3}$ cannot be trisected using a straight-edge-and-compasses construction.*

Proof. The question reduces to: what is that point W, and does it lie in Γ°? Writing $\mathrm{W} = (w_1, w_2)$, we see that $w_1 := \cos(\frac{\pi}{9})$, and $w_2 := \sin(\frac{\pi}{9})$. If $\mathrm{W} \in \Gamma^\circ$, then w_1 and w_2 must each be constructible as well!

Recall that for all $0 < \alpha \in \mathbb{R}$, we have that

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

Applying this with $\alpha = \frac{\pi}{9}$, and recalling that $\cos(3\alpha) = \cos(\frac{\pi}{3}) = \frac{1}{2}$, we see that w_1 must satisfy

$$\frac{1}{2} = \cos(3w_1) = 4w_1^3 - 3w_1,$$

or equivalently, w_1 must be a root of the polynomial

$$q(x) = 8x^3 - 6x - 1.$$

We claim that $q(x)$ is irreducible over $\mathbb{Q} = \Gamma_{\mathcal{P}^\circ}$. Indeed, suppose that $q(x)$ were reducible over $\mathbb{Q}$. Then we could write $q(x) = g(x)h(x)$ where $\deg(g(x)) = 1$ and $\deg(h(x)) = 2$. By Theorem 7.3.8, if $(q(x))$ is reducible over $\mathbb{Q}$, then $q(x)$ is reducible over $\mathbb{Z}$, and we can find factors $g_1(x)$ and $h_1(x) \in \mathbb{Z}[x]$ with $\deg(g_1(x)) = 1$ and $\deg(h_1(x)) = 2$ such that $q(x) = g_1(x)h_1(x)$.

Set $g_1(x) = a_1x + a_0$ and $h_1(x) = b_2x^2 + b_1x + b_0 \in \mathbb{Z}[x]$. We must therefore solve

$$8x^3 - 6x - 1 = (a_1b_2)x^3 + (a_1b_1 + a_0b_2)x^2 + (a_1b_0 + a_0b_1)x + (a_0b_0),$$

where all a_i, b_j's lie in $\mathbb{Z}$.

The equation $a_0b_0 = -1$ implies that $(a_0, b_0) = (1, -1)$ or $(a_0, b_0) = (-1, 1)$. By multiplying both $g_1(x)$ and $h_1(x)$ by -1 if necessary, we may assume without loss of generality that $a_0 = 1, b_0 = -1$.

Thus

$$\begin{aligned}
-6 &= a_1b_0 + a_0b_1 = -a_1 + b_1, \\
0 &= a_1b_1 + a_0b_2 = a_1b_1 + b_2,
\end{aligned}$$

and

$$8 = a_1b_2.$$

At this stage of the course, it is safe to leave it to the reader to verify that this can not be done.

It follows that $q(x)$ is irreducible over $\mathbb{Q}$. But then, as in Paragraph 3.2, we find that $\mathbb{Q}(w_1) \simeq \mathbb{Q}[x]/\langle q(x)\rangle$ is a field extension of degree 3 over $\mathbb{A}$, and so by Corollary 2.10, $w_1 \notin \Gamma^\circ$.

We conclude that the angle $\frac{\pi}{9}$ cannot be constructed, and so the angle $\frac{\pi}{3}$ can not be trisected.

□

Appendix

A10.1. Let A and B be distinct points in $\mathbb{R}^2$. Here we describe how to construct a line L passing through A which is perpendicular to the line passing through A and B.

1. Let L_1 denote the line passing through A and B, and let $\varrho = d(A, B)$.
2. Otherwise, let C be a circle of radius ϱ centred at A. The circle C will intersect L_0 at two points, namely B and D. Then $d(B, D) = 2\varrho$.
3. Let C_B be the circle of radius 2ϱ centred at B, and C_D be the circle of radius 2ϱ centred at D. Then C_B and C_D will intersect at two points, namely Y_1 and Y_2.
4. Let L denote the line passing through Y_1 and Y_2. Observe that the line segments S_1 from Y_1 to Y_2 and S_2 from D to B are the diagonals of the rhombus determined by the vertices D, B, Y_1 and Y_2, and as such, they are perpendicular bisectors. Then L is the desired line.

(In the case where $A = P_0 = (0, 0)$ and $B = P_1 = (1, 0)$, we find that $Y_1 = (0, \sqrt{2})$ and $Y_2 = (0, -\sqrt{2})$. The line L connecting these two points is then the y-axis, which is clearly perpendicular to the x-axis and passes through $A = (0, 0)$.)

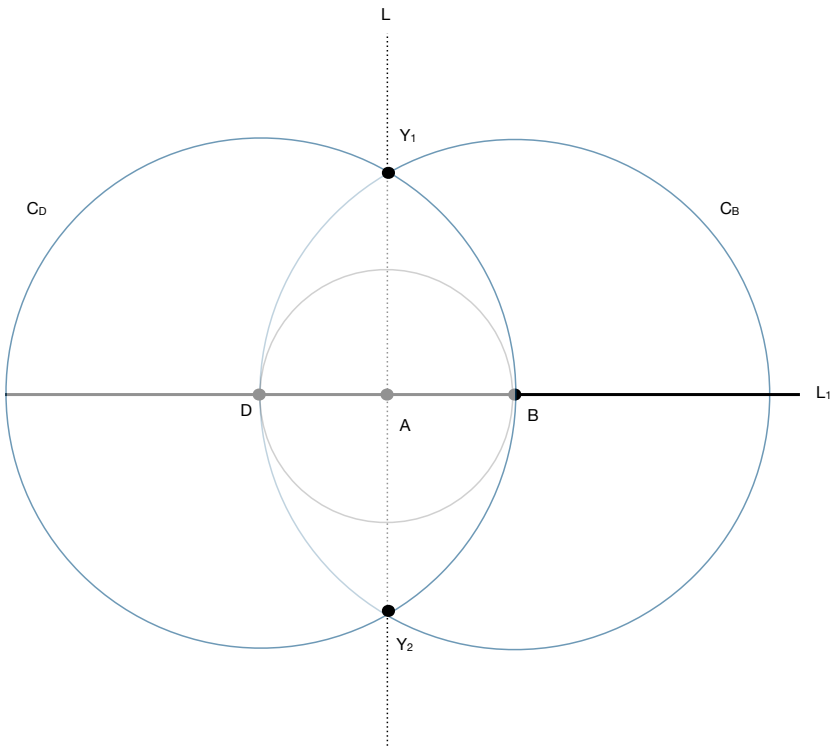

Figure 10.1. The line L through A perpendicular to the line through A and B.

