A Mathematical Introduction to Signals and Systems

Volume I. Introductory analysis and algebra

Andrew D. Lewis

This version: 2022/03/07

Preface for series

The subject of signals and systems, particularly linear systems, is by now an entrenched part of the curriculum in many engineering disciplines, particularly electrical engineering. Furthermore, the offshoots of signals and systems theory—e.g., control theory, signal processing, and communications theory—are themselves well-developed and equally basic to many engineering disciplines. As many a student will agree, the subject of signals and systems is one with a reliance on tools from many areas of mathematics. However, much of this mathematics is not revealed to undergraduates, and necessarily so. Indeed, a complete accounting of what is involved in signals and systems theory would take one, at times quite deeply, into the fields of linear algebra (and to a lesser extent, algebra in general), real and complex analysis, measure and probability theory, and functional analysis. Indeed, in signals and systems theory, many of these topics are woven together in surprising and often spectacular ways. The existing texts on signals and systems theory, and there is a true abundance of them, all share the virtue of presenting the material in such a way that it is comprehensible with the bare minimum background.

Should I bother reading these volumes?

This virtue comes at a cost, as it must, and the reader must decide whether this cost is worth paying. Let us consider a concrete example of this, so that the reader can get an idea of the sorts of matters the volumes in this text are intended to wrestle with. Consider the function of time

$$f(t) = \begin{cases} \mathrm{e}^{-t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

In the text (Example IV-6.1.3–2) we shall show that, were one to represent this function in the frequency domain with frequency represented by ν, we would get

$$\hat{f}(\nu) = \int_{\mathbb{R}} f(t)\mathrm{e}^{-2\mathrm{i}\pi\nu t}\,\mathrm{d}t = \frac{1}{1 + 2\mathrm{i}\pi\nu}.$$

The idea, as discussed in Chapter IV-2, is that $\hat{f}(\nu)$ gives a representation of the "amount" of the signal present at the frequency ν. Now, it is desirable to be able to reconstruct f from $\hat{f}$, and we shall see in Section IV-6.2 that this is done via the formula

$$f(t) \text{``=''} \int_{\mathbb{R}} \hat{f}(\nu)\mathrm{e}^{2\mathrm{i}\pi\nu t}\,\mathrm{d}\nu. \tag{FT}$$

The easiest way to do the integral is, of course, using a symbolic manipulation program. I just tried this with Mathematica®, and I was told it could not do the computation. Indeed, the integral *does not converge*! Nonetheless, in many tables of

Fourier transforms (that is what the preceding computations are about), we are told that the integral in (FT) does indeed produce $f(t)$. Are the tables wrong? Well, no. But they are only correct when one understands exactly what the right-hand side of (FT) means. What it means is that the integral converges, *in* $\mathsf{L}^2(\mathbb{R};\mathbb{C})$ to f. Let us say some things about the story behind this that are of a general nature, and apply to many ideas in signal and system theory, and indeed to applied mathematics as a whole.

1. The story—it is the story of the L^2-Fourier transform—is not completely trivial. It requires *some* delving into functional analysis at least, and some background in integration theory, if one wishes to understand that "L" stands for "Lebesgue," as in "Lebesgue integration." At its most simple-minded level, the theory is certainly understandable by many undergraduates. Also, at its most simple-minded level, it raises more questions than it answers.
2. The story, even at the most simple-minded level alluded to above, takes some time to deliver. The full story takes *a lot* of time to deliver.
3. It is not necessary to fully understand the story, perhaps even the most simple-minded version of it, to be a user of the technology that results.
4. By understanding the story well, one is led to new ideas, otherwise completely hidden, that are practically useful. In control theory, quadratic regulator theory, and in signal processing, the Kalman filter, are examples of this.
5. The full story of the L^2-Fourier transform, and the issues stemming from it, directly or otherwise, is beautiful.

The nature of the points above, as they relate to this series, are as follows. Points 1 and 2 indicate why the story cannot be told to all undergraduates, or even most graduate students. Point 3 indicates why it is okay that the story not be told to everyone. Point 4 indicates why it is important that the story be told to someone. Point 5 should be thought of as a sort of benchmark as to whether the reader should bother with understanding what is in this series. Here is how to apply it. If one reads the assertion that this is a beautiful story, and their reaction is, "Okay, but there better be a payoff," or, "So what?" or, "Beautiful to who?" then perhaps they should steer clear of this series. If they read the assertion that this is a beautiful story, and respond with, "Really? Tell me more," then I hope they enjoy these books. They were written for such readers. Of course, most readers' reactions will fall somewhere in between the above extremes. Such readers will have to sort out for themselves whether the volumes in this series lie on the right side, for them, of being worth reading. For these readers I will say that this series is *heavily* biased towards readers who react in an unreservedly positive manner to the assertions of intrinsic beauty.

For readers skeptical of assertions of the usefulness of mathematics, an interesting pair of articles concerning this is [Wigner 1960] and [Hamming 1980].

What is the best way of getting through this material?

Now that a reader has decided to go through with understanding what is in these volumes, they are confronted with actually doing so: a possibly nontrivial matter, depending on their starting point. Let us break down our advice according to the background of the reader.

I look at the tables of contents, and very little seems familiar. Clearly if nothing seems familiar at all, then a reader should not bother reading on until they have acquired an at least passing familiarity with some of the topics in the book. This can be done by obtaining an undergraduate degree in electrical engineering (or similar), or pure or applied mathematics.

If a reader already possess an undergraduate degree in mathematics or engineering, then certainly some of the following topics will appear to be familiar: linear algebra, differential equations, some transform analysis, Fourier series, system theory, real and/or complex analysis. However, it is possible that they have not been taught in a manner that is sufficiently broad or deep to quickly penetrate the texts in this series. That is to say, relatively inexperienced readers will find they have some work to do, even to get into topics with which they have some familiarity. The best way to proceed in these cases depends, to some extent, on the nature of one's background.

I am familiar with some or all of the applied topics, but not with the mathematics. For readers with an engineering background, even at the graduate level, the depth with which topics are covered in these books is perhaps a little daunting. The best approach for such readers is to select the applied topic they wish to learn more about, and then use the text as a guide. When a new topic is initiated, it is clearly stated what parts of the book the reader is expected to be familiar with. The reader with a more applied background will find that they will not be able to get far without having to unravel the mathematical background almost to the beginning. Indeed, readers with a typical applied background will normally be lacking a good background in linear algebra and real analysis. Therefore, they will need to invest a good deal of effort acquiring some quite basic background. At this time, they will quickly be able to ascertain whether it is worth proceeding with reading the books in this series.

I am familiar with some or all of the mathematics, but not with the applied topics. Readers with an undergraduate degree in mathematics will fall into this camp, and probably also some readers with a graduate education in engineering, depending on their discipline. They may want to skim the relevant background material, just to see what they know and what they don't know, and then proceed directly to the applied topics of interest.

I am familiar with most of the contents. For these readers, the series is one of reference books.

Comments on organisation

In the current practise of teaching areas of science and engineering connected with mathematics, there is much emphasis on "just in time" delivery of mathematical ideas and techniques. Certainly I have employed this idea myself in the classroom, without thinking much about it, and so apparently I think it a good thing. However, the merits of the "just in time" approach in written work are, in my opinion, debatable. The most glaring difficulty is that the same mathematical ideas can be "just in time" for multiple non-mathematical topics. This can even happen in a single one semester course. For example—to stick to something germane to this series—are differential equations "just in time" for general system theory? for modelling? for feedback control theory? The answer is, "For all of them," of course. However, were one to choose one of these topics for a "just in time" written delivery of the material, the presentation would immediately become awkward, especially in the case where that topic were one that an instructor did not wish to cover in class.

Another drawback to a "just in time" approach in written work is that, when combined with the corresponding approach in the classroom, a connection, perhaps unsuitably strong, is drawn between an area of mathematics and an area of application of mathematics. Given that one of the strengths of mathematics is to facilitate the connecting of seemingly disparate topics, inside and outside of mathematics proper, this is perhaps an overly simplifying way of delivering mathematical material. In the "just simple enough, but not too simple" spectrum, we fall on the side of "not too simple."

For these reasons and others, the material in this series is generally organised according to its mathematical structure. That is to say, mathematical topics are treated independently and thoroughly, reflecting the fact that they have life independent of any specific area of application. We do not, however, slavishly follow the Bourbaki[1] ideals of logical structure. That is to say, we do allow ourselves the occasional forward reference when convenient. However, we are certainly careful to maintain the standards of deductive logic that currently pervade the subject of "mainstream" mathematics. We also do not slavishly follow the Bourbaki dictum of starting with the most general ideas, and proceeding to the more specific. While there is something to be said for this, we feel that for the subject and intended readership of this series, such an approach would be unnecessarily off-putting.

Andrew D. Lewis Kingston, ON, Canada

[1] Bourbaki refers to "Nicolas Bourbaki," a pseudonym given (by themselves) to a group of French mathematicians who, beginning in mid-1930's, undertook to rewrite the subject of mathematics. Their dictums include presenting material in a completely logical order, where no concept is referred to before being defined, and starting developments from the most general, and proceeding to the more specific. The original members include Henri Cartan, André Weil, Jean Delsarte, Jean Dieudonné, and Claude Chevalley, and the group later counted such mathematicians as Roger Godement, Jean-Pierre Serre, Laurent Schwartz, Emile Borel, and Alexander Grothendieck among its members. They have produced eight books on fundamental subjects of mathematics.

Preface for Volume 1

In this, the first volume of five volumes in this series, we shall introduce elementary mathematics, many parts of which comprise an undergraduate degree in mathematics, either in explicit or implicit form. For readers not having a strong mathematics background—such as students in engineering or the physical sciences—the material here may look familiar, but are covered in depth, breadth, and detail in ways that far exceed the mathematics such readers will have previously encountered. The material covered is, for the main part, required at various points in subsequent volumes. However, we also present some material for the purpose of providing context, historical background, or topics of general interest related to the core material.

Our presentation begins in Chapter 1 with basic a presentation of set theory and basic mathematical notation. This should be regarded as lighter fare, but required reading for students with a weaker mathematical background. Oftentimes mathematics is regarded as a toolbox into which one can reach for the hammer one needs to smash the nail with which one is presently confronted. This is a rather dull view of what mathematics can accomplish. Indeed, one way of thinking about mathematics is that it provides a certain way of approaching problems, an approach where care, precision, and rigour are as important as methodology. To be able to take advantage of this way of approaching problems, one must be able to express problems and solutions using the language that has been especially adapted to the approach. One way to view Chapter 1 is that it provides the most elementary parts of the required vocabulary.

The next two chapters, Chapters 2 and 3, deal with a systematic development of the set of real numbers and functions of a single real variable. The material in these chapters, then, covers what often comprises a pair of courses in introductory calculus. It does so, however, in a comprehensive and rigorous way. Also, unlike in a typical introductory course in calculus, the emphasis is not as much on computing things like derivatives and integrals as on understanding what derivatives and integrals "really are," and on proving some of their useful properties. The computational aspects of the subject can all be gleaned from what is presented in Chapters 2 and 3, but to be really proficient in these things requires learning from a more standard introductory level course or textbook.

In Chapter 4 we present the background in algebra that we will subsequently require. A background in algebra is one thing that is typically deficient for students in engineering or the physical sciences. However, a decent background in algebra is essential in these volumes, and we shall often encounter somewhat advanced algebraic ideas outside the scope of topics that belong to algebra, proper. The material in this chapter resembles, in some ways, material that might be covered as part of an undergraduate curriculum in algebra. For example, the high-level list of topics is what one might cover in an undergraduate programme. However, there are substantial differences as well, and these resemble the differences between

Chapters 2 and 3; the topics are covered in greater depth and generality, with an emphasis on structure over specific examples and computation. This is a decided drawback of our approach here, and a reader can anticipate needing to fill in some facility with some more elementary ideas to claim proficiency with basic algebraic topics.

The final topic in this volume is linear algebra. Unlike "mere algebra," linear algebra is often a part of the background of students in engineering and the physical sciences. We cover linear algebra in Chapter 5. Here the presentation differs radically from what one typically covers in the linear algebra portion of any undergraduate education. An undergraduate in engineering and physical sciences will have seen linear algebra with an emphasis on matrices. An undergraduate in mathematics will preferably have seen an abstract linear algebra course, where the emphasis is on finite-dimensional vector spaces, and the structure of homomorphisms of such spaces, with a particular emphasis on the structure of endomorphisms. Our approach is infinite-dimensional from the start, and works with general (commutative) rings as well as fields. There is also a coverage of topics that simply do not appear in an undergraduate curriculum, such as multilinear algebra. That being said, we do also consider the usual players in linear algebra, such as matrices (though with possibly infinitely many rows and columns) and determinants. We also provide a comprehensive treatment of the structure of endomorphisms of finite-dimensional vector spaces.

Andrew D. Lewis Kingston, ON, Canada

Table of Contents

Chapter 1

Set theory and terminology

The principle purpose of this chapter is to introduce the mathematical notation and language that will be used in the remainder of these volumes. Much of this notation is standard, or at least the notation we use is generally among a collection of standard possibilities. In this respect, the chapter is a simple one. However, we also wish to introduce the reader to some elementary, although somewhat abstract, mathematics. The secondary objective behind this has three components.

1. We aim to provide a somewhat rigorous foundation for what follows. This means being fairly clear about defining the (usually) somewhat simple concepts that arise in the chapter. Thus "intuitively clear" concepts like sets, subsets, maps, etc., are given a fairly systematic and detailed discussion. It is at least interesting to know that this can be done. And, if it is not of interest, it can be sidestepped at a first reading.
2. This chapter contains some results, and many of these require very simple proofs. We hope that these simple proofs might be useful to readers who are new to the world where everything is proved. Proofs in other chapters in these volumes may not be so useful for achieving this objective.
3. The material is standard mathematical material, and should be known by anyone purporting to love mathematics.

Do I need to read this chapter? Readers who are familiar with standard mathematical notation (e.g., who understand the symbols $\in$, $\subseteq$, $\cup$, $\cap$, $\times$, $f\colon S \to T$, $\mathbb{Z}_{>0}$, and $\mathbb{Z}$) can simply skip this chapter in its entirety. Some ideas (e.g., relations, orders, Zorn's Lemma) may need to be referred to during the course of later chapters, but this is easily done.

Readers not familiar with the above standard mathematical notation will have some work to do. They should certainly read Sections 1.1, 1.2, and 1.3 closely enough that they understand the language, notation, and main ideas. And they should read enough of Section 1.4 that they know what objects, familiar to them from their being human, the symbols $\mathbb{Z}_{>0}$ and $\mathbb{Z}$ refer to. The remainder of the material can be overlooked until it is needed later. •

Contents

Section 1.1

Sets

The basic ingredient in modern mathematics is the set. The idea of a set is familiar to everyone at least in the form of "a collection of objects." In this section, we shall not really give a definition of a set that goes beyond that intuitive one. Rather we shall accept this intuitive idea of a set, and move forward from there. This way of dealing with sets is called ***naïve set theory***. There are some problems with naïve set theory, as described in Section 1.8.1, and these lead to a more formal notion of a set as an object that satisfies certain axioms, those given in Section 1.8.2. However, these matters will not concern us much at the moment.

Do I need to read this section? Readers familiar with basic set theoretic notation can skip this section. Other readers should read it, since it contains language, notation, and ideas that are absolutely commonplace in these volumes. •

1.1.1 Definitions and examples

First let us give our working definition of a set. A ***set*** is, for us, a well-defined collection of objects. Thus one can speak of everyday things like "the set of red-haired ladies who own yellow cars." Or one can speak of mathematical things like "the set of even prime numbers." Sets are therefore defined by describing their ***members*** or ***elements***, i.e., those objects that are in the set. When we are feeling less formal, we may refer to an element of a set as a ***point*** in that set. The set with no members is the ***empty set***, and is denoted by $\varnothing$. If S is a set with member x, then we write $x \in S$. If an object x is *not* in a set S, then we write $x \notin S$.

1.1.1 Examples (Sets)

1. If S is the set of even prime numbers, then $2 \in S$.
2. If S is the set of even prime numbers greater than 3, then S is the empty set.
3. If S is the set of red-haired ladies who own yellow cars and if $x =$ Ghandi, then $x \notin S$. •

If it is possible to write the members of a set, then they are usually written between braces $\{\ \}$. For example, the set of prime numbers less that 10 is written as $\{2, 3, 5, 7\}$ and the set of physicists to have won a Fields Prize as of 2005 is {Edward Witten}.

A set S is a ***subset*** of a set T if $x \in S$ implies that $x \in T$. We shall write $S \subseteq T$, or equivalently $T \supseteq S$, in this case. If $x \in S$, then the set $\{x\} \subseteq S$ with one element, namely x, is a ***singleton***. Note that x and $\{x\}$ are different things. For example, $x \in S$ and $\{x\} \subseteq S$. If $S \subseteq T$ and if $T \subseteq S$, then the sets S and T are ***equal***, and we write $S = T$. If two sets are not equal, then we write $S \neq T$. If $S \subseteq T$ and if $S \neq T$, then S

is a ***proper*** or ***strict*** subset of T, and we write $S \subset T$ if we wish to emphasise this fact.

1.1.2 Notation (Subsets and proper subsets) We adopt a particular convention for denoting subsets and proper subsets. That is, we write $S \subseteq T$ when S is a subset of T, allowing for the possibility that $S = T$. When $S \subseteq T$ and $S \neq T$ we write $S \subset T$. In this latter case, many authors will write $S \subsetneq T$. We elect not to do this. The convention we use is consistent with the convention one normally uses with inequalities. That is, one normally writes $x \leq y$ and $x < y$. It is not usual to write $x \lneq y$ in the latter case. •

Some of the following examples may not be perfectly obvious, so may require sorting through the definitions.

1.1.3 Examples (Subsets)

1. For any set S, $\varnothing \subseteq S$ (see Exercise 1.1.1).
2. $\{1, 2\} \subseteq \{1, 2, 3\}$.
3. $\{1, 2\} \subset \{1, 2, 3\}$.
4. $\{1, 2\} = \{2, 1\}$.
5. $\{1, 2\} = \{2, 1, 2, 1, 1, 2\}$. •

A common means of defining a set is to define it as the subset of an existing set that satisfies conditions. Let us be slightly precise about this. A ***one-variable predicate*** is a statement which, in order that its truth be evaluated, needs a single argument to be specified. For example, $P(x) =$ "x is blue" needs the single argument x in order that it be decided whether it is true or not. We then use the notation

$$\{x \in S \mid P(x)\}$$

to denote the members x of S for which the predicate P is true when evaluated at x. This is read as something like, "the set of x's in S such that $P(x)$ holds."

For sets S and T, the ***relative complement*** of T in S is the set

$$S - T = \{x \in S \mid x \notin T\}.$$

Note that for this to make sense, we do not require that T be a subset of S. It is a common occurrence when dealing with complements that one set be a subset of another. We use different language and notation to deal with this. If S is a set and if $T \subseteq S$, then $S \setminus T$ denotes the ***absolute complement*** of T in S, and is defined by

$$S \setminus T = \{x \in S \mid x \notin T\}.$$

Note that, if we forget that T is a subset of S, then we have $S \setminus T = S - T$. Thus $S - T$ is the more general notation. Of course, if $A \subseteq T \subseteq S$, one needs to be careful when using the words "absolute complement of A," since one must say whether one is taking the complement in T or the larger complement in S. For this reason, we prefer the notation we use rather the commonly encountered notation A^{C} or A' to refer to the absolute complement. Note that one should not talk about the absolute complement to a set, without saying within which subset the complement is being taken. To do so would imply the existence of "a set containing all sets," an object that leads one to certain paradoxes (see Section 1.8).

A useful set associated with every set S is its ***power set***, by which we mean the set

$$\mathbf{2}^S = \{A \mid A \subseteq S\}.$$

The reader can investigate the origins of the peculiar notation in Exercise 1.1.3.

1.1.2 Unions and intersections

In this section we indicate how to construct new sets from existing ones.

Given two sets S and T, the ***union*** of S and T is the set $S \cup T$ whose members are members of S *or* T. The ***intersection*** of S and T is the set $S \cap T$ whose members are members of S *and* T. If two sets S and T have the property that $S \cap T = \emptyset$, then S and T are said to be ***disjoint***. For sets S and T their ***symmetric complement*** is the set

$$S \triangle T = (S - T) \cup (T - S).$$

Thus $S \triangle T$ is the set of objects in union $S \cup T$ that do not lie in the intersection $S \cap T$. The symmetric complement is so named because $S \triangle T = T \triangle S$. In Figure 1.1 we give Venn diagrams describing union, intersection, and symmetric complement.

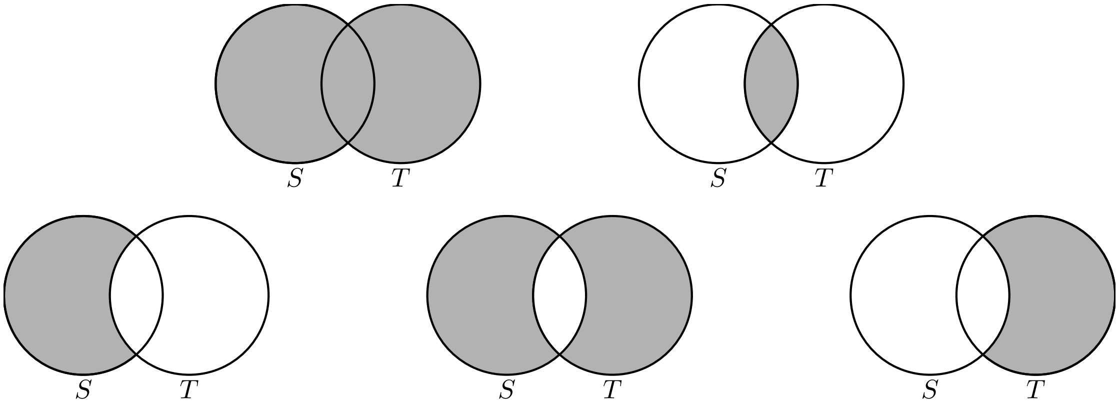


Figure 1.1 $S \cup T$ (top left), $S \cap T$ (top right), $S - T$ (bottom left), $S \triangle T$ (bottom middle), and $T - S$ (bottom right)

The following result gives some simple properties of pairwise unions and intersections of sets. We leave the straightforward verification of some or all of these to the reader as Exercise 1.1.5.

1.1.4 Proposition (Properties of unions and intersections) *For sets* S *and* T*, the following statements hold:*

(i) $\mathrm{S} \cup \emptyset = \mathrm{S}$*;*

(ii) $\mathrm{S} \cap \emptyset = \emptyset$*;*

(iii) $\mathrm{S} \cup \mathrm{S} = \mathrm{S}$*;*

(iv) $\mathrm{S} \cap \mathrm{S} = \mathrm{S}$*;*

(v) $\mathrm{S} \cup \mathrm{T} = \mathrm{T} \cup \mathrm{S}$ *(**commutativity**);*

(vi) $\mathrm{S} \cap \mathrm{T} = \mathrm{T} \cap \mathrm{S}$ *(**commutativity**);*

(vii) $\mathrm{S} \subseteq \mathrm{S} \cup \mathrm{T}$*;*

(viii) $\mathrm{S} \cap \mathrm{T} \subseteq \mathrm{S}$*;*

(ix) $\mathrm{S} \cup (\mathrm{T} \cup \mathrm{U}) = (\mathrm{S} \cup \mathrm{T}) \cup \mathrm{U}$ *(**associativity**);*

(x) $\mathrm{S} \cap (\mathrm{T} \cap \mathrm{U}) = (\mathrm{S} \cap \mathrm{T}) \cap \mathrm{U}$ *(**associativity**);*

(xi) $\mathrm{S} \cap (\mathrm{T} \cup \mathrm{U}) = (\mathrm{S} \cap \mathrm{T}) \cup (\mathrm{S} \cap \mathrm{U})$ *(**distributivity**);*

(xii) $\mathrm{S} \cup (\mathrm{T} \cap \mathrm{U}) = (\mathrm{S} \cup \mathrm{T}) \cap (\mathrm{S} \cup \mathrm{U})$ *(**distributivity**).*

We may more generally consider not just two sets, but an arbitrary collection $\mathscr{S}$ of sets. In this case we *posit* the existence of a set, called the ***union*** of the sets $\mathscr{S}$, with the property that it contains each element of each set $S \in \mathscr{S}$. Moreover, one can specify the subset of this big set to *only* contain members of sets from $\mathscr{S}$. This set we will denote by $\cup_{S\in\mathscr{S}} S$. We can also perform a similar construction with intersections of an arbitrary collection $\mathscr{S}$ of sets. Thus we denote by $\cap_{S\in\mathscr{S}} S$ the set, called the ***intersection*** of the sets $\mathscr{S}$, having the property that $x \in \cap_{S\in\mathscr{S}} S$ if $x \in S$ for every $S \in \mathscr{S}$. Note that we do not need to posit the existence of the intersection.

Let us give some properties of general unions and intersections as they relate to complements.

1.1.5 Proposition (De Morgan's[1] Laws) *Let* T *be a set and let* $\mathscr{S}$ *be a collection of subsets of* T*. Then the following statements hold:*

(i) $\mathrm{T} \setminus (\cup_{\mathrm{S}\in\mathscr{S}} \mathrm{S}) = \cap_{\mathrm{S}\in\mathscr{S}} (\mathrm{T} \setminus \mathrm{S})$*;*

(ii) $\mathrm{T} \setminus (\cap_{\mathrm{S}\in\mathscr{S}} \mathrm{S}) = \cup_{\mathrm{S}\in\mathscr{S}} (\mathrm{T} \setminus \mathrm{S})$*.*

Proof (i) Let $x \in T \setminus (\cup_{S\in\mathscr{S}})$. Then, for each $S \in \mathscr{S}$, $x \notin S$, or $x \in T \setminus S$. Thus $x \in \cap_{S\in\mathscr{S}} (T \setminus S)$. Therefore, $T \setminus (\cup_{S\in\mathscr{S}}) \supseteq \cap_{S\in\mathscr{S}} (T \setminus S)$. Conversely, if $x \in \cap_{S\in\mathscr{S}} (T \setminus S)$, then, for each $S \in \mathscr{S}$, $x \notin S$. Therefore, $x \notin \cup_{S\in\mathscr{S}}$. Therefore, $x \in T \setminus (\cup_{S\in\mathscr{S}})$, thus showing that $\cap_{S\in\mathscr{S}} (T \setminus S) \subseteq T \setminus (\cup_{S\in\mathscr{S}})$. It follows that $T \setminus (\cup_{S\in\mathscr{S}}) = \cap_{S\in\mathscr{S}} (T \setminus S)$.

(ii) This follows in much the same manner as part (i), and we leave the details to the reader. ■

[1] Augustus De Morgan (1806–1871) was a British mathematician whose principal mathematical contributions were to analysis and algebra.

1.1.6 Remark (Showing two sets are equal) Note that in proving part (i) of the preceding result, we proved two things. First we showed that $T \setminus (\cup_{S\in\mathscr{S}}) \subseteq \cap_{S\in\mathscr{S}}(T \setminus S)$ and then we showed that $\cap_{S\in\mathscr{S}}(T \setminus S) \subseteq T \setminus (\cup_{S\in\mathscr{S}})$. This is the standard means of showing that two sets are equal; first show that one is a subset of the other, and then show that the other is a subset of the one. •

For general unions and intersections, we also have the following generalisation of the distributive laws for unions and intersections. We leave the straightforward proof to the reader (Exercise 1.1.6)

1.1.7 Proposition (Distributivity laws for general unions and intersections) *Let* T *be a set and let* $\mathscr{S}$ *be a collection of sets. Then the following statements hold:*

(i) $\mathrm{T} \cap (\cup_{\mathrm{S}\in\mathscr{S}}\mathrm{S}) = \cup_{\mathrm{S}\in\mathrm{S}}(\mathrm{T} \cap \mathrm{S})$;

(ii) $\mathrm{T} \cup (\cap_{\mathrm{S}\in\mathscr{S}}\mathrm{S}) = \cap_{\mathrm{S}\in\mathrm{S}}(\mathrm{T} \cup \mathrm{S})$.

There is an alternative notion of the union of sets, one that retains the notion of membership in the original set. The issue that arises is this. If $S = \{1, 2\}$ and $T = \{2, 3\}$, then $S \cup T = \{1, 2, 3\}$. Note that we lose with the usual union the fact that 1 is an element of S only, but that 2 is an element of both S and T. Sometimes it is useful to retain these sorts of distinctions, and for this we have the following definition.

1.1.8 Definition (Disjoint union) For sets S and T, their ***disjoint union*** is the set

$$S \overset{\circ}{\cup} T = \{(S, x) \mid x \in S\} \cup \{(T, y) \mid y \in T\}. \quad \bullet$$

Let us see how the disjoint union differs from the usual union.

1.1.9 Example (Disjoint union) Let us again take the simple example $S = \{1, 2\}$ and $T = \{2, 3\}$. Then $S \cup T = \{1, 2, 3\}$ and

$$S \overset{\circ}{\cup} T = \{(S, 1), (S, 2), (T, 2), (T, 3)\}.$$

We see that the idea behind writing an element in the disjoint union as an ordered pair is that the first entry in the ordered pair simply keeps track of the set from which the element in the disjoint union was taken. In this way, if $S \cap T \neq \emptyset$, we are guaranteed that there will be no "collapsing" when the disjoint union is formed. •

1.1.3 Finite Cartesian products

As we have seen, if S is a set and if $x_1, x_2 \in S$, then $\{x_1, x_2\} = \{x_2, x_1\}$. There are times, however, when we wish to keep track of the order of elements in a set. To accomplish this and other objectives, we introduce the notion of an ordered pair. First, however, in order to make sure that we understand the distinction between ordered and unordered pairs, we make the following definition.

1.1.10 Definition (Unordered pair) If S is a set, an ***unordered pair*** from S is any subset of S with two elements. The collection of unordered pairs from S is denoted by $S^{(2)}$. •

Obviously one can talk about unordered collections of more than two elements of a set, and the collection of subsets of a set S comprised of k elements is denoted by $S^{(k)}$ and called the set of ***unordered* k*-tuples***.

With the simple idea of an unordered pair, the notion of an ordered pair is more distinct.

1.1.11 Definition (Ordered pair and Cartesian product) Let S and T be sets, and let $x \in S$ and $y \in T$. The ***ordered pair*** of x and y is the set $(x, y) = \{\{x\}, \{x, y\}\}$. The ***Cartesian product*** of S and T is the set

$$S \times T = \{(x, y) \mid \ x \in S, \ y \in T\}. \qquad \bullet$$

The definition of the ordered pair seems odd at first. However, it is as it is to secure the objective that if two ordered pairs (x_1, y_1) and (x_2, y_2) are equal, then $x_1 = x_2$ and $y_1 = y_2$. The reader can check in Exercise 1.1.8 that this objective is in fact achieved by the definition. It is also worth noting that the form of the ordered pair as given in the definition is seldom used after its initial introduction.

Clearly one can define the Cartesian product of any finite number of sets $S_1, \dots, S_k$ inductively. Thus, for example, $S_1 \times S_2 \times S_3 = (S_1 \times S_2) \times S_3$. Note that, according to the notation in the definition, an element of $S_1 \times S_2 \times S_3$ should be written as $((x_1, x_2), x_3)$. However, it is immaterial that we define $S_1 \times S_2 \times S_3$ as we did, or as $S_1 \times S_2 \times S_3 = S_1 \times (S_2 \times S_3)$. Thus we simply write elements in $S_1 \times S_2 \times S_3$ as (x_1, x_2, x_3), and similarly for a Cartesian product $S_1 \times \dots \times S_k$. The Cartesian product of a set with itself k-times is denoted by S^k. That is,

$$S^k = \underbrace{S \times \cdots \times S}_{k\text{-times}}.$$

In Section 1.6.2 we shall indicate how to define Cartesian products of more than finite collections of sets.

Let us give some simple examples.

1.1.12 Examples (Cartesian products)

1. If S is a set then note that $S \times \emptyset = \emptyset$. This is because there are no ordered pairs from S and $\emptyset$. It is just as clear that $\emptyset \times S = \emptyset$. It is also clear that, if $S \times T = \emptyset$, then either $S = \emptyset$ or $T = \emptyset$.
2. If $S = \{1, 2\}$ and $T = \{2, 3\}$, then

$$S \times T = \{(1, 2), (1, 3), (2, 2), (2, 3)\}. \qquad \bullet$$

Cartesian products have the following properties.

1.1.13 Proposition (Properties of Cartesian product) *For sets* S, T, U, *and* V, *the following statements hold:*

(i) $(\mathrm{S} \cup \mathrm{T}) \times \mathrm{U} = (\mathrm{S} \times \mathrm{U}) \cup (\mathrm{T} \times \mathrm{U})$;

(ii) $(\mathrm{S} \cap \mathrm{U}) \times (\mathrm{T} \cap \mathrm{V}) = (\mathrm{S} \times \mathrm{T}) \cap (\mathrm{U} \times \mathrm{V})$;

(iii) $(\mathrm{S} - \mathrm{T}) \times \mathrm{U} = (\mathrm{S} \times \mathrm{U}) - (\mathrm{T} \times \mathrm{U})$.

Proof Let us prove only the first identity, leaving the remaining two to the reader. Let $(x, u) \in (S \cup T) \times U$. Then $x \in S \cup T$ and $u \in U$. Therefore, x is an element of at least one of S and T. Without loss of generality, suppose that $x \in S$. Then $(x, u) \in S \times U$ and so $(x, u) \in (S \times U) \cup (T \times U)$. Therefore, $(S \cup T) \times U = (S \times U) \cup (T \times U)$. Conversely, suppose that $(x, u) \in (S \times U) \cup (T \times U)$. Without loss of generality, suppose that $(x, u) \in S \times U$. Then $x \in S \subseteq S \cup T$ and $u \in U$. Therefore, $(x, u) \in (S \cup T) \times U$. Thus $(S \times U) \cup (T \times U) \subseteq (S \cup T) \times U$, giving the result. ■

1.1.14 Remark ("Without loss of generality") In the preceding proof, we twice employed the expression "without loss of generality." This is a commonly encountered expression, and is frequently used in one of the following two contexts. The first, as above, indicates that one is making an arbitrary selection, but that were another arbitrary selection to have been made, the same argument holds. This is a more or less straightforward use of "without loss of generality." A more sophisticated use of the expression might indicate that one is making a simplifying assumption, and that this is okay, because it can be shown that the general case follows easily from the simpler one. The trick is to then understand *how* the general case follows from the simpler one, and this can sometimes be nontrivial, depending on the willingness of the writer to describe this process. •

Exercises

1.1.1 Prove that the empty set is a subset of every set.
Hint: *Assume the converse and arrive at an absurdity.*

1.1.2 Let S be a set, let $A, B, C \subseteq S$, and let $\mathscr{A}, \mathscr{B} \subseteq \mathbf{2}^S$.

(a) Show that $A \triangle \varnothing = A$.

(b) Show that $(S \setminus A) \triangle (S \setminus B) = A \triangle B$.

(c) Show that $A \triangle C \subseteq (A \triangle B) \cup (B \triangle C)$.

(d) Show that

$$\begin{gathered}(\cup_{A \in \mathscr{A}} A) \triangle (\cup_{B \in \mathscr{B}} B) \subseteq \cup_{(A,B) \in \mathscr{A} \times \mathscr{B}} (A \triangle B), \\ (\cap_{A \in \mathscr{A}} A) \triangle (\cap_{B \in \mathscr{B}} B) \subseteq \cap_{(A,B) \in \mathscr{A} \times \mathscr{B}} (A \triangle B), \\ \cap_{(A,B) \in \mathscr{A} \times \mathscr{B}} (A \triangle B) \subseteq (\cap_{A \in \mathscr{A}} A) \triangle (\cup_{B \in \mathscr{B}} B) .\end{gathered}$$

1.1.3 If S is a set with n members, show that $\mathbf{2}^S$ is a set with 2^n members.

1.1.4 Let S be a set with m elements. Show that the number of subsets of S having k distinct elements is $\binom{m}{k} = \frac{m!}{k!(m-k)!}$.

1.1.5 Prove as many parts of Proposition 1.1.4 as you wish.

1.1.6 Prove Proposition 1.1.7.

1.1.7 Let S be a set with n members and let T be a set with m members. Show that $S \overset{\circ}{\cup} T$ is a set with nm members.

1.1.8 Let S and T be sets, let $x_1, x_2 \in S$, and let $y_1, y_2 \in T$. Show that $(x_1, y_1) = (x_2, y_2)$ if and only if $x_1 = x_2$ and $y_1 = y_2$.

Section 1.2

Relations

Relations are a fundamental ingredient in the description of many mathematical ideas. One of the most valuable features of relations is that they allow many useful constructions to be explicitly made only using elementary ideas from set theory.

Do I need to read this section? The ideas in this section will appear in many places in the series, so this material should be regarded as basic. However, readers looking to proceed with minimal background can skip the section, referring back to it when needed. •

1.2.1 Definitions

We shall describe in this section "binary relations," or relations between elements of two sets. It is possible to define more general sorts of relations where more sets are involved. However, these will not come up for us.

1.2.1 Definition (Relation) A ***binary relation from* S *to* T** (or simply a ***relation from* S *to* T**) is a subset of $S \times T$. If $R \subseteq S \times T$ and if $(x, y) \in R$, then we shall write $x \, R \, y$, meaning that x and y are related by R. A relation from S to S is a ***relation in* S**. •

The definition is simple. Let us give some examples to give it a little texture.

1.2.2 Examples (Relations)

1. Let S be the set of husbands and let T be the set of wives. Define a relation R from S to T by asking that $(x, y) \in R$ if x is married to y. Thus, to say that x and y are related in this case means to say that x is married to y.
2. Let S be a set and consider the relation R in the power set $\mathbf{2}^S$ of S given by

$$R = \{(A, B) \mid A \subseteq B\}.$$

Thus A is related to B if A is a subset of B.

3. Let S be a set and define a relation R in S by

$$R = \{(x, x) \mid x \in S\}.$$

Thus, under this relation, two members in S are related if and only if they are equal.

4. Let S be the set of integers, let k be a positive integer, and define a relation R_k in S by

$$R_k = \{(n_1, n_2) \mid n_1 - n_2 = k\}.$$

Thus, if $n \in S$, then all integers of the form $n + mk$ for an integer m are related to n. •

1.2.3 Remark ("If" versus "if and only if") In part 3 of the preceding example we used the expression "if and only if" for the first time. It is, therefore, worth saying a few words about this commonly used terminology. One says that statement A holds "if and only if" statement B holds to mean that statements A and B are exactly equivalent. Typically, this language arises in theorem statements. In proving such theorems, it is important to note that one must prove *both* that statement A implies statement B *and* that statement B implies statement A.

To confuse matters, when stating a definition, the convention is to use "if" rather than "if and only if". It is not uncommon to see "if and only if" used in definitions, the thinking being that a definition makes the thing being defined as equivalent to what it is defined to be. However, there is a logical flaw here. Indeed, suppose one is defining "X" to mean that "Proposition A applies". If one writes "X if and only if Proposition A applies" then this makes no sense. Indeed the "only if" part of this statement says that the statement "Proposition A applies" if "X" holds. But "X" is undefined except by saying that it holds when "Proposition A applies". •

In the next section we will encounter the notion of the inverse of a function; this idea is perhaps known to the reader. However, the notion of inverse also applies to the more general setting of relations.

1.2.4 Definition (Inverse of a relation) If $R \subseteq S \times T$ is a relation from S to T, then the ***inverse*** of R is the relation R^{-1} from T to S defined by

$$R^{-1} = \{(y, x) \in T \times S \mid (x, y) \in R\}. \qquad \bullet$$

There are a variety of properties that can be bestowed upon relations to ensure they have certain useful attributes. The following is a partial list of such properties.

1.2.5 Definition (Properties of relations) Let S be a set and let R be a relation in S. The relation R is:

(i) ***reflexive*** if $(x, x) \in R$ for each $x \in S$;

(ii) ***irreflexive*** if $(x, x) \notin R$ for each $x \in S$;

(iii) ***symmetric*** if $(x_1, x_2) \in R$ implies that $(x_2, x_1) \in R$;

(iv) ***antisymmetric*** if $(x_1, x_2) \in R$ and $(x_2, x_1) \in R$ implies that $x_1 = x_2$;

(v) ***transitive*** if $(x_1, x_2) \in R$ and $(x_2, x_3) \in R$ implies that $(x_1, x_3) \in R$. •

1.2.6 Examples (Example 1.2.2 cont'd)

1. The relation of inclusion in the power set $\mathbf{2}^S$ of a set S is reflexive, antisymmetric, and transitive.
2. The relation of equality in a set S is reflexive, symmetric, antisymmetric, and transitive.
3. The relation R_k in the set S of integers is reflexive, symmetric, and transitive. •

1.2.2 Equivalence relations

In this section we turn our attention to an important class of relations, and we indicate why these are important by giving them a characterisation in terms of a decomposition of a set.

1.2.7 Definition (Equivalence relation, equivalence class) An ***equivalence relation*** in a set S is a relation R that is reflexive, symmetric, and transitive. For $x \in S$, the set of elements of S related to x is denoted by $[x]$, and is the ***equivalence class*** of x with respect to R. An element x' in an equivalence class $[x]$ is a ***representative*** of that equivalence class. The set of equivalence classes is denoted by S/R (typically pronounced as **S *modulo* R**). •

It is common to denote that two elements $x_1, x_2 \in S$ are related by an equivalence relation by writing $x_1 \sim x_2$. Of the relations defined in Example 1.2.2, we see that those in parts 3 and 4 are equivalence relations, but that in part 2 is not.

Let us now characterise equivalence relations in a more descriptive manner. We begin by defining a (perhaps seemingly unrelated) notion concerning subsets of a set.

1.2.8 Definition (Partition of a set) A ***partition*** of a set S is a collection $\mathscr{A}$ of subsets of S having the properties that

(i) two distinct subsets in $\mathscr{A}$ are disjoint and

(ii) $S = \cup_{A \in \mathscr{A}} A$. •

We now prove that there is an exact correspondence between equivalence classes associated to an equivalence relation.

1.2.9 Proposition (Equivalence relations and partitions) *Let* S *be a set and let* R *be an equivalence relation in* S*. Then the set of equivalence classes with respect to* R *is a partition of* S*.*

Conversely, if $\mathscr{A}$ *is a partition of* S*, then the relation*

$$\{(x_1, x_2) \mid x_1, x_2 \in A \textit{ for some } A \in \mathscr{A}\}$$

is an equivalence relation in S*.*

Proof We first claim that two distinct equivalence classes are disjoint. Thus we let $x_1, x_2 \in S$ and suppose that $[x_1] \neq [x_2]$. Suppose that $x \in [x_1] \cap [x_2]$. Then $x \sim x_1$ and $x \sim x_2$, or, by transitivity of R, $x_1 \sim x$ and $x \sim x_2$. By transitivity of R, $x_1 \sim x_2$, contradicting the fact that $[x_1] \neq [x_2]$. To show that S is the union of its equivalence classes, merely note that, for each $x \in S$, $x \in [x]$ by reflexivity of R.