A10.2. Let A, B and D be three non-collinear points in $\mathbb{R}^2$. Here we describe how to construct a line through D which is perpendicular to the line passing through A and B.

1. Let L_1 denote the line passing through A and B. If the line L_2 through D and B is perpendicular to L_1, then we are done.
2. Otherwise, let C_D be a circle of radius $\varrho := d(B, D)$ centred at D. Since ϱ is greater than the distance from D to L_1, the circle C_D will intersect L_1 at two points, namely B and X.
3. Let C_B be the circle of radius ϱ centred at B, and C_X be the circle of radius ϱ centred at X. Then C_B and C_X will intersect at two points, namely D and Y.
4. Let L_2 denote the line passing through D and Y.

We claim that L_2 is perpendicular to L_1. (Obviously L_2 contains D.) Indeed, the figure (BDXY) is a rhombus, and the segments $\overline{BX}$ and $\overline{DY}$ are the diagonals of that rhombus, which are perpendicular to each other.

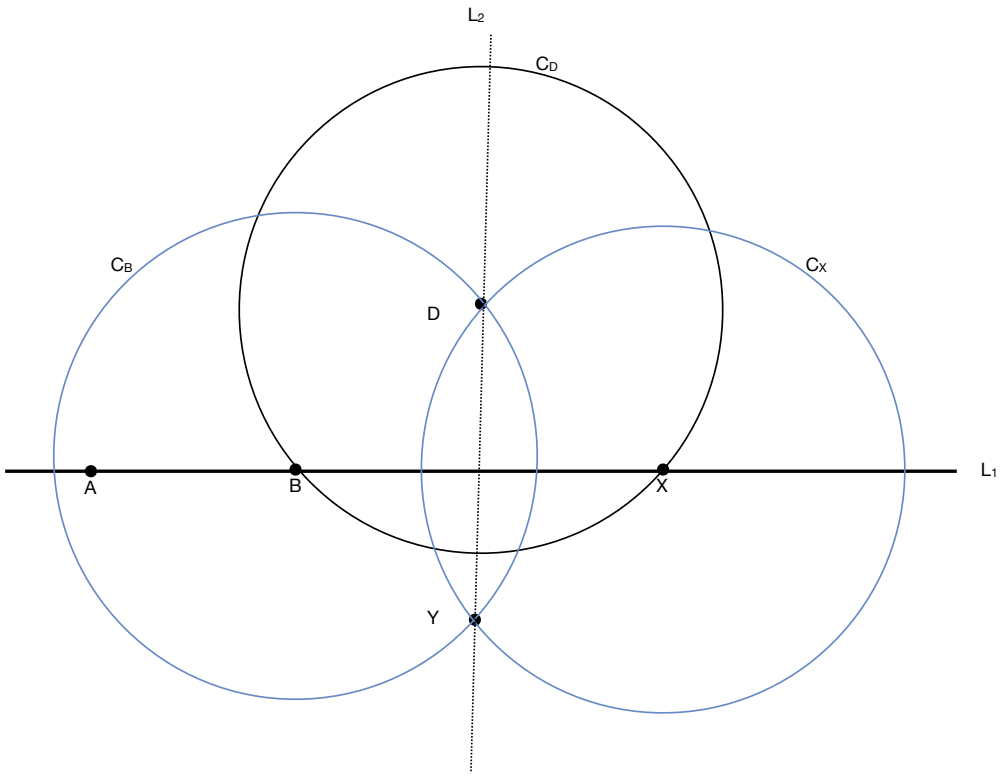

Figure 10.2. The line through D and Y is perpendicular to the line through A and B.

A10.3. Let L_1 be a line in $\mathbb{R}^2$ passing through the points A and B, and let $D \in \mathbb{R}^2$ be a point not on L_1. Here we describe how to construct a line L_2 through D which is parallel to L_1.

1. Observe that L_1 be the (unique!) line passing through A and B.
2. By Paragraph A10.2, we can construct a line L_2 which passes through D and is perpendicular to L_1.
3. By Paragraph A10.1, we can construct a line L_3 which is perpendicular to L_2 and passes through D.

Then L_3 is parallel to L_1 (since both are perpendicular to L_2) and L_3 passes through D, as claimed.

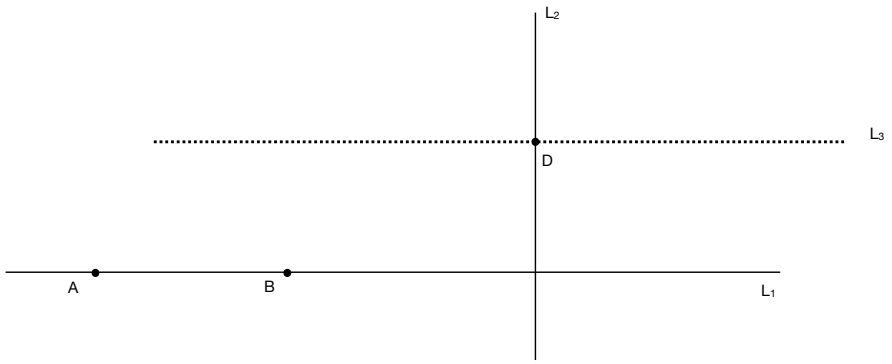

Figure 10.3. The line L_3 is parallel to the line L_1 and passes through D.

A10.4. Suppose that we have three non-collinear points A, B and C. It is always possible to bisect the angle $\angle ABC$ formed by these three points.

Exercises for Chapter 10

Exercise 10.1.
Prove that $\sqrt{3+\sqrt{36+25}}$ is constructible.

Exercise 10.2.
Prove that it is always possible to bisect a constructible angle.

Exercise 10.3.
Prove that the angle θ can be trisected using the plane method if and only if the polynomial

$$q(x) = 4x^3 - 3x - \cos\theta$$

is reducible over $\mathbb{Q}(\cos\theta)$.

Exercise 10.4.
Prove that one may *approximate* the angle $\frac{\pi}{9}$ with any degree of precision using the plane method, starting from $\mathcal{P}^\circ = \{\mathrm{P}_0, \mathrm{P}_1\}$.

Exercise 10.5.
Show that it is possible to construct a regular pentagon using the plane method, starting from $\mathcal{P}^\circ = \{\mathrm{P}_0, \mathrm{P}_1\}$.

Exercise 10.6.
Prove that if $\alpha, \beta \in \Gamma^\circ$, then $\sqrt{\alpha^2+\beta^2} \in \Gamma^\circ$. Is the converse true?