Now let $\mathscr{A}$ be a partition and defined R as in the statement of the proposition. Let $x \in S$ and let A be the element of $\mathscr{A}$ that contains x. Then clearly we see that $(x, x) \in R$ since $x \in A$. Thus R is reflexive. Next let $(x_1, x_2) \in R$ and let A be the element of $\mathscr{A}$ such that $x_1, x_2 \in A$. Clearly then, $(x_2, x_1) \in R$, so R is symmetric. Finally, let $(x_1, x_2), (x_2, x_3 \in R$. Then there are elements $A_{12}, A_{23} \in \mathscr{A}$ such that $x_1, x_2 \in A_{12}$ and such that $x_2, x_3 \in A_{23}$. Since A_{12} and A_{23} have the point x_2 in common, we must have $A_{12} = A_{23}$. Thus $(x_1, x_3 \in A_{12} = A_{23}$, giving transitivity of R. ■

Exercises

1.2.1 In a set S define a relation $R = \{(x, y) \in S \times S \mid \ x = y\}$.

(a) Show that R is an equivalence relation.

(b) Show that $S/R = S$.

Section 1.3

Maps

Another basic concept in all of mathematics is that of a map between sets. Indeed, many of the interesting objects in mathematics are maps of some sort. In this section we review the notation associated with maps, and give some simple properties of maps.

Do I need to read this section? The material in this section is basic, and will be used constantly throughout the series. Unless you are familiar already with maps and the notation associated to them, this section is essential reading. •

1.3.1 Definitions and notation

We begin with the definition.

1.3.1 **Definition (Map)** For sets S and T, a ***map*** from S to T is a relation R from S to T having the property that, for each $x \in S$, there exists a unique $y \in T$ such that $(x, y) \in R$. The set S is the ***domain*** of the map and the set T is the ***codomain*** of the map. The set of maps from S to T is denoted by T^S.[2] •

By definition, a map is a relation. This is not how one most commonly thinks about a map, although the definition serves to render the concept of a map in terms of concepts we already know. Suppose one has a map from S to T defined by a relation R. Then, given $x \in S$, there is a single $y \in T$ such that x and y are related. Denote this element of T by $f(x)$, since it is defined by x. When one refers to a map, one more typically refers to the assignment of the element $f(x) \in T$ to $x \in S$. Thus one refers to the map as f, leaving aside the baggage of the relation as in the definition. Indeed, this is how we from now on will think of maps. The definition above does, however, have some use, although we alter our language, since we are now thinking of a map as an "assignment." We call the set

$$\operatorname{graph}(f) = \{(x, f(x)) \mid x \in S\} \subseteq S \times T$$

(which we originally called the map in Definition 1.3.1) the ***graph*** of the map $f: S \to T$.

If one wishes to indicate a map f with domain S and codomain T, one typically writes $f: S \to T$ to compactly express this. If one wishes to *define* a map by saying what it does, the notation

$$\begin{aligned} f\colon S &\to T \\ x &\mapsto \text{what } x \text{ gets mapped to} \end{aligned}$$

[2] The idea behind this notation is the following. A map from S to T assigns to each point in S a point in T. If S and T are finite sets with k and l elements, respectively, then there are l possible values that can be assigned to each of the k elements of S. Thus the set of maps has l^k elements.

is sometimes helpful. Sometimes we shall write this in the text as $f\colon x \mapsto$ "what x gets mapped to". Note the distinct uses of the symbols "$\rightarrow$" and "$\mapsto$".

1.3.2 Notation (f versus f(x)) Note that a map is denoted by "f". It is quite common to see the expression "consider the map $f(x)$". Taken literally, these words are difficult to comprehend. First of all, x is unspecified. Second of all, even if x were specified, $f(x)$ is an element of T, not a map. Thus it is considered bad form mathematically to use an expression like "consider the map $f(x)$". However, there are times when it is quite convenient to use this poor notation, with an understanding that some compromises are being made. For instance, in this volume, we will be frequently dealing simultaneously with functions of both time (typically denoted by t) and frequency (typically denoted by ν). Thus it would be convenient to write "consider the map $f(t)$" when we wish to write a map that we are considering as a function of time, and similarly for frequency. Nonetheless, we shall refrain from doing this, and shall consistently use the mathematically precise language "consider the map f". •

The following is a collection of examples of maps. Some of these examples are not just illustrative, but also define concepts and notation that we will use throughout the series.

1.3.3 Examples (Maps)

1. There are no maps having $\varnothing$ as a domain or codomain since there are no elements in the empty set.
2. If S is a set and if $T \subseteq S$, then the map $i_T\colon T \rightarrow S$ defined by $i_T(x) = x$ is called the ***inclusion*** of T in S.
3. The inclusion map $i_S\colon S \rightarrow S$ of a set S into itself (since $S \subseteq S$) is the ***identity map***, and we denote it by id_S.
4. If $f\colon S \rightarrow T$ is a map and if $A \subseteq S$, then the map from A to T which assigns to $x \in A$ the value $f(x) \in T$ is called the ***restriction*** of f to A, and is denoted by $f|A\colon A \rightarrow T$.
5. If S is a set with $A \subseteq S$, then the map χ_A from S to the integers defined by

$$\chi_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A, \end{cases}$$

   is the ***characteristic function*** of A.
6. If $S_1, \dots, S_k$ are sets, if $S_1 \times \cdots \times S_k$ is the Cartesian product, and if $j \in \{1, \dots, k\}$, then the map

$$\begin{aligned} \mathrm{pr}_j\colon S_1 \times \cdots \times S_j \times \cdots \times S_k &\rightarrow S_j \\ (x_1, \dots, x_j, \dots, x_k) &\mapsto x_j \end{aligned}$$

   is the ***projection onto the* j*th factor***.

7. If R is an equivalence relation in a set S, then the map $\pi_R\colon S \to S/R$ defined by $\pi_R(x) = [x]$ is called the ***canonical projection*** associated to R.
8. If S, T, and U are sets and if $f\colon S \to T$ and $g\colon T \to U$ are maps, then we define a map $g \circ f\colon S \to U$ by $g \circ f(x) = g(f(x))$. This is the ***composition*** of f and g.
9. If S and $T_1, \dots, T_k$ are sets then a map $f\colon S \to T_1 \times \cdots \times T_k$ can be written as

$$f(x) = (f_1(x), \dots, f_k(x))$$

for maps $f_j\colon S \to T_j$, $j \in \{1, \dots, k\}$. In this case we will write $f = f_1 \times \cdots \times f_k$. •

Next we introduce the notions of images and preimages of points and sets.

1.3.4 Definition (Image and preimage) Let S and T be sets and let $f\colon S \to T$ be a map.

(i) If $A \subseteq S$, then $f(A) = \{f(x) \mid x \in A\}$.

(ii) The ***image*** of f is the set $\mathrm{image}(f) = f(S) \subseteq T$.

(iii) If $B \subseteq T$, then $f^{-1}(B) = \{x \in S \mid f(x) \in B\}$ is the ***preimage*** of B under f. If $B = \{y\}$ for some $y \in T$, then we shall often write $f^{-1}(y)$ rather that $f^{-1}(\{y\})$. •

Note that one can think of f as being a map from $\mathbf{2}^S$ to $\mathbf{2}^T$ and of f^{-1} as being a map from $\mathbf{2}^T$ to $\mathbf{2}^S$. Here are some elementary properties of f and f^{-1} thought of in this way.

1.3.5 Proposition (Properties of images and preimages) *Let* S *and* T *be sets, let* $\mathrm{f}\colon \mathrm{S} \to \mathrm{T}$ *be a map, let* $\mathrm{A} \subseteq \mathrm{S}$ *and* $\mathrm{B} \subseteq \mathrm{T}$, *and let* $\mathscr{A}$ *and* $\mathscr{B}$ *be collections of subsets of* S *and* T, *respectively. Then the following statements hold:*

(i) $\mathrm{A} \subseteq \mathrm{f}^{-1}(\mathrm{f}(\mathrm{A}))$;

(ii) $\mathrm{f}(\mathrm{f}^{-1}(\mathrm{B})) \subseteq \mathrm{B}$;

(iii) $\cup_{\mathrm{A}\in\mathscr{A}}\mathrm{f}(\mathrm{A}) = \mathrm{f}(\cup_{\mathrm{A}\in\mathscr{A}}\mathrm{A})$;

(iv) $\cup_{\mathrm{B}\in\mathscr{B}}\mathrm{f}^{-1}(\mathrm{B}) = \mathrm{f}^{-1}(\cup_{\mathrm{B}\in\mathscr{B}}\mathrm{B})$;

(v) $\cap_{\mathrm{A}\in\mathscr{A}}\mathrm{f}(\mathrm{A}) = \mathrm{f}(\cap_{\mathrm{A}\in\mathscr{A}}\mathrm{A})$;

(vi) $\cap_{\mathrm{B}\in\mathscr{B}}\mathrm{f}^{-1}(\mathrm{B}) = \mathrm{f}^{-1}(\cap_{\mathrm{B}\in\mathscr{B}}\mathrm{B})$.

Proof We shall prove only some of these, leaving the remainder for the reader to complete.

(i) Let $x \in A$. Then $x \in f^{-1}(f(x))$ since $f(x) = f(x)$.

(iii) Let $y \in \cup_{A\in\mathscr{A}} f(A)$. Then $y = f(x)$ for some $x \in \cup_{A\in\mathscr{A}} A$. Thus $y \in f(\cup_{A\in\mathscr{A}} A)$. Conversely, let $y \in f(\cup_{A\in\mathscr{A}} A)$. Then, again, $y = f(x)$ for some $x \in \cup_{A\in\mathscr{A}} A$, and so $y \in \cup_{A\in\mathscr{A}} f(A)$.

(vi) Let $x \in \cap_{B\in\mathscr{B}} f^{-1}(B)$. Then, for each $B \in \mathscr{B}$, $x \in f^{-1}(B)$. Thus $f(x) \in B$ for all $B \in \mathscr{B}$ and so $f(x) \in \cap_{B\in\mathscr{B}} B$. Thus $x \in f^{-1}(\cap_{B\in\mathscr{B}} B)$. Conversely, if $x \in f^{-1}(\cap_{B\in\mathscr{B}} B)$, then $f(x) \in B$ for each $B \in \mathscr{B}$. Thus $x \in f^{-1}(B)$ for each $B \in \mathscr{B}$, or $x \in \cap_{B\in\mathscr{B}} f^{-1}(B)$. ■

1.3.2 Properties of maps

Certain basic features of maps will be of great interest.

1.3.6 Definition (Injection, surjection, bijection) Let S and T be sets. A map $f\colon S \to T$ is:

(i) ***injective***, or an ***injection***, if $f(x) = f(y)$ implies that $x = y$;

(ii) ***surjective***, or a ***surjection***, if $f(S) = T$;

(iii) ***bijective***, or a ***bijection***, if it is both injective and surjective. •

1.3.7 Remarks (One-to-one, onto, 1–1 correspondence)

1. It is not uncommon for an injective map to be said to be ***1–1*** or ***one-to-one***, and that a surjective map be said to be ***onto***. In this series, we shall exclusively use the terms injective and surjective, however. These words appear to have been given prominence by their adoption by Bourbaki (see footnote on page iv).
2. If there exists a bijection $f\colon S \to T$ between sets S and T, it is common to say that there is a ***1–1 correspondence*** between S and T. This can be confusing if one is familiar with the expression "1–1" as referring to an injective map. The words "1–1 correspondence" mean that there is a bijection, not an injection. In case S and T are in 1–1 correspondence, we shall also say that S and T are ***equivalent***. •

Closely related to the above concepts, although not immediately obviously so, are the following notions of inverse.

1.3.8 Definition (Left-inverse, right-inverse, inverse) Let S and T be sets, and let $f\colon S \to T$ be a map. A map $g\colon T \to S$ is:

(i) a ***left-inverse*** of f if $g \circ f = \mathrm{id}_S$;

(ii) a ***right-inverse*** of f if $f \circ g = \mathrm{id}_T$;

(iii) an ***inverse*** of f if it is both a left- and a right-inverse. •

In Definition 1.2.4 we gave the notion of the inverse of a relation. Functions, being relations, also possess inverses in the sense of relations. We ask the reader to explore the relationships between the two concepts of inverse in Exercise 1.3.7.

The following result relates these various notions of inverse to the properties of injective, surjective, and bijective.

1.3.9 Proposition (Characterisation of various inverses) *Let* S *and* T *be sets and let* $\mathrm{f}\colon \mathrm{S} \to \mathrm{T}$ *be a map. Then the following statements hold:*

(i) f *is injective if and only if it possesses a left-inverse;*

(ii) f *is surjective if and only if it possess a right-inverse;*

(iii) f *is bijective if and only if it possesses an inverse;*

(iv) there is at most one inverse for f*;*

(v) if f *possesses a left-inverse and a right-inverse, then these necessarily agree.*

Proof (i) Suppose that f is injective. For $y \in \mathrm{image}(f)$, define $g(y) = x$ where $f^{-1}(y) = \{x\}$, this being well-defined since f is injective. For $y \notin \mathrm{image}(f)$, define $g(y) = x_0$ for some $x_0 \in S$. The map g so defined is readily verified to satisfy $g \circ f = \mathrm{id}_S$, and so is a left-inverse. Conversely, suppose that f possesses a left-inverse g, and let $x_1, x_2 \in S$ satisfy $f(x_1) = f(x_2)$. Then $g \circ f(x_1) = g \circ f(x_2)$, or $x_1 = x_2$. Thus f is injective.

(ii) Suppose that f is surjective. For $y \in T$ let $x \in f^{-1}(y)$ and define $g(y) = x$.[3] With g so defined it is easy to see that $f \circ g = \mathrm{id}_T$, so that g is a right-inverse. Conversely, suppose that f possesses a right-inverse g. Now let $y \in T$ and take $x = g(y)$. Then $f(x) = f \circ g(y) = y$, so that f is surjective.

(iii) Since f is bijective, it possesses a left-inverse g_L and a right-inverse g_R. We claim that these are equal, and each is actually an inverse of f. We have

$$g_L = g_L \circ \mathrm{id}_T = g_L \circ f \circ g_R = \mathrm{id}_S \circ g_R = g_R,$$

showing equality of g_L and g_R. Thus each is a left- and a right-inverse, and therefore an inverse for f.

(iv) Let g_1 and g_2 be inverses for f. Then, just as in part (iii),

$$g_1 = g_1 \circ \mathrm{id}_T = g_1 \circ f \circ g_2 = \mathrm{id}_S \circ g_2 = g_2.$$

(v) This follows from the proof of part (iv), noting that there we only used the facts that g_1 is a left-inverse and that g_2 is a right-inverse. ■

In Figure 1.2 we depict maps that have various of the properties of injectivity, surjectivity, or bijectivity. From these cartoons, the reader may develop some

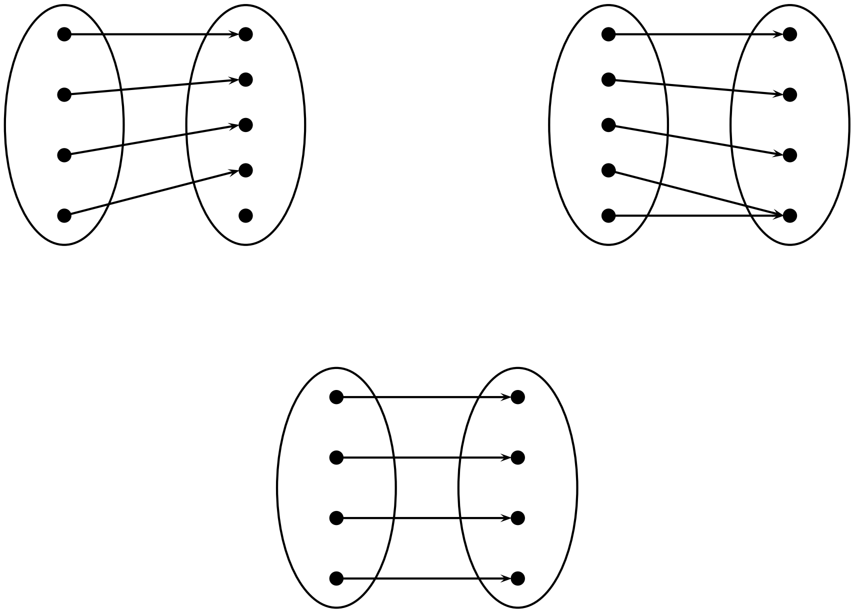


Figure 1.2 A depiction of maps that are injective but not surjective (top left), surjective but not injective (top right), and bijective (bottom)

[3] Note that the ability to choose an x from each set $f^{-1}(y)$ requires the Axiom of Choice (see Section 1.8.3).

intuition for Proposition 1.3.9. In the case that $f: S \to T$ is a bijection, we denote its unique inverse by $f^{-1}: T \to S$. The confluence of the notation f^{-1} introduced when discussing preimages is not a problem, in practice.

It is worth mentioning at this point that the characterisation of left- and right-inverses in Proposition 1.3.9 is not usually very helpful. Normally, in a given setting, one will want these inverses to have certain properties. For vector spaces, for example, one may want left- or right-inverses to be linear (see Proposition 5.4.46), and for topological spaces, for another example, one may want a left- or right-inverse to be continuous (see Chapter III-1).

1.3.3 Graphs and commutative diagrams

Often it is useful to be able to understand the relationship between a number of maps by representing them together in a diagram. We shall be somewhat precise about what we mean by a diagram by making it a special instance of a graph.

First the definitions for graphs.

1.3.10 Definition (Graph) A ***graph*** is a pair (V, E) where V is a set, an element of which is called a ***vertex***, and E is a subset of the set $V^{(2)}$ of unordered pairs from V, an element of which is called an ***edge***. If $\{v_1, v_2\} \in E$ is an edge, then the vertices v_1 and v_2 are the ***endvertices*** of this edge. •

In a graph, it is the way that vertices and edges are related that is of interest. To capture this structure, the following language is useful.

1.3.11 Definition (Adjacent and incident) Let (V, E) be a graph. Two vertices $v_1, v_2 \in V$ are ***adjacent*** if $\{v_1, v_2\} \in E$ and a vertex $v \in V$ and an edge $e \in E$ are ***incident*** if there exists $v' \in V$ such that $e = \{v, v'\}$. •

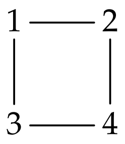
One typically represents a graph by placing the vertices in some sort of array on the page, and then drawing a line connecting two vertices if there is a corresponding edge associated with the two vertices. Some examples make this process clear.

1.3.12 Examples (Graphs)

1. Consider the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}.$$

There are many ways one can lay out the vertices on the page, but for this diagram, it is most convenient to arrange them in a square. Doing so gives rise to the following representation of the graph:

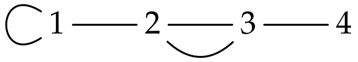


The vertices 1 and 2 are adjacent, but the vertices 1 and 4 are not. The vertex 1 and the edge $\{1,2\}$ are incident, but the vertex 1 and the edge $\{3,4\}$ are not.

2. For the graph (V, E) with

$$V = \{1,2,3,4\}, \quad E = \{\{1,2\},\{2,3\},\{2,3\},\{3,4\}\}$$

we have the representation



Note that we allow the same edge to appear twice, and we allow for an edge to connect a vertex to itself. We observe that the vertices 2 and 3 are adjacent, but the vertices 1 and 3 are not. Also, the vertex 3 and the edge $\{2,3\}$ are incident, but the vertex 4 and the edge $\{1,2\}$ are not. •

Often one wishes to attach "direction" to vertices. This is done with the following notion.

1.3.13 Definition (Directed graph) A ***directed graph***, or ***digraph***, is a pair (V, E) where V is a set an element of which is called a ***vertex*** and E is a subset of the set $V \times V$ of ordered pairs from V an element of which is called an ***edge***. If $e = (v_1, v_2) \in E$ is an edge, then v_1 is the ***source*** for e and v_2 is the ***target*** for e. •

Note that every directed graph is certainly also a graph, since one can assign an unordered pair to every ordered pair of vertices.

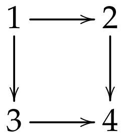
The examples above of graphs are easily turned into directed graphs, and we see that to represent a directed graph one needs only to put a "direction" on an edge, typically via an arrow.

1.3.14 Examples (Directed graphs)

1. Consider the directed graph (V, E) with

$$V = \{1,2,3,4\}, \quad E = \{(1,2),(1,3),(2,4),(3,4)\}.$$

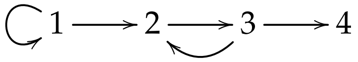
A convenient representation of this directed graph is as follows:



2. For the directed graph (V, E) with

$$V = \{1,2,3,4\}, \quad E = \{(1,1),(1,2),(2,3),(2,3),(3,4)\}$$

we have the representation



•

Of interest in graph theory is the notion of connecting two, perhaps nonadjacent, vertices with a sequence of edges (the notion of a sequence is familiar, but will be made precise in Section 1.6.3). This is made precise as follows.

1.3.15 Definition (Path)

(i) If (V, E) is a graph, a ***path*** in the graph is a sequence $(a_j)_{j\in\{1,\dots,k\}}$ in $V \cup E$ with the following properties:

(a) $a_1, a_k \in V$;

(b) for $j \in \{1, \dots, k-1\}$, if $a_j \in V$ (resp. $a_j \in E$), then $a_{j+1} \in E$ (resp. $a_{j+1} \in V$).

(ii) If (V, E) is a directed graph, a ***path*** in the graph is a sequence $(a_j)_{j\in\{1,\dots,k\}}$ in $V \cup E$ with the following properties:

(a) $(a_j)_{j\in\{1,\dots,k\}}$ is a path in the graph associated to (V, E);

(b) for $j \in \{2, \dots, k-1\}$, if $a_j \in E$, then $a_j = (a_{j-1}, a_{j+1})$.

(iii) If $(a_j)_{j\in\{1,\dots,k\}}$ is a path, the ***length*** of the path is the number of edges in the path.

(iv) For a path $(a_j)_{j\in\{1,\dots,k\}}$, the ***source*** is the vertex a_1 and the ***target*** is the vertex a_k. •

Let us give some examples of paths for graphs and for directed graphs.

1.3.16 Examples (Paths)

1. For the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\},$$

there are an infinite number of paths. Let us list a few:

(a) (1), (2), (3), and (4);

(b) $(4, \{3, 4\}, 3, \{1, 3\}, 1)$;

(c) $(1, \{1, 2\}, 2, \{2, 4\}, 4, \{3, 4\}, 3, \{1, 3\}, 1)$;

(d) $(1, \{1, 2\}, 2, \{1, 2\}, 1, \{1, 2\}, 2, \{1, 2\}, 1)$.

Note that for this graph there are infinitely many paths.

2. For the directed graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{(1, 2), (1, 3), (2, 4), (3, 4)\},$$

there are a finite number of paths:

(a) (1), (2), (3), and (4);

(b) $(1, (1, 2), 2)$;

(c) $(1, (1, 2), 2, (2, 4), 4)$;

(d) $(1, (1, 3), 3)$;

(e) $(1, (1, 3), 3, (2, 4), 4)$;

(f) $(2, (2, 4))$;

(g) $(3, (3, 4), 4)$.

3. For the graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{\{1, 2\}, \{2, 3\}, \{2, 3\}, \{3, 4\}\}$$

some examples of paths are:

(a) (1), (2), (3), and (4);

(b) $(1, \{1, 2\}, 2, \{2, 3\}, 3, \{2, 3\}, 2, \{1, 2\}, 1)$;

(c) $(4, \{3, 4\}, 3)$.

There are an infinite number of paths for this graph.

4. For the directed graph (V, E) with

$$V = \{1, 2, 3, 4\}, \quad E = \{(1, 1), (1, 2), (2, 3), (2, 3), (3, 4)\}$$

some paths include:

(a) (1), (2), (3), and (4);

(b) $(1, (1, 2), 2, (2, 3), 3, (3, 2), 2, (2, 3), 3, (3, 4), 4)$;

(c) $(3, (3, 4), 4)$.

This directed graph has an infinite number of paths by virtue of the fact that the path $(2, (2, 3), 3, (3, 2), 2)$ can be repeated an infinite number of times. •

1.3.17 Notation (Notation for paths of nonzero length) For paths which contain at least one edge, i.e., which have length at least 1, the vertices in the path are actually redundant. For this reason we will often simply write a path as the sequence of edges contained in the path, since the vertices can be obviously deduced. •

There is a great deal one can say about graphs, however, for our present purposes of defining diagrams, the notions at hand are sufficient. In the definition we employ Notation 1.3.17.

1.3.18 Definition (Diagram, commutative diagram) Let (V, E) be a directed graph.

(i) A ***diagram*** on (V, E) is a family $(S_v)_{v \in V}$ of sets associated with each vertex and a family $(f_e)_{e \in E}$ of maps associated with each edge such that, if $e = (v_1, v_2)$, then f_e has domain S_{v_1} and codomain S_{v_2}.

(ii) If $P = (e_j)_{j \in \{1, \dots, k\}}$ is a path of nonzero length in a diagram on (V, E), the ***composition*** along P is the map $f_{e_k} \circ \cdots \circ f_{e_1}$.

(iii) A diagram is ***commutative*** if, for every two vertices $v_1, v_2 \in V$ and any two paths P_1 and P_2 with source v_1 and target v_2, the composition along P_1 is equal to the composition along P_2. •

The notion of a diagram, and in particular a commutative diagram is straightforward.

1.3.19 Examples (Diagrams and commutative diagrams)

1. Let S_1, S_2, S_3, and S_4 be sets and consider maps $f_{21}\colon S_1 \to S_2$, $f_{31}\colon S_1 \to S_3$, $f_{42}\colon S_2 \to S_4$, and $f_{43}\colon S_3 \to S_4$. Note that if we assign set S_j to j for each $j \in \{1, 2, 3, 4\}$, then this gives a diagram on (V, E) where

   $$V = \{1, 2, 3, 4\}, \quad E = \{(1, 2), (1, 3), (2, 4), (3, 4)\}.$$

   This diagram can be represented by

   $$\begin{tikzcd} S_1 \arrow[r, "f_{21}"] \arrow[d, "f_{31}"'] & S_2 \arrow[d, "f_{42}"] \\ S_3 \arrow[r, "f_{43}"'] & 4 \end{tikzcd}$$

   The diagram is commutative if and only if $f_{42} \circ f_{21} = f_{43} \circ f_{31}$.

2. Let S_1, S_2, S_3, and S_4 be sets and let $f_{11}\colon S_1 \to S_1$, $f_{21}\colon S_1 \to S_2$, $f_{32}\colon S_2 \to S_3$, $f_{23}\colon S_3 \to S_2$, and $f_{43}\colon S_3 \to S_4$ be maps. This data then represents a commutative diagram on the directed graph (V, E) where

   $$V = \{1, 2, 3, 4\}, \quad E = \{(1, 1), (1, 2), (2, 3), (2, 3), (3, 4)\}.$$

   The diagram is represented as

   $$\begin{tikzcd} S_1 \arrow[loop left, "f_{11}"] \arrow[r, "f_{21}"] & S_2 \arrow[r, "f_{32}"] & S_3 \arrow[r, "f_{43}"] \arrow[l, bend left, "f_{23}"] & S_4 \end{tikzcd}$$

   While it is possible to write down conditions for this diagram to be commutative, there will be infinitely many such conditions. In practice, one encounters commutative diagrams with only finitely many paths with a given source and target. This example, therefore, is not so interesting as a commutative diagram, but is more interesting as a signal flow graph, which is interesting in feedback control theory. •

Exercises

1.3.1 Let S, T, U, and V be sets, and let $f\colon S \to T$, $g\colon T \to U$, and $h\colon U \to V$ be maps. Show that $h \circ (g \circ f) = (h \circ g) \circ f$.

1.3.2 Let S, T, and U be sets and let $f\colon S \to T$ and $g\colon T \to U$ be maps. Show that $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$ for every subset $C \subseteq U$.

1.3.3 Let S and T be sets, let $f\colon S \to T$, and let $B \subseteq T$. Show that $f^{-1}(T \setminus B) = S \setminus f^{-1}(B)$.

1.3.4 If S, T, and U are sets and if $f\colon S \to T$ and $g\colon T \to U$ are bijections, then show that $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

1.3.5 Let S, T and U be sets and let $f\colon S \to T$ and $g\colon T \to U$ be maps.

(a) Show that if f and g are injective, then so too is $g \circ f$.

(b) Show that if f and g are surjective, then so too is $g \circ f$.

1.3.6 Let S and T be sets, let $f\colon S \to T$ be a map, and let $A \subseteq S$ and $B \subseteq T$. Do the following:

(a) show that if f is injective then $A = f^{-1}(f(A))$;

(b) show that if f is surjective then $f(f^{-1}(B)) = B$.

1.3.7 Let S and T be sets and let $f\colon S \to T$ be a map.

(a) Show that if f is invertible as a map, then "the relation of its inverse is the inverse of its relation." (Part of the question is to precisely understand the statement in quotes.)

(b) Show that the inverse of the relation defined by f is itself the relation associated to a function if and only if f is invertible.

1.3.8 Show that equivalence of sets, as in Remark 1.3.7–2, is an "equivalence relation"[4] on collection of all sets.

[4] The quotes are present because the notion of equivalence relation, as we have defined it, applies to sets. However, there is no set containing all sets; see Section 1.8.1.

Section 1.4

Construction of the integers

It can be supposed that the reader has some idea of what the set of integers is. In this section we actually give the set of integers a *definition*. As will be seen, this is not overly difficult to do. Moreover, the construction has little bearing on what we do. We merely present it so that the reader can be comfortable with the fact that the integers, and so subsequently the rational numbers and the real numbers (see Section 2.1), have a formal definition.

Do I need to read this section? Much of this section is not of importance in the remainder of this series. The reader should certainly know what the sets $\mathbb{Z}_{>0}$ and $\mathbb{Z}$ are. However, the details of their construction should be read only when the inclination strikes. •

1.4.1 Construction of the natural numbers

The natural numbers are the numbers 1, 2, 3, and so on, i.e., the "counting numbers." As such, we are all quite familiar with them in that we can recognise, in the absence of trickery, when we are presented with 4 of something. However, what is 4? This is what we endeavour to define in this section.

The important concept in defining the natural numbers is the following.

1.4.1 Definition (Successor) Let S be a set. The ***successor*** of S is the set $S^+ = S \cup \{S\}$. •

Thus the successor is a set whose elements are the elements of S, plus an additional element which is the set S itself. This seems, and indeed is, a simple enough idea. However, it does make possible the following definition.

1.4.2 Definition (0, 1, 2, etc.)

(i) The number ***zero***, denoted by 0, is the set $\emptyset$.

(ii) The number ***one***, denoted by 1, is the set 0^+.

(iii) The number ***two***, denoted by 2, is the set 1^+.

(iv) The number ***three***, denoted by 3, is the set 2^+.

(v) The number ***four***, denoted by 4, is the set 3^+.

This procedure can be inductively continued to define any finite nonnegative integer. •

The procedure above is well-defined, and so gives meaning to the symbol "k" where k is any nonnegative finite number. Let us give the various explicit ways of writing the first few numbers:

$$\begin{aligned}
0 &= \varnothing, \\
1 &= 0^+ = \{0\} &&= \{\varnothing\}, \\
2 &= 1^+ = \{0,1\} &&= \{\varnothing,\{\varnothing\}\}, \\
3 &= 2^+ = \{0,1,2\} &&= \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}, \\
4 &= 3^+ = \{0,1,2,3\} &&= \{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\},\{\varnothing,\{\varnothing\},\{\varnothing,\{\varnothing\}\}\}\}.
\end{aligned}$$

This settles the matter of defining any desired number. We now need to indicate how to talk about the *set* of numbers. This necessitates an assumption. As we shall see in Section 1.8.2, this assumption is framed as an axiom in axiomatic set theory.

1.4.3 Assumption There exists a set containing $\varnothing$ and all subsequent successors. •

We are now almost done. The remaining problem is that the set guaranteed by the assumption may contain more than what we want. However, this is easily remedied as follows. Let S be the set whose existence is guaranteed by Assumption 1.4.3. Define a collection $\mathscr{A}$ of subsets of S by

$$\mathscr{A} = \{A \subseteq S \mid \varnothing \in A \text{ and } n^+ \in A \text{ if } n \in A\}.$$

Note that $S \in \mathscr{A}$ so that $\mathscr{A}$ is nonempty. The following simple result is now useful.

1.4.4 Lemma *With $\mathscr{A}$ as above, if $\mathscr{B} \subseteq \mathscr{A}$, then $(\cap_{\mathrm{B}\in\mathscr{B}}\mathrm{B}) \in \mathscr{A}$.*

Proof For each $B \in \mathscr{B}$, $\varnothing \in B$. Thus $\varnothing \in \cap_{B\in\mathscr{B}}B$. Also let $n \in \cap_{B\in\mathscr{B}}B$. Since $n^+ \in B$ for each $B \in \mathscr{B}$, $n^+ \in \cap_{B\in\mathscr{B}}B$. Thus $(\cap_{B\in\mathscr{B}}B) \in \mathscr{A}$, as desired. ■

The lemma shows that $\cap_{A\in\mathscr{A}}A \in \mathscr{A}$. Now we have the following definition of the *set* of numbers.

1.4.5 Definition (Natural numbers) Let S and $\mathscr{A}$ be as defined above.

(i) The set $\cap_{A\in\mathscr{A}}A$ is denoted by $\mathbb{Z}_{\geq 0}$, and is the set of ***nonnegative integers***.

(ii) The set $\mathbb{Z}_{\geq 0} \setminus \{0\}$ is denoted by $\mathbb{Z}_{>0}$, and is the set of ***natural numbers***. •

1.4.6 Remark (Convention concerning $\mathbb{Z}_{>0}$ and $\mathbb{Z}_{\geq 0}$) There are two standard conventions concerning notation for nonnegative and positive integers. Neither agree with our notation. The two more or less standard bits of notation are:

1. $\mathbb{N}$ is the set of natural numbers and something else, maybe $\mathbb{Z}_{\geq 0}$, denotes the set of nonnegative integers;
2. $\mathbb{N}$ is the set of nonnegative integers (these are called the natural numbers in this scheme) and something else, maybe $\mathbb{N}^*$, denotes the set of natural numbers (called the positive natural numbers in this scheme).

Neither of these schemes is optimal on its own, and since there is no standard here, we opt for notation that is more logical. This will not cause the reader problems we hope, and may lead some to adopt our entirely sensible notation. •

Next we turn to the definition of the usual operations of arithmetic with the set $\mathbb{Z}_{\geq 0}$. That is to say, we indicate how to "add" and "multiply." First we consider addition.

1.4.7 Definition (Addition in $\mathbb{Z}_{\geq 0}$) For $k \in \mathbb{Z}_{\geq 0}$, inductively define a map $a_k \colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$, called ***addition by*** $\mathbf{k}$, by

(i) $a_k(0) = k$;

(ii) $a_k(j^+) = (a_k(j))^+$, $j \in \mathbb{Z}_{>0}$.

We denote $a_k(j) = k + j$. •

Upon a moments reflection, it is easy to convince yourself that this formal definition of addition agrees with our established intuition. Roughly speaking, one defines $k + (j + 1) = (k + j) + 1$, where, by definition, the operation of adding 1 means taking the successor. With these definitions it is straightforward to verify such commonplace assertions as "$1 + 1 = 2$."

Now we define multiplication.

1.4.8 Definition (Multiplication in $\mathbb{Z}_{\geq 0}$) For $k \in \mathbb{Z}_{\geq 0}$, inductively define a map $m_k \colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}_{\geq 0}$, called ***multiplication by*** $\mathbf{k}$, by

(i) $m_k(0) = 0$;

(ii) $m_k(j^+) = m_k(j) + k$.

We denote $m_k(j) = k \cdot j$, or simply kj where no confusion can arise. •

Again, this definition of multiplication is in concert with our intuition. The definition says that $k \cdot (j + 1) = k \cdot j + k$. For $k, m \in \mathbb{Z}_{\geq 0}$, define k^m recursively by $k^0 = 1$, and $k^{m^+} = k^m \cdot k$. The element $k^m \in \mathbb{Z}_{\geq 0}$ is the mth ***power*** of k.

Let us verify that addition and multiplication in $\mathbb{Z}_{\geq 0}$ have the expected properties. In stating the properties, we use the usual order of operation rules one learns in high school; in this case, operations are done with the following precedence: (1) operations enclosed in parentheses, (2) multiplication, then (3) addition.

1.4.9 Proposition (Properties of arithmetic in $\mathbb{Z}_{\geq 0}$) *Addition and multiplication in $\mathbb{Z}_{\geq 0}$ satisfy the following rules:*

(i) $k_1 + k_2 = k_2 + k_1$, $k_1, k_2 \in \mathbb{Z}_{\geq 0}$ *(**commutativity** of addition);*

(ii) $(k_1 + k_2) + k_3 = k_1 + (k_2 + k_3)$, $k_1, k_2, k_3 \in \mathbb{Z}_{\geq 0}$ *(**associativity** of addition);*

(iii) $k + 0 = k$, $k \in \mathbb{Z}_{\geq 0}$ *(**additive identity**);*

(iv) $k_1 \cdot k_2 = k_2 \cdot k_1$, $k_1, k_2 \in \mathbb{Z}_{\geq 0}$ *(**commutativity** of multiplication);*

(v) $(k_1 \cdot k_2) \cdot k_3 = k_1 \cdot (k_2 \cdot k_3)$, $k_1, k_2, k_3 \in \mathbb{Z}_{\geq 0}$ *(**associativity** of multiplication);*

(vi) $k \cdot 1 = k$, $k \in \mathbb{Z}_{\geq 0}$ *(**multiplicative identity**);*

(vii) $j \cdot (k_1 + k_2) = j \cdot k_1 + j \cdot k_2$, $j, k_1, k_2 \in \mathbb{Z}_{\geq 0}$ *(**distributivity**);*

(viii) $j^{k_1} \cdot j^{k_2} = j^{k_1 + k_2}$, $j, k_1, k_2 \in \mathbb{Z}_{\geq 0}$*;*

(ix) if $j_1 + k = j_2 + k$ *then* $j_1 = j_2$, $j_1, j_2, k \in \mathbb{Z}_{\geq 0}$ *(**cancellation law for addition**);*

(x) if $j_1 \cdot k = j_2 \cdot k$ *then* $j_1 = j_2$, $j_1, j_2, k \in \mathbb{Z}_{>0}$ *(**cancellation law for multiplication**).*

Proof We shall prove these in logical order, rather than the order in which they are stated.

(ii) We prove this by induction on k_3. For $k_3 = 0$ we have $(k_1 + k_2) + 0 = k_1 + k_2$ and $k_1 + (k_2 + 0) = k_1 + k_2$, giving the result in this case. Now suppose that $(k_1 + k_2) + j = k_1 + (k_2 + j)$ for $j \in \{0, 1, \dots, k_3\}$. Then

$$(k_1 + k_2) + k_3^+ = ((k_1 + k_2) + k_3)^+ = (k_1 + (k_2 + k_3))^+ = k_1 + (k_2 + k_3)^+ = k_1 + (k_2 + k_3^+),$$

where we have used the definition of addition, the induction hypothesis, and then twice used the definition of addition.

(i) We first claim that $0 + k = k$ for all $k \in \mathbb{Z}_{\geq 0}$. It is certainly true, by definition, that $0 + 0 = 0$. Now suppose that $0 + j = j$ for $j \in \{0, 1, \dots, k\}$. Then

$$0 + k^+ = 0 + (k + 1) = (0 + k) + 1 = k + 1 = k^+.$$

We next claim that $k_1^+ + k_2 = (k_1 + k_2)^+$ for $k_1, k_2 \in \mathbb{Z}_{\geq 0}$. We prove this by induction on k_2. For $k_2 = 0$ we have $k_1^+ + 0 = k_1^+$ and $(k_1 + 0)^+ = k_1^+$, using the definition of addition. This gives the claim for $k_2 = 0$. Now suppose that $k_1^+ + j = (k_1 + j)^+$ for $j \in \{0, 1, \dots, k_2\}$. Then

$$k_1^+ + k_2^+ = k_1^+ + (k_2 + 1) = (k_1^+ + k_2) + 1 = (k_1^+ + k_2)^+,$$

as desired.

We now complete the proof of this part of the result by induction on k_1. For $k_1 = 0$ we have $0 + k_2 = k_2 = k_2 + 0$, using the first of our claims above and the definition of addition. Now suppose that $j + k_2 = k_2 + j$ for $j \in \{0, 1, \dots, k_1\}$. Then

$$k_1^+ + k_2 = (k_1 + k_2)^+ = (k_2 + k_1)^+ = k_2 + k_1^+,$$

using the second or our claims above and the definition of addition.

(iii) This is part of the definition of addition.

(vii) We prove the this by induction on k_2. First note that for $k_2 = 0$ we have $j \cdot (k_1 + 0) = j \cdot k_1$ and $j \cdot k_1 + j \cdot 0 = j \cdot k_1 + 0 = j \cdot k_1$, so the result holds when $k_2 = 0$. Now suppose that $j \cdot (k_1 + k) = j \cdot k_1 + j \cdot k$ for $k \in \{0, 1, \dots, k_2\}$. Then we have

$$\begin{aligned} j \cdot (k_1 + k_2^+) &= j \cdot (k_1 + k_2)^+ = j \cdot (k_1 + k_2) + j \\ &= (j \cdot k_1 + j \cdot k_2) + j = j \cdot k_1 + (j \cdot k_2 + j) \\ &= j \cdot k_1 + j \cdot k_2^+, \end{aligned}$$

as desired, where we have used, in sequence, the definition of addition, the definition of multiplication, the induction hypothesis, the associativity of addition, and the definition of multiplication.

(iv) We first prove by induction on k that $0 \cdot k = 0$ for $k \in \mathbb{Z}_{\geq 0}$. For $k = 0$ the claim holds by definition of multiplication. So suppose that $0 \cdot j = 0$ for $j \in \{0, 1, \dots, k\}$ and then compute $0 \cdot k^+ = 0 \cdot k + 0 = 0$, as desired.

We now prove the result by induction on k_2. For $k_2 = 0$ we have $k_1 \cdot 0 = 0$ by definition of multiplication. We also have $k_2 \cdot 0 = 0$ by the first part of the proof. So now suppose that $k_1 \cdot j = j \cdot k$ for $j \in \{0, 1, \dots, k_2\}$. We then have

$$k_1 \cdot k_2^+ = k_1 \cdot k_2 + k_1 = k_2 \cdot k_1 + k_1 = k_1 + k_2 \cdot k_1 = (1 + k_2) \cdot k_1 = k_2^+ \cdot k_1,$$

where we have used, in sequence, the definition of multiplication, the induction hypothesis, commutativity of addition, distributivity, commutativity of addition, and the definition of addition.

(v) We prove this part of the result by induction on k_3. For $k_3 = 0$ we have $(k_1 \cdot k_2) \cdot 0 = 0$ and $k_1 \cdot (k_2 \cdot 0) = k_1 \cdot 0 = 0$. Thus the result is true when $k_3 = 0$. Now suppose that $(k_1 \cdot k_2) \cdot j = k_1 \cdot (k_2 \cdot j)$ for $j \in \{0, 1, \dots, k_3\}$. Then

$$(k_1 \cdot k_2) \cdot k_3^+ = (k_1 \cdot k_2) \cdot k_3 + k_1 \cdot k_2 = k_1 \cdot (k_2 \cdot k_3) + k_1 \cdot k_2 = k_1 \cdot (k_2 \cdot k_3 + k_2) = k_1 \cdot (k_2 \cdot k_3^+),$$

where we have used, in sequence, the definition of multiplication, the induction hypothesis, distributivity, and the definition of multiplication.