Exercise 10.7.
Prove that it is possible to construct a regular dodecagon (i.e. a 12-sided polygon, all of whose sides are of equal length) using the plane method, starting from $\mathcal{P}^\circ = \{\mathrm{P}_0, \mathrm{P}_1\}$.

Exercise 10.8.
Prove that $[\mathbb{Q}(\sqrt{4+2\sqrt{3}}) : \mathbb{Q}] = 2$.

Exercise 10.9.
Let $a, b \in \Gamma^\circ$ with $a \neq 0$. Prove that

$$\frac{-b \pm \sqrt{b^2-4ac}}{2a} \in \Gamma^\circ.$$

Exercise 10.10.
Prove that it given two distinct points A and B in the plane, it is possible to construct the midpoint of the line segment from A to B using a straight-edge and compasses.

APPENDIX A

The Axiom of Choice

1. Introduction

1.1. The Axiom of Choice has been mentioned earlier in these course notes, as has Zorn's Lemma. In this note, we shall examine these two equivalent statements, as well as a third statement, the Well-Ordering Principle, which is also equivalent to these two.

1.2. Set theory is the study of specific collections, called... wait for it... *sets* of objects that are called *elements* of that set. *Pure set theory* deals with sets whose elements are again sets. Experts in set theory claim that the theory of *hereditarily-finite* sets – i.e. those finite sets whose elements are also finite sets, whose elements in turn are also finite sets, whose elements ... etc., is formally equivalent to arithmetic, which we shall neither define nor attempt to explain.

The form of set theory which interests us here arose out of attempts by mathematicians of the late 19th and early 20th centuries to try to understand infinite sets, and originated with the work of the German mathematician **Georg Cantor**. He had published a number of articles in number theory between 1867[1] and 1871[2], but began work on what was to become set theory by 1874.

In 1878, Cantor formulated was is now referred to as the *Continuum Hypothesis*(CH). Suppose that X is a subset of the real line. The Continuum Hypothesis asserts that either there exists a bijection between X and the set $\mathbb{N}$ of natural numbers, or there exists a bijection between X and the set $\mathbb{R}$ of real numbers.

Many well-known and leading mathematicians of his day attempted to prove this statement, including Cantor himself, as well as **David Hilbert**, who included it as the first of the twenty-three unsolved problems he presented at the Second International Congress of Mathematics in Paris, in 1900. Any attempt to prove this would require one to understand not only sets of real numbers, but sets in general.

Like mangoes, mathematical theories need time to ripen. Early attempts to define sets led to inconsistencies and paradoxes. Originally, it was thought that any property P could be used to define a set, namely the set of all sets (or even other objects) which satisfy property P. The mathematician-philosopher **Bertrand Russell** produced the following worrisome example:

[1]the year of Canada's birth

[2]the year that the Government of Canada promised B.C. a railway which they delivered in 1885, by the way

Russell's Paradox. Let P be the property of sets: $X \notin X$. That is, a set X satisfies property P if and only if X does not belong to X.

Let

$$\mathfrak{R} := \{X \text{ a set} : X \text{ satisfies property } P\} = \{X \text{ a set} : X \notin X\}.$$

The question becomes: does $\mathfrak{R}$ satisfy property P? If so, the $\mathfrak{R} \notin \mathfrak{R}$, and so $\mathfrak{R} \in \mathfrak{R}$. If not, then $\mathfrak{R} \in \mathfrak{R}$, so $\mathfrak{R} \notin \mathfrak{R}$.

This is not a good state of affairs, and the German mathematician **Ernst Zermelo**, who was also aware of this paradox, thought that set theory should be axiomatised in order to be rigorous and preclude paradoxes such as the one above. In 1908, Zermelo produced a first axiomitisation of set theory, although he was unable at that time to prove that his axiomatic system was consistent. There was also a second problem, insofar as some logician/set theorists were concerned, namely: Zermelo avoided Russell's Paradox by means of an axiom which he referred to as the *Separation Axiom*. This, however, he stated using *second order logic*, which is considered less desirable (in the sense that it requires stronger hypotheses) than the version offered by the Norwegian **Thoralf Skolem** and the German **Abraham Fraenkel**, which relied only on so-called *first-order logic*.

It occurs that the Hungarian mathematician **John von Neumann** also added to the list of so-called *Zermelo-Fraenkel Axioms* we shall enumerate below by introducing the *Axiom of Foundation*. We are not sufficiently versed in the history of this area to be able to explain why Skolem and von Neumann's names don't appear on the list of Axioms, although von Neumann's name appears in so many other mathematical contexts that even if he were alive today, he would not have the right to complain.

1.3. The Zermelo-Fraenkel Axioms (ZF).

- The Null Set Axiom.
  There exists a set $\varnothing$, called the *empty set*, which has no elements.
- The Axiom of Extension.
  Two sets A and B are equal if and only if they have the same elements.
- The Axiom of Regularity, aka the Axiom of Foundation.
  Every non-empty set A contains an element B such that no element of A belongs to B.
- The Axiom of Specification.
  If A is a set and P is a property, then there exists a set

$$B := \{b \in A : b \text{ satisfies property } P\}.$$

- The Axiom of Pairing.
  Given two sets A and B, there exists a set $\{A, B\}$ whose only elements are A and B. (Combined with the Null Set Axiom, we deduce that $\{A\}$ is also a set.)
- The Axiom of Union.
  If A is a set, then there exists a set $\cup A$, called the *union of* A, defined by
  $$\cup A = \{b : b \in B \text{ for some } B \in A\}.$$
- The Axiom of Power Sets.
  Given a set A, there exists a set $\mathcal{P}(A) = \{B : B \subseteq A\}$ whose elements are all of the subsets of A.
- The Axiom of infinity. There exists an infinite set. More specifically, there exists a set $\mathfrak{Z}$ such that $\varnothing \in \mathfrak{Z}$ and $A \in \mathfrak{Z}$ implies that $\cup\{A, \{A\}\} \in \mathfrak{Z}$.

1.4. Just a quick comment about the Axiom of Specification: why doesn't it lead right back to Russell's Paradox? Because the Axiom of Specification assumes that you *begin* with a set, and you look at the members of *that set* which satisfy a property P. In Russell's Paradox, you don't have a starting set – instead you *manufacture* a set out of thin air through a property P. This is the kind of object we now refer to as a *class*, which is an object more general than a set. In fact, a class is a collection of sets, which means that we can't apply Russell's paradigm to a class of classes.

1.5. There is one Axiom we have yet to specify, which – given the title of this note – we might want to do sooner rather than later.

- The Axiom of Choice. Let Λ be a non-empty set, and let $\{A_\lambda\}_{\lambda\in\Lambda}$ be a collection of non-empty sets. Then there exists a function $f : \Lambda \to \cup_{\lambda\in\Lambda} A_\lambda$ such that $f(\gamma) \in A_\gamma$ for each $\gamma \in \Lambda$.

In layman's terms, if you have a (non-empty) collection of (non-empty) sets, then you can choose an element from each set. In technical terms, we call such a function f a **choice function**.

We write (ZFC) to indicate (ZF) plus the Axiom of Choice.

1.6. Remark. There are a couple of remarks we should make here.

(a) First: what the heck is going on? This is so obviously true that it isn't even worth mentioning! How could it not be true?
(b) Doesn't it follow from the Axioms above? If not, what are they good for?

1.7. I sympathise with those who think it's obviously true. I really do. For one thing, the Axiom of Choice becomes an issue only if we have *a lot* of sets – in fact, only when we are dealing with infinitely many such sets. (Whether those sets are finite or infinite is not the issue – the issue is how many sets we have.)

The formal reason why this is so can be summarised in the following way: (ZFC) is a "*theory in first-order logic*", and our ability to "*choose*" an element from a (single) non-empty set is an application of a rule from that theory that carries the groovy moniker *existential instantiation*. We do not claim to be an expert in the theory of first-order logic, but the upshot that a hard-working, good-looking mathematician but non-logician like me can safely take away from this statement is that given a non-empty set, (ZF) allows you to pick an element of that set. If you are interested in the details of first-order logic, then be aware that treatments now exist to cure this, but also that some people exhibit these symptoms for years and still manage to lead productive lives, both inside and outside of mathematics.

Another thing to note is that existential instantiation is not a *constructive* argument. That is, we are not given a *method* of choosing an element from a non-empty set A – all we know is that it is possible. If we relabel A as A_λ, then saying that produces an element $a \in A$ is the same as saying that there exists a function $f : \{\lambda\} \to A_\lambda$ that satisfies $f(\lambda) = a$. That is, we have our choice function.

Having done this once, first-order logic allows us to repeat this process finitely often. That is, given *finitely many* non-empty sets $A_1, A_2, \ldots, A_n$, we can find a function
$f : \{1, 2, \ldots, n\} \to \cup_{k=1}^{n} A_k$ such that $f(k) \in A_k$, $1 \leq k \leq n$.

What first-order logic and the Zermelo-Fraenkel Axioms do not do, however, is to carry out this procedure infinitely often, willy-nilly. We shall try to make the expression "*willy-nilly*" more precise below.

Willy-Nilly, or not Willy-Nilly. That is the question.

1.8. While it is important for one to try to be original, it is also important for one to acknowledge the accomplishments of those who came before one. In particular, the same Bertrand Russell referenced above once said:

> *To choose one sock from each of infinitely many pairs of socks requires the Axiom of Choice, but for shoes the Axiom is not needed.*

It is interesting to try to figure out what exactly Russell meant by this, and what it says about his sartorial predilections.

The difference to which Bertrand Russell is alluding is related to the non-constructiveness of existential instantiation. Russell's underlying hypothesis while making this comment is that there is no way to distinguish between two socks in a pair. That is, he is assuming that the two socks in a given pair are for all intents and purposes identical. Thus, to pick a sock from a pair requires the rules of first-order logic to run (ZF) theory. This allows us to pick a sock from each pair for *finitely*

many pairs, but there is no way to do this for *infinitely many pairs at once*. Thus in order to pick a sock from amongst each of infinitely many pairs of socks, you need something more than (ZF) theory, and the Axiom of Choice grants you your wish.

A second underlying assumption of Russell's is that each pair of shoes includes a left shoe and a right shoe. This changes everything. We no longer require existential instantiation to select a shoe from a pair – instead, we just specify *the left* shoe. (We could have said the right.) Thus, to "pick a shoe from each of infinitely many pairs of shoes", we simply specify that we shall always pick the left shoe, thereby circumventing existential instantiation altogether. Having a method to pick an element of the non-empty sets we are dealing with allows us to avoid having recourse to the Axiom of Choice.

1.9. Let us consider how such a thing might arise in practice in a mathematical setting. Let $\{A_t : t \in \mathbb{R}\}$ be a collection of non-empty subsets $\varnothing \neq A_t \subseteq \mathbb{N}$, one such set for every real number $t \in \mathbb{R}$. We would like to choose one element from each set. Can we do this in (ZF), or do we require the Axiom of Choice?

If we had an explicit *method* to choose the element $a_t \in A_t$, then we could indeed avoid the Axiom of Choice. But how can we specify an element of A_t without knowing exactly what A_t is? The natural numbers have a very special and very useful property: they are *(well-) ordered*.

Before defining a well-order we remind the reader that a **relation** ρ on a non-empty set X is a subset of the set $X \times X = \{(x, y) : x, y \in X\}$. Often, we write $x \rho y$ to mean the ordered pair $(x, y) \in \rho$. This is especially true when dealing, for example, with the usual relation $\leq$ for real numbers: no one writes $(x, y) \in \leq$; we all write $x \leq y$. Incidentally, the notation $\leq$ is used not only for the relation "less than or equal to" for real numbers; it frequently appears to indicate a specific kind of a relation known as a *partial order* on an arbitrary set, which we now define.

1.10. Definition. *A relation $\leq$ on a non-empty set X is said to be a* ***partial order*** *if, given x, y, and $z \in X$,*

(a) $x \leq x$;
(b) *if* $x \leq y$ *and* $y \leq x$, *then* $x = y$; *and*
(c) *if* $x \leq y$ *and* $y \leq z$, *then* $x \leq z$.

The prototypical example of a partial order is the partial order of inclusion on the power set $\mathcal{P}(A)$ of a non-empty set A. That is, set $\mathcal{P}(A) = \{B : B \subseteq A\}$ and for $B, C \in \mathcal{P}(A)$, we set $B \subseteq C$ to mean that every member of B is a member of C. It is easy to see that $\subseteq$ is a partial order on $\mathcal{P}(A)$.

The word *partial* refers to the fact that given two elements B and C in $\mathcal{P}(A)$, they might not be comparable. For example, if $A = \{x, y\}$, $B = \{x\}$ and $C = \{y\}$, then it is not the case that $B \subseteq C$, nor is it the case that $C \subseteq B$. Only *some* subsets of $\mathcal{P}(A)$ are comparable.