(vi) This follows from the definition of multiplication.

(viii) We prove the result by induction on k_1. The result is obviously true for $k_2 = 0$, so suppose that $j^{k_1+l} = j^{k_1} \cdot j^l$ for $l \in \{1, \dots, k_2\}$. Then

$$j^{k_1+k_2^+} = j^{(k_1+k_2)^+} = j^{k_1+k_2} \cdot j = j^{k_1} \cdot j^{k_2} \cdot j = j^{k_1} \cdot j^{k_2^+},$$

as desired.

(ix) We prove the result by induction on k. Since

$$j_1 + 0 = j_1, \quad j_2 + 0 = j_2,$$

the assertion holds for all $j_1, j_2 \in \mathbb{Z}_{\geq 0}$ and for $k = 0$. Now suppose the result holds for all $j_1, j_2 \in \mathbb{Z}_{\geq 0}$ and for $k \in \{0, 1, \dots, m\}$. Then

$$j_1 + (m + 1) = (j_1 + m) + 1, \quad j_2 + (m + 1) = (j_2 + m) + 1$$

and so

$$(j_1 + m) + 1 = (j_2 + m) + 1 \quad \Longrightarrow \quad j_1 + m = j_2 + m \quad \Longrightarrow \quad j_1 = j_2,$$

using the induction hypotheses. Thus the result holds for $k = m + 1$, completing our proof by induction.

(x) We prove this result by induction on j_1. First take $j_1 = 1$ and assume that $1 \cdot k = j_2 \cdot k$ for all $j_2, k \in \mathbb{Z}_{>0}$. If $j_2 = 1$ then we conclude that the assertion holds. If $j_2 \neq 1$, then $j_2 = j_2' + 1$ for some $j_2' \in \mathbb{Z}_{>0}$ and so we have

$$1 \cdot k = (j_2' + 1) \cdot k = j_2' \cdot k + 1 \cdot k,$$

giving $j_2' \cdot k = 0$ using the cancellation rule for addition. But the definition of multiplication by j_2' implies that we must have $k = 0$, which is not the case since we are assuming that $k \in \mathbb{Z}_{>0}$. Thus the assertion holds for $j_1 = 1$ and for all $j_2, k \in \mathbb{Z}_{>0}$. Now assume that the assertion holds for $j_2 \in \{1, \dots, m\}$ and assume that $(m + 1) \cdot k = j_2 \cdot k$ for all $j_2, k \in \mathbb{Z}_{>0}$. We first assert that $j_2 \neq 1$. Indeed, if $j_2 = 1$ we have $m \cdot k = 0$ using the cancellation law for addition, and, as above, this cannot be since $k \in \mathbb{Z}_{>0}$. Therefore, $j_2 = j_2' + 1$ for some $j_2' \in \mathbb{Z}_{>0}$ and so

$$(m + 1) \cdot k = (j_2' + 1) \cdot k \quad \Longrightarrow \quad m \cdot k = j_2' \cdot k$$

by the cancellation law for addition. Thus, by the induction hypothesis, $m = j_2'$ and so $j_2 = m + 1$, which gives this part of the lemma. ■

1.4.2 Two relations on $\mathbb{Z}_{\geq 0}$

Another property of the naturals that we would all agree they ought to have is an "order." Thus we should have a means of saying when one natural number is less than another. To get started at this, we have the following result.

1.4.10 Lemma *For* j, k $\in \mathbb{Z}_{\geq 0}$*, exactly one of the following possibilities holds:*

(i) j $\subset$ k*;*

(ii) k $\subset$ j*;*

(iii) j = k.

Proof For $k \in \mathbb{Z}_{\geq 0}$ define

$$S(k) = \{j \in \mathbb{Z}_{>0} \mid j \subset k, k \subset j, \text{or } j = k\}.$$

We shall prove by induction that $S(k) = \mathbb{Z}_{\geq 0}$ for each $k \in \mathbb{Z}_{\geq 0}$.

First take the case of $k = 0$. Since $\varnothing$ is a subset of every set, $0 \in S(0)$. Now suppose that $j \in S(0)$ for $j \in \mathbb{Z}_{\geq 0}$. We have the following cases.

1. $j \in 0$: This is impossible since 0 is the empty set.
2. $0 \in j$: In this case $0 \in j^+$.
3. $0 = j$: In this case $0 \in j^+$.

Thus $j \in S(0)$ implies that $j^+ \in S(0)$, and so $S(0) = \mathbb{Z}_{\geq 0}$.

Now suppose that $S(m) = \mathbb{Z}_{\geq 0}$ for $m \in \{0, 1, \dots, k\}$. We will show that $S(k^+) = \mathbb{Z}_{\geq 0}$. Clearly $0 \in S(k^+)$. So suppose that $j \in S(k^+)$. We again have three cases.

1. $j \in k^+$: We have the following two subcases.
   (a) $j = k$: Here we have $j^+ = k^+$.
   (b) $j \in k$: Since $j^+ \in S(k)$ by the induction hypothesis, we have the following three cases.
      (i) $k \in j^+$: This is impossible since $j \in k$.
      (ii) $j^+ \in k$: Here $j^+ \in k^+$.
      (iii) $j^+ = k$: Here again, $j^+ \in n^+$.
2. $k^+ \in j$: In this case $k^+ \in j^+$.
3. $k^+ = j$: In this case $k^+ \in j^+$.

In all cases we conclude that $j^+ \in S(k^+)$, and this completes the proof. ■

It is easy to show that $j \in k$ if and only if $j \subseteq k$, and that, if $j \in k$ but $j \neq k$, then $j \subset k$ (see Exercise 1.4.2). With this result, it is now comparatively easy to prove the following.

1.4.11 Proposition (Order[5] on $\mathbb{Z}_{\geq 0}$) *On $\mathbb{Z}_{\geq 0}$ define two relations $<$ and $\leq$ by*

$$\begin{aligned} \mathrm{j} < \mathrm{k} \quad &\iff \quad \mathrm{j} \subset \mathrm{k}, \\ \mathrm{j} \leq \mathrm{k} \quad &\iff \quad \mathrm{j} \subseteq \mathrm{k}. \end{aligned}$$

Then

(i) *$<$ and $\leq$ are transitive,*

(ii) *$<$ is irreflexive;*

(iii) *$\leq$ is reflexive and antisymmetric.*

Furthermore, for any $\mathrm{j}, \mathrm{k} \in \mathbb{Z}_{\geq 0}$, either $\mathrm{j} \leq \mathrm{k}$ or $\mathrm{k} \leq \mathrm{j}$.

The following rewording of the final part of the result is distinguished.

1.4.12 Corollary (Trichotomy Law for $\mathbb{Z}_{\geq 0}$) *For $\mathrm{j}, \mathrm{k} \in \mathbb{Z}_{\geq 0}$, exactly one of the following possibilities holds:*

(i) $\mathrm{j} < \mathrm{k}$;

(ii) $\mathrm{k} < \mathrm{j}$;

(iii) $\mathrm{j} = \mathrm{k}$.

Of course, the symbols "$<$" and "$\leq$" have their usual meaning, which is "less than" and "less than or equal to," respectively. We shall explore such matters in more depth and generality in Section 1.5.

We shall also sometimes write "$j > k$" (resp. "$j \geq k$") for "$k < j$" (resp. "$k \leq j$"). The symbols "$>$" and "$\geq$" then have their usual meaning as "greater than" and "greater than or equal to," respectively.

The relations $<$ and $\leq$ satisfy some natural properties with respect to addition and multiplication in $\mathbb{Z}_{\geq 0}$. Let us record these, leaving their proof as Exercise 1.4.3.

1.4.13 Proposition (Relation between addition and multiplication and <) *For $\mathrm{j}, \mathrm{k}, \mathrm{m} \in \mathbb{Z}_{\geq 0}$, the following statements hold:*

(i) *if $\mathrm{j} < \mathrm{k}$ then $\mathrm{j} + \mathrm{m} < \mathrm{k} + \mathrm{m}$;*

(ii) *if $\mathrm{j} < \mathrm{k}$ and if $\mathrm{m} \neq 0$ then $\mathrm{m} \cdot \mathrm{j} < \mathrm{m} \cdot \mathrm{k}$.*

1.4.3 Construction of the integers from the natural numbers

Next we construct negative numbers to arrive at a definition of the integers. The construction renders the integers as the set of equivalence classes under a prescribed equivalence relation in $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$. The equivalence relation is defined formally as follows:

$$(j_1, k_1) \sim (j_2, k_2) \quad \iff \quad j_1 + k_2 = k_1 + j_2. \tag{1.1}$$

It is a simple exercise to check that this is indeed an equivalence relation.

We now define the integers.

[5] We have not introduced the notion of order yet, but refer the reader to Section 1.5.

1.4.14 Definition (Integers) The set of ***integers*** is the set $\mathbb{Z} = (\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0})/\sim$, where $\sim$ is the equivalence relation in (1.1). •

Now let us try to understand this definition by understanding the equivalence classes under the relation of (1.1). Key to this is the following result.

1.4.15 Lemma *Let* Z *be the subset of* $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ *defined by*

$$\mathrm{Z} = \{(\mathrm{k}, 0) \mid \mathrm{k} \in \mathbb{Z}_{>0}\} \cup \{(0, \mathrm{k}) \mid \mathrm{k} \in \mathbb{Z}_{>0}\} \cup \{(0, 0)\},$$

and define a map $\mathrm{f_Z} \colon \mathrm{Z} \to \mathbb{Z}$ *by* $\mathrm{f_Z}(\mathrm{j}, \mathrm{k}) = [(\mathrm{j}, \mathrm{k})]$. *Then* $\mathrm{f_Z}$ *is a bijection.*

Proof First we show that f_Z is injective. Suppose that $f_Z(j_1, k_1) = f_Z(j_2, k_2)$. This means that $(j_1, k_1) \sim (j_2, k_2)$, or that $j_1 + k_2 = k_1 + j_2$. If $(j_1, k_1) = (0, 0)$, then this means that $k_2 = j_2$, which means that $(j_2, k_2) = (0, 0)$ since this is the only element of Z whose entries agree. If $j_1 = 0$ and $k_1 > 0$, then we have $k_2 = k_1 + j_2$. Since at least one of j_2 and k_2 must be zero, we then deduce that it must be that j_2 is zero (or else the equality $k_2 = k_1 + j_2$) cannot hold. This then also gives $k_2 = k_1$. A similar argument holds if $j_1 > 0$ and $k_1 = 0$. This shows injectivity of f_Z.

Next we show that f_Z is surjective. Let $[(j, k)] \in \mathbb{Z}$. By the Trichotomy Law, we have three cases.

1. $j = k$: We claim that $[(j, j)] = f_Z(0, 0)$. Indeed, we need only note that $(0, 0) \sim (j, j)$ since $0 + j = 0 + j$.
2. $j < k$: Let $m \in \mathbb{Z}_{>0}$ be defined such that $j + m = k$. (Why can this be done?) We then claim that $f_Z(0, m) = [(j, k)]$. Indeed, since $0 + k = m + j$, this is so.
3. $k < j$: Here we let $m \in \mathbb{Z}_{>0}$ satisfy $k + m = j$, and, as in the previous case, we can easily check that $f_Z(m, 0) = [(j, k)]$. ■

With this in mind, we introduce the following notation to denote an integer.

1.4.16 Notation (Notation for integers) Let $[(j, k)] \in \mathbb{Z}$.

(i) If $f_Z^{-1}[(j, k)] = [(0, 0)]$ then we write $[(j, k)] = 0$.

(ii) If $[(j, k)] = [(m, 0)]$, $m > 0$, then we write $[(j, k)] = m$. Such integers are ***positive***.

(iii) If $[(j, k)] = [(0, m)]$, $m > 0$, then we write $[(j, k)] = -m$. Such integers are ***negative***.

An integer is ***nonnegative*** if it is either positive or zero, and an integer is ***nonpositive*** if it is either negative or zero. •

This then relates the equivalence class definition of integers to the notion we are more familiar with: positive and negative numbers. We can also define the familiar operations of addition and multiplication of integers.

1.4.17 Definition (Addition and multiplication in $\mathbb{Z}$) Define the operations of ***addition*** and ***multiplication*** in $\mathbb{Z}$ by

(i) $[(j_1, k_1)] + [(j_2, k_2)] = [(j_1 + j_2, k_1 + k_2)]$ and

(ii) $[(j_1, k_1)] \cdot [(j_2, k_2)] = [(j_1 \cdot j_2 + k_1 \cdot k_2, j_1 \cdot k_2 + k_1 \cdot j_2)]$, respectively, for $[(j_1,k_1)],[(j_2,k_2)] \in \mathbb{Z}$. As with multiplication in $\mathbb{Z}_{\geq 0}$, we shall sometimes omit the "$\cdot$". •

These definitions do not *a priori* make sense; this needs to be verified.

1.4.18 Lemma *The definitions for addition and multiplication in $\mathbb{Z}$ a well-defined in that they do not depend on the choice of representative.*

Proof Let $(j_1,k_1) \sim (\tilde{j}_1,\tilde{k}_1)$ and $(j_2,k_2) \sim (\tilde{j}_2,\tilde{k}_2)$. Thus

$$j_1 + \tilde{k}_1 = k_1 + \tilde{j}_1, \quad j_2 + \tilde{k}_2 = k_2 + \tilde{j}_2.$$

It therefore follows that

$$(\tilde{j}_1 + \tilde{j}_2) + (k_1 + k_2) = (\tilde{k}_1 + \tilde{k}_2) + (j_1 + j_2),$$

which gives the independence of addition on representative.

To verify the well-definedness of multiplication, we first see that

$$\begin{aligned} j_2 \cdot (j_1 + \tilde{k}_1) + k_2 \cdot (\tilde{j}_1 + k_1) + \tilde{j}_1 \cdot (j_2 + \tilde{k}_2) + \tilde{k}_1 \cdot (\tilde{j}_2 + k_2) \\ = j_2 \cdot (k_1 + \tilde{j}_1) + k_2 \cdot (j_1 + \tilde{k}_1) + \tilde{j}_1 \cdot (k_2 + \tilde{j}_2) + \tilde{k}_1 \cdot (j_2 + \tilde{k}_2), \end{aligned}$$

and expanding this and rearranging gives

$$\begin{aligned} (j_1 \cdot j_2 + k_1 \cdot k_2 + \tilde{k}_1 \cdot \tilde{j}_2 + \tilde{j}_1 \cdot \tilde{k}_2) + (\tilde{k}_1 \cdot j_2 + \tilde{j}_1 \cdot k_2 + \tilde{j}_1 \cdot j_2 + \tilde{k}_1 \cdot k_2) \\ = (k_1 \cdot j_2 + j_1 \cdot k_2 + \tilde{j}_1 \cdot \tilde{j}_2 + \tilde{k}_1 \cdot \tilde{k}_2) + (\tilde{k}_1 \cdot j_2 + \tilde{j}_1 \cdot k_2 + \tilde{j}_1 \cdot j_2 + \tilde{k}_1 \cdot k_2). \end{aligned}$$

Using the cancellation law for addition we then have

$$(\tilde{j}_1 \cdot \tilde{j}_2 + \tilde{k}_1 \cdot \tilde{k}_2) + (j_1 \cdot k_2 + k_1 \cdot j_2) = (\tilde{j}_1 \cdot \tilde{k}_2 + \tilde{k}_1 \cdot \tilde{j}_2) + (j_1 \cdot j_2 + k_1 \cdot k_2),$$

which gives the independence of multiplication on representative. ■

As with elements of $\mathbb{Z}_{\geq 0}$, we can define powers for integers. Let $k \in \mathbb{Z}$ and $m \in \mathbb{Z}_{\geq 0}$. We define k^m recursively as follows. We take $k^0 = 1$ and define $k^{m^+} = k^m \cdot k$. We call k^m the mth ***power*** of k. Note that, at this point, k^m only makes sense for $m \in \mathbb{Z}_{\geq 0}$.

Finally, we give the properties of addition and multiplication in $\mathbb{Z}$. Some of these properties are as for $\mathbb{Z}_{\geq 0}$. However, there is a useful new feature that arises in $\mathbb{Z}$ that mirrors our experience with negative numbers. In the statement of the result, it is convenient to denote an integer as in Notation 1.4.16, rather than as in the definition.

1.4.19 Proposition (Properties of addition and multiplication in $\mathbb{Z}$) *Addition and multiplication in $\mathbb{Z}$ satisfy the following rules:*

(i) $\mathrm{k}_1 + \mathrm{k}_2 = \mathrm{k}_2 + \mathrm{k}_1$, $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}$ *(**commutativity** of addition);*

(ii) $(\mathrm{k}_1 + \mathrm{k}_2) + \mathrm{k}_3 = \mathrm{k}_1 + (\mathrm{k}_2 + \mathrm{k}_3)$, $\mathrm{k}_1, \mathrm{k}_2, \mathrm{k}_3 \in \mathbb{Z}$ *(**associativity** of addition);*

(iii) $\mathrm{k} + 0 = \mathrm{k}$, $\mathrm{k} \in \mathbb{Z}$ *(**additive identity**);*

(iv) $\mathrm{k} + (-1 \cdot \mathrm{k}) = 0$, $\mathrm{k} \in \mathbb{Z}$ *(**additive inverse**);*

(v) $\mathrm{k}_1 \cdot \mathrm{k}_2 = \mathrm{k}_2 \cdot \mathrm{k}_1$, $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}$ *(**commutativity** of multiplication);*

(vi) $(\mathrm{k}_1 \cdot \mathrm{k}_2) \cdot \mathrm{k}_3 = \mathrm{k}_1 \cdot (\mathrm{k}_2 \cdot \mathrm{k}_3)$, $\mathrm{k}_1, \mathrm{k}_2, \mathrm{k}_3 \in \mathbb{Z}$ *(**associativity** of multiplication);*

(vii) $\mathrm{k} \cdot 1 = \mathrm{k}$, $\mathrm{k} \in \mathbb{Z}$ *(**multiplicative identity**);*

(viii) $\mathrm{j} \cdot (\mathrm{k}_1 + \mathrm{k}_2) = \mathrm{j} \cdot \mathrm{k}_1 + \mathrm{j} \cdot \mathrm{k}_2$, $\mathrm{j}, \mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}$ *(**distributivity**);*

(ix) $\mathrm{j}^{\mathrm{k}_1} \cdot \mathrm{j}^{\mathrm{k}_2} = \mathrm{j}^{\mathrm{k}_1 + \mathrm{k}_2}$, $\mathrm{j} \in \mathbb{Z}$, $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{\geq 0}$.

Moreover, if we define $\mathrm{i}_{\mathbb{Z}_{\geq 0}} \colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}$ *by* $\mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}) = [(\mathrm{k}, 0)]$, *then addition and multiplication in* $\mathbb{Z}$ *agrees with that in* $\mathbb{Z}_{\geq 0}$*:*

$$\mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}_1) + \mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}_2) = \mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}_1 + \mathrm{k}_2), \quad \mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}_1) \cdot \mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}_2) = \mathrm{i}_{\mathbb{Z}_{\geq 0}}(\mathrm{k}_1 \cdot \mathrm{k}_2).$$

Proof These follow easily from the definitions of addition and multiplication, using the fact that the corresponding properties hold for $\mathbb{Z}_{\geq 0}$. We leave the details to the reader as Exercise 1.4.4. We therefore only prove the new property (iv). For this, we suppose without loss of generality that $k \in \mathbb{Z}_{\geq 0}$, i.e., $k = [(k, 0)]$. Then $-k = [(0, k)]$ so that

$$k + (-k) = [(k + 0, 0 + k)] = [(k, k)] = [(0, 0)] = 0,$$

as claimed. ■

We shall make the convention that $-1 \cdot k$ be written as $-k$, whether k be positive or negative. We shall also, particularly as we move along to things of more substance, think of $\mathbb{Z}_{\geq 0}$ as a subset of $\mathbb{Z}$, without making explicit reference to the map $i_{\mathbb{Z}_{\geq 0}}$.

1.4.4 Two relations in $\mathbb{Z}$

Finally we introduce in $\mathbb{Z}$ two relations that extend the relations $<$ and $\leq$ for $\mathbb{Z}_{\geq 0}$. The following result is the analogue of Proposition 1.4.11.

1.4.20 Proposition (Order on $\mathbb{Z}$) *On* $\mathbb{Z}$ *define two relations* $<$ *and* $\leq$ *by*

$$\begin{aligned} [(\mathrm{j}_1, \mathrm{k}_1)] < [(\mathrm{j}_2, \mathrm{k}_2)] \quad &\iff \quad \mathrm{j}_1 + \mathrm{k}_2 < \mathrm{k}_1 + \mathrm{j}_2, \\ [(\mathrm{j}_1, \mathrm{k}_1)] \leq [(\mathrm{j}_2, \mathrm{k}_2)] \quad &\iff \quad \mathrm{j}_1 + \mathrm{k}_2 \leq \mathrm{k}_1 + \mathrm{j}_2. \end{aligned}$$

Then

(i) $<$ *and* $\leq$ *are transitive,*

(ii) $<$ *is irreflexive, and*

(iii) $\leq$ *is reflexive.*

Furthermore, for any $\mathrm{j}, \mathrm{k} \in \mathbb{Z}$, *either* $\mathrm{j} \leq \mathrm{k}$ *or* $\mathrm{k} \leq \mathrm{j}$.

Proof First one must show that the relations are well-defined in that they do not depend on the choice of representative. Thus let $[(j_1, k_1)] \sim [(\tilde{j}_1, \tilde{k}_1)]$ and $[(j_2, k_2)] \sim [(\tilde{j}_2, \tilde{k}_2)]$, so that

$$j_1 + \tilde{k}_1 = k_1 + \tilde{j}_1, \quad j_2 + \tilde{k}_2 = k_2 + \tilde{j}_2.$$

Now suppose that the relation $j_1 + k_2 < k_1 + j_2$ holds. Now perform the following steps:

1. add $\tilde{j}_1 + k_1 + j_2 + \tilde{k}_2 + j_1 + \tilde{k}_1 + k_2 + \tilde{j}_2$ to both sides of the relation;
2. observe that $j_1 + k_2 + k_1 + j_2$ appears on both sides of the relation;
3. observe that $j_1 + \tilde{k}_1$ appears on one side of the relation and that $\tilde{j}_1 + k_1$ appears on the other;
4. observe that $k_2 + \tilde{j}_2$ appears on one side of the relation and that $j_2 + \tilde{k}_2$ appears on the other.

After simplification using the above observations, and using Proposition 1.4.13, we note that the relation $\tilde{j}_1 + \tilde{k}_2 < \tilde{k}_1 + \tilde{j}_2$ holds, which gives independence of the definition of $<$ on the choice of representative. The same argument works for the relation $\leq$.

The remainder of the proof follows in a fairly straightforward manner from the corresponding assertions for $\mathbb{Z}_{\geq 0}$, and we leave the details to the reader as Exercise 1.4.6.

■

As with the natural numbers, the last assertion of the previous result has a standard restatement.

1.4.21 Corollary (Trichotomy Law for $\mathbb{Z}$) *For $j, k \in \mathbb{Z}$, exactly one of the following possibilities holds:*

(i) $j < k$;

(ii) $k < j$;

(iii) $j = k$.

Similarly with $\mathbb{Z}_{\geq 0}$, we shall also write "$j > k$" for "$k < j$" and "$j \geq k$" for "$k \leq j$." It is also easy to directly verify that the relations $<$ and $\leq$ have the expected properties with respect to positive and negative integers. These are given in Exercise 1.4.7, for the interested reader.

We also have the following extension of Proposition 1.4.13 that relates addition and multiplication to the relations $<$ and $\leq$. We again leave these to the reader to verify in Exercise 1.4.8.

1.4.22 Proposition (Relation between addition and multiplication and <) *For $j, k, m \in \mathbb{Z}$, the following statements hold:*

(i) if $j < k$ *then* $j + m < k + m$;

(ii) if $j < k$ *and if* $m > 0$ *then* $m \cdot j < m \cdot k$;

(iii) if $j < k$ *and if* $m < 0$ *then* $m \cdot k < m \cdot j$;

(iv) if $0 < j, k$ *then* $0 < j \cdot k$.

1.4.5 The absolute value function

On the set of integers there is an important map that assigns a nonnegative integer to each integer.

1.4.23 Definition (Integer absolute value function) The ***absolute value function*** on $\mathbb{Z}$ is the map from $\mathbb{Z}$ to $\mathbb{Z}_{\geq 0}$, denoted by $k \mapsto |k|$, defined by

$$|k| = \begin{cases} k, & 0 < k, \\ 0, & k = 0, \\ -k, & k < 0. \end{cases}$$ •

The absolute value has the following properties.

1.4.24 Proposition (Properties of absolute value on $\mathbb{Z}$) *The following statements hold:*

(i) $|\mathrm{k}| \geq 0$ *for all* $\mathrm{k} \in \mathbb{Z}$*;*

(ii) $|\mathrm{k}| = 0$ *if and only if* $\mathrm{k} = 0$*;*

(iii) $|\mathrm{j} \cdot \mathrm{k}| = |\mathrm{j}| \cdot |\mathrm{k}|$ *for all* $\mathrm{j}, \mathrm{k} \in \mathbb{Z}$*;*

(iv) $|\mathrm{j} + \mathrm{k}| \leq |\mathrm{j}| + |\mathrm{k}|$ *for all* $\mathrm{j}, \mathrm{k} \in \mathbb{Z}$ *(**triangle inequality**).*

Proof Parts (i) and (ii) follow directly from the definition of $|\cdot|$.

(iii) We first note that $|-k| = |k|$ for all $k \in \mathbb{Z}$. Now, if $0 \leq j, k$, then the result is clear. If $j < 0$ and $k \geq 0$, then

$$|j \cdot k| = |-1 \cdot (-j) \cdot k| = |(-j) \cdot k| = |-j| \cdot |k| = |j| \cdot |k|.$$

A similar argument holds when $k < 0$ and $j \geq 0$.

(iv) We consider various cases.

1. $|j| \leq |k|$:
   (a) $j, k \geq 0$: Here $|j + k| = j + k$, and $|j| = j$ and $|k| = k$. So the result is obvious.
   (b) $j < 0, k \geq 0$: Here one can easily argue, using the definition of addition, that $0 < j + k$. From Proposition 1.4.22 we have $j + k < 0 + k = k$. Therefore, $|j + k| < |k| < |j| + |k|$, again by Proposition 1.4.22.
   (c) $k < 0, j \geq 0$: This follows as in the preceding case, swapping j and k.
   (d) $j, k < 0$: Here $|j + k| = |-j + (-k)| = |-(j + k)| = -(j + k)$, and $|j| = -j$ and $|k| = -k$, so the result follows immediately.
2. $|k| \leq |j|$: The argument here is the same as the preceding one, but swapping j and k. ■

Exercises

1.4.1 Let $k \in \mathbb{Z}_{>0}$. Show that $k \subseteq \mathbb{Z}_{>0}$; thus k is both an element of $\mathbb{Z}_{>0}$ and a subset of $\mathbb{Z}_{>0}$.

1.4.2 Let $j, k \in \mathbb{Z}_{\geq 0}$. Do the following:

(a) show that $j \in k$ if and only if $j \subseteq k$;

(b) show that if $j \subset k$, then $k \notin j$ (and so $j \in k$ by the Trichotomy Law).

1.4.3 Prove Proposition 1.4.13.

1.4.4 Complete the proof of Proposition 1.4.19.

1.4.5 For $j_1, j_2, k \in \mathbb{Z}$, prove the distributive rule $(j_1 + j_2) \cdot k = j_1 \cdot k + j_2 \cdot k$.

1.4.6 Complete the proof of Proposition 1.4.20.

1.4.7 Show that the relations $<$ and $\leq$ on $\mathbb{Z}$ have the following properties:

1. $[(0, j)] < [(0, 0)]$ for all $j \in \mathbb{Z}_{>0}$;
2. $[(0, j)] < [(k, 0)]$ for all $j, k \in \mathbb{Z}_{>0}$;
3. $[(0, j)] < [(0, k)]$, $j, k, \in \mathbb{Z}_{\geq 0}$, if and only if $k < j$;
4. $[(0, 0)] < [(j, 0)]$ for all $j \in \mathbb{Z}_{>0}$;
5. $[(j, 0)] < [(k, 0)]$, $j, k \in \mathbb{Z}_{\geq 0}$, if and only if $j < k$;
6. $[(0, j)] \leq [(0, 0)]$ for all $j \in \mathbb{Z}_{\geq 0}$;
7. $[(0, j)] \leq [(k, 0)]$ for all $j, k \in \mathbb{Z}_{\geq 0}$;
8. $[(0, j)] \leq [(0, k)]$, $j, k, \in \mathbb{Z}_{\geq 0}$, if and only if $k \leq j$;
9. $[(0, 0)] \leq [(j, 0)]$ for all $j \in \mathbb{Z}_{\geq 0}$;
10. $[(j, 0)] \leq [(k, 0)]$, $j, k \in \mathbb{Z}_{\geq 0}$, if and only if $j \leq k$.

1.4.8 Prove Proposition 1.4.22.

Section 1.5

Orders of various sorts

In Section 1.4 we defined two relations, denoted by $<$ and $\leq$, on both $\mathbb{Z}_{\geq 0}$ and $\mathbb{Z}$. Here we see that these relations have additional properties that fall into a general class of relations called orders. There are various classes or orders, having varying degrees of "strictness," as we shall see.

Do I need to read this section? Much of the material in this section is not used widely in the series, so perhaps can be overlooked until it is needed. •

1.5.1 Definitions

Let us begin by defining the various types of orders we consider.

1.5.1 Definition (Partial order, total order, well order) Let S be a set and let R be a relation in S.

(i) R is a ***partial order*** in S if it is reflexive, transitive, and antisymmetric.

(ii) A ***partially ordered set*** is a pair (S, R) where R is a partial order in S.

(iii) R is a ***strict partial order*** in S if it is irreflexive and transitive.

(iv) A ***strictly partially ordered set*** is a pair (S, R) where R is a strict partial order in S.

(v) R is a ***total order*** in S if it is a partial order and if, for each $x_1, x_2 \in S$, either $(x_1, x_2) \in R$ or $(x_2, x_1) \in R$.

(vi) A ***totally ordered set*** is a pair (S, R) where R is a total order in S.

(vii) R is a ***well order*** in S if it is a partial order and if, for every nonempty subset $A \subseteq S$, there exists an element $x \in A$ such that $(x, x') \in R$ for every $x' \in A$.

(viii) A ***well ordered set*** is a pair (S, R) where R is a well order in S. •

1.5.2 Remark (Mathematical structures as ordered pairs) In the preceding definitions we see four instances of an "X set," where X is some property, e.g., a partial order. In such cases, it is common practice to do as we have done and write the object as an ordered pair, in the cases above, as (S, R). The practice dictates that the first element in the ordered pair be the name of the set, and that the second specifies the structure.

In many cases one simply wishes to refer to the set, with the structure being understood. For example, one might say, "Consider the partially ordered set S..." and not make explicit reference to the partial order. Both pieces of language are in common use by mathematicians, and in mathematical texts. •

Let us consider some simple examples of partial and strict partial orders.

1.5.3 Examples (Partial orders)

1. Consider the relation $R = \{(k_1, k_2) \mid k_1 \leq k_2\}$ in either $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$. Then one verifies that R is a partial order. In fact, it is both a total order and a well order.
2. Consider the relation $R = \{(k_1, k_2) \mid k_1 \leq k_2\}$ in either $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$. Here one can verify that R is a strict partial order.
3. Let S be a set and consider the relation R in $\mathbf{2}^S$ defined by $R = \{(A, B) \mid A \subseteq B\}$. Here one can see that R is a partial order, but it is generally neither a total order nor a well order (cf. Exercise 1.5.2).
4. Let S be a set and consider the relation R in $\mathbf{2}^S$ defined by $R = \{(A, B) \mid A \subset B\}$. In this case R can be verified to be a strict partial order.
5. A well order R is a total order. Indeed, for $(x_1, x_2) \in R$, there exists an element $x \in \{x_1, x_2\}$ such that $(x, x') \in R$ for every $x' \in \{x_1, x_2\}$. But this implies that either $(x_1, x_2) \in R$ or $(x_2, x_1) \in R$, meaning that R is a total order. •

Motivated by the first and second of these examples, we utilise the following more or less commonplace notation for partial orders.

1.5.4 Notation ($\preceq$ and $\prec$) If R is a partial order in S, we shall normally write $x_1 \preceq x_2$ for $(x_1, x_2) \in R$, and shall refer to $\preceq$ as the partial order. In like manner, if R is a strict partial order in S, we shall write $x_1 \prec x_2$ for $(x_1, x_2) \in R$. We shall also use $x_1 \succeq x_2$ and $x_1 \succ x_2$ to stand for $x_2 \preceq x_1$ and $x_2 \prec x_1$, respectively. •

There is a natural way of associating to every partial order a strict partial order, and vice versa.

1.5.5 Proposition (Relationship between partial and strict partial orders) *Let* S *be a set.*

(i) *If $\preceq$ is a partial order in* S*, then the relation $\prec$ defined by*

$$x_1 \prec x_2 \quad \Longleftrightarrow \quad x_1 \preceq x_2 \text{ and } x_1 \neq x_2$$

is a strict partial order in S*.*

(ii) *If $\prec$ is a strict partial order in* S*, then the relation $\preceq$ defined by*

$$x_1 \preceq x_2 \quad \Longleftrightarrow \quad x_1 \prec x_2 \text{ or } x_1 = x_2$$

is a partial order in S*.*

Proof This is a straightforward matter of verifying that the definitions are satisfied. ■

When talking about a partial order $\preceq$, the symbol $\prec$ will always refer to the strict partial order as in part (i) of the preceding result. Similarly, given a strict partial order $\prec$, the symbol $\preceq$ will always refer to the partial order as in part (ii) of the preceding result.

1.5.6 Examples (Example 1.5.3 cont'd)

1. One can readily verify that $<$ is the strict partial order associated with the partial order $\leq$ in either $\mathbb{Z}_{\geq 0}$ or $\mathbb{Z}$, and that $\leq$ is the partial order associated to $<$.
2. It is also easy to verify that, for a set S, $\subset$ is the strict partial order in $\mathbf{2}^S$ associated to the partial order $\subseteq$, and that $\subseteq$ is the partial order associated to $\subset$. •

1.5.2 Subsets of partially ordered sets

Surrounding subsets of a partially ordered set $(S, \preceq)$ there is some useful language. For the following definition, it is helpful to think of an order, be it partial, strictly partial, or whatever, as a relation, and to use the notation of a relation. Thus we refer to an order as R, and not as $\preceq$.

1.5.7 Definition (Restriction of an order) Let S be a set and let R be a partial order, (resp. strict partial order, total order, well order) in S. For a subset $T \subseteq S$, the ***restriction*** of R to T is the partial order (resp. strict partial order, total order, well order) in T defined by

$$R|T = R \cap \{(x_1, x_2) \in S \times S \mid x_1, x_2 \in T\}.$$ •

It is a trivial matter to see that if R is an order, then its restriction to T is an order having the same properties as R, as is tacitly assumed in the definition. The notion of the restriction of an order allows us to talk unambiguously about the order on a subset of a given set, and we shall do this freely in this section.

Since most of this section is language, let us begin with some simple language associated with points.

1.5.8 Definition (Comparing elements in a partially ordered set) Let $(S, \preceq)$ be a partially ordered set.

(i) A point $x_1 \in S$ is ***less*** than or ***smaller*** than x_2, or equivalently is a ***predecessor*** of x_2, if $x_1 \preceq x_2$.

(ii) A point $x_1 \in S$ is ***greater*** than or ***larger*** than x_2, or equivalently is a ***successor*** of x_2, if $x_1 \succeq x_2$.

(iii) A point x' is ***between*** x_1 and x_2 if $x_1 \preceq x'$ and if $x' \preceq x_2$.

Similarly, let $(S, \prec)$ be a strictly partially ordered set.

(iv) A point $x_1 \in S$ is ***strictly less*** than or ***strictly smaller*** than x_2, or equivalently is a ***strict predecessor*** of x_2, if $x_1 \prec x_2$.

(v) A point $x_1 \in S$ is ***strictly greater*** than or ***strictly larger*** than x_2, or equivalently is a ***strict successor*** of x_2, if $x_1 \succ x_2$.

(vi) A point x' is ***strictly between*** x_1 and x_2 if $x_1 \prec x'$ and if $x' \prec x_2$.

(vii) If $x_1 < x_2$ and there exists no $x' \in S$ that is strictly between x_1 and x_2, then x_1 is the ***immediate predecessor*** of x_2. •

Next we talk about some language attached to subsets of a partially ordered set.

1.5.9 Definition (Segment, least, greatest, minimal, maximal) Let $(S, \preceq)$ be a partially ordered set.

(i) The ***initial segment*** determined by $x \in S$ is the set $\underline{\mathrm{seg}}(x) = \{x' \in S \mid x' \preceq S\}$.

(ii) A ***least***, ***smallest***, or ***first*** element in S is an element $x \in S$ with the property that $x \preceq x'$ for every $x' \in S$.

(iii) A ***greatest***, ***largest***, or ***last*** element in S is an element $x \in S$ with the property that $x' \preceq x$ for every $x' \in S$.

(iv) A ***minimal*** element of S is an element $x \in S$ with the property that $x \preceq x'$ implies that $x' = x$.

(v) A ***maximal*** element of S is an element $x \in S$ with the property that $x \prec x'$ implies that $x' = x$.

Now let $(S, \preceq)$ be a partially ordered set.

(vi) The ***strict initial segment*** determined by $x \in S$ is the set $\mathrm{seg}(x) = \{x' \in S \mid x' \prec S\}$. •

The least and greatest elements of a set, if they exist, are unique. This is easy to prove (Exercise 1.5.4).

Let us give an example that distinguishes between least and minimal.

1.5.10 Example (Least and minimal are different) Let S be a set and consider the partially ordered set $(\mathbf{2}^S \setminus \varnothing, \subseteq)$. Then any singleton is a minimal element of $\mathbf{2}^S \setminus \varnothing$. However, unless S is itself a set with only one member, then $\mathbf{2}^S$ has no least element, i.e., there is no subset which is contained in every other subset. •

Next we turn to two important concepts related to partial orders.

1.5.11 Definition (Greatest lower bound and least upper bound) Let $(S, \preceq)$ be a partially ordered set and let $A \subseteq S$.

(i) An element $x \in S$ is a ***lower bound*** for A if $x \preceq x'$ for every $x' \in A$.

(ii) An element $x \in S$ is an ***upper bound*** for A if $x' \preceq x$ for every $x' \in A$.

(iii) If, in the set of lower bounds for A, there is a greatest element, this is the ***greatest lower bound***, or the ***infimum***, of E. This is denoted by $\inf(A)$.

(iv) If, in the set of upper bounds for A, there is a least element, this is the ***least upper bound***, or the ***supremum***, of E. This is denoted by $\sup(A)$.

Now let $(S, \prec)$ be a strictly partially ordered set and let $A \subseteq S$.

(v) An element $x \in S$ is a ***strict lower bound*** for A if $x \prec x'$ for every $x' \in A$.

(vi) An element $x \in S$ is a ***strict upper bound*** for A if $x' \prec x$ for every $x' \in A$. •

Let us give some examples that illustrate the various possibilities arising from the preceding definitions. The examples will be given for lower bounds, but similar examples can be conjured to give similar conclusions for upper bounds.

1.5.12 Examples (Greatest lower bounds)

1. A subset $A \subseteq S$ may have no lower bounds. For example, the set of negative integers has no lower bound if we use the standard partial order in $\mathbb{Z}$.
2. A subset $A \subseteq S$ may have a greatest lower bound in A. For example, the set of nonnegative integers has as lower bounds all nonpositive integers. The greatest of these lower bounds is 0, which is itself a nonnegative integer.
3. A subset $A \subseteq S$ may have a greatest lower bound that is not an element of A. To see this, let S be the set of nonpositive integers, let A be the set of negative integers, and define a partial order $\preceq$ in S by

$$k_1 \preceq k_2 \qquad \iff \qquad \begin{cases} k_1 \le k_2,\ k_1, k_2 \in A, & \text{or} \\ k_1 = k_2 = 0, & \text{or} \\ k_1 = 0,\ k_2 \in A. \end{cases}$$

Thus this is the usual partial order in $A \subseteq S$, and one declares 0 to be less than all elements of A. In this case, 0 is the only lower bound for A, and so is, therefore, the greatest lower bound. But $0 \notin A$. •

1.5.3 Zorn's Lemma

Zorn's[6] Lemma comes up frequently in mathematics during the course of nonconstructive existence proofs. Since some of these proofs appear in this series and are important, we state Zorn's Lemma.

1.5.13 Theorem (Zorn's Lemma) *Every partially ordered set* $(S, \preceq)$ *in which every totally ordered subset has an upper bound contains at least one maximal member.*

Proof Suppose that every totally ordered subset has an upper bound, but that S has no maximal member. By assumption, if $A \subseteq S$ is a totally ordered subset, then there exists an upper bound x for A. Since S has no maximal element, there exists $x' \in S$ such that $x \prec x'$. Therefore, x' is a strict upper bound for A. Thus we have shown that every totally ordered subset possesses a strict upper bound. Let b be a function from the collection of totally ordered subsets into S having the property that $b(A)$ is a strict upper bound for A.[7]

A ***b-set*** is a subset B of S that is well ordered and has the property that, for every $x \in B$, we have $x = b(\mathrm{seg}_B(x))$, where $\mathrm{seg}_B(x)$ denotes the strict initial segment of x in B.

1 Lemma *If* B_1 *and* B_2 *are unequal* b*-sets, then one of the following statements holds:*

(i) there exists $x_1 \in B_1$ *such that* $B_2 = \mathrm{seg}_{B_1}(x_1)$*;*

(ii) there exists $x_2 \in B_2$ *such that* $B_1 = \mathrm{seg}_{B_2}(x_2)$.

[6] Max August Zorn (1906–1993) was a German mathematician who did work in the areas of set theory, algebra, and topology.

[7] The existence of the function b relies on the Axiom of Choice (see Section 1.8.3).

Proof If $B_2 \subset B_1$, then we claim that (i) holds. Take x_1 to be the least member of $B_1 - B_2$. We claim that $B_2 = \operatorname{seg}_{B_1}(x_1)$. First of all, if $x \in B_2$, then $x < x_1$ since x_1 is the least member of $B_1 - B_2$. Therefore, $B_2 \subseteq \operatorname{seg}_{B_1}(x_1)$. Now suppose that $\operatorname{seg}_{B_1}(x_1) - B_2 \neq \emptyset$, and let x be the least member of this set. Note that for any $x' \in B_2$ we therefore have $x' < x$, contradicting the fact that x_1 is the least member of $B_1 - B_2$. Thus we must have $\operatorname{seg}_{B_1}(x_1) - B_2 = \emptyset$, and so $B_2 = \operatorname{seg}_{B_1}(x_1)$.