In dealing with the natural numbers, we observe that they come equipped with a partial order which we typically denote by $\leq$. In this setting, however, *any* two natural numbers are comparable. If (X, ρ) is a partial ordered set, and if $x, y \in X$ implies that either $x \rho y$ or $y \rho x$, then we say that ρ is a **total order** on X.

Thus $(\mathbb{N}, \leq)$ is a **totally ordered set**.

1.11. It gets even better (and yes, you should seriously ask yourself whether you deserve it). The most useful property of $\mathbb{N}$ for our current purposes is that it possesses one more *very* striking property:

> *given any non-empty subset* $\varnothing \neq H \subseteq \mathbb{N}$*, it admits a* **minimum element**;

that is, there exists an element $m \in H$ such that $m \leq h$ for all $h \in H$.

If (X, ρ) is a partially ordered set with the property that every non-empty subset Y of X admits a minimum element, then we say that (X, ρ) is **well-ordered**.

The observation above is that $(\mathbb{N}, \leq)$ is well-ordered.

1.12. What is the relevance of this to the Axiom of Choice? Returning to the example in paragraph 1.9, we were given a collection $\{A_t : t \in \mathbb{R}\}$ of non-empty subsets of $\mathbb{N}$. Since we have just seen that $(\mathbb{N}, \leq)$ is well-ordered, we can specify a choice function

$$\begin{array}{rccc} f: & \mathbb{R} & \to & \cup_{t \in \mathbb{R}} A_t \\ & t & \mapsto & \min A_t. \end{array}$$

That is, we have a *rule* for specifying which element of A_t we are choosing – we are choosing the minimum element of A_t which we know exists (and is unique). We did not need to know in advance what A_t was, and we did not need existential instantiation to define $f(t)$.

Thus we don't need the Axiom of Choice to pick an element from each subset A_t simultaneously.

1.13. What if we had a collection $\{B_t : t \in \mathbb{R}\}$ of non-empty subsets of $\mathbb{R}$? That is, $\varnothing \neq B_t \subseteq \mathbb{R}$ for each $t \in \mathbb{R}$? Do we need the Axiom of Choice to choose an element from each set?

It is easy to check that $(\mathbb{R}, \leq)$ is a totally ordered set, as was $(\mathbb{N}, \leq)$. On the other hand, the open interval $(0, 1) = \{x \in \mathbb{R} : 0 < x < 1\}$ certainly doesn't have a minimum element, so that $(\mathbb{R}, \leq)$ is not a well-ordered set.

How would we specify an element of B_t without first knowing what B_t is? Unless we first know something about the set B_t, there *is no way* of specifying which element of B_t we are choosing. It follows that we *do need* the Axiom of Choice to do this.

(Of course, every subset of $\mathbb{N}$ is a subset of $\mathbb{R}$, so if we had originally chosen $B_t = A_t$ for all t, then we could have just used the "minimum" element method above. The point here is that if you have no information about B_t other than the fact that it is non-empty, then you can't use that argument here.)

1.14. We have not forgotten the second question raised in Remark 1.6, namely: doesn't the Axiom of Choice follow from the Axioms of (ZF)? In fact, we have already partially addressed this – (ZF) and first-order logic do *not* allow us to choose an element from each of infinitely-many sets at a time, in large part because of the non-constructive nature of existential instantiation. (For those who are wondering, yes, your humble author does love that phrase.)

There remains the question: does it contradict any of the (ZF) axioms? The simple answer is "No", although the proof that it does not is anything but simple. That (ZFC) is *consistent* (i.e. doesn't contain any inherent contradictions) was first demonstrated by the Austro-Hungarian mathematician **Kurt Gödel** in 1938. Does that mean that we can't do mathematics without the Axiom of Choice? Interestingly enough, in 1962, the American mathematician **Paul Cohen** demonstrated that one can assume all of the Axioms of Zermelo-Fraenkel Theory, and also assume that the Axiom of Choice is FALSE(!!!), and *still* arrive at a consistent mathematical system. For this and for his work on the Continuum Hypothesis (he showed that (CH) is independent of (ZFC)), he was awarded the Fields medal, which is generally considered to be the mathematical equivalent of the Nobel Prize. Given how prestigious and merited the Nobel Peace prize invariably is, a prize in mathematics is... (we leave it as an exercise for the reader to complete this sentence).

2. An apology for the Axiom of Choice.

2.1. As important as it is to choose one's socks wisely, if the only thing that every resulted from the Axiom of Choice was our ability to choose a sock from each pair in an infinite collections of pairs of socks, then we would not be talking about it today.

The issue is that this seemingly innocuous assumption has major implications. In particular, it is known that to imply that every vector space (over any field) admits a vector-space basis. Given how useful vector space bases are when studying linear algebra, it would be beyond tragic to lose them. Of course, one might argue that one should instead drop the Axiom of Choice and just assume that every vector space admits a basis. But the joke is then upon the one who argued this: as it so happens, if we assume that every vector space admits a basis, then one can also prove that the Axiom of Choice holds.

In other words, the Axiom of Choice is equivalent to the statement that every vector space admits a basis. One of the statements is true if and only if the second is true.

2.2. The number of statements that are equivalent to the Axiom of Choice makes is huge. Entire books are written on this subject - including the book *Equivalents of the Axiom of Choice, II* by Herman Rubin and Jean Rubin, which includes 250 statements, each equivalent to the Axiom of Choice. Below we shall mention three such equivalent statements which are among the most important.

2.3. The Axiom of Choice II. This version is a slight modification of the Axiom of Choice, and you might want to think about how you would prove that it is equivalent to the Axiom of Choice:

> Let $\{A_\lambda\}_{\lambda\in\Lambda}$ be a non-empty collection of non-empty *disjoint* sets. That is, $\lambda_1 \neq \lambda_2 \in \Lambda$ implies that $A_{\lambda_1} \cap A_{\lambda_2} = \varnothing$.
>
> Then there exists a function $f : \Lambda \to \cup_{\lambda\in\Lambda} A_\lambda$ such that $f(\lambda) \in A_\lambda$ for all $\lambda \in \Lambda$.

2.4. The Well-Ordering Principle.

> Let $\varnothing \neq X$ be a non-empty set. Then there exists a relation ρ on X such that (X, ρ) is well-ordered.

If $(X, \leq)$ is a partially ordered set, then a subset $C \subseteq X$ is said to be a **chain** in X if $c, d \in C$ implies that either $c \leq d$ or $d \leq c$. In other words, C is totally ordered using the relation inherited from X.

2.5. Zorn's Lemma.

> Let $(X, \leq)$ be a partially ordered set. Suppose that for each chain C in X, there exists an element $\mu_C \in X$ such that $c \leq \mu_C$ for all $c \in C$. (We say that C is **bounded above** in X.)
>
> Then $(X, \leq)$ admits a **maximal element** m. That is, there exists $m \in X$ such that if $x \in X$ and $m \leq x$, then $x = m$.

2.6. A note about maximal elements of partially ordered sets. Our decision to use "*maximal*" instead of "*maximum*" is not just fancy. They usually mean different things.

Let $\mathcal{A} = \{\varnothing, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}\}$ be the set of *proper* subsets of $X = \{x, y, z\}$, and partially order $\mathcal{A}$ by inclusion. Thus $\{x\} \leq \{x, y\}$ because $\{x\} \subseteq \{x, y\}$.

Observe that $\{x, y\}$ is a maximal element of $\mathcal{A}$. *Nothing is bigger than* $\{x, y\}$ in $\mathcal{A}$. But $\{z\} \notin \{x, y\}$, and so $\{x, y\}$ is not a maximum element of $\mathcal{A}$. Note that $\{x, z\}$ is also a maximal elements. Maximal elements need not be unique. (Are maximum elements unique when they exist? Think about it!)

The issue is that a maximal element doesn't have to be comparable to every element in the set, while a maximum element does. In other words, a maximal element only has to be bigger than or equal to those things that you can actually compare it to, while a maximum element has to be bigger than or equal to everything in the set! In this example, $\varnothing$ would be a mimimum, and thus also a minimal element of $(\mathcal{A}, \leq)$.

That the Well-Ordering Principle implies the Axiom of Choice is definitely within reach of the interested reader. We leave it as an exercise.

2.7. The following quote is worth keeping in mind:

> *The Axiom of Choice is obviously true, the Well-Ordering Principle obviously false, and who can tell about Zorn's lemma?*
>
> *Jerry Bona*

It's funny people like Professor Bona who put the world of mathematics into perspective and allow us to laugh at ourselves for days at a time by revealing some heretofore hidden lacuna in our intuition. It is also worth keeping in mind that the Well-Ordering Principle is *not* constructive. It does not suggest any method of actually finding a well-order on a given set X; it simply says that one exists.

The question of how you might try to well-order the complex numbers is certainly worth thinking about.

2.8. Although your humble author has certainly not had a quote that compares in notoriety to that mentioned above, it strikes your humble author that the Well-Ordering Principle is not as "obviously false" as a somewhat different consequence of the Axiom of Choice, namely the Banach-Tarski Paradox. This we shall address in the next section.

3. The prosecution rests its case.

3.1. Given how "obviously true" the Axiom of Choice is, and how perfectly intuitive it may seem, it would seem outrageous not to include it amongst the axioms of set theory.

Consider then the following story, which by all accounts is true (see J. Mycielski, Notices of the AMS **53**, no. 2, page 209): **Alfred Tarski**, a Polish logician, proved that the Axiom of Choice was equivalent to the following statement.

> *Given any infinite set X, there is a bijection from X to $X \times X := \{(x, y) : x, y \in X\}$.*

He then submitted this result to the *Comptes Rendus de l'Académie des Sciences*, where it was refereed by two famous French mathematicians, **Maurice René Fréchet**, and **Henri Lebesgue**. Both referees rejected Tarski's paper. Fréchet wrote that the equivalence of two well-known truths is not a new result. Lebesgue claimed that both statements were false, and so their equivalence was of no interest.

Tarski is said to have never submitted another paper to *Comptes Rendus*. (Feel free to revise the sentence you completed at the end of Section 1.)

3.2. Perhaps the most damning evidence one can provide against the Axiom of Choice is the following result which the Axiom of Choice implies, and was alluded to above.

3.3. The Banach-Tarski Paradox. *Let* $S \coloneqq \{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq 1\}$, *so that* S *denotes the* (*solid*) *ball in* $\mathbb{R}^3$ *centred at the origin. It is possible to partition the ball* S *into finitely many disjoint pieces in such a way, that by translating and rotating them, they can be rearranged into two disjoint identical balls, each having the same volume as* S.

Note that stretching, bending, and twisting pieces is not allowed. Once you have determined the pieces, you can only translate and rotate them.

If you are very thirsty, you might want to try this out on a bag of oranges at home. Note that the inordinate amount of time you spend cleaning up your mess, as well as your abject failure in accomplishing this task, might result from the fact that most knives on the market today produce so-called "measurable" cuts of your oranges. The same applies to grapefruit, and need we say (?), most other food products.

The Banach-Tarski Paradox is, to coin a phrase, "paradoxical". It is infuriating, in large measure due to the fact that it is so highly counterintuitive. On the other hand, once you accept the Axiom of Choice, this result follows as a consequence, and you have no choice but to accept it and move on with your life.

4. So what to do?

4.1. Zermelo-Fraenkel theory plus Choice (ZFC) is popular because it works. It works well. The Banach-Tarski Paradox and the existence of non-measurable sets aside, it produces so many great results that the vast majority of mathematicians use it without hesitation. Having said that – the Axiom of Choice is considered *special*, and unlike the other (ZF) axioms. In general, when it is not required to prove a result, a proof that avoids it is considered better than a proof that uses it. Also, it is often considered "good form" to indicate when it is being used, although some uses of the Axiom of Choice have become so commonplace that mathematicians expect the reader to realise on their own that the Axiom of Choice has been used in the argument.

There is a group of mathematicians who feel that to prove that a mathematical object exists, one must construct it. They are referred to as *constructivists* (at least this is how they are referred to in polite company). Your humble author has not met a constructivist mathematician himself, and it occurs to your humble author as he writes this that if we were to use constructivist logic, then we could not conclude that one exists, could we? [How do I know that the articles that refer to them aren't all made up?] But this author is not a constructivist, and so is willing to accept that they do exist.

4.2. We conclude by indicating that the Continuum Hypothesis, which was also shown by Gödel and by Cohen to be independent of (ZFC), is not universally accepted at all. Any argument that uses the Continuum Hypothesis should mention where it is being used, and should be avoided if there is any way at all of obtaining the same result without it. Good results have been proven that require it, and are usually stated as: *assuming the continuum hypothesis, we show that ...* .