We now suppose that $B_2 - B_1 \neq \emptyset$. Let x_2 be the least member of $B_2 - B_1$. If $x \in \operatorname{seg}_{B_2}(x_2)$ then $x < x_2$ and x must therefore be an element of B_1, or else this contradicts the definition of x_2. Now suppose that $B_1 \setminus \operatorname{seg}_{B_2}(x_2) \neq \emptyset$ and let y_1 be the least member of this set. If $y \in \operatorname{seg}_{B_1}(y_1)$ and $y' \in B_2$ satisfies $y' < y$, then $y' \in \operatorname{seg}_{B_1}(y_1)$. If z is the least member of $B_2 \setminus \operatorname{seg}_{B_1}(y_1)$, we then have $\operatorname{seg}_{B_2}(z) = \operatorname{seg}_{B_1}(y_1)$. Therefore

$$z = b(\operatorname{seg}_{B_2}(z)) = b(\operatorname{seg}_{B_1}(y_1)) = y_1.$$

Since $y_1 \in B_1$, $z = y_1 \neq x_2$. Since $z \leq x_2$, it follows that $z < x_2$. Thus $y_1 = z \in \operatorname{seg}_{B_2}(x_2)$. This, however, contradicts the choice of y_1, so we conclude that $B_1 \setminus \operatorname{seg}_{B_2}(x_2) = \emptyset$, and so that $B_1 = \operatorname{seg}_{B_2}(x_2)$. Thus (ii) holds.

A swapping of the rôles of B_1 and B_2 will complete the proof. ▼

2 Lemma *The union of all* b*-sets is a* b*-set.*

Proof Let U denote the union of all b-sets. First we must show that U is well ordered. Let $A \subseteq U$ and let $x \in A$. Then there is a b-set B such that $x \in B$. We claim that $\operatorname{seg}_A(x) \subseteq B$. Indeed, if $x' < x$ then, by Lemma 1, either $x' \in B$ or x' does not lie in any b-set. Since A lies in the union of all b-sets, it must be the case that $x' \in B$. Thus $\operatorname{seg}_A(x)$ is a subset of the well ordered set B, and as such has a least element x_0. This is clearly also a least element for A, so U is well ordered.

Next, let $x \in U$ and let B be a b-set such that $x \in B$. Our above argument shows that $\operatorname{seg}_U(x) \subseteq B$ so that $\operatorname{seg}_U(x) = \operatorname{seg}_B(x)$. Therefore, $x = b(\operatorname{seg}_B(x)) = b(\operatorname{seg}_U(x))$. This completes the proof. ▼

To complete the proof, let U be the union of all b-sets and let $x = b(U)$. Then we claim that $U \cup \{x\}$ is a b-set. That $U \cup \{x\}$ is well ordered follows since U is well ordered and since x is an upper bound for U. Since U is the union of all b-sets, it must hold that $x \in U$. However, this contradicts the fact that x is a strict upper bound for U. ■

1.5.4 Induction and recursion

In some of the proofs we have given in this section, and in our definition of $\mathbb{Z}_{\geq 0}$, we have used the idea of induction. This idea is an eminently reasonable one. One starts with a fact or a definition that applies to the element $0 \in \mathbb{Z}_{\geq 0}$, and a rule for extending this from the jth number to the $(j + 1)$st number, and then asserts that the fact or definition applies to all elements of $\mathbb{Z}_{\geq 0}$. In this section we formulate this principle in a more general setting that the set $\mathbb{Z}_{\geq 0}$, namely for a well ordered set.

Since the result will have to do with a property being true for the elements of a well ordered set, let us formally say that a ***property*** defined in a set S is a map $P\colon S \to \{\text{true}, \text{false}\}$. A property is ***true***, or ***holds***, at x if $P(x) = \text{true}$.

1.5.14 Theorem (Principle of Transfinite Induction) *Let* $(\mathrm{W}, \preceq)$ *be a well ordered set and let* P *be a property defined in* W*. Suppose that, for every* $\mathrm{w} \in \mathrm{W}$*, the fact that* $\mathrm{P}(\mathrm{w}')$ *is true for every* $\mathrm{w}' \prec \mathrm{w}$ *implies that* $\mathrm{P}(\mathrm{w})$ *is true. Then* $\mathrm{P}(\mathrm{w})$ *is true for every* $\mathrm{w} \in \mathrm{W}$*.*

Proof Suppose that the hypothesis is true, but the conclusion is false. Then

$$F = \{w \in W \mid P(w) = \text{false}\} \neq \emptyset.$$

Let w be the least element of F Therefore, for $w' < w$ it must hold that $P(w') = \text{true}$. But then the hypotheses imply that $P(w) = \text{true}$, so that $w \in W \setminus F$. This is a contradiction. ■

Next we turn to the process of defining something using recursion. As we did for induction, let us first consider doing this for $\mathbb{Z}_{\geq 0}$. What we wish to define is a map $f\colon \mathbb{Z}_{\geq 0} \to S$. The idea for doing this is that, if, for each $k \in \mathbb{Z}_{\geq 0}$, one knows the value of f on the first k elements of $\mathbb{Z}_{\geq 0}$, and if one knows a rule for then giving the value of f at $k+1$, then the f extends uniquely to a function on all of $\mathbb{Z}_{\geq 0}$. To give a concrete example, if $S = \mathbb{Z}$ and if we define $f(k+1) = 2 \cdot f(k)$, then the resulting function $f\colon \mathbb{Z}_{\geq 0} \to \mathbb{Z}$ is determined by its value at 0: $f(k) = 2^k \cdot f(0)$.

To state the general theorem requires some notation. We let W be a well ordered set and let S be a set. For $w \in W$, we let $\mathrm{seq}_S(w)$ be the set of maps from $\mathrm{seg}(w)$ into S. We then let $\mathrm{Seq}_S(W)$ be the set of all maps of the form $g\colon \mathrm{seq}_S(w) \to S$. The idea is that an element of $\mathrm{S}_S(W)$ tells us how to extend a map from $\mathrm{seg}(w)$ to give its value at w.

The desired result is now the following.

1.5.15 Theorem (Transfinite recursion) *Let* $(\mathrm{W}, \preceq)$ *be a well ordered set and let* S *be a set. Given a member* $\mathrm{g} \in \mathrm{Seq}_{\mathrm{S}}(\mathrm{W})$*, there exists a unique map* $\mathrm{f}_{\mathrm{g}}\colon \mathrm{W} \to \mathrm{S}$ *such that* $\mathrm{f}_{\mathrm{g}}(\mathrm{w}) = \mathrm{g}(\mathrm{f}|\,\mathrm{seg}(\mathrm{w}))$*.*

Proof That there can be only one map f_g as in the theorem statement follows from the Principle of Transfinite Induction (take $P(w) = \text{true}$ if and only if $f_g(w) = g(f_g|\,\mathrm{seg}(w))$).

So we shall prove the existence of f_g. Define

$$\begin{aligned}\mathscr{C}_g = \{A \subseteq W \times S|\\ &w \in W, h \in \mathrm{seq}_S(w), (w', h(w')) \in A \text{ for all } w' \in \mathrm{seg}(w) \implies (w, g(h)) \in A\}.\end{aligned}$$

Note that $W \times S \in \mathscr{C}_g$, so that C_g is not empty. It is easy to check that the intersection of members of $\mathscr{C}_g$ is also a member of $\mathscr{C}_g$. Therefore we let $F_g = \cap_{A \in \mathscr{C}_g} A$, and note that $F_g \in \mathscr{C}_g$. We shall show that F_g is the graph of a function f_g that satisfies the conditions in the theorem statement.

First we need to show that, for each $w \in W$, there exists exactly one $x \in S$ such that $(w, x) \in F_g$. Define

$$A_g = \{w \in W \mid \text{ there exists exactly one } x \in S \text{ such that } (w, x) \in F_g\}.$$

For $w \in W$, we claim that if $\mathrm{seg}(w) \subseteq A_g$, then $w \in A_g$. Indeed, if $\mathrm{seg}(w) \subseteq A_g$, define $h \in \mathrm{seq}_S(w)$ by $h(w') = x'$ where $x' \in S$ is the unique element such that $(w', x') \in A_g$. Since $F_g \in \mathscr{C}_g$, there exists some $x \in S$ such that $(w, x) \in F_g$. Suppose that $x \neq g(h)$. We claim that $F_g - \{(w,x)\} \in \mathscr{C}_g$. Let $w' \in W$ and let $h' \in \mathrm{seg}_S(w')$ satisfy $(w'', h'(w'')) \in F_g - \{(w,x)\}$ for all $w'' \in \mathrm{seg}(w')$. If $w' = w$ then $h' = h$ by the uniqueness assertion of the theorem, and therefore $(w', g(h')) \in F_g - \{(w,x)\}$ since $x \neq g(h) = g(h')$. On the other hand, if $w' \neq w$ then $(w', g(h')) \in F_g - \{(w,x)\}$ since $F_g \in \mathscr{C}_g$. Thus, indeed, $F_g - \{(w,x)\} \in \mathscr{C}_g$, contradicting the fact that F_g is the intersection of all sets in $\mathscr{C}_g$. Thus we can conclude that $x = g(h)$, and therefore that there is exactly one $x \in S$ such that $(w,x) \in F_g$. By the Principle of Transfinite Induction, we can then conclude that for *every* $w \in W$, there is exactly one $x \in S$ such that $(w,x) \in F_g$. Thus F_g is the graph of a map $f_g \colon W \to S$.

It remains to verify that $f_g(w) = g(f_g|\mathrm{seg}(w))$. This, however, follows easily from the definition of F_g. ■

One of the features of transfinite induction and transfinite recursion that requires some getting used to is that, unlike the usual induction with natural numbers as the well ordered set, one does not begin the induction or recursion by starting at 0 (or, in the case of a well ordered set, the least element), and proceeding element by element. Rather, one deals with initial segments. The reason for this is that in a well ordered set one may not have an immediate predecessor for every element, so that cannot be part of the induction/recursion; so the initial segment serves this purpose instead.

1.5.5 Zermelo's Well Ordering Theorem

The final topic in this section is a somewhat counterintuitive one. It says that every set possesses as well order.

1.5.16 Theorem (Zermelo's[8] Well Ordering Theorem) *For every set* S*, there is a well order in* S*.*

Proof Define

$$\mathscr{W} = \{(W, \preceq_W) \mid W \subseteq S \text{ and } \preceq_W \text{ is a well order on } W\}.$$

Since $\emptyset \in \mathscr{W}$, $\mathscr{W}$ is nonempty. Define a partial order $\preceq$ on $\mathscr{W}$ by

$$W_1 \preceq W_2 \quad \iff \quad W_2 \text{ is similar to a segment of } W_1.$$

Suppose that $\mathscr{T}$ is a totally ordered subset of $\mathscr{W}$.

1 Lemma *The set* $\cup_{\mathrm{A}\in\mathscr{T}}\mathrm{A}$ *has a unique well ordering, denoted by* $\precsim$*, such that* $\mathrm{A}' \precsim \cup_{\mathrm{A}\in\mathscr{T}}$ *for all* $\mathrm{A}' \in \mathscr{T}$*.*

Proof Let $x_1, x_2 \in \cup_{A\in\mathscr{T}}A$, and let $W_1, W_2 \in \mathscr{T}$ have the property that $x_1 \in W_1$ and $x_2 \in W_2$. Note that since either $W_1 = W_2$, $W_1 \preceq W_2$, or $W_2 \preceq W_1$, it must be the case that x_1 and x_2 lie in the same set from $\mathscr{C}$, let us call this W. The order in $\cup_{A\in\mathscr{T}}A$ is then defined by giving to the points x_1 and x_2 their order in W. This is unambiguous since $\mathscr{T}$ is totally ordered. It is then a simple exercise, left to the reader, that this is a well order. ▼

[8] Ernst Friedrich Ferdinand Zermelo (1871–1953) was a German mathematician whose mathematical contributions were mainly in the area of set theory.

The lemma ensures that the hypotheses of Zorn's Lemma apply to the totally ordered subsets of $\mathscr{W}$, and therefore the conclusions of Zorn's Lemma ensure that there is a maximal element W in $\mathscr{W}$. We claim that this maximal element is S. Suppose this is not the case, and that $x \in S - W$. We claim that $W \cup \{x\} \in \mathscr{W}$. To see this, simply define a well order on $W \cup \{x\}$ by asking that points in W have their usual order, and that x be greater that all points in W. The result is easily verified to be a well order on $W \cup \{x\}$, so contradiction the maximality of W. This completes the proof. ■

It might be surprising that it should be possible to well order any set. A well order can be thought of as allowing an arranging of the elements in a set, starting from the least element, and moving upwards in order:

$$x_0 < x_1 < x_2 < \cdots .$$

The complicated thing to understand here are the "$\cdots$," since they only mean "and so on" with an appropriate interpretation of these words (this is entirely related to the idea of ordinal numbers discussed in Section 1.7.1). As an example, the reader might want to imagine trying to order the real numbers (which we define in Section 2.1). It might seem absurd that it is possible to well order the real numbers. However, this is one of the many counterintuitive consequences arising from set theory, in this case directly related to the Axiom of Choice (Section 1.8.3).

1.5.6 Similarity

Between partially ordered sets, there are classes of maps that are distinguished by their preserving of the order relation. In this section we look into these and some of their properties, particularly with respect to well orders.

1.5.17 Definition (Similarity) If $(S, \preceq_S)$ and $(T, \preceq_T)$ are partially ordered sets, a bijection $f\colon S \to T$ is a ***similarity***, and $(S, \preceq_S)$ and $(T, \preceq_T)$ are said to be ***similar***, if $f(x_1) \preceq_T f(x_2)$ if and only if $x_1 \preceq_S x_2$. •

Now we prove a few results relating to similarities between well ordered sets. These shall be useful in our discussion or ordinal numbers in Section 1.7.1.

1.5.18 Proposition (Similarities of a well ordered set with itself) *If* $(\mathrm{S}, \preceq)$ *is a well ordered set and if* $\mathrm{f}\colon \mathrm{S} \to \mathrm{S}$ *is a similarity, then* $\mathrm{x} \preceq \mathrm{f(x)}$ *for each* $\mathrm{x} \in \mathrm{S}$.

Proof Define $A = \{x \in S \mid f(x) \prec x\}$ and let x be the least element of A. Then, for any $x' < x$, we have $x/ \preceq f(x')$. In particular, $f(x) \preceq f \circ f(x)$. But $f(x) \prec x$ implies that $f \circ f(x) < f(x)$, giving a contradiction. Thus $A = \emptyset$. ■

1.5.19 Proposition (Well ordered sets are similar in at most one way) *If* $\mathrm{f}, \mathrm{g}\colon \mathrm{S} \to \mathrm{T}$ *are similarities between well ordered sets* $(\mathrm{S}, \preceq_{\mathrm{S}})$ *and* $(\mathrm{T}, \preceq_{\mathrm{T}})$, *then* $\mathrm{f} = \mathrm{g}$.

Proof Let $h = f^{-1} \circ g$, and note that h is a similarity from S to itself. By Proposition 1.5.18 this implies that $x \preceq_S h(x)$ for each $x \in S$. Thus

$$\begin{aligned} & x \preceq_S f^{-1} \circ g(x), \qquad x \in S \\ \implies\ & f(x) \preceq_T g(x), \qquad x \in S. \end{aligned}$$

Reversing the argument gives $g(x) \preceq_T f(x)$ for every $x \in S$. This gives the result. ■

1.5.20 Proposition (Well ordered sets are not similar to their segments) *If* $(\mathrm{S}, \prec)$ *is a well ordered set and if* $\mathrm{x} \in \mathrm{S}$*, then* S *is not similar to* $\mathrm{seg}(\mathrm{x})$*.*

Proof If $f(x) \in \mathrm{seg}(x)$ then $f(x) < x$, contradiction Proposition 1.5.18. ■

The final result is the deepest of the results we give here, because it gives a rather simple structure to the collection of all well ordered sets.

1.5.21 Proposition (Comparing well ordered sets) *If* $(\mathrm{S}, \preceq_\mathrm{S})$ *and* $(\mathrm{T}, \preceq_\mathrm{T})$ *are well ordered sets, then one of the following statements holds:*

(i) S *and* T *are similar;*

(ii) there exists $\mathrm{x} \in \mathrm{S}$ *such that* $\mathrm{seg}(\mathrm{x})$ *and* T *are similar;*

(iii) there exists $\mathrm{y} \in \mathrm{T}$ *such that* $\mathrm{seg}(\mathrm{y})$ *and* S *are similar.*

Proof Define

$$S_0 = \{x \in S \mid \text{ there exists } y \in T \text{ such that } \mathrm{seg}(x) \text{ is similar to } \mathrm{seg}(y)\},$$

noting that S_0 is nonempty, since the segment of the least element in S is similar to the segment of the least element in T. Define $f: S_0 \to T$ by $f(x) = y$ where $\mathrm{seg}(x)$ is similar to $\mathrm{seg}(y)$. Note that this uniquely defines f by Propositions 1.5.19 and 1.5.20. We then take $T_0 = \mathrm{image}(f)$. If $S_0 = S$, then the result immediately follows. If $S_0 \subset S$, then we claim that $S_0 = \mathrm{seg}(x_0)$ for some $x_0 \in S$. Indeed, we simply take x_0 to be the least strict upper bound for S_0, and then apply the definition of S_0 to see that $S_0 = \mathrm{seg}(x_0)$. We next claim that $T_0 = T$. Indeed, suppose that $T_0 \subset T$, let y_0 be the least strict upper bound for T_0, and let x_0 be the least strict upper bound for S_0. We claim that $\mathrm{seg}(x_0)$ is similar to $\mathrm{seg}(y_0)$. Indeed, if this is not the case, then there exists $y < y_0$ such that $\mathrm{seg}(y)$ is not similar to a segment in S. However, this contradicts the definition of T_0. ■

1.5.7 Notes

The proof of Zorn's Lemma we give is from the paper of [Lewin 1991].

Exercises

1.5.1 Show that any set S possesses a partial order.

1.5.2 Give conditions on S under which the partial order $\subseteq$ on $\mathbf{2}^S$ is

(a) a total order or

(b) a well-order.

1.5.3 Given two partially ordered sets $(S, \preceq_S)$ and $(T, \preceq_T)$, we define a relation $\preceq_{S\times T}$ in $S \times T$ by

$$(x_1, y_1) \preceq_{S\times T} (x_2, y_2) \qquad \Longleftrightarrow \qquad (x_1 \prec_S x_2) \text{ or } (x_1 = x_2 \text{ and } y_1 \preceq_T y_2).$$

This is called the ***lexicographic order*** on $S \times T$. Show the following:

(a) the lexicographic order is a partial order;

(b) if $\preceq_S$ and $\preceq_T$ are total orders, then the lexicographic order is a total order.

1.5.4 Show that a partially ordered set $(S, \preceq)$ possesses at most one least element and/or at most one greatest element.

Section 1.6

Indexed families of sets and general Cartesian products

In this section we discuss general collections of sets, and general collections of members of sets. In Section 1.1.3 we considered Cartesian products of a finite collection of sets. In this section, we wish to extend this to allow for an arbitrary collection of sets. The often used idea of an index set is introduced here, and will come up on many occasions in the text.

Do I need to read this section? The idea of a general family of sets, and notions related to it, do not arise in a lot of places in these volumes. But they do arise. The ideas here are simple (although the notational nuances can be confusing), and so perhaps can be read through. But the reader in a rush can skip the material, knowing they can look back on it if necessary. •

1.6.1 Indexed families and multisets

Recall that when talking about sets, a set is determined only by the concept of membership. Therefore, for example, the sets $\{1,2,2,1,2\}$ and $\{1,2\}$ are the same since they have the same members. However, what if one wants to consider a set with two 1's and three 2's? The way in which one does this is by the use of an index to label the members of the set.

1.6.1 Definition (Indexed family of elements) Let A and S be sets. An ***indexed family of elements*** of S with ***index set*** A is a map $f\colon A\to S$. The element $f(a)\in S$ is sometimes denoted as x_a and the indexed family is denoted as $(x_a)_{a\in A}$. •

With the notion of an indexed family we can make sense of "repeated entries" in a set, as is shown in the first of these examples.

1.6.2 Examples (Indexed family)

1. Consider the two index sets $A_1=\{1,2,3,4,5\}$ and $A_2=\{1,2\}$ and let S be the set of natural numbers. Then the functions $f_1\colon A_1\to S$ and $f_2\colon A_2\to S$ defined by

$$f_1(1)=1,\ f_1(2)=2,\ f_1(3)=2,\ f_1(4)=1,\ f_1(5)=2,$$
$$f_2(1)=1,\ f_2(2)=2,$$

give the indexed families $(x_1=1,x_2=2,x_3=2,x_4=1,x_5=2)$ and $(x_1=1,x_2=2)$, respectively. In this way we can arrive at a set with two 1's and three 2's, as desired. Moreover, each of the 1's and 2's is assigned a specific place in the list $(x_1,\dots,x_5)$.

2. Any set S gives rise in a natural way to an indexed family of elements of S indexed by S itself: $(x)_{x\in S}$. •

We can then generalise this notion to an indexed family of sets as follows.

1.6.3 Definition (Indexed family of sets) Let A and S be sets. An ***indexed family of subsets*** of S with ***index set*** A is an indexed family of elements of $\mathbf{2}^S$ with index set A. Thus an indexed family of subsets of S is denoted by $(S_a)_{a\in A}$ where $S_a \subseteq S$ for $a \in A$. •

We use the notation $\cup_{a\in A}S_a$ and $\cap_{a\in A}S_a$ to denote the union and intersection of an indexed family of subsets indexed by A. Similarly, when considering the disjoint union of an indexed family of subsets indexed by A, we define this to be

$$\overset{\circ}{\bigcup_{a\in A}} S_a = \cup_{a\in A}(\{a\} \times S_a).$$

Thus an element in the disjoint union has the form (a, x) where $x \in S_a$. Just as with the disjoint union of a pair of sets, the disjoint union of a family of sets keeps track of the set that element belongs to, now labelled by the index set A, along with the element. A family of sets $(S_a)_{a\in A}$ is ***pairwise disjoint*** if, for every distinct $a_1, a_2 \in A$, $S_{a_1} \cap S_{a_2} = \varnothing$.

Often when one writes $(S_a)_{a\in A}$, one omits saying that the family is "indexed by A," this being understood from the notation. Moreover, many authors will say things like, "Consider the family of sets $\{S_a\}$," so omitting any reference to the index set. In such cases, the index set is usually understood (often it is $\mathbb{Z}_{>0}$). However, we shall not use this notation, and will always give a symbol for the index set.

Sometimes we will simply say something like, "Consider a family of sets $(S_a)_{a\in A}$." When we say this, we tacitly suppose there to be a set S which contains each of the sets S_a as a subset; the union of the sets S_a will serve to give such a set.

There is an alternative way of achieving the objective of allowing sets where the same member appears multiple times.

1.6.4 Definition (Multiset, submultiset) A ***multiset*** is an ordered pair (S, ϕ) where S is a set and $\phi\colon S \to \mathbb{Z}_{\geq 0}$ is a map. A multiset (T, ψ) is a ***submultiset*** of (S, ϕ) if $T \subseteq S$ and if $\psi(x) \leq \phi(x)$ for every $x \in T$. •

This is best illustrated by examples.

1.6.5 Examples (Multisets)

1. The multiset alluded to at the beginning of this section is (S, ϕ) with $S = \{1, 2\}$, and $\phi(1) = 2$ and $\phi(2) = 3$. Note that some information is lost when considering the multiset (S, ϕ) as compared to the indexed family $(1, 2, 2, 1, 2)$; the order of the elements is now immaterial and only the number of occurrences is accounted for.
2. Any set S can be thought of as a multiset (S, ϕ) where $\phi(x) = 1$ for each $x \in S$.
3. Let us give an example of how one might use the notion of a multiset. Let $P \subseteq \mathbb{Z}_{>0}$ be the set of prime numbers and let S be the set $\{2, 3, 4, \dots\}$ of integers greater than 1. As we shall prove in Corollary 4.2.73, every element $n \in S$ can

be written in a unique way as $n = p_1^{k_1} \cdots p_m^{k_m}$ for distinct primes $p_1, \dots, p_m$ and for $k_1, \dots, k_m \in \mathbb{Z}_{>0}$. Therefore, for every $n \in S$ there exists a unique multiset (P, ϕ_n) defined by

$$\phi_n(p) = \begin{cases} k_j, & p = p_j, \\ 0, & \text{otherwise,} \end{cases}$$

understanding that $k_1, \dots, k_m$ and $p_1, \dots, p_m$ satisfy $n = p_1^{k_1} \cdots p_m^{k_m}$. •

1.6.6 Notation (Sets and multisets from indexed families of elements) Let A and S be sets and let $(x_a)_{a\in A}$ be an indexed family of elements of S. If for each $x \in S$ the set $\{a \in A \mid x_a = x\}$ is finite, then one can associate to $(x_a)_{a\in A}$ a multiset (S, ϕ) by

$$\phi(x) = \operatorname{card}\{a \in A \mid x_a = x\}.$$

This multiset is denoted by $\{x_a\}_{a\in A}$. One also has a subset of S associated with the family $(x_a)_{a\in A}$. This is simply the set

$$\{x \in S \mid x = x_a \text{ for some } a \in A\}.$$

This set is denoted by $\{x_a \mid a \in A\}$. Thus we have three potentially quite different objects:

$$(x_a)_{a\in A}, \quad \{x_a\}_{a\in A}, \quad \{x_a \mid a \in A\},$$

arranged in decreasing order of information prescribed (be sure to note that the multiset in the middle is only defined when the sets $\{a \in A \mid x_a = x\}$ are finite). This is possibly confusing, although there is not much in it, really.

For example, the indexed family $(1, 2, 2, 1, 2)$ gives the multiset denoted $\{1, 1, 2, 2, 2\}$ and the set $\{1, 2\}$. Now, this is truly confusing since there is no notational discrimination between the *set* $\{1, 1, 2, 2, 2\}$ (which is simply the set $\{1, 2\}$) and the *multiset* $\{1, 1, 2, 2, 2\}$ (which is not the set $\{1, 2\}$). However, the notation is standard, and the hopefully the intention will be clear from context.

If the map $a \mapsto x_a$ is injective, i.e., the elements in the family $(x_a)_{a\in A}$ are distinct, then the three objects are in natural correspondence with one another. For this reason we can sometimes be a bit lax in using one piece of notation over another. •

1.6.2 General Cartesian products

Before giving general definitions, it pays to revisit the idea of the Cartesian product $S_1 \times S_2$ of sets S_1 and S_2 as defined in Section 1.1.3 (the reason for our change from S and T to S_1 and S_2 will become clear shortly). Let $A = \{1, 2\}$, and let $f \colon A \to S_1 \cup S_2$ be a map satisfying $f(1) \in S_1$ and $f(2) \in S_2$. Then $(f(1), f(2)) \in S_1 \times S_2$. Conversely, given a point $(x_1, x_2) \in S_1 \times S_2$, we define a map $f \colon A \to S_1 \cup S_2$ by $f(1) = x_1$ and $f(2) = x_2$, noting that $f(1) \in S_1$ and $f(2) \in S_2$.

The punchline is that, for a pair of sets S_1 and S_2, their Cartesian product is in 1–1 correspondence with maps f from $A = \{1, 2\}$ to $S_1 \cup S_1$ having the property that

$f(x_1) \in S_1$ and $f(x_2) \in S_2$. There are two things to note here: (1) the use of the set A to label the sets S_1 and S_2 and (2) the alternative characterisation of the Cartesian product.

Now we generalise the Cartesian product to families of sets.

1.6.7 Definition (Cartesian product) The ***Cartesian product*** of a family of sets $(S_a)_{a\in A}$ is the set

$$\prod_{a\in A} S_a = \{f\colon A \to \cup_{a\in A} S_a \mid f(a) \in S_a\}.$$

For $b \in A$, the **b*th projection*** for the Cartesian product $\prod_{a\in A} S_a$ is the map $\mathrm{pr}_b\colon \prod_{a\in A} S_a \to S_b$ defined by $\mathrm{pr}_b(f) = f(b)$. •

Note that the analogue to the ordered pair in a general Cartesian product is simply the set $f(A)$ for some $f \in \prod_{a\in A} S_a$. The reader should convince themselves that this is indeed the appropriate generalisation.

1.6.3 Sequences

The notion of a sequence is very important for us, and we give here a general definition for sequences in arbitrary sets.

1.6.8 Definition (Sequence, subsequence) Let S be a set.

(i) A ***sequence*** in S is an indexed family $(x_j)_{j\in\mathbb{Z}_{>0}}$ of elements of S with index set $\mathbb{Z}_{>0}$.

(ii) A ***subsequence*** of a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in S is a map $f\colon A \to S$ where

(a) $A \subseteq \mathbb{Z}_{>0}$ is a nonempty set with no upper bound and

(b) $f(k) = x_k$ for all $k \in A$.

If the elements in the set A are ordered as $j_1 < j_2 < j_3 < \cdots$, then the subsequence may be written as $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$. •

Note that in a sequence the location of the elements is important, and so the notation $(x_j)_{j\in\mathbb{Z}_{>0}}$ is the correct choice. It is, however, not uncommon to see sequences denoted $\{x_j\}_{j\in\mathbb{Z}_{>0}}$. According to Notation 1.6.6 this would imply that the same element in S could only appear in the list $(x_j)_{j\in\mathbb{Z}_{>0}}$ a finite number of times. However, this is often not what is intended. However, there is seldom any real confusion induced by this, but the reader should simply be aware that our (not uncommon) notational pedantry is not universally followed.

1.6.4 Directed sets and nets

What we discuss in this section is a generalisation of the notion of a sequence. A sequence is a collection of objects where there is a natural order to the objects inherited from the total order of $\mathbb{Z}_{>0}$.

First we define the index sets for this more general type of sequence.

1.6.9 Definition (Directed set) A ***directed set*** is a partially ordered set $(D, \preceq)$ with the property that, for $x, y \in D$, there exists $z \in D$ such that $x \preceq z$ and $y \preceq z$. •

Thus for any two elements in a directed set D it is possible to find an element greater than either, relative to the specified partial order. Let us give some examples to clarify this.

1.6.10 Examples (Directed sets)

1. The set $(\mathbb{Z}_{>0}, \leq)$ is a directed set since clearly one can find a natural number exceeding any two specified natural numbers.
2. The partially ordered set $([0, \infty), \leq)$ is similarly a directed set.
3. The partially ordered set $((0, 1], \geq)$ is also a directed set since, given $x, y \in (0, 1]$, one can find an element of $(0, 1]$ which is smaller than either x or y.
4. Next take $D = \mathbb{R} \setminus \{x_0\}$ and consider the partial order $\preceq$ on D defined by $x \preceq y$ if $|x - x_0| \leq |y - y_0|$. This may be shown to be a directed set since, given two elements $x, y \in \mathbb{R} \setminus \{x_0\}$, one can find another element of $\mathbb{R} \setminus \{x_0\}$ which is closer to x_0 than either x or y.
5. Let S be a set with more than one element and consider the partially ordered set $(\mathbf{2}^S \setminus \{\emptyset\}, \preceq)$ specified by $A \preceq B$ if $A \supseteq B$. This is readily verified to be a partial order. However, this order does not make $(S, \supseteq)$ a directed set. Indeed, suppose that $A, B \in \mathbf{2}^S \setminus \{\emptyset\}$ are disjoint. Since the only set contained in both A and B is the empty set, it follows that there is no element $T \in \mathbf{2}^S \setminus \{\emptyset\}$ for which $A \supseteq T$ and $B \supseteq T$. •

The next definition is of the generalisation of sequences built on the more general notion of index set given by a directed set.

1.6.11 Definition (Net) Let $(D, \preceq)$ be a directed set. A ***net*** in a set S defined on D is a map $\phi\colon D \to S$ from D into S. •

As with a sequence, it is convenient to instead write $\{x_\alpha\}_{\alpha \in D}$ where $x_\alpha = \phi(\alpha)$ for a net. The idea here is that a net generalises the notion of a sequence to the case where the index set may not be countable and where the order is more general than the total order of $\mathbb{Z}$.

Exercises

1.6.1

Section 1.7

Ordinal numbers, cardinal numbers, cardinality

The notion of cardinality has to do with the "size" of a set. For sets with finite numbers of elements, there is no problem with "size." For example, it is clear what it means for one set with a finite number of elements to be "larger" or "smaller" than another set with a finite number of elements. However, for sets with infinite numbers of elements, can one be larger than another? If so, how can this be decided? In this section we see that there is a set, called the ***cardinal numbers***, which exactly characterises the "size" of all sets, just as natural numbers characterise the "size" if finite sets.

Do I need to read this section? The material in this section is used only slightly, so it can be thought of as "cultural," and hopefully interesting. Certainly the details of constructing the ordinal numbers, and then the cardinal numbers, plays no essential rôle in these volumes. The idea of cardinality comes up, but only in the simple sense of Theorem 1.7.12. •

1.7.1 Ordinal numbers

Ordinal numbers generalise the natural numbers. Recall from Section 1.4.1 that a natural number is a set, and moreover, from Section 1.4.2, a well ordered set. Indeed, the number $k \in \mathbb{Z}_{\geq 0}$ is, by definition,

$$k = \{0, 1, \dots, k-1\}.$$

Moreover, note that, for every $j \in k$, $j = \mathrm{seg}(j)$. This motivates our definition of the ordinal numbers.

1.7.1 Definition (Ordinal number) An ***ordinal number*** is a well ordered set $(o, \leq)$ with the property that, for each $x \in o$, $x = \mathrm{seg}(x)$. •

Let us give some examples of ordinal numbers. The examples we give are all of "small" ordinals. We begin our constructions in a fairly detailed way, and then we omit the details as we move on, since the idea becomes clear after the initial constructions.

1.7.2 Examples (Ordinal numbers)

1. As we saw before we stated Definition 1.7.1, each nonnegative integer is an ordinal number.
2. The set $\mathbb{Z}_{\geq 0}$ is an ordinal number. This is easily verified, but discomforting. We are saying that the set of numbers is itself a new kind of number, an ordinal number. Let us call this ordinal number ω. Pressing on. . .

3. The successor $\mathbb{Z}_{\geq 0}^{+} = \mathbb{Z}_{\geq 0} \cup \{\mathbb{Z}_{\geq 0}\}$ is also an ordinal number, in just the same manner as a natural number is an ordinal number. This ordinal number is denoted by $\omega + 1$.
4. One carries on in this way defining ordinal numbers $\omega + (k+1) = (\omega + k)^{+}$.
5. Next we assume that there is a set containing ω and all of its successors. In axiomatic set theory, this follows from a construction like that justifying Assumption 1.4.3, along with another axiom (the Axiom of Substitution; see Section 1.8.2) saying, essentially, that we can repeat the process. Just as we did with the definition of $\mathbb{Z}_{\geq 0}$, we take the smallest of these sets of successors to arrive at a net set that is to ω as ω is to 0. As was $\omega = \mathbb{Z}_{\geq 0}$, we well order this set by the partial order $\subseteq$. This set is then clearly an ordinal number, and is denoted by $\omega 2$.
6. One now proceeds to construct the successors $\omega 2 + 1 = \omega 2^{+}$, $\omega 2 + 2 = (\omega 2 + 1)^{+}$, and so on. These new sets are also ordinal numbers.
7. The preceding process yields ordinal numbers $\omega, \omega 2, \omega 3$, and so on.
8. We now again apply the same procedure to define an ordinal number that is contains ω, $\omega 2$, etc. This set we denote by ω^2.
9. One then defines $\omega^2 + 1 = (\omega^2)^{+}$, $\omega^2 + 2 = (\omega^2 + 1)^{+}$, etc., noting that these two are all ordinal numbers.
10. Next comes $\omega^2 + \omega$, which is the set containing all ordinal numbers $\omega^2 + 1$, $\omega^2 + 2$, etc.
11. Then comes $\omega^2 + \omega + 1$, $\omega^2 + \omega + 2$, etc.
12. Following these is $\omega^2 + \omega 2$, $\omega^2 + \omega 2 + 1$, and so on.
13. Then comes $\omega^2 + \omega 3$, $\omega^2 + \omega 3 + 1$, and so on.
14. After ω^2, $\omega^2 + \omega$, $\omega^2 + \omega 2$, and so on, we arrive at $\omega^2 2$.
15. One then arrives at $\omega^2 2 + 1, \dots, \omega^2 2 + \omega, \dots, \omega^2 2 + \omega 2$, etc.
16. After $\omega^2 2$, $\omega^2 3$, and so on comes ω^3.
17. After ω, ω^2, ω^3, etc., comes ω^{ω}.
18. After ω, ω^{ω}, $\omega^{\omega^{\omega}}$, etc., comes ϵ_0. The entire construction starts again from ϵ_0. Thus we get to $\epsilon_0 + 1$, $\epsilon_0 + 2$, and so on reproducing all of the above steps with an ϵ_0 in front of everything.
19. Then we get $\epsilon_0 2$, $\epsilon_0 3$, and so on up to $\epsilon_0 \omega$.
20. These are followed by $\epsilon_0 \omega^2$, $\epsilon_0 \omega^3$ and so on up to $\epsilon_0 \omega^{\omega}$.
21. Then comes $\epsilon_0 \omega^{\omega^{\omega}}$, etc.
22. These are followed by ϵ_0^2.
23. We hope the reader is getting the point of these constructions, and can produce more such ordinals derived from the natural numbers. •

The above constructions of examples of ordinal numbers suggests that there are a lot of them. However, the concrete constructions do not really do justice to the number of ordinals. The ordinals that are elements of $\mathbb{Z}_{\geq 0}$ are called ***finite*** ordinals, and all other ordinals are ***transfinite***. All of the ordinals we have named above are called "enumerable" (see Definition 1.7.13). There are many other ordinals not included in the above list, but before we can appreciate this, we first have to describe some properties of ordinals.

First we note that ordinals are exactly defined by similarity. More precisely, we have the following result.

1.7.3 Proposition (Similar ordinals are equal) *If* o_1 *and* o_2 *are similar ordinal numbers then* $\mathrm{o}_1 = \mathrm{o}_2$.

Proof Let $f\colon o_1 \to o_2$ be a similarity and define

$$S = \{x \in o_1 \mid f(x) = x\}.$$

We wish to show that $S = o_1$. Suppose that $\mathrm{seg}(x) \subseteq S$ for $x \in o_1$. Then x is the least element of $\mathrm{seg}(x)$ and, since f is a similarity, $f(x)$ is the least element of $f(\mathrm{seg}(x))$. Therefore, x and $f(x)$ both have $\mathrm{seg}(x)$ as their strict initial segment, by definition of S. Thus, by the definition of ordinal numbers, $x = f(x)$. The result now follows by the Principle of Transfinite Induction. ■

The next result gives a rather rigid structure to any set of ordinal numbers.

1.7.4 Proposition (Sets of ordinals are always well ordered) *If* O *is a set of ordinal numbers, then this set is well ordered by* $\subseteq$*.*

Proof First we claim that O is totally ordered. Let $o_1, o_2 \in O$ and note that these are both well ordered sets. Therefore, by Proposition 1.5.21, either $o_1 = o_2$, o_1 is similar to a strict initial segment in o_2, or o_2 is similar to a strict initial segment in o_1. In either of the last two cases, it follows from Proposition 1.7.3 that either o_1 is *equal* to a strict initial segment in o_2, or vice versa. Thus, either $o_1 \leq o_2$ or $o_2 \leq o_1$. Thus O is totally ordered, a fact we shall assume in the remainder of the proof.

Let $o \in O$. If $o \leq o'$ for every $o' \in O$, then o is the least member of O, and so O has a least member, namely o. If o is not the least member of O, then there exists $o' \in O$ such that $o' < o$. Thus $o' \in o$ and so the set $o \cap E$ is nonempty. Let o_0 be the least element of o. We claim that o_0 is also the least element of O. Indeed, let $o' \in O$. If $o' < o$ then $o' \in o \cap E$ and so $o_0 \leq o'$. If $o \leq o'$ then $o_0 < o'$, so showing that o_0 is indeed the least element of O. ■

Our constructions in Example 1.7.2, and indeed the definition of an ordinal number, suggest the true fact that every ordinal number has a successor that is an ordinal number. However, it may not be the case that an ordinal number has an immediate predecessor. For example, each of the ordinals that are natural numbers has an immediate predecessor, but the ordinal ω does not have an immediate predecessor. That is to say, there is no largest ordinal number strictly less ω.

Recall that the set $\mathbb{Z}_{\geq 0}$ was defined by being the smallest set, having a certain property, that contains all nonnegative integers. One can then ask, "Is there a set containing all ordinal numbers?" It turns out the definition of the ordinal numbers prohibits this.

1.7.5 Proposition (Burali-Forti[9] Paradox) *There is no set $\mathbb{O}$ having the property that, if o is an ordinal number, then $\mathrm{o} \in \mathbb{O}$.*

Proof Suppose that such a set $\mathbb{O}$ exists. We claim that $\sup \mathbb{O}$ exists and is an ordinal number. Indeed, we claim that $\sup \mathbb{O} = \cup_{o \in \mathbb{O}} o$. Note that the set $\cup_{o \in \mathbb{O}} o$ is well ordered by inclusion by Proposition 1.7.4. Clearly, $\cup_{o \in \mathbb{O}}$ is the smallest such set containing each $o \in \mathbb{O}$. Moreover, it is also clear from Proposition 1.7.4 that if $o' \in \cup_{o \in \mathbb{O}}$, then $o' = \operatorname{seg}(o')$. Thus $\sup \mathbb{O}$ exists, and is an ordinal number. Moreover, this order number is greater than all those in $\mathbb{O}$, thus showing that $\mathbb{O}$ cannot exist. ■

For our purposes, the most useful feature of the ordinal numbers is the following.

1.7.6 Theorem (Ordinal numbers can count the size of a set) *If $(\mathrm{S}, \preceq)$ is a well ordered set, then there exists a unique ordinal number o_{S} with the property that S and o_{S} are similar.*

Proof The uniqueness follows from Proposition 1.7.3. Let $x_0 \in S$ have the property that if $x < x_0$ then $\operatorname{seg}(x)$ is similar to some (necessarily unique) ordinal. (Why does x_0 exist?) Now let $P(x, o)$ be the proposition "o is an ordinal number similar to $\operatorname{seg}(x)$". Then define the set of ordinal numbers

$$o_0 = \{o \mid \text{ for each } x \in \operatorname{seg}(x_0), \text{ there exists } o \text{ such that } P(x, o) \text{ holds}\}.$$

One can easily verify that o_0 is itself an ordinal number that is similar to $\operatorname{seg}(x_0)$. Therefore, the Principle of Transfinite Induction can be applied to show that S is similar to an ordinal number. ■

This theorem is important, because it tells us that the ordinal numbers are the same, essentially, as the well ordered sets. Thus one can use the two concepts interchangeably; this is not obvious from the definition of an ordinal number.

It is also possible to define addition and multiplication of ordinal numbers. Since we will not make use of this, let us merely sketch how this goes. For ordinal numbers o_1 and o_2, let $(S_1, \preceq_1)$ and $(S_2, \preceq_2)$ be well ordered sets similar to o_1 and o_2, respectively. Define a partial order in $S_1 \overset{\circ}{\cup} S_2$ by

$$(i_1, x_1) \preceq_+ (i_2, x_2) \quad \iff \quad \begin{cases} i_1 = i_2,\ x_1 \preceq_{i_1}, & \text{or} \\ i_1 < i_2. \end{cases}$$

One may verify that this is a well order. Then define $o_1 + o_2$ as the unique ordinal number equivalent to the well ordered set $(S_1 \overset{\circ}{\cup} S_2, \preceq_+)$. To define product of o_1 and o_2, on the Cartesian product $S_1 \times S_2$ consider the partial order

$$(x_1, x_2) \preceq_\times (y_1, y_2) \quad \iff \begin{cases} x_2 \prec_2 y_2, & \text{or} \\ x_2 = y_2,\ x_1 \prec_1 y_1. \end{cases}$$

[9] Cesare Burali-Forti (1861–1931) was an Italian mathematician who made contributions to mathematical logic.

Again, this is verifiable as being a well order. One then defines $o_1 \cdot o_2$ to be the unique ordinal number similar to the well ordered set $(S_1 \times S_2, \preceq_\times)$. One must exercise care when dealing with addition and multiplication of ordinals, since, for example, neither addition nor multiplication are commutative. For example, $1 + \omega \neq \omega + 1$ (why?). However, since we do not make use of this arithmetic, we shall not explore this further. It is worth noting that the notation in Example 1.7.2 is derived from ordinal arithmetic. Thus, for example, $\omega 2 = \omega \cdot 2$, etc.

1.7.2 Cardinal numbers

The cardinal numbers, as mentioned at the beginning of this section, are intended to be measures of the size of a set. If one combines the Zermelo's Well Ordering Theorem (Theorem 1.5.16) and Theorem 1.7.6, one might be inclined to say that the ordinal numbers are suited to this task. Indeed, simply place a well order on the set of interest by Theorem 1.5.16, and then use the associated ordinal number, given by Theorem 1.7.6, to define "size." The problem with this construction is that this notion of the "size" of a set would depend on the choice of well ordering. As an example, let us take the set $\mathbb{Z}_{\geq 0}$. We place two well orderings on $\mathbb{Z}_{\geq 0}$, one being the natural well ordering $\leq$ and the other being defined by

$$k_1 \preceq k_2 \qquad \Longleftrightarrow \qquad \begin{cases} k_1 \leq k_2,\ k_1, k_2 \in \mathbb{Z}_{>0}, & \text{or} \\ k_1 = k_2 = 0, & \text{or} \\ k_1 = 0,\ k_2 \in \mathbb{Z}_{>0}. \end{cases}$$

Thus, for the partial order $\preceq$, one places 0 after all other natural numbers. One then verifies that $(\mathbb{Z}_{\geq 0}, \leq)$ is similar to the ordinal number ω and that $(\mathbb{Z}_{\geq 0}, \preceq)$ is similar to the ordinal number $\omega + 1$. Thus, even in a fairly simple example of a non-finite set, we see that the well order can change the size, if we go with size being determined by ordinals.

Therefore, we introduce a special subset of ordinals.

1.7.7 Definition (Cardinal number) A ***cardinal number*** is an ordinal number c with the property that, for all ordinal numbers o for which there exists a bijection from c to o, we have $c \leq o$. •

In other words, a cardinal number is the least ordinal number in a collection of ordinal numbers that are equivalent. Note that finite ordinals are only equivalent with a single ordinal, namely themselves. However, transfinite ordinals may be equivalent to different transfinite ordinals. The following example illustrates this.

1.7.8 Example (Equivalent transfinite ordinals) We claim that there is a 1–1 correspondence between ω and $\omega + 1$. We can establish this correspondence explicitly by defining a map $f \colon \omega \to \omega + 1$ by

$$f(x) = \begin{cases} \omega, & x = 0, \\ x - 1, & x \in \mathbb{Z}_{>0}, \end{cases}$$

where $x - 1$ denotes the immediate predecessor of $x \in \mathbb{Z}_{>0}$.

One can actually check that *all* of the ordinal numbers presented in Example 1.7.2 are equivalent to ω! This is a consequence of Proposition 1.7.16 below. Accepting this as fact for the moment, we see that the only ordinals from Example 1.7.2 that are cardinal numbers are the elements of $\mathbb{Z}_{\geq 0}$ along with ω. •

Certain of the facts about ordinal numbers translate directly to equivalent facts about cardinal numbers. Let us record these

1.7.9 Proposition (Properties of cardinal numbers) *The following statements hold:*

(i) if c_1 *and* c_2 *are similar cardinal numbers then* $\mathrm{c}_1 = \mathrm{c}_2$*;*

(ii) if C *is a set of cardinal numbers, then this set is well ordered by* $\subseteq$*;*

(iii) there is no set $\overline{\mathrm{C}}$ *having the property that, if* c *is a cardinal number, then* $\mathrm{c} \in \overline{\mathrm{C}}$ *(**Cantor's paradox**).*[10]

Proof The only thing that does not follow immediately from the corresponding results for ordinal numbers is Cantor's Paradox. The proof of this part of the result goes exactly as does that of Proposition 1.7.5. One only needs to verify that, if $\mathbb{C}$ is any set of cardinal numbers, then there exists a cardinal number greater or equal to $\operatorname{supp}\mathbb{C}$. This, however, is clear since $\operatorname{supp}\mathbb{C}$ is an ordinal number strictly greater than any element of $\mathbb{C}$, meaning that there is a corresponding cardinal number c equivalent to $\operatorname{supp}\mathbb{C}$. Thus $c \geq \operatorname{supp}\mathbb{C}$. ■

1.7.3 Cardinality

Cardinality is the measure of the "size" of a set that we have been after. The following result sets the stage for the definition.

1.7.10 Lemma *For a set* S *there exists a unique cardinal number* $\mathrm{card}(\mathrm{S})$ *such that* S *and* $\mathrm{card}(\mathrm{S})$ *are equivalent.*

Proof By Theorem 1.7.6 there exists an ordinal number o_S that is similar to S, and therefore equivalent to S. Any ordinal equivalent to o_S is therefore also equivalent to S, since equivalence of sets is an "equivalence relation" (Exercise 1.3.8). Therefore, the result follows by choosing the unique least element in the set of ordinals equivalent to o_S. ■

With this fact at hand, the following definition makes sense.

1.7.11 Definition (Cardinality) The ***cardinality*** of a set S is the unique cardinal number $\mathrm{card}(S)$ that is equivalent to S. •

The next result indicates how one often deals with cardinality in practice. The important thing to note is that, provided one is interested only in *comparing* cardinalities of sets, then one need not deal with the complication of cardinal numbers.

[10] Georg Ferdinand Ludwig Philipp Cantor (1845–1918) was born in Denmark, grew up in St. Petersburg, and lived much of his mathematical life in Germany. He made many important contributions to set theory and logic. He is regarded as the founder of set theory as we now know it.

1.7.12 Theorem (Cantor–Schröder–Bernstein[11] Theorem) *For sets* S *and* T*, the following statements are equivalent:*

(i) $\operatorname{card}(\mathsf{S}) = \operatorname{card}(\mathsf{T})$*;*

(ii) there exists a bijection $\mathsf{f}\colon \mathsf{S} \to \mathsf{T}$*;*

(iii) there exists injections $\mathsf{f}\colon \mathsf{S} \to \mathsf{T}$ *and* $\mathsf{g}\colon \mathsf{T} \to \mathsf{S}$*;*

(iv) there exists surjections $\mathsf{f}\colon \mathsf{S} \to \mathsf{T}$ *and* $\mathsf{g}\colon \mathsf{T} \to \mathsf{S}$*.*

Proof It is clear from Lemma 1.7.10 that (i) and (ii) are equivalent. It is also clear that (ii) implies both (iii) and (iv).

(iii) $\Longrightarrow$ (ii) We start with a lemma.

1 Lemma *If* $\mathsf{A} \subseteq \mathsf{S}$ *and if there exists an injection* $\mathsf{f}\colon \mathsf{S} \to \mathsf{A}$*, then there exists a bijection* $\mathsf{g}\colon \mathsf{S} \to \mathsf{A}$*.*

Proof Define $B_0 = S \setminus A$ and then inductively define B_j, $j \in \mathbb{Z}_{>0}$, by $B_{j+1} = f(B_j)$. We claim that the sets $(B_j)_{j\in\mathbb{Z}_{\geq 0}}$ (this notation for a family of sets will be made clear in Section 1.6.1) are pairwise disjoint. Suppose not and let $(j,k) \in \mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ be the least pair, with respect to the lexicographic ordering (see Exercise 1.5.3), for which $B_j \cap B_k \neq \varnothing$. Since clearly $B_0 \cap B_j = \varnothing$ for $j \in \mathbb{Z}_{>0}$, we can assume that $j = \tilde{j} + 1$ and $k = \tilde{k} + 1$ for $\tilde{j}, \tilde{k} \in \mathbb{Z}_{\geq 0}$, and so therefore that $B_j = f(B_{\tilde{j}})$ and $B_k = f(B_{\tilde{k}})$. Thus $f(B_{\tilde{j}} \cap B_{\tilde{k}}) \neq \varnothing$ by Proposition 1.3.5, and so $B_{\tilde{j}} \cap B_{\tilde{k}} \neq \varnothing$. Since $(\tilde{j}, \tilde{k})$ is less that (j,k) with respect to the lexicographic order, we have a contradiction.

Now let $B = \cup_{j\in\mathbb{Z}_{\geq 0}} B_j$ and define $g\colon S \to A$ by

$$g(x) = \begin{cases} f(x), & x \in B, \\ x, & x \notin B. \end{cases}$$

For $x \in B$, $g(x) = f(x) \in A$. For $x \notin B$, we have $x \in A$ by definition of B_0, so that g indeed takes values in A. By definition g is injective. Also, let $x \in A$. If $x \notin B$ then $g(x) = x$. If $x \in B$ then $x \in B_{j+1}$ for some $j \in \mathbb{Z}_{\geq 0}$. Since $B_{j+1} = f(B_j)$, $x \in \operatorname{image}(g)$, so showing that g is surjective. ▼

We now continue with the proof of this part of the theorem. Note that $g \circ f\colon S \to g(T)$ is injective (cf. Exercise 1.3.5). Therefore, by the preceding lemma, there exists a bijection $h\colon S \to g(T)$. Since g is injective, $g\colon T \to g(T)$ is bijective, and let us denote the inverse by, abusing notation, $g^{-1}\colon g(T) \to T$. We then define $b\colon S \to T$ by $b = g^{-1} \circ h$, and leave it to the reader to perform the easy verification that b is a bijection.

(iv) $\Longrightarrow$ (iii) Since f is surjective, by Proposition 1.3.9 there exists a right inverse $f_R\colon T \to S$. Thus $f \circ f_R = \mathrm{id}_T$. Thus f is a *left*-inverse for f_R, implying that f_R is injective, again by Proposition 1.3.9. In like manner, g being surjective implies that there is an injective map from S to T, namely a right-inverse for g. ■

Distinguished names are given to certain kinds of sets, based on their cardinality. Recall that ω is the cardinal number corresponding to the set of natural numbers.

[11] Friedrich Wilhelm Karl Ernst Schröder (1814–1902) was a German mathematician whose work was in the area of mathematical logic. Felix Bernstein (1878–1956) was born in Germany. Despite his name being attached to a basic result in set theory, Bernstein's main contributions were in the areas of statistics, mathematical biology, and actuarial mathematics.

1.7.13 Definition (Finite, countable, enumerable, uncountable) A set S is:

(i) ***finite*** if $\operatorname{card}(S) \in \mathbb{Z}_{\geq 0}$;

(ii) ***infinite*** if $\operatorname{card}(S) \geq \omega$;

(iii) ***countable*** if $\operatorname{card}(S) \leq \omega$;

(iv) ***enumerable*** if $\operatorname{card}(S) = \omega$;

(v) ***uncountable*** if $\operatorname{card}(S) > \omega$. •

Let us give some examples illustrating the distinctions between the various notions of set size.

1.7.14 Examples (Cardinality)

1. All elements of $\mathbb{Z}_{\geq 0}$ are, of course, finite sets.
2. The set $\mathbb{Z}_{\geq 0}$ is countably infinite. Indeed, $\operatorname{card}(\mathbb{Z}_{\geq 0}) = \omega$.
3. We claim that $\mathbf{2}^{\mathbb{Z}_{\geq 0}}$ is uncountable. More generally, we claim that, for any set S, $\operatorname{card}(S) < \operatorname{card}(\mathbf{2}^S)$. To see this, we shall show that any map $f\colon S \to \mathbf{2}^S$ is not surjective. For such a map, let

$$A_f = \{x \in S \mid x \notin f(x)\}.$$

We claim that $A_f \notin \operatorname{image}(f)$. Indeed, suppose that $A_f = f(x)$. If $x \in A_f$ then $x \notin f(x) = A_f$ by definition of A_f; a contradiction. On the other hand, if $x \notin A_f$, then $x \in f(x) = A_f$; again a contradiction. We thus conclude that $A_f \notin \operatorname{image}(f)$. Thus there is no surjective map from S to $\mathbf{2}^S$. There is, however, a surjective map from $\mathbf{2}^S$ to S; for example, for any $x_0 \in S$, the map

$$g(A) = \begin{cases} x, & A = \{x\}, \\ x_0, & \text{otherwise} \end{cases}$$

is surjective. Thus S is "smaller than" $\mathbf{2}^S$, or $\operatorname{card}(S) < \operatorname{card}(\mathbf{2}^S)$. •

1.7.15 Remark (Uncountable sets exist, Continuum Hypothesis) A consequence of the last of the preceding examples is that fact that uncountable sets exist since $\mathbf{2}^{\mathbb{Z}_{\geq 0}}$ has a cardinality strictly greater than that of $\mathbb{Z}_{\geq 0}$.

It is usual to denote the enumerable ordinal by $\aleph_0$ (pronounced "aleph zero" or "aleph naught"). The smallest uncountable ordinal is then denoted by $\aleph_1$. An easy way to characterise $\aleph_1$ is as follows. Note that the cardinal $\aleph_0$ has the property that each of its initial segments is finite. In like manner, $\aleph_1$ has the property that each of its segments is enumerable. This does not *define* $\aleph_1$, but perhaps gives the reader some idea what it is.

It is conjectured that there are no cardinal numbers between $\aleph_0$ and $\aleph_1$; this conjecture is called the ***Continuum Hypothesis***. For readers prepared to accept the existence of the real numbers (or to look ahead to Section 2.1), we comment that card($\mathbb{R}$) = card($\mathbf{2}^{\mathbb{Z}_{\geq 0}}$) (see Exercise 1.7.5). From this follows a slightly more concrete statement of the Continuum Hypothesis, namely the conjecture that card($\mathbb{R}$) = $\aleph_1$. Said yet otherwise, the Continuum Hypothesis is the conjecture that, among the subsets of $\mathbb{R}$, the only possibilities are (1) countable sets and (2) sets having the same cardinality as $\mathbb{R}$. •

It is clear the finite union of finite sets is finite. The following result, however, is less clearly true.

1.7.16 Proposition (Countable unions of countable sets are countable) *Let* $(\mathsf{S}_\mathsf{j})_{\mathsf{j}\in\mathbb{Z}_{\geq 0}}$ *be a family of sets, each of which is countable. Then* $\cup_{\mathsf{j}\in\mathbb{Z}_{\geq 0}}\mathsf{S}_\mathsf{j}$ *is countable.*

Proof Let us explicitly enumerate the elements in the sets S_j, $j \in \mathbb{Z}_{\geq 0}$. Thus we write $S_j = (x_{jk})_{k\in\mathbb{Z}_{\geq 0}}$. We now indicate how one constructs a surjective map f from $\mathbb{Z}_{\geq 0}$ to $\cup_{j\in\mathbb{Z}_{\geq 0}} S_j$:

$$f(0) = x_{00},\ f(1) = x_{01},\ f(2) = x_{10},\ f(3) = x_{02},\ f(4) = x_{11},\ f(5) = x_{20},\\ f(6) = x_{03},\ f(7) = x_{12},\ f(8) = x_{21},\ f(9) = x_{30},\ f(10) = x_{04}, \dots.$$

We leave it to the reader to examine this definition and convince themselves that, if it were continued indefinitely, it would include every element of the set $\cup_{j\in\mathbb{Z}_{>0}} S_j$ in the domain of f. ■

For cardinal numbers one can define arithmetic in a manner similar to, but not the same as, that for ordinal numbers. Given cardinal numbers c_1 and c_2 we let S_1 and S_2 be sets equivalent to (not necessarily similar to, note) c_1 and c_2, respectively. We then define $c_1 + c_2 = \mathrm{card}(S_1 \overset{\circ}{\cup} S_2)$ and $c_1 \cdot c_2 = \mathrm{card}(S_1 \times S_2)$. Note that cardinal number arithmetic is not just ordinal number arithmetic restricted to the cardinal numbers. That is to say, for example, the sum of two cardinal numbers is *not* the ordinal sum of the cardinal numbers thought of as ordinal numbers. It is easy to see this with an example. If S and T are two countably infinite sets, then so too is $S \overset{\circ}{\cup} T$ a countably infinite set (this is Proposition 1.7.16). Therefore, $\mathrm{card}(S) + \mathrm{card}(T) = \mathrm{card}(S \overset{\circ}{\cup} T) = \omega = \mathrm{card}(S) = \mathrm{card}(T)$. We can also define exponentiation of cardinal numbers. For cardinal numbers c_1 and c_2 we, as above, let S_1 and S_2 be sets equivalent to c_1 and c_2, respectively. We then define $c_1^{c_2} = \mathrm{card}(S_1^{S_2})$, where we recall that $S_1^{S_2}$ denotes the set of maps from S_2 to S_1.

The only result that we shall care about concerning cardinal arithmetic is the following.

1.7.17 Theorem (Sums and products of infinite cardinal numbers) *If* c *is an infinite cardinal number then*

(i) $\mathrm{c} + \mathrm{k} = \mathrm{c}$ *for every finite cardinal number* k*,*

(ii) $\mathrm{c} = \mathrm{c} + \mathrm{c}$*, and*

(iii) $\mathrm{c} = \mathrm{c} \cdot \mathrm{c}$*.*

Proof (i) Let S and T be disjoint sets such that $\operatorname{card}(S) = c$ and $\operatorname{card}(T) = k$. Let $g\colon T \to \{1, \dots, k\}$ be a bijection. Since S is infinite, we may suppose that S contains $\mathbb{Z}_{>0}$ as a subset. Define $f\colon S \cup T \to S$ by

$$f(x) = \begin{cases} g(x), & x \in T, \\ x + k, & x \in \mathbb{Z}_{>0} \subseteq S, \\ x, & x \in S \setminus \mathbb{Z}_{>0}. \end{cases}$$

This is readily seen to be a bijection, and so gives the result by definition of cardinal addition.

(ii) Let S be a set such that $\operatorname{card}(S) = c$ and define

$$G(S) = \{(f, A) \mid A \subseteq S,\ f\colon A \times \{0, 1\} \to A \text{ is a bijection}\}.$$

If $A \subseteq S$ is countably infinite, then $\operatorname{card}(A \times \{0, 1\}) = \operatorname{card}(A)$, and so $G(S)$ is not empty. Place a partial order $\preceq$ on $G(S)$ by $(f_1, A_1) \preceq (f_2, A_2)$ if $A_1 \subseteq A_2$ and if $f_2|A_1 = f_1$. This is readily verified to be a partial order. Moreover, if $\{(f_j, A_j) \mid j \in J\}$ is a totally ordered subset, then we define an upper bound (f, A) as follows. We take $A = \cup_{j \in J} A_j$ and $f(x, k) = f_j(x, k)$ where $j \in J$ is defined such that $x \in A_j$. One can now use Zorn's Lemma to assert the existence of a maximal element of $G(S)$ which we denote by (f, A). We claim that $S \setminus A$ is finite. Indeed, if $S \setminus A$ is infinite, then there exists a countably infinite subset B of $S \setminus A$. Let g be a bijection from $B \times \{0, 1\}$ to B and note that the map $f \times g\colon (A \cup B) \times \{0, 1\} \to A \cup B$ defined by

$$f \times g(x, k) = \begin{cases} f(x, k), & x \in A, \\ g(x, k), & x \in B \end{cases}$$

if then a bijection, thus contradicting the maximality of (f, A). Thus $S \setminus A$ is indeed finite. Finally, since $(f, A) \in G(S)$, we have $\operatorname{card}(A) + \operatorname{card}(A) = \operatorname{card}(A)$. Also, $\operatorname{card}(S) = \operatorname{card}(A) + \operatorname{card}(A \setminus S)$. Since $\operatorname{card}(S \setminus A)$ is finite, by part (i) this part of the theorem follows.

(iii) Let S be a set such that $\operatorname{card}(S) = c$ and define

$$F(S) = \{(f, A) \mid A \subseteq S,\ f\colon A \times A \to Af \text{ is a bijection}\}.$$

If $A \subseteq S$ is countably infinite, then $\operatorname{card}(A \times A) = \operatorname{card}(A)$ and so there exists a bijection from $A \times A$ to A. Thus $F(S)$ is not empty. Place a partial order $\preceq$ on $F(S)$ by asking that $(f_1, A_1) \preceq (f_2, A_2)$ if $A_1 \subseteq A_2$ and $f_2|A_1 \times A_1 = f_1$; we leave to the reader the straightforward verification that this is a partial order. Moreover, if $\{(f_j, A_j) \mid j \in J\}$ is a totally ordered subset, it is easy to define an upper bound (f, A) for this set as follows. Take $A = \cup_{j \in J} A_j$ and define $f(x, y) = f_j(x, y)$ where $j \in J$ is defined such that $(x, y) \in A_j \times A_j$. Thus, by Zorn's Lemma, there exists a maximal element (f, A) of $F(S)$. By definition of $F(S)$ we have $\operatorname{card}(A)\operatorname{card}(A) = \operatorname{card}(A)$. We now show that $\operatorname{card}(A) = \operatorname{card}(S)$.

Clearly $\operatorname{card}(A) \leq \operatorname{card}(S)$ since $A \subseteq S$. Thus suppose that $\operatorname{card}(A) < \operatorname{card}(S)$. We now use a lemma.

1 **Lemma** *If* c_1 *and* c_2 *are cardinal numbers at least one of which is infinite, and if* c_3 *is the larger of* c_1 *and* c_2*, then* $\mathrm{c}_1 + \mathrm{c}_2 = \mathrm{c}_3$*.*

Proof Let S_1 and S_2 be disjoint sets such that $\mathrm{card}(S_1) = c_1$ and $\mathrm{card}(S_2) = c_2$. Since $c_1 \leq c_3$ and $c_2 \leq c_3$ it follows that $c_1 + c_2 = c_3 + c_3$. Also, $\mathrm{card}(c_3) \leq \mathrm{card}(c_1) + \mathrm{card}(c_2)$. The lemma now follows from part (ii). ▼

From the lemma we know that $\mathrm{card}(S)$ is the larger of $\mathrm{card}(A)$ and $\mathrm{card}(S \setminus A)$, i.e., that $\mathrm{card}(S) = \mathrm{card}(S \setminus A)$. Therefore $\mathrm{card}(A) < \mathrm{card}(S \setminus A)$. Thus there exists a subset $B \subseteq (S \setminus A)$ such that $\mathrm{card}(B) = \mathrm{card}(A)$. Therefore,

$$\mathrm{card}(A \times B) = \mathrm{card}(B \times A) = \mathrm{card}(B \times B) = \mathrm{card}(A) = \mathrm{card}(B).$$

Therefore,

$$\mathrm{card}((A \times B) \cup (B \times A) \cup (B \times B)) = \mathrm{card}(B)$$

by part (ii). Therefore, there exists a bijection g from $(A \times B) \cup (B \times A) \cup (B \times B)$ to B. Thus we can define a bijection $f \times g$ from

$$(A \cup B) \times (A \cup B) = (A \times A) \cup (A \times B) \cup (B \times A) \cup (B \times B)$$

to $A \cup B$ by

$$f \times g(x, y) = \begin{cases} f(x, y), & (x, y) \in A \times A, \\ g(x, y), & \text{otherwise.} \end{cases}$$

Since $A \subseteq (A \cup B)$ and since $f \times g|(A \times A) = f$, this contradicts the maximality of (f, A). Thus our assumption that $\mathrm{card}(A) < \mathrm{card}(S)$ is invalid. ■

The following corollary will be particularly useful.

1.7.18 Corollary (Sum and product of a countable cardinal and an infinite cardinal) *If* c *is an infinite cardinal number then*

(i) $\mathrm{c} \leq \mathrm{c} + \mathrm{card}(\mathbb{Z}_{>0})$ *and*

(ii) $\mathrm{c} \leq \mathrm{c} \cdot \mathrm{card}(\mathbb{Z}_{>0})$*.*

Proof This follows from Theorem 1.7.17 since $\mathrm{card}(\mathbb{Z}_{>0})$ is the smallest infinite cardinal number, and so $\mathrm{card}(\mathbb{Z}_{>0}) \leq c$. ■

Exercises

1.7.1 Show that every element of an ordinal number is an ordinal number.

1.7.2 Show that any finite union of finite sets is finite.

1.7.3 Show that the Cartesian product of a finite number of countable sets is countable.

1.7.4 For a set S, as per Definition 1.3.1, let 2^S denote the collection of maps from the set S to the set 2. Show that $\mathrm{card}(2^S) = \mathrm{card}(\mathbf{2}^S)$, so justifying the notation $\mathbf{2}^S$ as the collection of subsets of S.
Hint: *Given a subset* $\mathrm{A} \subseteq \mathrm{S}$*, think of a natural way of assigning a map from* S *to* 2*.*

In the next exercise you will show that $\operatorname{card}(\mathbb{R}) = \operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$. We refer to Section 2.1 for the definition of the real numbers. There the reader can also find the definition of the rational numbers, as these are also used in the next exercise.

1.7.5 Show that $\operatorname{card}(\mathbb{R}) = \operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$ by answering the following questions.

Define $f_1\colon \mathbb{R} \to \mathbf{2}^{\mathbb{Q}}$ by

$$f_1(x) = \{q \in \mathbb{Q} \mid \; q \le x\}.$$

(a) Show that f_1 is injective to conclude that $\operatorname{card}(\mathbb{R}) \le \operatorname{card}(\mathbf{2}^{\mathbb{Q}})$.

(b) Show that $\operatorname{card}(\mathbf{2}^{\mathbb{Q}}) = \operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$, and conclude that $\operatorname{card}(\mathbb{R}) \le \operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$.

Let $\{0,2\}^{\mathbb{Z}_{>0}}$ be the set of maps from $\mathbb{Z}_{>0}$ to $\{0,2\}$, and regard $\{0,2\}^{\mathbb{Z}_{>0}}$ as a subset of $[0,1]$ by thinking of $\{0,2\}^{\mathbb{Z}_{>0}}$ as being a sequence representing a decimal expansion in base 3. That is, to $f\colon \mathbb{Z}_{>0} \to \{0,2\}$ assign the real number

$$f_2(f) = \sum_{j=1}^{\infty} \frac{f(j)}{3^j}.$$

Thus f_2 is a map from $\{0,2\}^{\mathbb{Z}_{>0}}$ to $[0,1]$.

(c) Show that f_2 is injective so that $\operatorname{card}(\{0,2\}^{\mathbb{Z}_{>0}}) \le \operatorname{card}([0,1])$.

(d) Show that $\operatorname{card}([0,1]) \le \operatorname{card}(\mathbb{R})$.

(e) Show that $\operatorname{card}(\{0,2\}^{\mathbb{Z}_{>0}}) = \operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$, and conclude that $\operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}}) \le \operatorname{card}(\mathbb{R})$.

Hint: *Use Exercise 1.7.4.*

This shows that $\operatorname{card}(\mathbb{R}) = \operatorname{card}(\mathbf{2}^{\mathbb{Z}_{>0}})$, as desired.

Section 1.8

Some words on axiomatic set theory

The account of set theory in this chapter is, as we said at the beginning of Section 1.1, called "naïve set theory." It turns out that the lack of care in saying what a set *is* in naïve set theory causes some problems. We indicate the nature of these problems in Section 1.8.1. To get around these problems, the presently accepted technique is the define a set as an element of a collection of objects satisfying certain axioms. This is called ***axiomatic set theory***, and we refer the reader to the notes at the end of the chapter for references. The most commonly used such axioms are those of Zermelo–Fränkel set theory, and we give these in Section 1.8.2. There are alternative collections of axioms, some equivalent to the Zermelo–Fränkel axioms, and some not. We shall not discuss this here. An axiom commonly, although not incontroversially, accepted is the Axiom of Choice, which we discuss in Section 1.8.3. We also discuss the Peano Axioms in Section 1.8.4, as these are the axioms of arithmetic. We close with a discussion of some of the issues in set theory, since these are of at least cultural interest.

Do I need to read this section? The material in this section is used exactly nowhere else in the texts. However, we hope the reader will find the informal presentation, and historical slant, interesting. •

1.8.1 Russell's Paradox

Russell's Paradox[12] is the following. Let S be the set of all sets that are not members of themselves. For example, the set P of prime numbers is in S since the set of prime numbers is not a prime number. However, the set N of all things that are not prime numbers is in S since the set of all things that are not prime numbers is not a prime number. Now argue as follows. Suppose that $S \in S$. Then S is a set that does not contain itself as a member; that is, $S \notin S$. Now suppose that $S \notin S$. Then S is a set that does not contain itself as a member; that is, $S \in S$. This is clearly absurd, so the set S cannot exist, although there seems to be nothing wrong with its definition. That a contradiction can be derived from the naïve version of set theory means that it is ***inconsistent***.

A consequence of Russell's Paradox is that there is no set containing all sets. Indeed, let S be any set. Then define

$$T = \{x \in S \mid \ x \notin x\}.$$

[12] So named for Bertrand Arthur William Russell (1872–1970), who was a British philosopher and mathematician. Russell received a Nobel prize for literature in recognition of his popular writings on philosophy.

We claim that $T \notin S$. Indeed, suppose that $T \in S$. Then either $T \in T$ or $T \notin T$. In the first instance, since $T \in S$, $T \notin T$. In the second instance, again since $T \in S$, we have $T \notin T$. This is clearly a contradiction, and so we have concluded that, for every set S, there exists something that is not in A. Thus there can be no "set of sets."

Another consequence of Russell's Paradox is the ridiculous conclusion that everything is true. This is a simply logical consequence of the fact that, if a contradiction holds, then all statements hold. Here a contradiction means that a proposition P and its negation $\neg P$ both hold. The argument is as follows. Consider a proposition P'. Then P or P' holds, since P holds. However, since $\neg P$ holds and either P or P' holds, it must be the case that P' holds, no matter what P' is!

Thus the contradiction arising from Russell's Paradox is unsettling since it now calls into question any conclusions that might arise from our discussion of set theory. Various attempts were made to eliminate the eliminate the inconsistency in the naïve version of set theory. The presently most widely accepted of these attempts is the collection of axioms forming Zermelo–Fränkel set theory.

1.8.2 The axioms of Zermelo–Fränkel set theory

The axioms we give here are the culmination of the work of Ernst Friedrich Ferdinand Zermelo (1871–1953) and Adolf Abraham Halevi Fränkel (1891–1965).[13] The axioms were constructed in an attempt to arrive at a basis for set theory that was free of inconsistencies. At present, it is unknown whether the axioms of Zermelo–Fränkel set theory, abbreviated ***ZF***, are consistent.

Here we shall state the axioms, give a slight discussion of them, and indicate some of the places in the chapter where the axioms were employed.

The first axiom merely says that two sets are equal if they have the same elements. This is not controversial, and we have used this axiom out of hand throughout the chapter.

Axiom of Extension For sets S and T, if $x \in S$ if and only if $x \in T$, then $S = T$. •

The next axiom indicates that one can form the set of elements for which a certain property holds. Again, this is not controversial, and is an axiom we have used throughout the chapter.

Axiom of Separation For a set S and a property P defined in S, there exists a set A such that $x \in A$ if and only if $x \in S$ and $P(x) = \text{true}$. •

We also have an axiom which says that one can extract two members from two sets, and think of these as members of another set. This is another uncontroversial axiom that we have used without much fuss.

Axiom of the Unordered Pair For sets S_1 and S_2 and for $x_1 \in S_1$ and $x_2 \in S_2$, there exists a set T such that $x \in T$ if and only if $x = x_1$ or $x = x_2$. •

[13] Fränkel was a German mathematician who worked primarily in the areas of set theory and mathematical logic.

To form the union of two sets, one needs an axiom asserting that the union exists. This is natural, and we have used it whenever we use the notion of union, i.e., frequently.

Axiom of Union For sets S_1 and S_2 there exists a set T such that $x \in T$ if and only if $x \in S_1$ or $x \in S_2$. •

The existence of the power set is also included in the axioms. It is natural and we have used it frequently.

Axiom of the Power Set For a set S there exists a set T such that $A \in T$ if and only if $A \subseteq S$. •

When we constructed the set of natural numbers, we needed an axiom to ensure that this set existed (cf. Assumption 1.4.3). This axiom is the following.

Axiom of Infinity There exists a set S such that

(i) $\varnothing \in S$ and

(ii) for each $x \in S$, $x^+ \in S$. •

When we constructed a large number of ordinal numbers in Example 1.7.2, we repeatedly used an axiom, the essence of which was, “The same principle used to assert the existence of $\mathbb{Z}_{\geq 0}$ can be applied to this more general setting.” Let us now state this idea more formally.

Axiom of Substitution For a set S, if for all $x \in S$ there exists a unique y such that $P(x, y)$ holds, then there exists a set T and a map $f\colon S \to T$ such that $f(x) = y$ where $P(x, y) = \text{true}$. •

The idea is that, for each $x \in S$, the collection of objects y for which $P(x, y)$ holds forms a set. Let us illustrate how the Axiom of Substitution can be used to define the ordinal number $\omega 2$, as in Example 1.7.2. For $k \in \mathbb{Z}_{\geq 0}$ we define

$$P(k, y) = \begin{cases} \text{true}, & y = \omega + k, \\ \text{false}, & \text{otherwise}. \end{cases}$$

The Axiom of Substitution then says that there is a set T and a map $f\colon \mathbb{Z}_{\geq 0} \to T$ such that $f(k) = \omega + k$. The ordinal number $\omega 2$ is then simply the image of the map f.

The final axiom in ZF is the one whose primary purpose is to eliminate inconsistencies such as those arising from Russell’s Paradox.

Axiom of Regularity For each nonempty set S there exists $x \in S$ such that $x \cap S = \varnothing$. •

The Axiom of Regularity rules out sets like $S = \{S\}$ whose only members are themselves. It is no great loss having to live without such sets.

1.8.3 The Axiom of Choice

The Axiom of Choice has its origins in Zermelo's proof of his theorem that every set can be well ordered. In order to prove the theorem, he had to introduce a new axiom in addition to those accepted at the time to characterise sets. The new axiom is the following.

Axiom of Choice For each family $(S_a)_{a\in A}$ of nonempty sets, there exists a function, $f\colon A \to \cup_{a\in A} S_a$, called a ***choice function***, having the property that $f(a) \in S_a$. •

The combination of the axioms of ZF with the Axiom of Choice is sometimes called ***ZF with Choice***, or ***ZFC***. Work of Cohen[14] shows that the Axiom of Choice is independent of the axioms of ZF. Thus, when one adopts ZFC, the Axiom of Choice is really something additional that one is adding to one's list of assumptions of set theory.

At first glance, the Axiom of Choice, at least in the form we give it, does not seem startling. It merely says that, from any collection of sets, it is possible to select an element from each set. A trivial rephrasing of the Axiom of Choice is that, for any family $(S_a)_{a\in A}$ of nonempty sets, the Cartesian product $\prod_{a\in A} S_a$ is nonempty.

What is less settling about the Axiom of Choice is that it can lead to some non-intuitive conclusions. For example, as mentioned above, Zermelo's Well Ordering Theorem follows from the Axiom of Choice. Indeed, the two are equivalent. Let us, in fact, list the equivalence of the Axiom of Choice with two other important results from the chapter, one of which is Zermelo's Well Ordering Theorem.

1.8.1 Theorem (Equivalents of the Axiom of Choice) *If the axioms of ZF hold, then the following statements are equivalent:*

(i) the Axiom of Choice holds;

(ii) Zorn's Lemma holds;

(iii) Zermelo's Well Ordering Theorem holds.

Proof Let us suppose that the proofs we give of Theorems 1.5.13 and 1.5.16 are valid using the axioms of ZF. This is true, and can be verified, if tediously. One only needs to check that no constructions, other than those allowed by the axioms of ZF were used in the proofs. Assuming this, the implications (i) $\Longrightarrow$ (ii) and (ii) $\Longrightarrow$ (iii) hold, since these are what is used in the proofs of Theorems 1.5.13 and 1.5.16. It only remains to prove the implication (iii) $\Longrightarrow$ (i). However, this is straightforward. Let $(S_a)_{a\in A}$ be a family of sets. By Zermelo's Well Ordering Theorem, well order each of these sets, and then define a choice function by assigning to $a \in A$ the least member of S_a. ■

There are, in fact, many statements that are equivalent to the Axiom of Choice. For example, the fact that a surjective map possesses a right-inverse is equivalent to the Axiom of Choice. In Exercise 1.8.1 we give a few of the more easily proved

[14] Paul Joseph Cohen was born in the United States in 1934, and has made outstanding contributions to the foundations of mathematics and set theory.

equivalents of the Axiom of Choice. At the time of its introduction, the equivalence of the Axiom of Choice with Zermelo's Well Ordering Theorem led many mathematicians to reject the validity of the Axiom of Choice. Zermelo, however, countered that many mathematicians implicitly used the Axiom of Choice without saying so. This then led to much activity in mathematics along the lines of deciding which results *required* the Axiom of Choice for their proof. Results can then be divided into three groups, in ascending order of "goodness," where the Axiom of Choice is deemed "bad":

1. results that are equivalent to the Axiom of Choice;
2. results that are not equivalent to the Axiom of Choice, but can be shown to require it for their proof;
3. results that are true, whether or not the Axiom of Choice holds.

Somewhat more startling is that, if one accepts the Axiom of Choice, then it is possible to derive results which seem absurd. Perhaps the most famous of these is the ***Banach–Tarski Paradox***,[15] which says, very roughly, that it is possible to divide a sphere into a finite number of pieces and then reassemble them, while maintaining their shape, into two spheres of equal volume. Said in this way, the result seems impossible. However, if one looks at the result carefully, the nature of the pieces into which the sphere is divided is, obviously, extremely complicated. In the language of Chapter III-2, they are nonmeasurable sets. Such sets correspond poorly with our intuition, and indeed require the Axiom of Choice to assert their existence. We shall give a proof of the Banach–Tarski Paradox in Section III-2.5.6.

On the flip side of this is the fact that there are statements that seem like they *must* be true, and that are equivalent to the Axiom of Choice. One such statement is the Trichotomy Law for the real numbers, which says that, given two real numbers x and y, either $x < y$, $y < x$, or $x = y$. If rejecting the Axiom of Choice means rejecting the Trichotomy Law for real numbers, then many mathematicians would have to rethink the way they do mathematics!

Indeed, there is a branch of mathematics that is dedicated to just this sort of rethinking, and this is called ***constructivism***; see the notes at the end of the chapter for references. The genesis of this branch of mathematics is the dissatisfaction, often arising from applications of the Axiom of Choice, with nonconstructive proofs in mathematics (for example, our proof that a surjective map possesses a right-inverse).

In this book, we will unabashedly assume the validity of the Axiom of Choice. In doing so, we follow in the mainstream of contemporary mathematics.

[15] Stefan Banach (1892–1945) was a well-known Polish mathematician who made significant and foundational contributions to functional analysis. Alfred Tarski (1902–1983) was also Polish, and his main contributions were to set theory and mathematical logic.

1.8.4 Peano's axioms

Peano's axioms[16] were derived in order to establish a basis for arithmetic. They essentially give those properties of the set of "numbers" that allow the establishment of the usual laws for addition and multiplication of natural numbers. ***Peano's axioms*** are these:

1. $0 = \varnothing$ is a number;
2. if k is a number, the successor of k is a number;
3. there is no number for which 0 is a successor;
4. if $j^+ = k^+$ then $j = k$ for all numbers j and k;
5. if S is a set of numbers containing 0 and having the property that the successor of every element of S is in S, then S contains the set of numbers.

Peano's axioms, since they led to the integers, and so there to the rational and real numbers (as in Section 2.1), were once considered as the basic ingredient from which all the rest of mathematics stemmed. This idea, however, received a blow with the publication of a paper by Kurt Gödel[17]. Gödel showed that in any logical system sufficiently general to include the Peano axioms, there exist statements whose truth cannot be validated within the axioms of the system. Thus, this showed that any system built on arithmetic could not possibly be self-contained.

1.8.5 Discussion of the status of set theory

In this section, we have painted a picture of set theory that suggests it is something of a morass of questionable assumptions and possibly unverifiable statements. There is some validity in this, in the sense that there are many fundamental questions unanswered. However, we shall not worry much about these matters as we proceed onto more concrete topics.

1.8.6 Notes

There are many general references for axiomatic set theory. We cite [Suppes 1960].

The independence of the Axiom of Choice from the ZF axioms was proved in [Cohen 1963]. An interesting book on the Axiom of Choice is that of Moore [1982]. Constructivism is discussed by [Bridges and Richman 1987], for example. It is the paper of Gödel [1931] where the incompleteness of axiomatic systems which contain the Peano axioms is proved.

[16] Named after Giuseppe Peano (1858–1932), an Italian mathematician who did work with differential equations and set theory.

[17] Kurt Gödel (1906–1978) was born in a part of the Austro-Hungarian Empire that is now Czechoslovakia. He made outstanding contributions to the subject of mathematical logic.

Exercises

1.8.1 Prove the following result.

Theorem *If the axioms of ZF hold, then the following statements are equivalent:*

(i) the Axiom of Choice holds;

(ii) for any family $(S_a)_{a \in A}$ *of sets, the Cartesian product* $\prod_{a \in A} S_a$ *is nonempty;*

(iii) every surjective map possesses a right inverse.

Section 1.9

Some words about proving things

Rigour is an important part of the presentation in this series, and if you are so unfortunate as to be using these books as a text, then hopefully you will be asked to prove some things, for example, from the exercises. In this section we say a few (almost uselessly) general things about techniques for proving things. We also say some things about poor proof technique, much (but not all) of which is delivered with tongue in cheek. The fact of the matter is that the best way to become proficient at proving things is to (1) read a lot of (needless to say, good) proofs, and (2) most importantly, get lots of practice. What is certainly true is that it much easier to begin your theorem-proving career by proving simple things. In this respect, the proofs and exercises in this chapter are good ones. Similarly, many of the proofs and exercises in Chapters 4 and 5 provide a good basis for honing one's theorem-proving skills. By contrast, some of the results in Chapter 2 are a little more sophisticated, while still not difficult. As we progress through the preparatory material, we shall increasingly encounter material that is quite challenging, and so proofs that are quite elaborate. The neophyte should not be so ambitious as to tackle these early on in their mathematical development.

Do I need to read this section? Go ahead, read it. It will be fun. •

1.9.1 Legitimate proof techniques

The techniques here are the principle ones use in proving simple results. For very complicated results, many of which appear in this series, one is unlikely to get much help from this list.