4.3. Quiz. Can you avoid the Axiom of choice to choose an element from

(a) a finite set?
(b) each member of an infinite set of sets, each of which has only one element?
(c) each member of a finite set of sets if each of the members is infinite?
(d) each member of an infinite set of sets of rationals?
(e) each member of an infinite set of finite sets of reals?
(f) each member of a denumerable set of sets if each of the members is denumerable?
(g) each member of an infinite set of sets of reals?
(h) each member of an infinite set of sets of integers?

5. Equivalences

Let us now provide a proof of the equivalence of the Axiom of Choice, Zorn's Lemma and the Well-Ordering Principle. We remind the reader that a **poset** is a partially ordered set, as defined above. We begin with the definition of an **initial segment**, which will be required in the proof.

We also remind the reader that the statement that, given a collection $\{X_\lambda\}_{\lambda\in\Lambda}$ of non-empty sets, the existence of a **choice function** $f : \Lambda \to \cup_{\lambda\in\Lambda} X_\lambda$ such that $f(\lambda) \in X_\lambda$ for all $\lambda \in \Lambda$ is the statement that

$$\prod_{\lambda\in\Lambda} X_\lambda \neq \varnothing,$$

for the simple reason that

$$\prod_{\lambda\in\Lambda} X_\lambda := \{f : \Lambda \to \cup_{\lambda\in\Lambda} X_\lambda \text{ a function such that } f(\lambda) \in X_\lambda \text{ for all } \lambda \in \Lambda\}.$$

5.1. Definition. *Let $(X, \leq)$ be a poset, $C \subseteq X$ be a chain in X and $d \in C$. We define*

$$P(C, d) = \{c \in C : c < d\}.$$

*An **initial segment** of C is a subset of the form $P(C, d)$ for some $d \in C$.*

5.2. Example.

(a) For each $r \in \mathbb{R}$, $(-\infty, r)$ is an initial segment of $(\mathbb{R}, \leq)$.
(b) For each $n \in \mathbb{N}$, $\{1, 2, ..., n\}$ is an initial segment of $\mathbb{N}$.

5.3. Theorem. *The following are equivalent:*

(i) *The Axiom of Choice (AC): given a non-empty collection* $\{X_\lambda\}_{\lambda\in\Lambda}$ *of non-empty sets,* $\prod_{\lambda\in\Lambda} X_\lambda \neq \varnothing$.

(ii) *Zorn's Lemma (ZL): Let* $(Y,\leq)$ *be a poset. Suppose that every chain* $\mathcal{C}\subseteq Y$ *has an upper bound. Then* Y *has a maximal element.*

(iii) *The Well-Ordering Principle (WO): Every non-empty set* Z *admits a well-ordering.*

Proof.

(i) implies (ii): This is the most delicate of the three implications. We shall argue by contradiction.

Suppose that $(X,\leq)$ is a poset such that every chain in X is bounded above, but that X no maximal elements. Given a chain $C\subseteq X$, we can find an upper bound u_C for C. Since u_C is not a maximal element, we can find $v_C\in X$ with $u_C < v_C$. We shall refer to such an element v_C as a **strict upper bound** for C.

By the Axiom of Choice, for each chain C in X, we can choose a strict upper bound $f(C)$. If $C=\varnothing$, we arbitrarily select $x_0\in X$ and set $f(\varnothing)=x_0$.

We shall say that a subset $A\subseteq X$ satisfies **property L** if

(I) The partial order $\leq$ on X when restricted to A is a well-ordering of A, and

(II) for all $x\in A$, $x=f(P(A,x))$.

• **Claim 1:** *if* $A,B\subseteq X$ *satisfy property L and* $A\neq B$*, then either* A *is an initial segment of* B*, or* B *is an initial segment of* A.

Without loss of generality, we may assume that $A\setminus B\neq\varnothing$. Let

$$x=\min\{a\in A : a\notin B\}.$$

Note that x exists because A is well-ordered. Then $P(A,x)\subseteq B$. We shall argue that $B=P(A,x)$. If not, then $B\setminus P(A,x)\neq\varnothing$, and using the well-orderedness of B,

$$y=\min\{b\in B : b\notin P(A,x)\}$$

exists. Thus $P(B,y)\subseteq P(A,x)$.

Let $z=\min(A\setminus P(B,y))$. Then $z\leq x=\min(A\setminus B)$.

- **Subclaim 1:** $P(A,z)=P(B,y)$.
  By definition, $P(A,z)\subseteq P(B,y)$.
  To obtain the reverse inclusion, we first argue that if $t\in P(B,y)=A\cap P(B,y)$, then $P(A,t)\cup\{t\}\subseteq P(B,y)$. By hypothesis, $t\in P(B,y)$, so suppose that $u\in P(A,t)$. Now $t\in P(B,y)\subseteq P(A,x)$, so $u<t<x$ implies that $u\in P(A,x)$. In other words, $P(A,t)\subseteq P(A,x)\subseteq B$. But then $u\in B$ and $u<t<y$ implies that $u\in P(B,y)$.

We now have that if $s \in P(B,y)$, then $P(A,s) \cup \{s\} \subseteq P(B,y) \subseteq P(A,x) \subseteq A$. This forces $s < z := \min(A \setminus P(B,y))$, so that $s \in P(A,z)$. Together, we find that $P(B,y) \subseteq P(A,z) \subseteq P(B,y)$, which proves the subclaim.

Returning to the proof of the claim, we now have that $z = f(P(A,z)) = f(P(B,y)) = y$. But $y \in B$, so $y \neq x$. Hence $z < x$. Thus $y = z \in P(A,x)$, contradicting the definition of y. We deduce that $P(A,x) = B$, and hence that B is an initial segment of A, thereby proving our claim.

Suppose that $A \subseteq X$ has property L, and let $x \in A$. It follows from the above argument that given $y < x$, either $y \in A$ or y does not belong to any set B with property L.

Let $V = \cup\{A \subseteq X : A \text{ has property } L\}$.

• **Claim 2:** *We claim that if $w = f(V)$, then $V \cup \{w\}$ has property L.*

Suppose that we can show this. Then $V \cup \{w\} \subseteq V$, so $w \in V$, a contradiction. This will complete the proof.

• **Subclaim 2a:** *First we show that V itself has property L.* We must show that V is well-ordered, and that for all $x \in V$, $x = f(P(V,x))$.

(a) V is well-ordered.
Let $\varnothing \neq B \subseteq V$. Then there exists $A_0 \subseteq X$ so that A_0 has property L and $B \cap A_0 \neq \varnothing$. Since A_0 is well-ordered and $\varnothing \neq B \cap A_0 \subseteq A_0$, $m := \min(B \cap A_0)$ exists. We claim that $m = \min(B)$.
Suppose that $y \in B$. Then there exists $A_1 \subseteq X$ so that A_1 has property L and $y \in A_1$. Now, both A_0 and A_1 have property L:
- ◇ if $A_0 = A_1$, then $m = \min(B \cap A_1)$, so $m \leq y$.
- ◇ if $A_0 \neq A_1$, then either
  - A_0 is an initial segment of A_1, so $A_0 = P(A_1,d)$ for some $d \in A_1$. Then

    $$m = \min(B \cap A_0) = \min(B \cap A_1),$$

    since $r \in A_1 \setminus A_0$ implies that $m < d \leq r$. Hence $m \leq y$;, or
  - A_1 is an initial segment of A_0, say $A_1 = P(A_0,d) \subseteq A_0$ for some $d \in A_0$. Then

    $$m = \min(B \cap A_0) \leq \min(B \cap A_1).$$

    Hence $m \leq y$.

In both cases we see that $m \leq y$. Since $y \in B$ was arbitrary, $m = \min(B)$.
Thus, any non-empty subset B of V has a minimum element, and so V is well-ordered.

(b) Let $x \in V$. Then there exists $A_2 \subseteq X$ with property L so that $x \in A_2$. Then $x = P(A_2,x)$. Suppose that $y \in V$ and $y < x$. Then there exists

$A_3 \subseteq X$ with property L so that $y \in A_3$. Since A_2 and A_3 both have property L, either

- $A_2 = A_3$, and so $y \in A_2$; or
- $A_2 = P(A_3, d)$ for some $d \in A_3$. Since $x \in A_2$, $P(A_2, x) = P(A_3, x)$ and therefore $y \in A_2$; or
- $A_3 = P(A_2, d)$ for some $d \in A_2$. Then $y \in A_3$ implies that $y \in A_2$.

In any of these three cases, $y \in A_2$. Hence $P(V, x) \subseteq P(A_2, x)$. Since $A_2 \subseteq V$, we have that $P(A_2, x) \subseteq P(V, x)$, whence $P(A_2, x) = P(V, x)$. But then

$$x = f(P(A_2, x)) = f(P(V, x)).$$

By (a) and (b), V has property L.

We now return to the proof of Claim 2. That is, we prove that if $w = f(V)$, then $V \cup \{w\}$ has property L.

(I) $V \cup \{w\}$ is well-ordered.
We know that V is well-ordered by part (a) above. Suppose that $\varnothing \neq B \subseteq V \cup \{w\}$. If $B \cap V \neq \varnothing$, then by (a) above, $m := \min(B \cap V)$ exists. Clearly $m \in V$ implies $m \leq f(V) = w$, so $m = \min(B \cap (V \cup \{w\}))$. If $\varnothing \neq B \subseteq V \cup \{w\}$ and $B \cap V = \varnothing$, then $B = \{w\}$, and so $w = \min(B)$ exists.
Hence $V \cup \{w\}$ is well-ordered.

(II) Let $x = V \cup \{w\}$. If $x \in V$, then $x = f(P(V, x))$ by part (a). If $x = w$, then

$$P(V \cup \{w\}, x) = P(V \cup \{w\}, w) = V,$$

so $x = w = f(V) = f(P(V \cup \{w\}, x))$.

By (I) and (II), $V \cup \{w\}$ has property L. As we saw in the statement following Claim 2, this completes the proof that the Axiom of Choice implies Zorn's Lemma. Now let us never speak of this again.

(ii) implies (iii): Let $X \neq \varnothing$ be a set. It is clear that every finite subset $F \subseteq X$ can be well-ordered. Let $\mathcal{A}$ denote the collection of pairs $(Y, \leq_Y)$, where $Y \subseteq X$ and $\leq_Y$ is a well-ordering of Y. For $(A, \leq_A)$, $(B, \leq_B) \in \mathcal{A}$, observe that A is an **initial segment** of B if the following two conditions are met:

- $A \subseteq B$ and $a_1 \leq_A a_2$ implies that $a_1 \leq_B a_2$;
- if $b \in B \setminus A$, then $a \leq_B b$ for all $a \in A$.

Let us partially order $\mathcal{A}$ by setting $(A, \leq_A) \leq (B, \leq_B)$ if A is an initial segment of B. Let $\mathcal{C} = \{C_\lambda\}_{\lambda \in \Lambda}$ be a chain in $\mathcal{A}$.

Then (**exercise**): $\cup_{\lambda \in \Lambda} C_\lambda$ is an upper bound for $\mathcal{C}$.

By Zorn's Lemma, $\mathcal{A}$ admits a maximal element, say $(M, \leq_M)$. We claim that $M = X$. Suppose otherwise. Then we can choose $x_0 \in X \setminus M$ and set $M_0 = M \cup \{x_0\}$. define a partial order on M_0 via: $x \leq_{M_0} y$ if either (a) $x, y \in M$ and $x \leq_M y$, or (b) x is arbitrary and $y = x_0$. Then $(M_0, \leq_{M_0})$ is a well-ordered set and $(M, \leq_M) < (M_0, \leq_{M_0})$, a contradiction

of the maximality of $(M, \leq_M)$. Thus $M = X$ and $\leq_M$ is a well- ordering of X.

(iii) implies (i):

Suppose that $\{X_\lambda\}_{\lambda \in \Lambda}$ is a non-empty collection of non-empty sets. Let $X = \cup_{\lambda \in \Lambda} X_\lambda$. By hypothesis, X admits a well-ordering $\leq_X$. Since each $\varnothing \neq X_\lambda \subseteq X$, it has a minimum element relative to the ordering on X. Define a choice function f by setting $f(\lambda)$ to be this minimum element of X_λ for each $\lambda \in \Lambda$.

□

Bibliography

[Her69] I.N. Herstein. *Topics in Ring Theory*. Chicago Lecture Notes in Mathematics. University of Chicago Press, Chicago and London, 1969.

[Her75] I.N. Herstein. *Topics in Algebra, 2nd Edition*. Chicago Lecture Notes in Mathematics. John Wiley and Sons, New York, Chichester, Brisbane, Toronto and Singapore, 1975.

[Ste04] I. Stewart. *Galois Theory*. CRC Mathematics Texts. Chapman and Hall, Boca Raton, London, New York, Washington, D.C., 2004.

[vL82] C.L.F. von Lindemann. Über die Zahl π. *Mathematische Annalen*, 20:213–225, 1882.

Index