1. *Proof by definition:* Show that the desired proposition follows directly from the given definitions and assumptions. Theorems that have already been proven to follow from the definitions and assumptions may also be used. Proofs of this sort are often abbreviated by "This is obvious." While this may well be true, it is better to replace this hopelessly vague assertion with something more meaningful like "This follows directly from the definition."
2. *Proof by contradiction:* Assume that the hypotheses of the desired proposition hold, but that the conclusions are false, and make no other assumption. Show that this leads to an impossible conclusion. This implies that the assumption must be false, meaning the desired proposition is true.
3. *Proof by induction:* In this method one wishes to prove a proposition for an enumerable number of cases, say $1, 2, \dots, n, \dots$. One first proves the proposition for case 1. Then one proves that, if the proposition is true for the nth case, it is true for the $(n+1)$st case.

4. *Proof by exhaustion:* One proves the desired proposition to be true for all cases. This method only applies when there is a *finite* number of cases.
5. *Proof by contrapositive:* To show that proposition A implies proposition B, one shows that proposition B *not* being true implies that proposition A is *not* true. It is common to see newcomers get proof by contrapositive and proof by contradiction confused.
6. *Proof by counterexample:* This sort of proof is typically useful in showing that some general assertion *does not* hold. That is to say, one wishes to show that a certain conclusion does not follow from certain hypotheses. To show this, it suffices to come up with a single example for which the hypotheses hold, but the conclusion does not. Such an example is called a ***counterexample***.

1.9.2 Improper proof techniques

Many of these seem so simple that a first reaction is, "Who would be dumb enough to do something so obviously incorrect." However, it is easy, and sometimes tempting, to hide one of these incorrect arguments inside something complicated.

1. *Proof by reverse implication:* To prove that A implies B, shows that B implies A.
2. *Proof by half proof:* One is required to show that A and B are equivalent, but one only shows that A implies B. Note that the appearance of "if and only if" means that you have two implications to prove!
3. *Proof by example:* Show only a single case among many. Assume that only a single case is sufficient (when it is not) or suggest that the proof of this case contains most of the ideas of the general proof.
4. *Proof by picture:* A more convincing form of proof by example. Pictures can provide nice illustrations, but suffice in no part of a rigorous argument.
5. *Proof by special methods:* You are allowed to divide by zero, take wrong square roots, manipulate divergent series, etc.
6. *Proof by convergent irrelevancies:* Prove a lot of things related to the desired result.
7. *Proof by semantic shift:* Some standard but inconvenient definitions are changed for the statement of the result.
8. *Proof by limited definition:* Define (or implicitly assume) a set S, for which all of whose elements the desired result is true, then announce that in the future only members of the set S will be considered.
9. *Proof by circular cross-reference:* Delay the proof of a lemma until many theorems have been derived from it. Use one or more of these theorems in the proof of the lemma.
10. *Proof by appeal to intuition:* Cloud-shaped drawings frequently help here.
11. *Proof by elimination of counterexample:* Assume the hypothesis is true. Then show that a counterexample cannot exist. (This is really just a well-disguised proof by

reverse implication.) A common variation, known as “begging the question” involves getting deep into the proof and then using a step that assumes the hypothesis.

12. *Proof by obfuscation:* A long plotless sequence of true and/or meaningless syntactically related statements.
13. *Proof by cumbersome notation:* Best done with access to at least four alphabets and special symbols. Can help make proofs by special methods look more convincing.
14. *Proof by cosmology:* The negation of a proposition is unimaginable or meaningless.
15. *Proof by reduction to the wrong problem:* To show that the result is true, compare (reduce/translate) the problem (in)to another problem. This is valid if the other problem is then solvable. The error lies in comparing to an unsolvable problem.

Exercises

1.9.1 Find the flaw in the following inductive “proof” of the fact that, in any class, if one selects a subset of students, they will have received the same grade.

> Suppose that we have a class with students $S = \{S_1, \dots, S_m\}$. We shall prove by induction on the size of the subset that any subset of students receive the same grade. For a subset $\{S_{j_1}\}$, the assertion is clearly true. Now suppose that the assertion holds for all subsets of S with k students with $k \in \{1, \dots, l\}$, and suppose we have a subset $\{S_{j_1}, \dots, S_{j_l}, S_{j_{l+1}}\}$ of $l+1$ students. By the induction hypothesis, the students from the set $\{S_{j_1}, \dots, S_{j_l}\}$ all receive the same grade. Also by the induction hypothesis, the students from the set $\{S_2, \dots, S_{j_l}, S_{j_{l+1}}\}$ all receive the same grade. In particular, the grade received by student $S_{j_{l+1}}$ is the same as the grade received by student S_{j_l}. But this is the same as the grade received by students $S_{j_1}, \dots, S_{j_{l-1}}$, and so, by induction, we have proved that all students receive the same grade.

In the next exercise you will consider one of Zeno’s paradoxes. Zeno[18] is best known for having developed a collection of paradoxes, some of which touch surprisingly deeply on mathematical ideas that were not perhaps fully appreciated until the 19th century. Many of his paradoxes have a flavour similar to the one we give here, which may be the most commonly encountered during dinnertime conversations.

1.9.2 Consider the classical problem of the Achilles chasing the tortoise. A tortoise starts off a race T seconds before Achilles. Achilles, of course, is faster than the tortoise, but we shall argue that, despite this, Achilles will actually never overtake the tortoise.

[18] Zeno of Elea (~490BC–~425BC) was an Italian born philosopher of the Greek school.

> At time T when Achilles starts after the tortoise, the tortoise will be some distance d_1 ahead of Achilles. Achilles will reach this point after some time t_1. But, during the time it took Achilles to travel distance d_1, the tortoise will have moved along to some point d_2 ahead of d_1. Achilles will then take a time t_2 to travel the distance d_2. But by then the tortoise will have travelled another distance d_3. This clearly will continue, and when Achilles reaches the point where the tortoise was at some moment before, the tortoise will have moved inexorably ahead. Thus Achilles will never actually catch up to the tortoise.

What is the flaw in the argument?

Chapter 2

Real numbers and their properties

Real numbers and functions of real numbers form an integral part of mathematics. Certainly all students in the sciences receive basic training in these ideas, normally in the form of courses on calculus and differential equations. In this chapter we establish the basic properties of the set of real numbers and of functions defined on this set. In particular, using the construction of the integers in Section 1.4 as a starting point, we *define* the set of real numbers in Section 2.1, thus providing a fairly firm basis on which to develop the main ideas in these volumes. We follow this by discussing various structural properties of the set of real numbers. These cover both algebraic properties (Section 2.2) and topological properties (Section 2.5). We also talk in a little details about sequences and series of real numbers.

Do I need to read this chapter? Yes you do, unless you already know its contents. While the construction of the real numbers in Section 2.1 is perhaps a little bit of an extravagance, it does set the stage for the remainder of the material. Moreover, the material in the remainder of the chapter is, in some ways, the backbone of the mathematical presentation. We say this for two reasons.

1. The technical material concerning the structure of the real numbers is, very simply, assumed knowledge for reading everything else in the series.
2. The *ideas* introduced in this chapter will similarly reappear constantly throughout the volumes in the series. But here, many of these ideas are given their most concrete presentation and, as such, afford the inexperienced reader the opportunity to gain familiarity with useful techniques (e.g., the $\epsilon-\delta$ formalism) in a setting where they presumably possess some degree of comfort. This will be crucial when we discuss more abstract ideas in Chapters 5, III-1, and III-6, to name a few. •

Contents

Section 2.1

Construction of the real numbers

In this section we undertake to define the set of real numbers, using as our starting point the set $\mathbb{Z}$ of integers constructed in Section 1.4. The construction begins by building the rational numbers, which are defined, loosely speaking, as fractions of integers. We know from our school days that every real number can be arbitrarily well approximated by a rational number, e.g., using a decimal expansion. We use this intuitive idea as our basis for defining the set of real numbers from the set of rational numbers.

Do I need to read this section? If you feel comfortable with your understanding of what a real number is, then this section is optional reading. However, it is worth noting that in Section 2.1.2 we first use the $\epsilon - \delta$ formalism that is so important in the analysis featured in this series. Readers unfamiliar/uncomfortable with this idea may find this section a good place to get comfortable with this idea. It is also worth mentioning at this point that the $\epsilon - \delta$ formalism is one with which it is difficult to become fully comfortable. Indeed, PhD theses have been written on the topic of how difficult it is for students to fully assimilate this idea. We shall not adopt any unusual pedagogical strategies to address this matter. However, students are well-advised to spend some time understanding $\epsilon - \delta$ language, and instructors are well-advised to appreciate the difficulty students have in coming to grips with it. •

2.1.1 Construction of the rational numbers

The set of rational numbers is, roughly, the set of fractions of integers. However, we do not know what a fraction is. To define the set of rational numbers, we introduce an equivalence relation $\sim$ in $\mathbb{Z} \times \mathbb{Z}_{>0}$ by

$$(j_1, k_1) \sim (j_2, k_2) \qquad \Longleftrightarrow \qquad j_1 \cdot k_2 = j_2 \cdot k_1.$$

We leave to the reader the straightforward verification that this is an equivalence relation. Using this relation we define the rational numbers as follows.

2.1.1 Definition (Rational numbers) A ***rational number*** is an element of $(\mathbb{Z} \times \mathbb{Z}_{>0})/\sim$. The set of rational numbers is denoted by $\mathbb{Q}$. •

2.1.2 Notation (Notation for rationals) For the rational number $[(j,k)]$ we shall typically write $\frac{j}{k}$, reflecting the usual fraction notation. We shall also often write a typical rational number as "q" when we do not care which equivalence class it comes from. We shall denote by 0 and 1 the rational numbers $[(0,1)]$ and $[(1,1)]$, respectively •

The set of rational numbers has many of the properties of integers. For example, one can define addition and multiplication for rational numbers, as well as a total order in the set of rationals. However, there is an important construction that can be made for rational numbers that cannot generally be made for integers, namely that of division. Let us see how this is done.

2.1.3 Definition (Addition, multiplication, and division in $\mathbb{Q}$) Define the operations of ***addition***, ***multiplication***, and ***division*** in $\mathbb{Q}$ by

(i) $[(j_1, k_1)] + [(j_2, k_2)] = [(j_1 \cdot k_2 + j_2 \cdot k_1, k_1 \cdot k_2)]$,

(ii) $[(j_1, k_1)] \cdot [(j_2, k_2)] = [(j_1 \cdot j_2, k_1 \cdot k_2)]$, and

(iii) $[(j_1, k_1)]/[(j_2, k_2)] = [(j_1 \cdot k_2, k_1 \cdot j_2)]$ (we will also write $\frac{[(j_1,k_1)]}{[(j_2,k_2)]}$ for $[(j_1, k_1)]/[(j_2, k_2)]$),

respectively, where $[(j_1, k_1)], [(j_2, k_2)] \in \mathbb{Q}$ and where, in the definition of division, we require that $j_2 \neq 0$. We will sometimes omit the "$\cdot$" when in multiplication. •

We leave to the reader as Exercise 2.1.1 the straightforward task of showing that these definitions are independent of choice of representatives in $\mathbb{Z} \times \mathbb{Z}_{>0}$. We also leave to the reader the assertion that, with respect to Notation 2.1.2, the operations of addition, multiplication, and division of rational numbers assume the familiar form:

$$\frac{j_1}{k_1} + \frac{j_2}{k_2} = \frac{j_1 \cdot k_2 + j_2 \cdot k_1}{k_1 \cdot k_2}, \quad \frac{j_1}{k_1} \cdot \frac{j_2}{k_2} = \frac{j_1 \cdot j_2}{k_2 \cdot k_2}, \quad \frac{\frac{j_1}{k_1}}{\frac{j_2}{k_2}} = \frac{j_1 \cdot k_2}{k_1 \cdot j_2}.$$

For the operation of division, it is convenient to introduce a new concept. Given $[(j, k)] \in \mathbb{Q}$ with $j \neq 0$, we define $[(j, k)]^{-1} \in \mathbb{Q}$ by $[(k, j)]$. With this notation, division then can be written as $[(j_1, k_1)]/[(j_2, k_2)] = [(j_1, k_1)] \cdot [(j_2, k_2)]^{-1}$. Thus division is really just multiplication, as we already knew. Also, if $q \in \mathbb{Q}$ and if $k \in \mathbb{Z}_{\geq 0}$, then we define $q^k \in \mathbb{Q}$ inductively by $q^0 = 1$ and $q^{k^+} = q^k \cdot q$. The rational number q^k is the kth ***power*** of q.

Let us verify that the operations above satisfy the expected properties. Note that there are now some new properties, since we have the operation of division, or multiplicative inversion, to account for. As we did for integers, we shall write $-q$ for $-1 \cdot q$.

2.1.4 Proposition (Properties of addition and multiplication in $\mathbb{Q}$) *Addition and multiplication in $\mathbb{Q}$ satisfy the following rules:*

(i) $q_1 + q_2 = q_2 + q_1$, $q_1, q_2 \in \mathbb{Q}$ *(**commutativity** of addition);*

(ii) $(q_1 + q_2) + q_3 = q_1 + (q_2 + q_3)$, $q_1, q_2, q_3 \in \mathbb{Q}$ *(**associativity** of addition);*

(iii) $q + 0 = q$, $q \in \mathbb{Q}$ *(**additive identity**);*

(iv) $q + (-q) = 0$, $q \in \mathbb{Q}$ *(**additive inverse**);*

(v) $q_1 \cdot q_2 = q_2 \cdot q_1$, $q_1, q_2 \in \mathbb{Q}$ *(**commutativity** of multiplication);*

(vi) $(q_1 \cdot q_2) \cdot q_3 = q_1 \cdot (q_2 \cdot q_3)$, $q_1, q_2, q_3 \in \mathbb{Q}$ *(**associativity** of multiplication);*

(vii) $q \cdot 1 = q$, $q \in \mathbb{Q}$ *(**multiplicative identity**);*

(viii) $q \cdot q^{-1} = 1$, $q \in \mathbb{Q} \setminus \{0\}$ ***(multiplicative inverse)****;*

(ix) $r \cdot (q_1 + q_2) = r \cdot q_1 + r \cdot q_2$, $r, q_1, q_2 \in \mathbb{Q}$ ***(distributivity)****;*

(x) $q^{k_1} \cdot q^{k_2} = q^{k_1+k_2}$, $q \in \mathbb{Q}$, $k_1, k_2 \in \mathbb{Z}_{\geq 0}$.

Moreover, if we define $i_{\mathbb{Z}} \colon \mathbb{Z} \to \mathbb{Q}$ *by* $i_{\mathbb{Z}}(k) = [(k, 1)]$*, then addition and multiplication in* $\mathbb{Q}$ *agrees with that in* $\mathbb{Z}$*:*

$$i_{\mathbb{Z}}(k_1) + i_{\mathbb{Z}}(k_2) = i_{\mathbb{Z}}(k_1 + k_2), \quad i_{\mathbb{Z}}(k_1) \cdot i_{\mathbb{Z}}(k_2) = i_{\mathbb{Z}}(k_1 \cdot k_2).$$

Proof All of these properties follow directly from the definitions of addition and multiplication, using Proposition 1.4.19. ■

Just as we can naturally think of $\mathbb{Z}_{\geq 0}$ as being a subset of $\mathbb{Z}$, so too can we think of $\mathbb{Z}$ as a subset of $\mathbb{Q}$. Moreover, we shall very often do so without making explicit reference to the map $i_{\mathbb{Z}}$.

Next we consider on $\mathbb{Q}$ the extension of the partial order $\leq$ and the strict partial order $<$.

2.1.5 Proposition (Order on $\mathbb{Q}$) *On* $\mathbb{Q}$ *define two relations* $<$ *and* $\leq$ *by*

$$\begin{aligned} [(j_1, k_1)] < [(j_2, k_2)] \quad &\iff \quad j_1 \cdot k_2 < k_1 \cdot j_2, \\ [(j_1, k_1)] \leq [(j_2, k_2)] \quad &\iff \quad j_1 \cdot k_2 \leq k_1 \cdot j_2. \end{aligned}$$

Then $\leq$ *is a total order and* $<$ *is the corresponding strict partial order.*

Proof First let us show that the relations defined make sense, in that they are independent of choice of representative. Thus we suppose that $[(j_1, k_1)] = [(\tilde{j}_1, \tilde{k}_1)]$ and that $[(j_2, k_2)] = [(\tilde{j}_2, \tilde{k}_2)]$. Then

$$\begin{aligned} & [(j_1, k_1)] \leq [(j_2, k_2)] \\ \iff \quad & j_1 \cdot k_2 \leq k_1 \cdot j_2 \\ \iff \quad & j_1 \cdot k_2 \cdot j_2 \cdot \tilde{k}_2 \cdot \tilde{j}_1 \cdot k_1 \leq k_1 \cdot j_2 \cdot \tilde{j}_2 \cdot k_1 \cdot j_1 \cdot \tilde{k}_1 \\ \iff \quad & (\tilde{j}_1 \cdot \tilde{k}_2) \cdot (j_1 \cdot j_2 \cdot k_1 \cdot k_2) \leq (\tilde{j}_2 \cdot \tilde{k}_1) \cdot (j_1 \cdot j_2 \cdot k_1 \cdot k_2) \\ \iff \quad & \tilde{j}_1 \cdot \tilde{k}_2 \leq \tilde{j}_2 \cdot \tilde{k}_1. \end{aligned}$$

This shows that the definition of $\leq$ is independent of representative. Of course, a similar argument holds for $<$.

That $\leq$ is a partial order, and that $<$ is its corresponding strict partial order, follow from a straightforward checking of the definitions, so we leave this to the reader.

Thus we only need to check that $\leq$ is a total order. Let $[(j_1, k_1)], [(j_2, k_2)] \in \mathbb{Q}$. Then, by the Trichotomy Law for $\mathbb{Z}$, either $j_1 \cdot k_2 < k_1 \cdot j_2$, $k_1 \cdot j_2 < j_1 \cdot k_2$, or $j_1 \cdot k_2 = k_1 \cdot j_2$. But this directly implies that either $[(j_1, k_1)] < [(j_2, k_2)]$, $[(j_2, k_2)] < [(j_1, k_1)]$, or $[(j_1, k_1)] = [(j_2, k_2)]$, respectively. ■

The total order on $\mathbb{Q}$ allows a classification of rational numbers as follows.

2.1.6 Definition (Positive and negative rational numbers) A rational number $q \in \mathbb{Q}$ is:

(i) ***positive*** if $0 < q$;

(ii) ***negative*** if $q < 0$;

(iii) ***nonnegative*** if $0 \leq q$;

(iv) ***nonpositive*** if $q \leq 0$.

The set of positive rational numbers is denoted by $\mathbb{Q}_{>0}$ and the set of nonnegative rational numbers is denoted by $\mathbb{Q}_{\geq 0}$. •

As we did with natural numbers and integers, we isolate the Trichotomy Law.

2.1.7 Corollary (Trichotomy Law for $\mathbb{Q}$) *For $\mathrm{q}, \mathrm{r} \in \mathbb{Q}$, exactly one of the following possibilities holds:*

(i) $\mathrm{q} < \mathrm{r}$;

(ii) $\mathrm{r} < \mathrm{q}$;

(iii) $\mathrm{q} = \mathrm{r}$.

The following result records the relationship between the order on $\mathbb{Q}$ and the arithmetic operations.

2.1.8 Proposition (Relation between addition and multiplication and <) *For $\mathrm{q}, \mathrm{r}, \mathrm{s} \in \mathbb{Q}$, the following statements hold:*

(i) if $\mathrm{q} < \mathrm{r}$ *then* $\mathrm{q} + \mathrm{s} < \mathrm{r} + \mathrm{s}$;

(ii) if $\mathrm{q} < \mathrm{r}$ *and if* $\mathrm{s} > 0$ *then* $\mathrm{s} \cdot \mathrm{q} < \mathrm{s} \cdot \mathrm{r}$;

(iii) if $\mathrm{q} < \mathrm{r}$ *and if* $\mathrm{s} < 0$ *then* $\mathrm{s} \cdot \mathrm{r} < \mathrm{s} \cdot \mathrm{q}$;

(iv) if $0 < \mathrm{q}, \mathrm{r}$ *then* $0 < \mathrm{q} \cdot \mathrm{r}$;

(v) if $\mathrm{q} < \mathrm{r}$ *and if either*

(a) $0 < \mathrm{q}, \mathrm{r}$ *or*

(b) $\mathrm{q}, \mathrm{r} < 0$,

then $\mathrm{r}^{-1} < \mathrm{q}^{-1}$.

Proof (i) Write $q = [(j_q, k_q)]$, $r = [(j_r, k_r)]$, and $s = [(j_s, k_s)]$. Since $q < r$, $j_q \cdot k_r \leq j_r \cdot k_q$. Therefore,

$$\begin{aligned} & j_q \cdot k_r \cdot k_s^2 < j_r \cdot k_q \cdot k_s^2 \\ \implies \quad & j_q \cdot k_r \cdot k_s^2 + j_s \cdot k_q \cdot k_r \cdot k_s < j_r \cdot k_q \cdot k_s^2 + j_2 \cdot k_q \cdot k_r \cdot k_s, \end{aligned}$$

using Proposition 1.4.22. This last inequality is easily seen to be equivalent to $q + s < r + s$.

(ii) Write $q = [(j_q, k_q)]$, $r = [(j_r, k_r)]$, and $s = [(j_s, k_s)]$. Since $s > 0$ it follows that $j_s > 0$. Since $q \leq r$ it follows that $j_q \cdot k_r \leq j_r \cdot k_q$. From Proposition 1.4.22 we then have

$$j_q \cdot j_s \cdot j_s \cdot k_s \leq j_r \cdot k_q \cdot j_s \cdot k_s,$$

which is equivalent to $s \cdot q \leq s \cdot r$ by definition of multiplication.

(iii) The result here follows, as does (ii), from Proposition 1.4.22, but now using the fact that $j_s < 0$.

(iv) This is a straightforward application of the definition of multiplication and $<$.

(v) This follows directly from the definition of $<$. ■

The final piece of structure we discuss for rational numbers is the extension of the absolute value function defined for integers.

2.1.9 Definition (Rational absolute value function) The ***absolute value function*** on $\mathbb{Q}$ is the map from $\mathbb{Q}$ to $\mathbb{Q}_{\geq 0}$, denoted by $q \mapsto |q|$, defined by

$$|q| = \begin{cases} q, & 0 < q, \\ 0, & q = 0, \\ -q, & q < 0. \end{cases}$$ •

The absolute value function on $\mathbb{Q}$ has properties like that on $\mathbb{Z}$.

2.1.10 Proposition (Properties of absolute value on $\mathbb{Q}$) *The following statements hold:*

(i) $|q| \geq 0$ *for all* $q \in \mathbb{Q}$;

(ii) $|q| = 0$ *if and only if* $q = 0$;

(iii) $|r \cdot q| = |r| \cdot |q|$ *for all* $r, q \in \mathbb{Q}$;

(iv) $|r + q| \leq |r| + |q|$ *for all* $r, q \in \mathbb{Q}$ *(**triangle inequality**)*;

(v) $|q^{-1}| = |q|^{-1}$ *for all* $q \in \mathbb{Q} \setminus \{0\}$.

Proof Parts (i), (ii), and (v), follow directly from the definition, and part (iii) follows in the same manner as the analogous statement in Proposition 1.4.24. Thus we have only to prove part (iv). We consider various cases.

1. $|r| \leq |q|$:

   (a) $0 \geq r, q$: Since $|r + q| = r + q$, and $|r| = r$ and $|q| = q$, this follows directly.

   (b) $r < 0, 0 \leq q$: Let $r = [(j_r, k_r)]$ and $q = [(j_q, k_q)]$. Then $r < 0$ gives $j_r < 0$ and $0 \leq q$ gives $j_q \geq 0$. We now have

$$|r + q| = \left| \frac{j_r \cdot k_q + j_q \cdot k_r}{k_r \cdot k_q} \right| = \frac{|j_r \cdot k_q + j_q \cdot k_r|}{k_r \cdot k_q}$$

   and

$$|r| + |q| = \frac{|j_r| \cdot k_q + |j_q| \cdot k_r}{k_r \cdot k_q}.$$

   Therefore,

$$\begin{aligned} |r + q| &= \frac{|j_r \cdot k_q + j_q \cdot k_r|}{k_r \cdot k_q} \\ &\leq \frac{|j_r| \cdot k_q + |j_q| \cdot k_r}{k_r \cdot k_q} \\ &= |r| + |q|, \end{aligned}$$

   where we have used Proposition 2.1.8.

(c) $r, q < 0$: Here $|r + q| = |-r + (-q)| = |-(r + q)| = -(r + q)$, and $|r| = -r$ and $|q| = -q$, so the result follows immediately.

2. $|q| \leq |r|$: This argument is the same as above, swapping r and q. ■

2.1.11 Remark Having been quite fussy about how we arrived at the set of integers and the set of rational numbers, and about characterising their important properties, we shall now use standard facts about these, some of which we may not have proved, but which can easily be proved using the definitions of $\mathbb{Z}$ and $\mathbb{Q}$. Some of the arithmetic properties of $\mathbb{Z}$ and $\mathbb{Q}$ that we use without comment are in fact proved in Section 4.2 in the more general setting of rings. However, we anticipate that most readers will not balk at the instances where we use unproved properties of integers and rational numbers. •

2.1.2 Construction of the real numbers from the rational numbers

Now we use the rational numbers as the building block for the real numbers. The idea of this construction, which was originally due to Cauchy[1], is the intuitive idea that the rational numbers may be used to approximate well a real number. For example, we learn in school that any real number is expressible as a decimal expansion (see Exercise 2.4.8 for the precise construction of a decimal expansion). However, any finite length decimal expansion (and even some infinite length decimal expansions) is a rational number. So one could *define* real numbers as a limit of decimal expansions in some way. The problem is that there may be multiple decimal expansions giving rise to the same real number. For example, the decimal expansions 1.0000 and 0.9999 . . . represent the same real number. The way one gets around this potential problem is to use equivalence classes, of course. But equivalence classes of what? This is where we begin the presentation, proper.

2.1.12 Definition (Cauchy sequence, convergent sequence) Let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{Q}$. The sequence:

(i) is a ***Cauchy sequence*** if, for each $\epsilon \in \mathbb{Q}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \epsilon$ for $j, k \geq N$;

(ii) ***converges to*** $\mathbf{q_0}$ if, for each $\epsilon \in \mathbb{Q}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_0| < \epsilon$ for $j \geq N$.

(iii) is ***bounded*** if there exists $M \in \mathbb{Q}_{>0}$ such that $|q_j| < M$ for each $j \in \mathbb{Z}_{>0}$. •

The set of Cauchy sequences in $\mathbb{Q}$ is denoted by $\mathrm{CS}(\mathbb{Q})$. A sequence converging to q_0 has q_0 as its ***limit***. •

The idea of a Cauchy sequence is that the terms in the sequence can be made arbitrarily close as we get to the tail of the sequence. A convergent sequence,

[1] The French mathematician Augustin Louis Cauchy (1789–1857) worked in the areas of complex function theory, partial differential equations, and analysis. His collected works span twenty-seven volumes.

however, gets closer and closer to its limit as we get to the tail of the sequence. Our instinct is probably that there is a relationship between these two ideas. One thing that is true is the following.

2.1.13 Proposition (Convergent sequences are Cauchy) *If a sequence $(q_j)_{j\in\mathbb{Z}_{>0}}$ converges to q_0, then it is a Cauchy sequence.*

Proof Let $\epsilon \in \mathbb{Q}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_0| < \frac{\epsilon}{2}$ for $j \geq N$. Then, for $j, k \geq N$ we have

$$|q_j - q_k| = |q_j - q_0 - q_k + q_0| = |q_j - q_0| + |q_k - q_0| < \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon,$$

using the triangle inequality of Proposition 2.1.10. ■

Cauchy sequences have the property of being bounded.

2.1.14 Proposition (Cauchy sequences are bounded) *If $(q_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, then it is bounded.*

Proof Choose $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < 1$ for $j, k \in \mathbb{Z}_{>0}$. Then take M_N to be the largest of the nonnegative rational numbers $|q_1|, \dots, |q_N|$. Then, for $j \geq N$ we have, using the triangle inequality,

$$|q_j| = |q_j - q_N + q_N| \leq |q_j - q_N| + |q_N| < 1 + M_N,$$

giving the result by taking $M = M_N + 1$. ■

The question as to whether there are nonconvergent Cauchy sequences is now the obvious one.

2.1.15 Example (Nonconvergent Cauchy sequences in $\mathbb{Q}$ exist) If one already knows the real numbers exist, it is somewhat easy to come up with Cauchy sequences in $\mathbb{Q}$. However, to fabricate one "out of thin air" is not so easy.

For $k \in \mathbb{Z}_{>0}$, since $2k + 5 > k + 4$, it follows that $2^{2k+5} - 2^{k+4} > 0$. Let m_k be the smallest nonnegative integer for which

$$m_k^2 \geq 2^{2k+5} - 2^{k+4}. \tag{2.1}$$

The following contains a useful property of m_k.

1 Lemma $m_k^2 \leq 2^{2k+5}$.

Proof First we show that $m_k \leq 2^{k+3}$. Suppose that $m_k > 2^{k+3}$. Then

$$(m_k - 1)^2 > (2^{k+3} - 1)^2 = 2^{2k+6} - 2^{k+4} + 1 = 2(2^{2k+5} - 2^{k+4}) + 1) > 2^{2k+5} - 2^{k+4},$$

which contradicts the definition of m_k.

Now suppose that $m_k^2 > 2^{2k+5}$. Then

$$(m_k - 1)^2 = m_k^2 - 2m_k + 1 > 2^{2k+5} - 2^{k+4} + 1 > 2^{2k+5} - 2^{k+4},$$

again contradicting the definition of m_k. ▼

Now define $q_k = \frac{m_k}{2^{k+2}}$.

2 Lemma $(q_k)_{k\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence.*

Proof By Lemma 1 we have

$$q_k^2 = \frac{m_k^2}{2^{2k+4}} \le \frac{2^{2k+5}}{2^{2k+4}} = 2, \qquad k \in \mathbb{Z}_{>0},$$

and by (2.1) we have

$$q_k^2 = \frac{m_k^2}{2^{2k+4}} \ge \frac{2^{2k+5}}{2^{2k+4}} - \frac{2^{k+4}}{2^{2k+4}} = 2 - \frac{1}{2k}, \qquad k \in \mathbb{Z}_{>0}.$$

Summarising, we have

$$2 - \frac{1}{2^k} \le q_k^2 \le 2, \qquad k \in \mathbb{Z}_{>0}. \tag{2.2}$$

Then, for $j, k \in \mathbb{Z}_{>0}$ we have

$$2 - \frac{1}{2^k} \le q_k^2 \le 2,\ 2 - \frac{1}{2^j} \le q_k^2 \le 2 \quad\Longrightarrow\quad -\frac{1}{2^j} \le q_j^2 - q_k^2 \le \frac{1}{2^k}.$$

Next we have, from (2.1),

$$q_k^2 = \frac{m_k^2}{2^{2k+4}} \ge \frac{2^{2k+5}}{2^{2k+4}} - \frac{2^{k+4}}{2^{2k+4}} = 2 - \frac{1}{2^k}, \qquad k \in \mathbb{Z}_{>0},$$

from which we deduce that $q_k^2 \ge 1$, which itself implies that $q_k \ge 1$. Next, using this fact and $(q_j - q_k)^2 = (q_j + q_k)(q_j - q_k)$ we have

$$-\frac{1}{2^j}\frac{1}{q_j + q_k} \le q_j - q_k \le \frac{1}{2^j}\frac{1}{q_j + q_k} \quad\Longrightarrow\quad -\frac{1}{2^{j+1}} \le q_j - q_k \le \frac{1}{2^{k+1}}, \qquad j, k \in \mathbb{Z}_{>0}. \tag{2.3}$$

Now let $\epsilon \in \mathbb{Q}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $\frac{1}{2^{N+1}} < \epsilon$. Then we immediately have $|q_j - q_k| < \epsilon$, $j, k \ge N$, using (2.3). ▼

The following result gives the character of the limit of the sequence $(q_k)_{k\in\mathbb{Z}_{>0}}$, were it to be convergent.

3 Lemma *If* q_0 *is the limit for the sequence* $(q_k)_{k\in\mathbb{Z}_{>0}}$*, then* $q_0^2 = 2$*.*

Proof We claim that if $(q_k)_{k\in\mathbb{Z}_{>0}}$ converges to q_0, then $(q_k^2)_{k\in\mathbb{Z}_{>0}}$ converges to q_0^2. Let $M \in \mathbb{Q}_{>0}$ satisfy $|q_k| < M$ for all $k \in \mathbb{Z}_{>0}$, this being possible by Proposition 2.1.14. Now let $\epsilon \in \mathbb{Q}_{>0}$ and take $N \in \mathbb{Z}_{>0}$ such that

$$|q_k - q_0| < \frac{\epsilon}{M + |q_0|}.$$

Then

$$|q_k^2 - q_0^2| = |q_k - q_0||q_k + q_0| < \epsilon,$$

giving our claim.

Finally, we prove the lemma by proving that $(q_k^2)_{k\in\mathbb{Z}_{>0}}$ converges to 2. Indeed, let $\epsilon \in \mathbb{Q}_{>0}$ and note that, if $N \in \mathbb{Z}_{>0}$ is chosen to satisfy $\frac{1}{2^N} < \epsilon$. Then, using (2.2), we have

$$|q_k^2 - 2| \leq \frac{1}{2^k} < \epsilon, \qquad k \geq N,$$

as desired. ▼

Finally, we have the following result, which is contained in the mathematical works of Euclid.

4 **Lemma** *There exists no* $q_0 \in \mathbb{Q}$ *such that* $q_0^2 = 2$*.*

Proof Suppose that $q_0^2 = [(j_0, k_0)]$ and further suppose that there is no integer m such that $q_0 = [(mj_0, mk_0)]$. We then have

$$q_0^2 = \frac{j_0^2}{k_0^2} = 2 \quad \Longrightarrow \quad j_0^2 = 2k_0^2.$$

Thus j_0^2 is even, and then so too is j_0 (why?). Therefore, $j_0 = 2\tilde{j}_0$ and so

$$q_0^2 = \frac{4\tilde{j}_0^2}{k_0^2} = 2 \quad \Longrightarrow \quad k_0^2 = 2\tilde{j}_0^2$$

which implies that k_0^2, and hence k_0 is also even. This contradicts our assumption that there is no integer m such that $q_0 = [(mj_0, mk_0)]$. ▼

With these steps, we have constructed a Cauchy sequence that does not converge. •

Having shown that there are Cauchy sequences that do not converge, the idea is now to define a real number to be, essentially, that to which a nonconvergent Cauchy sequence would converge if only it could. First we need to allow for the possibility, realised in practice, that different Cauchy sequences may converge to the same limit.

2.1.16 Definition (Equivalent Cauchy sequences) Two sequences $(q_j)_{j\in\mathbb{Z}_{>0}}, (r_j)_{j\in\mathbb{Q}} \in \mathrm{CS}(\mathbb{Q})$ are ***equivalent*** if the sequence $(q_j - r_j)_{j\in\mathbb{Z}_{>0}}$ converges to zero. We write $(q_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$ if the two sequences are equivalent. •

We should verify that this notion of equivalence of Cauchy sequences is indeed an equivalence relation.

2.1.17 Lemma *The relation $\sim$ defined in* CS($\mathbb{Q}$) *is an equivalence relation.*

Proof It is clear that the relation $\sim$ is reflexive and symmetric. To prove transitivity, suppose that $(q_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$ and that $(r_j)_{j\in\mathbb{Z}_{>0}} \sim (s_j)_{j\in\mathbb{Z}_{>0}}$. For $\epsilon \in \mathbb{Q}_{>0}$ let $N \in \mathbb{Z}_{>0}$ satisfy

$$|q_j - r_j| < \tfrac{\epsilon}{2},\ |r_j - s_j| < \tfrac{\epsilon}{2}, \qquad j \geq N.$$

Then, using the triangle inequality,

$$|q_j - s_j| = |q_j - r_j + r_j - s_j| \leq |q_j - r_j| + |r_j - s_j| < \epsilon, \qquad j \geq \mathbb{Z}_{>0},$$

showing that $(q_j)_{j\in\mathbb{Z}_{>0}} \sim (s_j)_{j\in\mathbb{Z}_{>0}}$. ■

We are now prepared to define the set of real numbers.

2.1.18 Definition (Real numbers) A ***real number*** is an element of CS($\mathbb{Q}$)/ $\sim$. The set of real numbers is denoted by $\mathbb{R}$. •

The definition encodes, in a precise way, our intuition about what a real number is. In the next section we shall examine some of the properties of the set $\mathbb{R}$.

Let us give the notation we will use for real numbers, since clearly we do not wish to write these explicitly as equivalence classes of Cauchy sequences.

2.1.19 Notation (Notation for reals) We shall frequently write a typical element in $\mathbb{R}$ as "x". We shall denote by 0 and 1 the real numbers associated with the Cauchy sequences $(0)_{j\in\mathbb{Z}_{>0}}$ and $(1)_{j\in\mathbb{Z}_{>0}}$. •

Exercises

2.1.1 Show that the definitions of addition, multiplication, and division of rational numbers in Definition 2.1.3 are independent of representative.

2.1.2 Show that the order and absolute value on $\mathbb{Q}$ agree with those on $\mathbb{Z}$. That is to say, show the following:

(a) for $j, k \in \mathbb{Z}$, $j < k$ if and only if $i_{\mathbb{Z}}(j) < i_{\mathbb{Z}}(k)$;

(b) for $k \in \mathbb{Z}$, $|k| = |i_{\mathbb{Z}}(k)|$.

(Note that we see clearly here the abuse of notation that follows from using $<$ for both the order on $\mathbb{Z}$ and $\mathbb{Q}$ and from using $|\cdot|$ as the absolute value both on $\mathbb{Z}$ and $\mathbb{Q}$. It is expected that the reader can understand where the notational abuse occurs.)

2.1.3 Show that the set of rational numbers is countable using an argument along the following lines.

1. Construct a doubly infinite grid in the plane with a point at each integer coordinate. Note that every rational number $q = \frac{n}{m}$ is represented by the grid point (n, m).
2. Start at the "centre" of the grid with the rational number 0 being assigned to the grid point $(0, 0)$, and construct a spiral which passes through each grid point. Note that this spiral should hit every grid point exactly once.

3. Use this spiral to infer the existence of a bijection from $\mathbb{Q}$ to $\mathbb{Z}_{>0}$.

The following exercise leads you through Cantor's famous "diagonal argument" for showing that the set of real numbers is uncountable.

2.1.4 Fill in the gaps in the following construction, justifying all steps.

1. Let $\{x_j \mid j \in \mathbb{Z}_{>0}\}$ be a countable subset of $(0, 1)$.
2. Construct a doubly infinite table for which the kth column of the jth row contains the kth term in the decimal expansion for x_j.
3. Construct $\bar{x} \in (0, 1)$ by declaring the kth term in the decimal expansion for $\bar{x}$ to be different from the kth term in the decimal expansion for x_k.
4. Show that $\bar{x}$ is not an element of the set $\{x_j \mid j \in \mathbb{Z}_{>0}\}$.

   Hint: *Be careful to understand that a real number might have different decimal expansions.*

2.1.5 Show that for any $x \in \mathbb{R}$ and $\epsilon \in \mathbb{R}_{>0}$ there exists $k \in \mathbb{Z}_{>0}$ and an odd integer j such that $|x - \frac{j}{2^k}| < \epsilon$.

Section 2.2

Properties of the set of real numbers

In this section we present some of the well known properties as the real numbers, both algebraic and (referring ahead to the language of Chapter III-1) topological.

Do I need to read this section? Many of the properties given in Sections 2.2.1, 2.2.2 and 2.2.3 will be well known to any student with a high school education. However, these may be of value as a starting point in understanding some of the abstract material in Chapters 4 and 5. Similarly, the material in Section 2.2.4 is "obvious." However, since this material will be assumed knowledge, it might be best for the reader to at least skim the section, to make sure there is nothing new in it for them. •

2.2.1 Algebraic properties of $\mathbb{R}$

In this section we define addition, multiplication, order, and absolute value for $\mathbb{R}$, mirroring the presentation for $\mathbb{Q}$ in Section 2.1.1. Here, however, the definitions and verifications are not just trivialities, as they are for $\mathbb{Q}$.

First we define addition and multiplication. We do this by defining these operations first on elements of $\mathrm{CS}(\mathbb{Q})$, and then showing that the operations depend only on equivalence class. The following is the key step in doing this.

2.2.1 Proposition (Addition, multiplication, and division of Cauchy sequences) *Let* $(q_j)_{j\in\mathbb{Z}_{>0}}, (r_j)_{j\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$*. Then the following statements hold.*

(i) The sequence $(q_j+r_j)_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence which we denote by* $(q_j)_{j\in\mathbb{Z}_{>0}}+(r_j)_{j\in\mathbb{Z}_{>0}}$*.*

(ii) The sequence $(q_j\cdot r_j)_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence which we denote by* $(q_j)_{j\in\mathbb{Z}_{>0}}\cdot(r_j)_{j\in\mathbb{Z}_{>0}}$*.*

(iii) If, for all $j\in\mathbb{Z}_{>0}$, $q_j\neq 0$ *and if the sequence* $(q_j)_{j\in\mathbb{Z}_{>0}}$ *does not converge to* 0*, then* $(q_j^{-1})_{j\in\mathbb{Z}_{>0}}$ *is a Cauchy sequence.*

Furthermore, if $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}, (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ *satisfy*

$$(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}, \quad (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}},$$

then

(iv) $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}+(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} = (q_j)_{j\in\mathbb{Z}_{>0}}+(r_j)_{j\in\mathbb{Z}_{>0}}$*,*

(v) $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}\cdot(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} = (q_j)_{j\in\mathbb{Z}_{>0}}\cdot(r_j)_{j\in\mathbb{Z}_{>0}}$*, and*

(vi) if, for all $j\in\mathbb{Z}_{>0}$, $q_j, \tilde{q}_j\neq 0$ *and if the sequences* $(q_j)_{j\in\mathbb{Z}_{>0}}, (\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}$ *do not converge to* 0*, then* $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}$*.*

Proof (i) Let $\epsilon\in\mathbb{Q}_{>0}$ and let $N\in\mathbb{Z}_{>0}$ have the property that $|q_j-q_k|, |r_j-r_k| < \frac{\epsilon}{2}$ for all $j,k\geq N$. Then, using the triangle inequality,

$$|(q_j+r_j)-(q_k+r_k)| \leq |q_j-q_k|+|r_j-r_k| = \epsilon, \qquad j,k\geq N.$$

(ii) Let $M \in \mathbb{Q}_{>0}$ have the property that $|q_j|, |r_j| < M$ for all $j \in \mathbb{Z}_{>0}$. For $\epsilon \in \mathbb{Q}_{>0}$ let $N \in \mathbb{Z}_{>0}$ have the property that $|q_j - q_k|, |r_j - r_k| < \frac{\epsilon}{2M}$ for all $j, k \geq N$. Then, using the triangle inequality,

$$\begin{aligned}|(q_j \cdot r_j) - (q_k \cdot r_k)| &= |q_j(r_j - r_k) - r_k(q_k - q_j)| \\ &\leq |q_j||r_j - r_k| + |r_k||q_k - q_j| < \epsilon, \qquad j, k \geq N.\end{aligned}$$

(iii) We claim that if $(q_j)_{j\in\mathbb{Z}_{>0}}$ satisfies the conditions stated, then there exists $\delta \in \mathbb{Q}_{>0}$ such that $|q_k| \geq \delta$ for all $k \in \mathbb{Z}_{>0}$. Indeed, since $(q_j)_{j\in\mathbb{Z}_{>0}}$ does not converge to zero, choose $\epsilon \in \mathbb{Q}_{>0}$ such that, for all $N \in \mathbb{Z}_{>0}$, there exists $j \geq N$ for which $|q_j| \geq \epsilon$. Next take $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \frac{\epsilon}{2}$ for $j, k \geq N$. Then there exists $\tilde{N} \geq N$ such that $|q_{\tilde{N}}| \geq \epsilon$. For any $j \geq N$ we then have

$$|q_j| = |q_{\tilde{N}} - (q_{\tilde{N}} - q_j)| \geq ||q_{\tilde{N}}| - |q_{\tilde{N}} - q_j|| \geq \epsilon - \tfrac{\epsilon}{2} = \tfrac{\epsilon}{2},$$

where we have used Exercise 2.2.8. The claim follows by taking δ to be the smallest of the numbers $\frac{\epsilon}{2}, |q_1|, \dots, |q_N|$.

Now let $\epsilon \in \mathbb{Q}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \delta^2\epsilon$ for $j, k \geq N$. Then

$$|q_j^{-1} - q_k^{-1}| = \left|\frac{q_k - q_j}{q_j q_k}\right| < \frac{\delta^2\epsilon}{\delta^2} = \epsilon, \qquad j, k \geq N.$$

(iv) For $\epsilon \in \mathbb{Q}_{>0}$ let $N \in \mathbb{Z}_{>0}$ have the property that $|\tilde{q}_j - q_j|, |\tilde{r}_j - r_j| < \frac{\epsilon}{2}$. Then, using the triangle inequality,

$$|(\tilde{q}_j + \tilde{r}_j) - (q_k + r_k)| \leq |\tilde{q}_j - q_k| + |\tilde{r}_k - r_k| < \epsilon, \qquad j, k \geq N.$$

(v) Let $M \in \mathbb{Q}_{>0}$ have the property that $|\tilde{q}_j|, |r_j| < M$ for all $j \in \mathbb{Z}_{>0}$. Then, for $\epsilon \in \mathbb{Q}_{>0}$, take $N \in \mathbb{Z}_{>0}$ such that $|\tilde{r}_j - r_k|, |\tilde{q}_j - q_k| < \frac{\epsilon}{2M}$ for $j, k \geq N$. We then use the triangle inequality to give

$$|(\tilde{q}_j \cdot \tilde{r}_j) - (q_k \cdot r_k)| = |\tilde{q}_j(\tilde{r}_j - r_k) - r_k(q_k - \tilde{q}_j)| < \epsilon, \qquad j, k \geq N.$$

(vi) Let $\delta \in \mathbb{Q}_{>0}$ satisfy $|q_j|, |\tilde{q}_j| \geq \delta$ for all $j \in \mathbb{Z}_{>0}$. Then, for $\epsilon \in \mathbb{Q}_{>0}$, choose $N \in \mathbb{Z}_{>0}$ such that $|\tilde{q}_j - q_j| < \delta^2\epsilon$ for $j \geq N$. Then we have

$$|\tilde{q}_j^{-1} - q_j^{-1}| = \left|\frac{q_j - \tilde{q}_j}{q_j \tilde{q}_j}\right| < \frac{\delta^2\epsilon}{\delta^2}, \qquad j \geq N,$$

so completing the proof. ■

The requirement, in parts (iii) and (vi), that the sequence $(q_j)_{j\in\mathbb{Z}_{>0}}$ have no zero elements is not really a restriction in the same way as is the requirement that the sequence not converge to zero. The reason for this is that, as we showed in the proof, if the sequence does not converge to zero, then there exists $\epsilon \in \mathbb{Q}_{>0}$ and $N \in \mathbb{Z}_{>0}$ such that $|q_j| > \epsilon$ for $j \geq N$. Thus the tail of the sequence is guaranteed to have no zero elements, and the tail of the sequence is all that matters for the equivalence class.

Now that we have shown how to add and multiply Cauchy sequences in $\mathbb{Q}$, and that this addition and multiplication depends only on equivalence classes under the notion of equivalence given in Definition 2.1.16, we can easily define addition and multiplication in $\mathbb{R}$.

2.2.2 Definition (Addition, multiplication, and division in $\mathbb{R}$) Define the operations of ***addition***, ***multiplication***, and ***division*** in $\mathbb{R}$ by

(i) $[(q_j)_{j\in\mathbb{Z}_{>0}}] + [(r_j)_{j\in\mathbb{Z}_{>0}}] = [(q_j)_{j\in\mathbb{Z}_{>0}} + (r_j)_{j\in\mathbb{Z}_{>0}}]$,

(ii) $[(q_j)_{j\in\mathbb{Z}_{>0}}] \cdot [(r_j)_{j\in\mathbb{Z}_{>0}}] = [(q_j)_{j\in\mathbb{Z}_{>0}} \cdot (r_j)_{j\in\mathbb{Z}_{>0}}]$,

(iii) $[(q_j)_{j\in\mathbb{Z}_{>0}}]/[(r_j)_{j\in\mathbb{Z}_{>0}}] = [(q_j/r_j)_{j\in\mathbb{Z}_{>0}} + (r_j)_{j\in\mathbb{Z}_{>0}}]$,

respectively, where, in the definition of division, we require that the sequence $(r_j)_{j\in\mathbb{Z}_{>0}}$ have no zero elements, and that it not converge to 0. We will sometimes omit the "$\cdot$" when writing multiplication. •

Similarly to what we have done previously with $\mathbb{Z}$ and $\mathbb{Q}$, we let $-x = [(-1)_{j\in\mathbb{Z}_{>0}}]\cdot x$. For $x \in \mathbb{R}\setminus\{0\}$, we also denote by x^{-1} the real number corresponding to a Cauchy sequence $(\frac{1}{q_j})_{j\in\mathbb{Z}_{>0}}$, where $x = [(q_j)_{j\in\mathbb{Z}_{>0}}]$.

As with integers and rational numbers, we can define powers of real numbers. For $x \in \mathbb{R}\setminus\{0\}$ and $k \in \mathbb{Z}_{\geq 0}$ we define $x^k \in \mathbb{R}$ inductively by $x^0 = 1$ and $x^{k^+} = x^k \cdot x$. As usual, we call x^k the kth ***power*** of x. For $k \in \mathbb{Z}\setminus\mathbb{Z}_{\geq 0}$, we take $x^k = (x^{-k})^{-1}$. For real numbers, the notion of the power of a number can be extended. Let us show how this is done. In the statement of the result, we use the notion of positive real numbers which are not defined until Definition 2.2.8. Also, in our proof, we refer ahead to properties of $\mathbb{R}$ that are not considered until Section 2.3. However, it is convenient to state the construction here.

2.2.3 Proposition ($\mathbf{x^{1/k}}$) *For* $x \in \mathbb{R}_{>0}$ *and* $k \in \mathbb{Z}_{>0}$, *there exists a unique* $y \in \mathbb{R}_{>0}$ *such that* $y^k = x$. *We denote the number* y *by* $x^{1/k}$.

Proof Let $S_x = \{y \in \mathbb{R} \mid y^k < x\}$. Since $x \geq 0$, $0 \in S$ so $S \neq \emptyset$. We next claim that $\max\{1, x\}$ is an upper bound for S_x. First suppose that $x < 1$. Then, for $y \in S_x$, $y^k < x < 1$, and so 1 is an upper bound for S_x. If $x \geq 1$ and $y \in S_x$, then we claim that $y \leq x$. Indeed, if $y > x$ then $y^k > x^k > x$, and so $y \notin S_x$. This shows that S_x is upper bounded by x in this case. Now we know that S_x has a least upper bound by Theorem 2.3.7. Let y denote this least upper bound.

We shall now show that $y^k = x$. Suppose that $y^k \neq x$. From Corollary 2.2.9 we have $y^k < x$ or $y^k > x$.

Suppose first that $y^k < x$. Then, for $\epsilon \in \mathbb{R}_{>0}$ we have

$$(y+\epsilon)^k = \epsilon^k + a_{k-1}y\epsilon^{k-1} + \cdots + a_1y^{k-1}\epsilon + y^k$$

for some numbers $a_1, \dots, a_{k-1}$ (these are the binomial coefficients of Exercise 2.2.1). If $\epsilon \leq 1$ then $\epsilon^k \leq \epsilon$ for $k \in \mathbb{Z}_{>0}$. Therefore, if $\epsilon \leq 1$ we have

$$(y+\epsilon)^k \leq \epsilon(1 + a_{k-1}y + \cdots + a_1y^{k-1}) + y^k.$$

Now, if $\epsilon < \min\{1, \frac{x-y^k}{1+a_{k-1}y+\cdots+a_a y^{k-1}}\}$, then $(y+\epsilon)^k < x$, contradicting the fact that y is an upper bound for S_x.

Now suppose that $y^k > x$. Then, for $\epsilon \in \mathbb{R}_{>0}$, we have

$$(y-\epsilon)^k = (-1)^k\epsilon^k + (-1)^{k-1}a_{k-1}y\epsilon^{k-1} + \cdots - a_1y^{k-1}\epsilon + y^k.$$

The sum on the right involves terms that are positive and negative. This sum will be greater than the corresponding sum with the positive terms involving powers of ϵ removed. That is to say,

$$(y-\epsilon)^k > y^k - a_1 y^{k-1}\epsilon - a_3 y^{k-3}\epsilon^3 + \cdots .$$

For $\epsilon \le 1$ we again gave $\epsilon^k \le \epsilon$ for $k \in \mathbb{Z}_{>0}$. Therefore

$$(y-\epsilon)^k > y^k - (a_1 y^{k-1} + a_3 y^{k-3} + \cdots)\epsilon.$$

Thus, if $\epsilon < \min\{1, \frac{y^k - x}{a_1 y^{k-1} + a_3 y^{k-3} + \ldots}\}$ we have $(y-\epsilon)^k > x$, contradicting the fact that y is the least upper bound for S_x.

We are forced to conclude that $y^k = x$, so giving the result. ■

If $x \in \mathbb{R}_{>0}$ and $q = \frac{j}{k} \in \mathbb{Q}$ with $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$, we define $x^q = (x^{1/k})^j$.

Let us record the basic properties of addition and multiplication, mirroring analogous results for $\mathbb{Q}$. The properties all follow easily from the similar properties for $\mathbb{Q}$, along with Proposition 2.2.1 and the definition of addition and multiplication in $\mathbb{R}$.

2.2.4 Proposition (Properties of addition and multiplication in $\mathbb{R}$) *Addition and multiplication in $\mathbb{R}$ satisfy the following rules:*

(i) $x_1 + x_2 = x_2 + x_1$, $x_1, x_2 \in \mathbb{R}$ *(**commutativity** of addition);*

(ii) $(x_1 + x_2) + x_3 = x_1 + (x_2 + x_3)$, $x_1, x_2, x_3 \in \mathbb{R}$ *(**associativity** of addition);*

(iii) $x + 0 = x$, $t \in \mathbb{R}$ *(**additive identity**);*

(iv) $x + (-x) = 0$, $x \in \mathbb{R}$ *(**additive inverse**);*

(v) $x_1 \cdot x_2 = x_2 \cdot x_1$, $x_1, x_2 \in \mathbb{R}$ *(**commutativity** of multiplication);*

(vi) $(x_1 \cdot x_2) \cdot x_3 = x_1 \cdot (x_2 \cdot x_3)$, $x_1, x_2, x_3 \in \mathbb{R}$ *(**associativity** of multiplication);*

(vii) $x \cdot 1 = x$, $x \in \mathbb{R}$ *(**multiplicative identity**);*

(viii) $x \cdot x^{-1} = 1$, $x \in \mathbb{R} \setminus \{0\}$ *(**multiplicative inverse**);*

(ix) $y \cdot (x_1 + x_2) = y \cdot x_1 + y \cdot x_2$, $y, x_1, x_2 \in \mathbb{R}$ *(**distributivity**);*

(x) $x^{k_1} \cdot x^{k_2} = x^{k_1 + k_2}$, $x \in \mathbb{R}$, $k_1, k_2 \in \mathbb{Z}_{\ge 0}$.

Moreover, if we define $i_{\mathbb{Q}} \colon \mathbb{Q} \to \mathbb{R}$ *by* $i_{\mathbb{Q}}(q) = [(q)_{j \in \mathbb{Z}_{>0}}]$*, then addition and multiplication in* $\mathbb{R}$ *agrees with that in* $\mathbb{Q}$*:*

$$i_{\mathbb{Q}}(q_1) + i_{\mathbb{Q}}(q_2) = i_{\mathbb{Q}}(q_1 + q_2), \quad i_{\mathbb{Q}}(q_1) \cdot i_{\mathbb{Q}}(q_2) = i_{\mathbb{Q}}(q_1 \cdot q_2).$$

As we have done in the past with $\mathbb{Z} \subseteq \mathbb{Q}$, we will often regard $\mathbb{Q}$ as a subset of $\mathbb{R}$ without making explicit mention of the inclusion $i_{\mathbb{Q}}$. Note that this also allows us to think of both $\mathbb{Z}_{\ge 0}$ and $\mathbb{Z}$ as subsets of $\mathbb{R}$, since $\mathbb{Z}_{\ge 0}$ is regarded as a subset of $\mathbb{Z}$, and since $\mathbb{Z} \subseteq \mathbb{Q}$. Of course, this is nothing surprising. Indeed, perhaps the more surprising thing is that it is not actually the case that the definitions do not precisely give $\mathbb{Z}_{\ge 0} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$!

Now is probably a good time to mention that an element of $\mathbb{R}$ that is not in the image of $i_{\mathbb{Q}}$ is called ***irrational***. Also, one can show that the set $\mathbb{Q}$ of rational numbers is countable (Exercise 2.1.3), but that the set $\mathbb{R}$ of real numbers is uncountable (Exercise 2.1.4). Note that it follows that the set of irrational numbers is uncountable, since an uncountable set cannot be a union of two countable sets.

2.2.2 The total order on $\mathbb{R}$

Next we define in $\mathbb{R}$ a natural total order. To do so requires a little work. The approach we take is this. On the set $\mathsf{CS}(\mathbb{Q})$ of Cauchy sequences in $\mathbb{Q}$ we define a partial order that is *not* a total order. We then show that, for any two Cauchy sequences, in each equivalence class in $\mathsf{CS}(\mathbb{Q})$ with respect to the equivalence relation of Definition 2.1.16, there exists representatives that can be compared using the order. In this way, while the order on the set of Cauchy sequences is not a total order, there is induced a total order on the set of equivalence classes.

First we define the partial order on the set of Cauchy sequences.

2.2.5 Definition (Partial order on CS($\mathbb{Q}$)) The partial order $\preceq$ on $\mathsf{CS}(\mathbb{Q})$ is defined by

$$(q_j)_{j\in\mathbb{Z}_{>0}} \preceq (r_j)_{j\in\mathbb{Z}_{>0}} \quad\iff\quad q_j \le r_j,\ j\in\mathbb{Z}_{>0}. \qquad \bullet$$

This partial order is clearly not a total order. For example, the Cauchy sequences $(\frac{1}{j})_{j\in\mathbb{Z}_{>0}}$ and $(\frac{(-1)^j}{j})_{j\in\mathbb{Z}_{>0}}$ are not comparable with respect to this order. However, what is true is that equivalence classes of Cauchy sequences *are* comparable. We refer the reader to Definition 2.1.16 for the definition of the equivalence relation we denote by $\sim$ in the following result.

2.2.6 Proposition *Let* $(q_j)_{j\in\mathbb{Z}_{>0}}, (r_j)_{j\in\mathbb{Z}_{>0}} \in \mathsf{CS}(\mathbb{Q})$ *and suppose that* $(q_j)_{j\in\mathbb{Z}_{>0}} \nsim (r_j)_{j\in\mathbb{Z}_{>0}}$. *The following two statements hold:*

(i) There exists $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}, (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \in \mathsf{CS}(\mathbb{Q})$ *such that*

(a) $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$, *and*

(b) either $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \prec (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}}$ *or* $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \prec (\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}$.

(ii) There does not exist $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}, (\bar{q}_j)_{j\in\mathbb{Z}_{>0}}, (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}}, (\bar{r}_j)_{j\in\mathbb{Z}_{>0}} \in \mathsf{CS}(\mathbb{Q})$ *such that*

(a) $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (\bar{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (\bar{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$, *and*

(b) one of the following two statements holds:

I. $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \prec (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(\bar{r}_j)_{j\in\mathbb{Z}_{>0}} \prec (\bar{q}_j)_{j\in\mathbb{Z}_{>0}}$;

II. $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \prec (\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(\bar{q}_j)_{j\in\mathbb{Z}_{>0}} \prec (\bar{r}_j)_{j\in\mathbb{Z}_{>0}}$.

Proof (i) We begin with a useful lemma.

1 **Lemma** *With the given hypotheses, there exists* $\delta \in \mathbb{Q}_{>0}$ *and* $\mathrm{N} \in \mathbb{Z}_{>0}$ *such that* $|\mathrm{q_j} - \mathrm{r_j}| \geq \delta$ *for all* $\mathrm{j} \geq \mathrm{N}$*.*

Proof Since $(q_j - r_j)_{j\in\mathbb{Z}_{>0}}$ does not converge to zero, choose $\epsilon \in \mathbb{Q}_{>0}$ such that, for all $N \in \mathbb{Z}_{>0}$, there exists $j \geq N$ such that $|q_j - r_j| \geq \epsilon$. Now take $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k|, |r_k - r_k| \leq \frac{\epsilon}{4}$ for $j, k \geq N$. Then, by our assumption about ϵ, there exists $\tilde{N} \geq N$ such that $|q_{\tilde{N}} - r_{\tilde{N}}| \geq \epsilon$. Then, for any $j \geq N$, we have

$$\begin{aligned} |q_j - r_j| &= |(q_{\tilde{N}} - r_{\tilde{N}}) - (q_{\tilde{N}} - r_{\tilde{N}}) - (q_j - r_j)| \\ &\geq ||q_{\tilde{N}} - r_{\tilde{N}}| - |(q_{\tilde{N}} - r_{\tilde{N}}) - (q_j - r_j)|| \geq \epsilon - \tfrac{\epsilon}{2}. \end{aligned}$$

The lemma follows by taking $\delta = \frac{\epsilon}{2}$. ▼

Now take N and δ as in the lemma. Then take $\tilde{N} \in \mathbb{Z}_{>0}$ such that $|q_j - q_k|, |r_j - r_k| < \frac{\delta}{2}$ for $j, k \geq \tilde{N}$. Then, using the triangle inequality,

$$|(q_j - r_j) - (q_k - r_k)| \leq \delta, \qquad j, k \geq \tilde{N}.$$

Now take K to be the larger of N and $\tilde{N}$. We then have either $q_K - r_K \geq \delta$ or $r_K - q_K \geq \delta$. First suppose that $q_K - r_K \geq \delta$ and let $j \geq K$. Either $q_j - r_j \geq \delta$ or $r_j - q_j \geq \delta$. If the latter, then

$$q_j - r_j \leq -\delta \quad \Longrightarrow \quad (q_j - r_k) - (q_K - r_K) \leq 2\delta,$$

contradicting the definition of K. Therefore, we must have $q_j - r_j \geq \delta$ for all $j \geq K$. A similar argument when $r_K - q_K \geq \delta$ shows that $r_j - q_j \geq \delta$ for all $j \geq K$. For $j \in \mathbb{Z}_{>0}$ we then define

$$\tilde{q}_j = \begin{cases} q_K, & j < K, \\ q_j, & j \geq K, \end{cases} \qquad \tilde{r}_j = \begin{cases} r_K, & j < K, \\ r_j, & j \geq K, \end{cases},$$

and we note that the sequences $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}$ and $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}}$ satisfy the required conditions.

(ii) Suppose that

1. $(q_j)_{j\in\mathbb{Z}_{>0}} \nsim (r_j)_{j\in\mathbb{Z}_{>0}}$,
2. $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (\bar{q}_j)_{j\in\mathbb{Z}_{>0}} \sim (q_j)_{j\in\mathbb{Z}_{>0}}$,
3. $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (\bar{r}_j)_{j\in\mathbb{Z}_{>0}} \sim (r_j)_{j\in\mathbb{Z}_{>0}}$, and
4. $(\tilde{q}_j)_{j\in\mathbb{Z}_{>0}} \prec (\tilde{r}_j)_{j\in\mathbb{Z}_{>0}}$.

From the previous part of the proof we know that there exists $\delta \in \mathbb{Q}_{>0}$ and $N \in \mathbb{Z}_{>0}$ such that $\tilde{q}_j - \tilde{r}_j \geq \delta$ for $j \geq N$. Then take $\tilde{N} \in \mathbb{Z}_{>0}$ such that $|\tilde{q}_j - \bar{q}_j|, |\tilde{r}_j - \bar{r}_j| < \frac{\delta}{4}$ for $j \geq \tilde{N}$. This implies that for $j \geq \tilde{N}$ we have

$$|(\tilde{q}_j - \tilde{r}_j) - (\bar{q}_j - \bar{r}_j)| < \tfrac{\delta}{2}.$$

Therefore,

$$(\bar{q}_j - \bar{r}_j) > (\tilde{q}_j - \tilde{r}_j) - \tfrac{\delta}{2}, \qquad j \geq \tilde{N}.$$

If additionally $j \geq N$, then we have

$$(\bar{q}_j - \bar{r}_j) > \delta - \tfrac{\delta}{2} = \tfrac{\delta}{2}.$$

This shows the impossibility of $(\bar{r}_j)_{j\in\mathbb{Z}_{>0}} \prec (\bar{q}_j)_{j\in\mathbb{Z}_{>0}}$. A similar argument shows that $(\tilde{r}_j)_{j\in\mathbb{Z}_{>0}} \prec (\tilde{q}_j)_{j\in\mathbb{Z}_{>0}}$ bars the possibility that $(\bar{q}_j)_{j\in\mathbb{Z}_{>0}} \prec (\bar{r}_j)_{j\in\mathbb{Z}_{>0}}$. ■

Using the preceding result, the following definition then makes sense.

2.2.7 Definition (Order on $\mathbb{R}$) The total order on $\mathbb{R}$ is defined by $x \leq y$ if and only if there exists $(q_j)_{j\in\mathbb{Z}_{>0}}, (r_j)_{j\in\mathbb{Z}_{>0}} \in \mathrm{CS}(\mathbb{Q})$ such that

(i) $x = [(q_j)_{j\in\mathbb{Z}_{>0}}]$ and $y = [(r_j)_{j\in\mathbb{Z}_{>0}}]$ and

(ii) $(q_j)_{j\in\mathbb{Z}_{>0}} \preceq (r_j)_{j\in\mathbb{Z}_{>0}}$. •

Note that we have used the symbol "$\leq$" for the total order on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. This is justified since, if we think of $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$, then the various total orders agree (Exercises 2.1.2 and 2.2.6).

We have the usual language and notation we associate with various kinds of numbers.

2.2.8 Definition (Positive and negative real numbers) A real number x is:

(i) ***positive*** if $0 < x$;

(ii) ***negative*** if $x < 0$;

(iii) ***nonnegative*** if $0 \leq x$;

(iv) ***nonpositive*** if $x \leq 0$.

The set of positive real numbers is denoted by $\mathbb{R}_{>0}$, the set of nonnegative real numbers is denoted by $\mathbb{R}_{\geq 0}$, the set of negative real numbers is denoted by $\mathbb{R}_{<0}$, and the set of nonpositive real numbers is denoted by $\mathbb{R}_{\leq 0}$. •

Now is a convenient moment to introduce some simple notation and concepts that are associated with the natural total order on $\mathbb{R}$. The ***signum function*** is the map $\mathrm{sign}\colon \mathbb{R} \to \{-1, 0, 1\}$ defined by

$$\mathrm{sign}(x) = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

For $x \in \mathbb{R}$, $\lceil x \rceil$ is the ***ceiling*** of x which is the smallest integer not less than x. Similarly, $\lfloor x \rfloor$ is the ***floor*** of x which is the largest integer less than or equal to x. In Figure 2.1 we show the ceiling and floor functions.

A consequence of our definition of order is the following extension of the Trichotomy Law to $\mathbb{R}$.

2.2.9 Corollary (Trichotomy Law for $\mathbb{R}$) *For* $\mathrm{x}, \mathrm{y} \in \mathbb{R}$, *exactly one of the following possibilities holds:*

(i) $\mathrm{x} < \mathrm{y}$;

(ii) $\mathrm{y} < \mathrm{x}$;

(iii) $\mathrm{x} = \mathrm{y}$.

As with integers and rational numbers, addition and multiplication of real numbers satisfy the expected properties with respect to the total order.

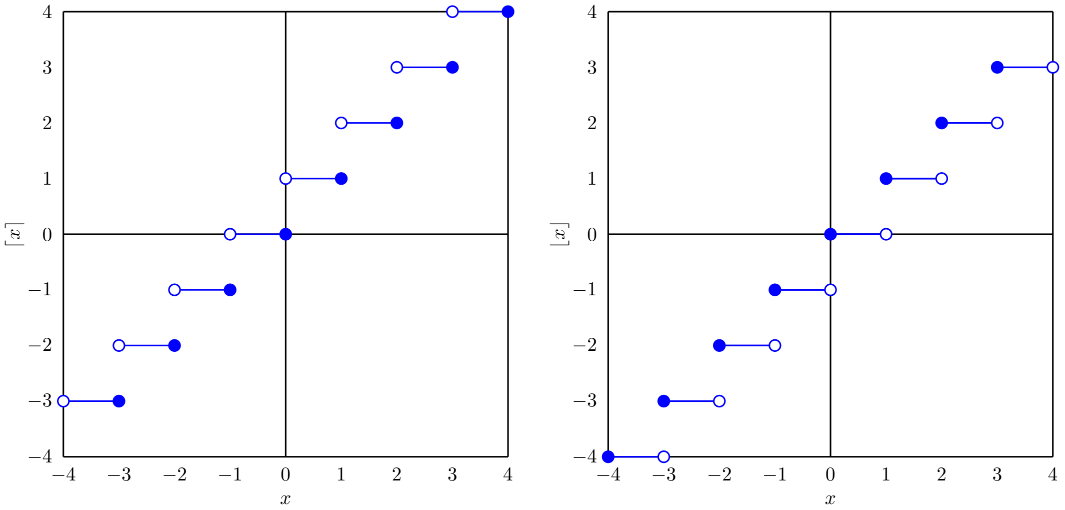


Figure 2.1 The ceiling function (left) and floor function (right)

2.2.10 Proposition (Relation between addition and multiplication and <) *For* $x, y, z \in \mathbb{R}$*, the following statements hold:*

(i) *if* $x < y$ *then* $x + z < y + z$*;*

(ii) *if* $x < y$ *and if* $z > 0$ *then* $z \cdot x < z \cdot y$*;*

(iii) *if* $x < y$ *and if* $z < 0$ *then* $z \cdot y < z \cdot x$*;*

(iv) *if* $0 < x, y$ *then* $0 < x \cdot y$*;*

(v) *if* $x < y$ *and if either*

(a) $0 < x, y$ *or*

(b) $x, y < 0$*,*

then $y^{-1} < x^{-1}$*.*

Proof These statements all follow from the similar statements for $\mathbb{Q}$, along with Proposition 2.2.6. We leave the straightforward verifications to the reader as Exercise 2.2.5. ■

2.2.3 The absolute value function on $\mathbb{R}$

In this section we generalise the absolute value function on $\mathbb{Q}$. As we shall see in subsequent sections, this absolute value function is essential for providing much of the useful structure of the set of real numbers.

The definition of the absolute value is given as usual.

2.2.11 Definition (Real absolute value function) The ***absolute value function*** on $\mathbb{R}$ is the map from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$, denoted by $x \mapsto |x|$, defined by

$$|x| = \begin{cases} x, & 0 < x, \\ 0, & x = 0, \\ -x, & x < 0. \end{cases} \qquad \bullet$$

Note that we have used the symbol "$|\cdot|$" for the absolute values on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. This is justified since, if we think of $\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$, then the various absolute value functions agree (Exercises 2.1.2 and 2.2.6).

The real absolute value function has the expected properties. The proof of the following result is straightforward, and so omitted.

2.2.12 Proposition (Properties of absolute value on $\mathbb{R}$) *The following statements hold:*

(i) $|x| \geq 0$ *for all* $x \in \mathbb{R}$;

(ii) $|x| = 0$ *if and only if* $x = 0$;

(iii) $|x \cdot y| = |x| \cdot |y|$ *for all* $x, y \in \mathbb{R}$;

(iv) $|x + y| \leq |x| + |y|$ *for all* $x, y \in \mathbb{R}$ *(**triangle inequality**)*;

(v) $|x^{-1}| = |x|^{-1}$ *for all* $x \in \mathbb{R} \setminus \{0\}$.

2.2.4 Properties of $\mathbb{Q}$ as a subset of $\mathbb{R}$

In this section we give some seemingly obvious, and indeed not difficult to prove, properties of the rational numbers as a subset of the real numbers.

The first property bears the name of Archimedes,[2] but Archimedes actually attributes this to Eudoxus.[3] In any case, it is an Ancient Greek property.

2.2.13 Proposition (Archimedean property of $\mathbb{R}$) *Let* $\epsilon \in \mathbb{R}_{>0}$. *Then, for any* $x \in \mathbb{R}$ *there exists* $k \in \mathbb{Z}_{>0}$ *such that* $k \cdot \epsilon > x$.

Proof Let $(q_j)_{j\in\mathbb{Z}_{>0}}$ and $(e_j)_{j\in\mathbb{Z}_{>0}}$ be Cauchy sequences in $\mathbb{Q}$ such that $x = [(q_j)_{j\in\mathbb{Z}_{>0}}]$ and $\epsilon = [(e_j)_{j\in\mathbb{Z}_{>0}}]$. By Proposition 2.1.14 there exists $M \in \mathbb{R}_{>0}$ such that $|q_j| < M$ for all $j \in \mathbb{Z}_{>0}$, and by Proposition 2.2.6 we may suppose that $e_j > \delta$ for $j \in \mathbb{Z}_{>0}$, for some $\delta \in \mathbb{Q}_{>0}$. Let $k \in \mathbb{Z}_{>0}$ satisfy $k > \frac{M+1}{\delta}$ (why is this possible?). Then we have

$$k \cdot e_j > \frac{M+1}{\delta} \cdot \delta = M + 1 \geq q_j + 1, \qquad j \in \mathbb{Z}_{>0}.$$

[2] Archimedes of Syracuse (287 BC–212 BC) was a Greek mathematician and physicist (although in that era such classifications of scientific aptitude were less rigid than they are today). Much of his mathematical work was in the area of geometry, but many of Archimedes' best known achievements were in physics (e.g., the Archimedean Principle in fluid mechanics). The story goes that when the Romans captured Syracuse in 212 BC, Archimedes was discovered working on some mathematical problem, and struck down in the act by a Roman soldier.

[3] Eudoxus of Cnidus (408 BC–355 BC) was a Greek mathematician and astronomer. His mathematical work was concerned with geometry and numbers.

Now consider the sequence $(k \cdot e_j - q_j)_{j\in\mathbb{Z}_{>0}}$. This is a Cauchy sequence by Proposition 2.2.1 since it is a sum of products of Cauchy sequences. Moreover, our computations show that each term in the sequence is larger than 1. Also, this Cauchy sequence has the property that $[(k \cdot e_j - q_j)_{j\in\mathbb{Z}_{>0}}] = k \cdot \epsilon - x$. This shows that $k \cdot \epsilon - x \in \mathbb{R}_{>0}$, so giving the result. ■

The Archimedean property roughly says that there are no real numbers which are greater all rational numbers. The next result says that there are no real numbers that are smaller than all rational numbers.

2.2.14 Proposition (There is no smallest positive real number) *If* $\epsilon \in \mathbb{R}_{>0}$ *then there exists* $\mathrm{q} \in \mathbb{Q}_{>0}$ *such that* $\mathrm{q} < \epsilon$*.*

Proof Since $\epsilon^{-1} \in \mathbb{R}_{>0}$ let $k \in \mathbb{Z}_{>0}$ satisfy $k \cdot 1 > \epsilon^{-1}$ by Proposition 2.2.13. Then taking $q = k^{-1} \in \mathbb{Q}_{>0}$ gives $q < \epsilon$. ■

Using the preceding two results, it is then easy to see that arbitrarily near any real number lies a rational number.

2.2.15 Proposition (Real numbers are well approximated by rational numbers I) *If* $\mathrm{x} \in \mathbb{R}$ *and if* $\epsilon \in \mathbb{R}_{>0}$*, then there exists* $\mathrm{q} \in \mathbb{Q}$ *such that* $|\mathrm{x} - \mathrm{q}| < \epsilon$*.*

Proof If $x = 0$ then the result follows by taking $q = 0$. Let us next suppose that $x > 0$. If $x < \epsilon$ then the result follows by taking $q = 0$, so we assume that $x \geq \epsilon$. Let $\delta \in \mathbb{Q}_{>0}$ satisfy $\delta < \epsilon$ by Proposition 2.2.14. Then use Proposition 2.2.13 to choose $k \in \mathbb{Z}_{>0}$ to satisfy $k \cdot \delta > x$. Moreover, since $x > 0$, we will assume that k is the smallest such number. Since $x \geq \epsilon$, $k \geq 2$. Thus $(k - 1) \cdot \delta \leq x$ since k is the smallest natural number for which $k \cdot \delta > x$. Now we compute

$$0 \leq x - (k-1) \cdot \delta < k \cdot \delta - (k-1) \cdot \delta = \delta < \epsilon.$$

It is now easy to check that the result holds by taking $q = (k-1) \cdot \delta$. The situation when $x < 0$ is easily shown to follow from the situation when $x > 0$. ■

The following stronger result is also useful, and can be proved along the same lines as Proposition 2.2.15, using the Archimedean property of $\mathbb{R}$. The reader is asked to do this as Exercise 2.2.4.

2.2.16 Corollary (Real numbers are well approximated by rational numbers II) *If* $\mathrm{x}, \mathrm{y} \in \mathbb{R}$ *with* $\mathrm{x} < \mathrm{y}$*, then there exists* $\mathrm{q} \in \mathbb{Q}$ *such that* $\mathrm{x} < \mathrm{q} < \mathrm{y}$*.*

One can also show that irrational numbers have the same property.

2.2.17 Proposition (Real numbers are well approximated by irrational numbers) *If* $\mathrm{x} \in \mathbb{R}$ *and if* $\epsilon \in \mathbb{R}_{>0}$*, then there exists* $\mathrm{y} \in \mathbb{R} \setminus \mathbb{Q}$ *such that* $|\mathrm{x} - \mathrm{y}| < \epsilon$*.*

Proof By Corollary 2.2.16 choose $q_1, q_2 \in \mathbb{Q}$ such that $x - \epsilon < q_1 < q_2 < x + \epsilon$. Then the number

$$y = q_1 + \frac{q_2 - q_1}{\sqrt{2}}$$

is irrational and satisfies $q_1 < y < q_2$. Therefore, $x - \epsilon < y < x + \epsilon$, or $|x - y| < \epsilon$. ■

It is also possible to state a result regarding the approximation of a collection of real numbers by rational numbers of a certain form. The following result gives one such result.

2.2.18 Theorem (Dirichlet Simultaneous Approximation Theorem) *If* $x_1, \dots, x_k \in \mathbb{R}$ *and if* $N \in \mathbb{Z}_{>0}$*, then there exists* $m \in \{1, \dots, N^k\}$ *and* $m_1, \dots, m_k \in \mathbb{Z}$ *such that*

$$\max\{|mx_1 - m_1|, \dots, |mx_k - m_k|\} < \frac{1}{N}.$$

Proof Let

$$C = [0,1)^k \subseteq \mathbb{R}^k$$

be the "cube" in $\mathbb{R}^k$. For $j \in \{1, \dots, N\}$ denote $I_j = [\frac{j-1}{N}, \frac{j}{N})$ and note that the sets

$$\{I_{j_1} \times \cdots \times I_{j_k} \subseteq C \mid j_1, \dots, j_k \in \{1, \dots, N\}\}$$

form a partition of the cube C into N^k "subcubes." Now consider the $N^k + 1$ points

$$\{(lx_1, \dots, lx_k) \mid l \in \{0, 1, \dots, N^k\}\}$$

in $\mathbb{R}^k$. If $\lfloor x \rfloor$ denotes the floor of $x \in \mathbb{R}$ (i.e., the largest integer less than or equal to x), then

$$\{(lx_1 - \lfloor lx_1 \rfloor, \dots, lx_k - \lfloor lx_k \rfloor) \mid l \in \{0, 1, \dots, N^k\}\}$$

is a collection of $N^k + 1$ numbers in C. Since C is partitioned into the N^k cubes, it must be that at least two of these $N^k + 1$ points lie in the same cube. Let these points correspond to $l_1, l_2 \in \{0, 1, \dots, n^k\}$ with $l_2 > l_1$. Then, letting $m = l_2 - l_2$ and $m_j = \lfloor l_2 x_j \rfloor - \lfloor l_1 x_j \rfloor$, $j \in \{1, \dots, k\}$, we have

$$|mx_j - m_j| = |l_2 - \lfloor l_2 x_j \rfloor - (l_1 x_j - \lfloor l_1 x_j \rfloor)| < \frac{1}{N}$$

for every $j \in \{1, \dots, k\}$, which is the result since $m \in \{1, \dots, N^k\}$. ■

2.2.19 Remark (Dirichlet's "pigeonhole principle") The proof of the preceding theorem is a clever application of the so-called "pigeonhole principle," whose use seems to have been pioneered by Dirichlet. The idea behind this principle is simple. One uses the problem data to define elements $x_1, \dots, x_m$ of some set S. One then constructs a partition $(S_1, \dots, S_k)$ of S with the property that, if any $x_{j_1}, x_{j_2} \in S_l$ for some $l \in \{1, \dots, k\}$ and some $j_1, j_2 \in \{1, \dots, m\}$, then the desired result holds. If $k > m$ this is automatically satisfied. •

Note that the previous result gives an arbitrarily accurate simultaneous approximation of the numbers $x_1, \dots, x_j$ by rational numbers with the same denominator since we have

$$\left| x_j - \frac{m_j}{m} \right| < \frac{1}{mN^k} \leq \frac{1}{N^{k+1}}.$$

By choosing N large, our simultaneous approximations can be made as good as desired.

Let us now ask a somewhat different sort of question. Given a fixed set $a_1, \dots, a_k \in \mathbb{R}$, what are the conditions on these numbers such that, given *any* set $x_1, \dots, x_k \in \mathbb{R}$, we can find another number $b \in \mathbb{R}$ such that the approximations $|ba_j - x_j|$, $j \in \{1, \dots, k\}$, are arbitrarily close to integer multiples of a certain number. The exact reason why this is interesting is not immediately clear, but becomes clear in Theorem II-3.2.7 when we talk about the geometry of the unit circle in the complex plane. In any event, the following result addresses this approximation question, making reference to the notion of linear independence which we discuss in Section 4.5.3. In the statement of the theorem, we think of $\mathbb{R}$ as being a $\mathbb{Q}$-vector space.

2.2.20 Theorem (Kronecker Approximation Theorem) *For* $\mathrm{a}_1, \dots, \mathrm{a}_k \in \mathbb{R}$ *and* $\Delta \in \mathbb{R}_{>0}$ *the following statements hold:*

(i) if $\{\mathrm{a}_1, \dots, \mathrm{a}_k\}$ *are linearly over* $\mathbb{Q}$ *then, for any* $\mathrm{x}_1, \dots, \mathrm{x}_k \in \mathbb{R}$*, for any* $\epsilon \in \mathbb{R}_{>0}$ *and for any* $\mathrm{N} \in \mathbb{Z}_{>0}$*, there exists* $\mathrm{b} \in \mathbb{R}$ *with* $\mathrm{b} > \mathrm{N}$ *and integers* $\mathrm{m}_1, \dots, \mathrm{m}_k$ *such that*

$$\max\{|\mathrm{b}\mathrm{a}_1 - \mathrm{x}_1 - \mathrm{m}_1\Delta|, \dots, |\mathrm{b}\mathrm{a}_k - \mathrm{x}_k - \mathrm{m}_k\Delta|\} < \epsilon;$$

(ii) if $\{\Delta, \mathrm{a}_1, \dots, \mathrm{a}_k\}$ *are linearly over* $\mathbb{Q}$ *then, for any* $\mathrm{x}_1, \dots, \mathrm{x}_k \in \mathbb{R}$*, for any* $\epsilon \in \mathbb{R}_{>0}$*, and for any* $\mathrm{N} \in \mathbb{Z}_{>0}$*, there exists* $\mathrm{b} \in \mathbb{Z}$ *with* $\mathrm{b} > \mathrm{N}$ *and integers* $\mathrm{m}_1, \dots, \mathrm{m}_k$ *such that*

$$\max\{|\mathrm{b}\mathrm{a}_1 - \mathrm{x}_1 - \mathrm{m}_1\Delta|, \dots, |\mathrm{b}\mathrm{a}_k - \mathrm{x}_k - \mathrm{m}_k\Delta|\} < \epsilon.$$

Proof Let us first suppose that $\Delta = 1$.

We prove the two assertions together, using induction on k.

First we prove (i) for $k = 1$. Thus suppose that $\{a_1\} \neq \{0\}$. Let $x_1 \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, and let $N \in \mathbb{Z}_{>0}$. If m_1 is an integer greater than N and if $b = a_1^{-1}(x_1 + m_1)$, then we have $ba_1 - x_1 - m_1 = 0$, giving the result in this case.

Next we prove that if (i) holds for $k = r$ then (ii) also holds for $k = r$. Thus suppose that $\{1, a_1, \dots, a_r\}$ are linearly independent over $\mathbb{Q}$. Let $x_1, \dots, x_r \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, and let $N \in \mathbb{Z}_{>0}$. By the Dirichlet Simultaneous Approximation Theorem, let $m, m'_1, \dots, m'_r \in \mathbb{Z}$ with $m \in \mathbb{Z}_{>0}$ be such that

$$|ma_j - m'_j| < \frac{\epsilon}{2}, \qquad j \in \{1, \dots, r\}.$$

We claim that $\{ma_1 - m'_1, \dots, ma_r - m'_r\}$ are linearly independent over $\mathbb{Q}$. Indeed, suppose that

$$q_1(ma_1 - m'_1) + \dots + q_r(ma_r - m'_r) = 0$$

for some $q_1, \dots, q_r \in \mathbb{Q}$. Then we have

$$(mq_1)a_1 + \dots + (mq_r)a_r) - (m'_1q_1 + \dots + m'_rq_r)1 = 0.$$

By linear independence of $\{1, a_1, \dots, a_r\}$ over $\mathbb{Q}$ it follows that $mq_j = 0$, $j \in \{1, \dots, r\}$, and so $q_j = 0$, $j \in \{1, \dots, r\}$, giving the desired linear independence. Since $\{ma_1 -$

$m'_1, \dots, ma_r - m'_r\}$ are linearly independent over $\mathbb{Q}$, we may use our assumption that (i) holds for $k = r$ to give the existence of $b' \in \mathbb{R}$ with $b' > N + 1$ and integers $m''_1, \dots, m''_r$ such that

$$|b'(ma_j - m'_j) - x_j - m''_j| < \frac{\epsilon}{2}, \qquad j \in \{1, \dots, r\}.$$

Now let $b = \lfloor b' \rfloor m > N$ and $m_j = m''_j + \lfloor b' \rfloor m'_j$, $j \in \{1, \dots, k\}$. Using the triangle inequality we have

$$\begin{aligned}|ba_j - x_j - m_j| &= |\lfloor b'm \rfloor a_j - x_j - (m''_j + \lfloor b' \rfloor m'_j)| \\ &= |\lfloor b' \rfloor (ma_j - m'_j) - x_j - m''_j| \\ &= |(\lfloor b' \rfloor - b')(ma_j - m'_j) + b'(ma_j - m'_j) - x_j - m''_j| \\ &\leq |(\lfloor b' \rfloor - b')(ma_j - m'_j)| + |b'(ma_j - m'_j) - x_j - m''_j| < \epsilon,\end{aligned}$$

as desired.

Now we prove that (ii) with $k = r$ implies (i) with $k = r + 1$. Thus let $a_1, \dots, a_{r+1}$ be linearly independent over $\mathbb{Q}$. Let $x_1, \dots, x_{r+1} \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, and let $N \in \mathbb{Z}_{>0}$. Note that linear independence implies that $a_{r+1} \neq 0$ (see Proposition 4.5.19(ii)). We claim that $\{1, \frac{a_1}{a_{r+1}}, \dots, \frac{a_r}{a_{r+1}}\}$ are linearly independent over $\mathbb{Q}$. Since (ii) holds for $k = r$ there exists $b' \in \mathbb{Z}$ with $b' > N$ and integers $m'_1, \dots, m'_r$ such that

$$\left| b' \frac{a_j}{a_{r+1}} - \left(x_j - x_{r+1} \frac{a_j}{a_{r+1}} \right) - m'_j \right| < \epsilon, \qquad j \in \{1, \dots, r\}.$$

Rewriting this as

$$\left| \left(\frac{b' + x_{r+1}}{a_{r+1}} \right) a_j - x_j - m'_j \right| < \epsilon, \qquad j \in \{1, \dots, r\},$$

and noting that

$$\left(\frac{b' + x_{r+1}}{a_{r+1}} \right) a_{r+1} - x_{r+1} - b' = 0,$$

which gives (i) by taking

$$b = \frac{b' + x_{r+1}}{a_{r+1}}, \; m_1 = m'_1, \; \dots, \; m_r = m'_r, \; m_{r+1} = b'.$$

The above induction arguments give the theorem with $\Delta = 1$. Now let us relax the assumption that $\Delta = 1$. Thus let $\Delta \in \mathbb{R}_{>0}$. Let us define $a'_j = \Delta^{-1} a_j$, $j \in \{1, \dots, k\}$. We claim that $\{a'_1, \dots, a'_k\}$ is linearly independent over $\mathbb{Q}$ if $\{a_1, \dots, a_k\}$ is linearly independent over $\mathbb{Q}$. Indeed, suppose that

$$q_1 a'_1 + \dots + q_k a'_k = 0$$

for some $q_1, \dots, q_k \in \mathbb{Q}$. Multiplying by Δ and using the linear independence of $\{a_1, \dots, a_k\}$ immediately gives $q_j = 0$, $j \in \{1, \dots, k\}$. We also claim that $\{1, a'_1, \dots, a'_k\}$ is linearly independent over $\mathbb{Q}$ if $\{\Delta, a_1, \dots, a_k\}$ is linearly independent over $\mathbb{Q}$. Indeed, suppose that

$$q_0 \, 1 + q_1 a'_1 + \dots + q_k a'_k = 0$$

for some $q_0, q_1, \dots, q_k \in \mathbb{Q}$. Multiplying by Δ and using the linear independence of $\{\Delta, a_1, \dots, a_k\}$ immediately gives $q_j = 0$, $j \in \{1, \dots, k\}$. Let $x_1, \dots, x_k \in \mathbb{R}$, $\epsilon \in \mathbb{R}_{>0}$, and $N \in \mathbb{Z}$. Define $x'_j = \Delta^{-1}x_j$, $j \in \{1, \dots, k\}$. Since the theorem holds for $\Delta = 1$, there exists $b > N$ (with $b \in \mathbb{R}$ for part (i) and $b \in \mathbb{Z}$ for part (ii)) such that

$$|ba'_j - x'_j - m_1| < \frac{\epsilon}{\Delta}, \qquad j \in \{1, \dots, k\}.$$

Multiplying the inequality by Δ gives the result. ■

2.2.5 The extended real line

It is sometimes convenient to be able to talk about the concept of “infinity” in a somewhat precise way. We do so by using the following idea.

2.2.21 Definition (Extended real line) The ***extended real line*** is the set $\mathbb{R} \cup \{-\infty\} \cup \{\infty\}$, and we denote this set by $\overline{\mathbb{R}}$. •

Note that in this definition the symbols “$-\infty$” and “∞” are to simply be thought of as labels given to the elements of the singletons $\{-\infty\}$ and $\{\infty\}$. That they somehow correspond to our ideas of what “infinity” means is a consequence of placing some additional structure on $\overline{\mathbb{R}}$, as we now describe.

First we define “arithmetic” in $\overline{\mathbb{R}}$. We can also define some rules for arithmetic in $\overline{\mathbb{R}}$.

2.2.22 Definition (Addition and multiplication in $\overline{\mathbb{R}}$) For $x, y \in \overline{\mathbb{R}}$, define

$$x + y = \begin{cases} x + y, & x, y \in \mathbb{R}, \\ \infty, & x \in \mathbb{R},\ y = \infty, \text{ or } x = \infty,\ y \in \mathbb{R}, \\ \infty, & x = y = \infty, \\ -\infty, & x = -\infty,\ y \in \mathbb{R} \text{ or } x \in \mathbb{R},\ y = -\infty, \\ -\infty, & x = y = -\infty. \end{cases}$$

The operations $\infty + (-\infty)$ and $(-\infty) + \infty$ are undefined. Also define

$$x \cdot y = \begin{cases} x \cdot y, & x, y \in \mathbb{R}, \\ \infty, & x \in \mathbb{R}_{>0},\ y = \infty, \text{ or } x = \infty,\ y \in \mathbb{R}_{>0}, \\ \infty, & x \in \mathbb{R}_{<0},\ y = -\infty, \text{ or } x = -\infty,\ y \in \mathbb{R}_{<0}, \\ \infty, & x = y = \infty, \text{ or } x = y = -\infty, \\ -\infty, & x \in \mathbb{R}_{>0},\ y = -\infty, \text{ or } x = -\infty,\ y \in \mathbb{R}_{>0}, \\ -\infty, & x \in \mathbb{R}_{<0},\ y = \infty, \text{ or } x = \infty,\ y \in \mathbb{R}_{<0}, \\ -\infty, & x = \infty,\ y = -\infty \text{ or } x = -\infty,\ y = \infty, \\ 0, & x = 0, y \in \{-\infty, \infty\} \text{ or } x \in \{-\infty, \infty\},\ y = 0. \end{cases}$$ •

2.2.23 Remarks (Algebra in $\overline{\mathbb{R}}$)

1. The above definitions of addition and multiplication on $\overline{\mathbb{R}}$ *do not* make this a field. Thus, in some sense, the operations are simply notation, since they do not have the usual properties we associate with addition and multiplication.
2. Note we *do* allow multiplication between 0 and $-\infty$ and ∞. This convention is not universally agreed upon, but it will be useful for us to do adopt this convention in Chapter III-2. •

2.2.24 Definition (Order on $\overline{\mathbb{R}}$) For $x, y \in \overline{\mathbb{R}}$, write

$$x \leq y \qquad \Longleftrightarrow \qquad \begin{cases} x = y, & \text{or} \\ x, y \in \mathbb{R},\ x \leq y, & \text{or} \\ x \in \mathbb{R},\ y = \infty, & \text{or} \\ x = -\infty,\ y \in \mathbb{R}, & \text{or} \\ x = -\infty,\ y = \infty. & \end{cases} \qquad \bullet$$

This is readily verified to be a total order on $\overline{\mathbb{R}}$, with $-\infty$ being the least element and ∞ being the greatest element of $\overline{\mathbb{R}}$. As with $\mathbb{R}$, we have the notation

$$\overline{\mathbb{R}}_{>0} = \{x \in \overline{\mathbb{R}} \mid x > 0\}, \quad \overline{\mathbb{R}}_{\geq 0} = \{x \in \overline{\mathbb{R}} \mid x \geq 0\}.$$

Finally, we can extend the absolute value on $\mathbb{R}$ to $\overline{\mathbb{R}}$.

2.2.25 Definition (Extended real absolute value function) The ***extended real absolute function*** is the map from $\overline{\mathbb{R}}$ to $\overline{\mathbb{R}}_{\geq 0}$, denoted by $x \mapsto |x|$, and defined by

$$|x| = \begin{cases} |x|, & x \in \mathbb{R}, \\ \infty, & x = \infty, \\ \infty, & x = -\infty. \end{cases} \qquad \bullet$$

2.2.6 sup and inf

We recall from Definition 1.5.11 the notation $\sup S$ and $\inf S$ for the least upper bound and greatest lower bound, respectively, associated to a partial order. This construction applies, in particular to the partially ordered set $(\overline{\mathbb{R}}, \leq)$. Note that if $A \subseteq \mathbb{R}$ then we might possibly have $\sup(A) = \infty$ and/or $\inf(A) = -\infty$. In brief section we give a few properties of sup and inf.

The following property of sup and inf is often useful.

2.2.26 Lemma (Property of sup and inf) *Let* $\mathrm{A} \subseteq \mathbb{R}$ *be such that* $\inf(\mathrm{A}), \sup(\mathrm{A}) \in \mathbb{R}$ *and let* $\epsilon \in \mathbb{R}_{>0}$*. Then there exists* $\mathrm{x}_+, \mathrm{x}_- \in \mathrm{A}$ *such that*

$$\mathrm{x}_+ + \epsilon > \sup(\mathrm{A}), \quad \mathrm{x}_- - \epsilon < \inf(\mathrm{A}).$$

Proof We prove the assertion for sup, as the assertion for inf follows along similar lines, of course. Suppose that there is no $x_+ \in A$ such that $x_+ + \epsilon > \sup(A)$. Then $x \le \sup(A) - \epsilon$ for every $x \in A$, and so $\sup(A) - \epsilon$ is an upper bound for A. But this contradicts $\sup(A)$ being the least upper bound. ■

Let us record and prove the properties of interest for sup.

2.2.27 Proposition (Properties of sup) *For subsets* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for* $\mathrm{a} \in \mathbb{R}_{>0}$*, the following statements hold:*

(i) *if* $\mathrm{A} + \mathrm{B} = \{\mathrm{x} + \mathrm{y} \mid \mathrm{x} \in \mathrm{A}, \ \mathrm{y} \in \mathrm{B}\}$*, then* $\sup(\mathrm{A} + \mathrm{B}) = \sup(\mathrm{A}) + \sup(\mathrm{B})$*;*

(ii) *if* $-\mathrm{A} = \{-\mathrm{x} \mid \mathrm{x} \in \mathrm{A}\}$*, then* $\sup(-\mathrm{A}) = -\inf(\mathrm{A})$*;*

(iii) *if* $\mathrm{aA} = \{\mathrm{ax} \mid \mathrm{x} \in \mathrm{A}\}$*, then* $\sup(\mathrm{aA}) = \mathrm{a}\sup(\mathrm{A})$*;*

(iv) *if* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval, if* $\mathrm{A} \subseteq \mathrm{I}$ *is such that* $\sup \mathrm{A} \in \mathrm{A}$*, if* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$ *is strictly monotonically (see Definition 3.1.27), and if* $\mathrm{f}(\mathrm{A}) = \{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in \mathrm{A}\}$*, then* $\sup(\mathrm{f}(\mathrm{A})) = \mathrm{f}(\sup(\mathrm{A}))$*.*

Proof (i) Let $x \in A$ and $y \in B$ so that $x + y \in A + B$. Then $x + y \le \sup A + \sup B$ which implies that $\sup A + \sup B$ is an upper bound for $A + B$. Since $\sup(A + B)$ is the least upper bound this implies that $\sup(A + B) \le \sup A + \sup B$. Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in A$ and $y \in B$ satisfy $\sup A - x < \frac{\epsilon}{2}$ and $\sup B - y < \frac{\epsilon}{2}$. Then

$$\sup A + \sup B - (x + y) < \epsilon.$$

Thus, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $x + y \in A + B$ such that $\sup A + \sup B - (x + y) < \epsilon$. Therefore, $\sup A + \sup B \le \sup(A + B)$.

(ii) Let $x \in -A$. Then $\sup(-A) \ge x$ or $-\sup(-A) \le -x$. Thus $-\sup(-A)$ is a lower bound for A and so $\inf(A) \ge -\sup(-A)$. Next let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in -A$ satisfy $x + \epsilon > \sup(-A)$. Then $-x - \epsilon < -\sup(-A)$. Thus, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $y \in A$ such that $y - (-\sup(-A)) < \epsilon$. Thus $-\sup(-A) \ge \inf(A)$, giving this part of the result.

(iii) Let $x \in A$ and note that since $\sup(A) \ge x$, we have $a\sup(A) \ge ax$. Thus $a\sup(A)$ is an upper bound for aA, and so we must have $\sup(aA) \le a\sup(A)$. Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in A$ be such that $x + \frac{\epsilon}{a} > \sup(A)$. Then $ax + \epsilon > a\sup(A)$. Thus, given $\epsilon \in \mathbb{R}_{>0}$ there exists $y \in aA$ such that $a\sup(A) - ax < \epsilon$. Thus $a\sup(A) \le \sup(aA)$.

(iv) Let $x \in A$. Since $x \le \sup A$ and since f is strictly monotonically increasing, $f(x) \le f(\sup(A))$. Thus $f(\sup(A))$ is an upper bound for $f(A)$. Since $\sup(f(A))$ is the least upper bound, $\sup(f(A)) \le f(\sup(A))$. Suppose that $\sup(f(A)) < f(\sup(A))$. Since $\sup(A) \in A$, we have $f(\sup(A)) \in f(A)$ and so $f(\sup(A)) \le \sup(f(A))$. ■

For inf the result is, of course, quite similar. We leave the proof, which mirrors the above proof for sup, to the reader.

2.2.28 Proposition (Properties of inf) *For subsets* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for* $\mathrm{a} \in \mathbb{R}_{\geq 0}$*, the following statements hold:*

(i) if $\mathrm{A} + \mathrm{B} = \{\mathrm{x} + \mathrm{y} \mid \mathrm{x} \in \mathrm{A},\ \mathrm{y} \in \mathrm{B}\}$*, then* $\inf(\mathrm{A} + \mathrm{B}) = \inf(\mathrm{A}) + \inf(\mathrm{B})$*;*

(ii) if $-\mathrm{A} = \{-\mathrm{x} \mid \mathrm{x} \in \mathrm{A}\}$*, then* $\inf(-\mathrm{A}) = -\sup(\mathrm{A})$*;*

(iii) if $\mathrm{aA} = \{\mathrm{ax} \mid \mathrm{x} \in \mathrm{A}\}$*, then* $\inf(\mathrm{aA}) = \mathrm{a}\inf(\mathrm{A})$*;*

(iv) if $\mathrm{I} \subseteq \mathbb{R}$ *is an interval, if* $\mathrm{A} \subseteq \mathrm{I}$ *is such that* $\inf(\mathrm{A}) \in \mathrm{A}$*, if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is strictly monotonically (see Definition 3.1.27), and if* $\mathrm{f}(\mathrm{A}) = \{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in \mathrm{A}\}$*, then* $\inf(\mathrm{f}(\mathrm{A})) = \mathrm{f}(\inf(\mathrm{A}))$.

If $S \subseteq \mathbb{R}$ is a *finite* set, then both $\sup S$ and $\inf S$ are elements of S. In this case we might denote $\max S = \sup S$ and $\min S = \inf S$.

2.2.7 Notes

The Archimedean property of $\mathbb{R}$ seems obvious. The lack of the Archimedean property would mean that there exists t for which $t > N$ for every natural number N. This property is actually possessed by certain fields used in so-called "nonstandard analysis," and we refer the interested reader to [Robinson 1974].

Theorem 2.2.18 is due to Dirichlet [1842], and the proof is a famous use of the "pigeonhole principle." Theorem 2.2.20 is due to [Kronecker 1899], and the proof we give is from [Kueh 1986].

Exercises

2.2.1 Prove the ***Binomial Theorem*** which states that, for $x, y \in \mathbb{R}$ and $k \in \mathbb{Z}_{>0}$,

$$(x + y)^k = \sum_{j=0}^{k} B_{k,j} x^j y^{k-j},$$

where

$$B_{k,j} = \binom{k}{j} \triangleq \frac{k!}{j!(k-j)!}, \qquad j, k \in \mathbb{Z}_{>0},\ j \leq k,$$

are the ***binomial coefficients***, and $k! = 1 \cdot 2 \cdot \dots \cdot k$ is the ***factorial*** of k. We take the convention that $0! = 1$.

2.2.2 Prove that, for $k \in \mathbb{Z}_{>0}$ and $j \in \{0, 1, \dots, k\}$,

$$\binom{k}{j} + \binom{k}{j-1} = \binom{k+1}{j}.$$

2.2.3 Let $q \in \mathbb{Q} \setminus \{0\}$ and $x \in \mathbb{R} \setminus \mathbb{Q}$. Show the following:

(a) $q + x$ is irrational;

(b) qx is irrational;

(c) $\frac{x}{q}$ is irrational;

(d) $\frac{q}{x}$ is irrational.

2.2.4 Prove Corollary 2.2.16.

2.2.5 Prove Proposition 2.2.10.

2.2.6 Show that the order and absolute value on $\mathbb{R}$ agree with those on $\mathbb{Q}$. That is to say, show the following:

(a) for $q, r \in \mathbb{Q}$, $q < r$ if and only if $i_{\mathbb{Q}}(q) < i_{\mathbb{Q}}(r)$;

(b) for $q \in \mathbb{Q}$, $|q| = |i_{\mathbb{Q}}(q)|$.

(Note that we see clearly here the abuse of notation that follows from using $<$ for both the order on $\mathbb{Z}$ and $\mathbb{Q}$ and from using $|\cdot|$ as the absolute value both on $\mathbb{Z}$ and $\mathbb{Q}$. It is expected that the reader can understand where the notational abuse occurs.)

2.2.7 Do the following:

(a) show that if $x \in \mathbb{R}_{>0}$ satisfies $x < 1$, then $x^k < x$ for each $k \in \mathbb{Z}_{>0}$ satisfying $k \geq 2$;

(b) show that if $x \in \mathbb{R}_{>0}$ satisfies $x > 1$, then $x^k > x$ for each $k \in \mathbb{Z}_{>0}$ satisfying $k \geq 2$.

2.2.8 Show that, for $t, s \in \mathbb{R}$, $||t| - |s|| \leq |t - s|$.

2.2.9 Show that if $s, t \in \mathbb{R}$ satisfy $s < t$, then there exists $q \in \mathbb{Q}$ such that $s < q < t$.

Section 2.3

Sequences in $\mathbb{R}$

In our construction of the real numbers, sequences played a key rôle, inasmuch as Cauchy sequences of rational numbers were integral to our definition of real numbers. In this section we study sequences of real numbers. In particular, in Theorem 2.3.5 we prove the result, absolutely fundamental in analysis, that $\mathbb{R}$ is "complete," meaning that Cauchy sequences of real numbers converge.

Do I need to read this section? If you do not already know the material in this section, then it ought to be read. It is also worth the reader spending some time over the idea that Cauchy sequences of real numbers converge, as compared to rational numbers where this is not the case. The same idea will arise in more abstract settings in Chapter III-6, and so it will pay to understand it well in the simplest case. •

2.3.1 Definitions and properties of sequences

In this section we consider the extension to $\mathbb{R}$ of some of the ideas considered in Section 2.1.2 concerning sequences in $\mathbb{Q}$. As we shall see, it is via sequences, and other equivalent properties, that the nature of the difference between $\mathbb{Q}$ and $\mathbb{R}$ is spelled out quite clearly.

We begin with definitions, generalising in a trivial way the similar definitions for $\mathbb{Q}$.

2.3.1 Definition (Cauchy sequence, convergent sequence, bounded sequence, monotone sequence) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}$. The sequence:

(i) is a ***Cauchy sequence*** if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|x_j - x_k| < \epsilon$ for $j, k \geq N$;

(ii) ***converges to*** $\mathbf{s_0}$ if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|x_j - s_0| < \epsilon$ for $j \geq N$;

(iii) ***diverges*** if it does not converge to any element in $\mathbb{R}$;

(iv) is ***bounded above*** if there exists $M \in \mathbb{R}$ such that $x_j < M$ for each $j \in \mathbb{Z}_{>0}$;

(v) is ***bounded below*** if there exists $M \in \mathbb{R}$ such that $x_j > M$ for each $j \in \mathbb{Z}_{>0}$;

(vi) is ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $|x_j| < M$ for each $j \in \mathbb{Z}_{>0}$;

(vii) is ***monotonically increasing*** if $x_{j+1} \geq x_j$ for $j \in \mathbb{Z}_{>0}$;

(viii) is ***strictly monotonically increasing*** if $x_{j+1} > x_j$ for $j \in \mathbb{Z}_{>0}$;

(ix) is ***monotonically decreasing*** if $x_{j+1} \leq x_j$ for $j \in \mathbb{Z}_{>0}$;

(x) is ***strictly monotonically decreasing*** if $x_{j+1} < x_j$ for $j \in \mathbb{Z}_{>0}$;

(xi) is ***constant*** if $x_j = x_1$ for every $j \in \mathbb{Z}_{>0}$;

(xii) is ***eventually constant*** if there exists $N \in \mathbb{Z}_{>0}$ such that $x_j = x_N$ for every $j \geq N$. •

Associated with the notion of convergence is the notion of a limit. We also, for convenience, wish to allow sequences with infinite limits. This makes for some rather subtle use of language, so the reader should pay attention to this.

2.3.2 Definition (Limit of a sequence) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence.

(i) If $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to s_0, then the sequence has s_0 as a ***limit***, and we write $\lim_{j\to\infty} x_j = s_0$.

(ii) If, for every $M \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $x_j > M$ (resp. $x_k < -M$) for $j \geq N$, then the sequence ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\lim_{j\to\infty} x_j = \infty$ (resp. $\lim_{j\to\infty} x_j = -\infty$);

(iii) If $\lim_{j\to\infty} x_j \in \mathbb{R}$, then the limit of the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ ***exists***.

(iv) If the limit of the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ does not exist, does not diverge to ∞, or does not diverge to $-\infty$, then the sequence is ***oscillatory***. •

The reader can prove in Exercise 2.3.1 that limits, if they exist, are unique.

That convergent sequences are Cauchy, and that Cauchy sequences are bounded follows in exactly the same manner as the analogous results, stated as Propositions 2.1.13 and 2.1.14, for $\mathbb{Q}$. Let us state the results here for reference.

2.3.3 Proposition (Convergent sequences are Cauchy) *If a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to x_0, then it is a Cauchy sequence.*

2.3.4 Proposition (Cauchy sequences are bounded) *If $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ then it is bounded.*

Moreover, what is true for $\mathbb{R}$, and that is not true for $\mathbb{Q}$, is that every Cauchy sequence converges.

2.3.5 Theorem (Cauchy sequences in $\mathbb{R}$ converge) *If $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ then there exists $s_0 \in \mathbb{R}$ such that $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to s_0.*

Proof For $j \in \mathbb{Z}_{>0}$ choose $q_j \in \mathbb{Q}_{>0}$ such that $|x_j - q_j| < \frac{1}{j}$, this being possible by Proposition 2.2.15. For $\epsilon \in \mathbb{R}_{>0}$ let $N_1 \in \mathbb{Z}_{>0}$ satisfy $|x_j - x_k| < \frac{\epsilon}{2}$ for $j, k \geq N_1$. By Proposition 2.2.13 let $N_2 \in \mathbb{Z}_{>0}$ satisfy $N_2 \cdot 1 > 4\epsilon^{-1}$, and let N be the larger of N_1 and N_2. Then, for $j, k \geq N$, we have

$$|q_j - q_k| = |q_j - x_j + x_j - x_k + x_k - q_k| \leq |x_j - q_j| + |x_j - x_k| + |x_k - q_k| < \tfrac{1}{j} + \tfrac{\epsilon}{2} + \tfrac{1}{k} < \epsilon.$$

Thus $(q_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, and so we define $s_0 = [(q_j)_{j\in\mathbb{Z}_{>0}}]$.

Now we show that $(q_j)_{j\in\mathbb{Z}_{>0}}$ converges to s_0. Let $\epsilon \in \mathbb{R}_{>0}$ and take $N \in \mathbb{Z}_{>0}$ such that $|q_j - q_k| < \frac{\epsilon}{2}$, $j, k \geq N$, and rewrite this as

$$\tfrac{\epsilon}{2} < q_j - q_k + \epsilon, \quad \tfrac{\epsilon}{2} < -q_k + q_k + \epsilon, \qquad j, k \geq N. \tag{2.4}$$

For $j_0 \geq N$ consider the sequence $(q_j - q_{j_0} + \epsilon)_{j \in \mathbb{Z}_{>0}}$. This is a Cauchy sequence by Proposition 2.2.1. Moreover, by Proposition 2.2.6, $[(q_j - q_{j_0} + \epsilon)_{j \in \mathbb{Z}_{>0}}] > 0$, using the first of the inequalities in (2.4). Thus we have $s_0 - q_{j_0} + \epsilon > 0$, or

$$-\epsilon < s_0 - q_{j_0}, \qquad j_0 \geq N.$$

Arguing similarly, but using the second of the inequalities (2.4), we determine that

$$s_0 - q_{j_0} < \epsilon, \qquad j_0 \geq N.$$

This gives $|s_0 - q_j| < \epsilon$ for $j \geq N$, so showing that $(q_j)_{j \in \mathbb{Z}_{>0}}$ converges to s_0.

Finally, we show that $(x_j)_{j \in \mathbb{Z}_{>0}}$ converges to s_0. Let $\epsilon \in \mathbb{R}_{>0}$ and take $N_1 \in \mathbb{Z}_{>0}$ such that $|s_0 - q_j| < \frac{\epsilon}{2}$ for $j \geq N_1$. Also choose $N_2 \in \mathbb{Z}_{>0}$ such that $N_2 \cdot 1 > 2\epsilon^{-1}$ by Proposition 2.2.13. If N is the larger of N_1 and N_2, then we have

$$|s_0 - x_j| = |s_0 - q_j + q_j - x_j| \leq |s_0 - q_j| + |q_j - x_j| < \tfrac{\epsilon}{2} + \tfrac{1}{j} < \epsilon,$$

for $j \geq N$, so giving the result. ■

2.3.6 Remark (Completeness of $\mathbb{R}$) The property of $\mathbb{R}$ that Cauchy sequences are convergent gives, in the more general setting of Section III-1.1.6, $\mathbb{R}$ the property of being ***complete***. Completeness is an extremely important concept in analysis. We shall say some words about this in Section III-3.3.2; for now let us just say that the subject of calculus would not exist, but for the completeness of $\mathbb{R}$. •

2.3.2 Some properties equivalent to the completeness of $\mathbb{R}$

Using the fact that Cauchy sequences converge, it is easy to prove two other important features of $\mathbb{R}$, both of which seem obvious intuitively.

2.3.7 Theorem (Bounded subsets of $\mathbb{R}$ have a least upper bound) *If $\mathsf{S} \subseteq \mathbb{R}$ is nonempty and possesses an upper bound with respect to the standard total order $\leq$, then S possesses a least upper bound with respect to the same total order.*

Proof Since S has an upper bound, there exists $y \in \mathbb{R}$ such that $x \leq y$ for all $x \in S$. Now choose some $x \in S$. We then define two sequences $(x_j)_{j \in \mathbb{Z}_{>0}}$ and $(y_j)_{j \in \mathbb{Z}_{>0}}$ recursively as follows:

1. define $x_1 = x$ and $y_1 = y$;
2. suppose that x_j and y_j have been defined;
3. if there exists $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$, take $x_{j+1} = z$ and $y_{j+1} = y_j$;
4. if there is no $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$, take $x_{j+1} = x_j$ and $y_{j+1} = \frac{1}{2}(x_j + y_j)$.

A lemma characterises these sequences.

1 Lemma *The sequences* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(y_j)_{j\in\mathbb{Z}_{>0}}$ *have the following properties:*

(i) $x_j \in S$ *for* $j \in \mathbb{Z}_{>0}$*;*

(ii) $x_{j+1} \geq x_j$ *for* $j \in \mathbb{Z}_{>0}$*;*

(iii) y_j *is an upper bound for* S *for* $j \in \mathbb{Z}_{>0}$*;*

(iv) $y_{j+1} \leq y_j$ *for* $j \in \mathbb{Z}_{>0}$*;*

(v) $0 \leq y_j - x_j \leq \frac{1}{2^j}(y - x)$ *for* $j \in \mathbb{Z}_{>0}$.

Proof We prove the result by induction on j. The result is obviously true for $= 0$. Now suppose the result true for $j \in \{1, \dots, k\}$.

First take the case where there exists $z \in S$ with $\frac{1}{2}(x_k + y_k) < z \leq y_k$, so that $x_{k+1} = z$ and $y_{k+1} = y_k$. Clearly $x_{k+1} \in S$ and $y_{k+1} \geq y_k$. Since $y_k \geq x_k$ by the induction hypotheses, $\frac{1}{2}(x_k + y_k) \geq x_k$ giving $x_{k+1} = z \geq x_k$. By the induction hypotheses, y_{k+1} is an upper bound for S. By definition of x_{k+1} and y_{k+1},

$$y_{k+1} - x_{k+1} = y_k - z \geq 0$$

and

$$y_{k+1} - x_{k+1} = y_k - z = y_k - \tfrac{1}{2}(y_k - x_k) = \tfrac{1}{2}(y_k - x_k),$$

giving $y_{k+1} - x_{k+1} \leq \frac{1}{2^{k+1}}(y - x)$ by the induction hypotheses.

Now we take the case where there is no $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$, so that $x_{k+1} = x_k$ and $y_{k+1} = \frac{1}{2}(x_k + y_k)$. Clearly $x_{k+1} \geq x_k$ and $x_{k+1} \in S$. If y_{k+1} were not an upper bound for S, then there exists $a \in S$ such that $a > y_{k+1}$. By the induction hypotheses, y_k is an upper bound for S so $a \leq y_k$. But this means that $\frac{1}{2}(y_k + x_k) < a \leq y_k$, contradicting our assumption concerning the nonexistence of $z \in S$ with $\frac{1}{2}(x_j + y_j) < z \leq y_j$. Thus y_{k+1} is an upper bound for S. Since $x_k \leq y_k$ by the induction hypotheses,

$$y_{k+1} = \tfrac{1}{2}(y_k + x_k) \leq y_k.$$

Also

$$y_{k+1} - x_{k+1} = \tfrac{1}{2}(y_k - x_k)$$

by the induction hypotheses. This completes the proof. ▼

The following lemma records a useful fact about the sequences $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $(y_j)_{j\in\mathbb{Z}_{>0}}$.

2 Lemma *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(y_j)_{j\in\mathbb{Z}_{>0}}$ *be sequences in* $\mathbb{R}$ *satisfying:*

(i) $x_{j+1} \geq x_j$, $j \in \mathbb{Z}_{>0}$*;*

(ii) $y_{j+1} \leq y_j$, $j \in \mathbb{Z}_{>0}$*;*

(iii) the sequence $(y_j - x_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* 0.

Then $(x_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(y_j)_{j\in\mathbb{Z}_{>0}}$ *converge, and converge to the same limit.*

Proof First we claim that $x_j \leq y_k$ for all $j, k \in \mathbb{Z}_{>0}$. Indeed, suppose not. Then there exists $j, k \in \mathbb{Z}_{>0}$ such that $x_j > y_k$. If N is the larger of j and k, then we have $y_N \leq y_k < x_j \leq x_N$. This implies that

$$x_m - y_m \geq x_j - y_m \geq x_j - y_k > 0, \qquad m \geq N,$$

which contradicts the fact that $(y_j - x_j)_{j\in\mathbb{Z}_{>0}}$ converges to zero.

Now, for $\epsilon \in \mathbb{R}_{>0}$ let $N \in \mathbb{Z}_{>0}$ satisfy $|y_j - x_j| < \epsilon$ for $j \geq N$, or, simply, $y_j - x_j < \epsilon$ for $j \geq N$. Now let $j, k \geq N$, and suppose that $j \geq k$. Then

$$0 \leq x_j - x_k \leq x_j - y_k < \epsilon.$$

Similarly, if $j \leq k$ we have $0 \leq x_k - x_j < \epsilon$. In other words, $|x_j - x_k| < \epsilon$ for $j, k \geq N$. Thus $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence. In like manner one shows that $(y_j)_{j\in\mathbb{Z}_{>0}}$ is also a Cauchy sequence. Therefore, by Theorem 2.3.5, these sequences converge, and let us denote their limits by s_0 and t_0, respectively. However, since $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $(y_j)_{j\in\mathbb{Z}_{>0}}$ are equivalent Cauchy sequences in the sense of Definition 2.1.16, it follows that $s_0 = t_0$. ▼

Using Lemma 1 we easily verify that the sequences $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $(y_j)_{j\in\mathbb{Z}_{>0}}$ satisfy the hypotheses of Lemma 2. Therefore these sequences converge to a common limit, which we denote by s. We claim that s is a least upper bound for S. First we show that it is an upper bound. Suppose that there is $x \in S$ such that $x > s$ and define $\epsilon = x - s$. Since $(y_j)_{j\in\mathbb{Z}_{>0}}$ converges to s, there exists $N \in \mathbb{Z}_{>0}$ such that $|s - y_j| < \epsilon$ for $j \geq N$. Then, for $j \geq N$,

$$y_j - s < \epsilon = x - s,$$

implying that $y_j < x$, and so contradicting Lemma 1.

Finally, we need to show that s is a least upper bound. To see this, let b be an upper bound for S and suppose that $b < s$. Define $\epsilon = s - b$, and choose $N \in \mathbb{Z}_{>0}$ such that $|s - x_j| < \epsilon$ for $j \geq N$. Then

$$s - x_j < \epsilon = s - b,$$

implying that $b < x_j$ for $j \geq N$. This contradicts the fact, from Lemma 1, that $x_j \in S$ and that b is an upper bound for S. ■

As we shall explain more fully in Aside 2.3.9, the least upper bound property of the real numbers as stated in the preceding theorem is actually *equivalent* to the completeness of $\mathbb{R}$. In fact, the least upper bound property forms the basis for an alternative definition of the real numbers using ***Dedekind cuts***.[4] Here the idea is that one defines a real number as being a splitting of the rational numbers into two halves, one corresponding to the rational numbers less than the real number one is defining, and the other corresponding to the rational numbers greater than the real number one is defining. Historically, Dedekind cuts provided the first rigorous construction of the real numbers. We refer to Section 2.3.9 for further discussion. We also comment, as we discuss in Aside 2.3.9, that any construction of the real numbers with the property of completeness, or an equivalent, will produce something that is "essentially" the real numbers as we have defined them.

Another consequence of Theorem 2.3.5 is the following.

[4] After Julius Wihelm Richard Dedekind (1831–1916), the German mathematician, did work in the areas of analysis, ring theory, and set theory. His rigorous mathematical style has had a strong influence on modern mathematical presentation.

2.3.8 Theorem (Bounded, monotonically increasing sequences in $\mathbb{R}$ converge) *If* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *is a bounded, monotonically increasing sequence in* $\mathbb{R}$*, then it converges.*

Proof The subset $(x_j)_{j\in\mathbb{Z}_{>0}}$ of $\mathbb{R}$ has an upper bound, since it is bounded. By Theorem 2.3.7 let b be the least upper bound for this set. We claim that $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to b. Indeed, let $\epsilon \in \mathbb{R}_{>0}$. We claim that there exists some $N \in \mathbb{Z}_{>0}$ such that $b - x_N < \epsilon$ since b is a least upper bound. Indeed, if there is no such N, then $b \geq x_j + \epsilon$ for all $j \in \mathbb{Z}_{>0}$ and so $b - \frac{\epsilon}{2}$ is an upper bound for $(x_j)_{j\in\mathbb{Z}_{>0}}$ that is smaller than b. Now, with N chosen so that $b - x_N < \epsilon$, the fact that $(x_j)_{j\in\mathbb{Z}_{>0}}$ is monotonically increasing implies that $|b - x_j| < \epsilon$ for $j \geq N$, as desired. ■

It turns out that Theorems 2.3.5, 2.3.7, and 2.3.8 are equivalent. But to make sense of this requires one to step outside the concrete representation we have given for the real numbers to a more axiomatic one. This can be skipped, so we present it as an aside.

2.3.9 Aside (Complete ordered fields) An ***ordered field*** is a field $\mathbb{F}$ (see Definition 4.3.1 for the definition of a field) equipped with a total order satisfying the conditions

1. if $x < y$ then $x + z < y + z$ for $x, y, z \in \mathbb{F}$ and
2. if $0 < x, y$ then $0 < x \cdot y$.

Note that in an ordered field one can define the absolute value exactly as we have done for $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. There are many examples of ordered fields, of which $\mathbb{Q}$ and $\mathbb{R}$ are two that we have seen. However, if one adds to the conditions for an ordered field an additional condition, then this turns out to essentially uniquely specify the set of real numbers. (We say "essentially" since the uniqueness is up to a bijection that preserves the field structure as well as the order.) This additional structure comes in various forms, of which three are as stated in Theorems 2.3.5, 2.3.7, and 2.3.8. To be precise, we have the following theorem.

Theorem *If* $\mathbb{F}$ *is an ordered field, then the following statements are equivalent:*

(i) *every Cauchy sequence converges;*

(ii) *each set possessing an upper bound possesses a least upper bound;*

(iii) *each bounded, monotonically increasing sequence converges.*

We have almost proved this theorem with our arguments above. To see this, note that in the proof of Theorem 2.3.7 we use the fact that Cauchy sequences converge. Moreover, the argument can easily be adapted from the special case of $\mathbb{R}$ to a general ordered field. This gives the implication (i) $\Longrightarrow$ (ii) in the theorem above. In like manner, the proof of Theorem 2.3.8 gives the implication (ii) $\Longrightarrow$ (iii), since the proof is again easily seen to be valid for a general ordered field. The argument for the implication (iii) $\Longrightarrow$ (i) is outlined in Exercise 2.3.4. An ordered field satisfying any one of the three equivalent conditions (i), (ii), and (iii) is called a ***complete ordered field***. Thus there is essentially only one complete ordered field, and it is $\mathbb{R}$. ♠

2.3.3 Tests for convergence of sequences

There is generally no algorithmic way, other than checking the definition, to ascertain when a sequence converges. However, there are a few simple results that are often useful, and here we state some of these.

2.3.10 Proposition (Squeezing Principle) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$, $(y_j)_{j\in\mathbb{Z}_{>0}}$, *and* $(z_j)_{j\in\mathbb{Z}_{>0}}$ *be sequences in* $\mathbb{R}$ *satisfying*

(i) $x_j \leq z_j \leq y_j$ *for all* $j \in \mathbb{Z}_{>0}$ *and*

(ii) $\lim_{j\to\infty} x_j = \lim_{j\to\infty} y_j = \alpha$.

Then $\lim_{j\to\infty} z_j = \alpha$.

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1, N_2 \in \mathbb{Z}_{>0}$ have the property that $|x_j - \alpha| < \frac{\epsilon}{3}$ for $j \geq N_1$ and $|y_j - \alpha| < \frac{\epsilon}{3}$. Then, for $j \geq \max\{N_1, N_2\}$,

$$|x_j - y_j| = |x_j - \alpha + \alpha - y_j| \leq |x_j - \alpha| + |y_j - \alpha| < \tfrac{2\epsilon}{3},$$

using the triangle inequality. Then, for $j \geq \max\{N_1, N_2\}$, we have

$$|z_j - \alpha| = |z_j - x_j + x_j - \alpha| \leq |z_j - x_j| + |x_j - \alpha| \leq |y_j - x_j| + |x_j - \alpha| = \epsilon,$$

again using the triangle inequality. ■

The next test for convergence of a series is sometimes useful.

2.3.11 Proposition (Ratio Test for sequences) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *be a sequence in* $\mathbb{R}$ *for which* $\lim_{j\to\infty} \left|\frac{x_{j+1}}{x_j}\right| = \alpha$. *If* $\alpha < 1$ *then the sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *converges to* 0, *and if* $\alpha > 1$ *then the sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *diverges.*

Proof For $\alpha < 1$, define $\beta = \frac{1}{2}(\alpha + 1)$. Then $\alpha < \beta < 1$. Now take $N \in \mathbb{Z}_{>0}$ such that

$$\left|\left|\frac{x_{j+1}}{x_j}\right| - \alpha\right| < \tfrac{1}{2}(1 - \alpha), \qquad j > N.$$

This implies that

$$\left|\frac{x_{j+1}}{x_j}\right| < \beta.$$

Now, for $j > N$,

$$|x_j| < \beta|x_{j-1}| < \beta^2|x_{j-1}| < \cdots < \beta^{j-N}|x_N|.$$

Clearly the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to 0 if and only if the sequence obtained by replacing the first N terms by 0 also converges to 0. If this latter sequence is denoted by $(y_j)_{j\in\mathbb{Z}_{>0}}$, then we have

$$0 \leq y_j \leq \frac{|x_N|}{\beta^N}\beta^j.$$

The sequence $(\frac{|x_N|}{\beta^N}\beta^j)_{j\in\mathbb{Z}_{>0}}$ converges to 0 since $\beta < 1$, and so this part of the result follows from the Squeezing Principle.

For $\alpha > 1$, there exists $N \in \mathbb{Z}_{>0}$ such that, for all $j \geq N$, $x_j \neq 0$. Consider the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ which is 0 for the first N terms, and satisfies $y_j = x_j^{-1}$ for the remaining terms. We then have $\left|\frac{y_{j+1}}{y_j}\right| < \alpha^{-1} < 1$, and so, from the first part of the proof, the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ converges to 0. Thus the sequence $(|y_j|)_{j\in\mathbb{Z}_{>0}}$ converges to ∞, which prohibits the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ from converging. ■

In Exercise 2.3.3 the reader can explore the various possibilities for the ratio test when $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$.

2.3.4 lim sup and lim inf

Recall from Section 2.2.6 the notions of sup and inf for subsets of $\mathbb{R}$. Associated with the least upper bound and greatest lower bound properties of $\mathbb{R}$ is a useful notion that weakens the usual idea of convergence. In order for us to make a sensible definition, we first prove a simple result.

2.3.12 Proposition (Existence of lim sup and lim inf) *For any sequence* $(\mathrm{x_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *in* $\mathbb{R}$, *the limits*

$$\lim_{\mathrm{N}\to\infty}\Big(\sup\{\mathrm{x_j} \mid \mathrm{j} \geq \mathrm{N}\}\Big), \quad \lim_{\mathrm{N}\to\infty}\Big(\inf\{\mathrm{x_j} \mid \mathrm{j} \geq \mathrm{N}\}\Big)$$

exist, diverge to ∞, *or diverge to* $-\infty$.

Proof Note that the sequences $(\sup\{x_j \mid j \geq N\})_{N\in\mathbb{Z}_{>0}}$ and $(\inf\{x_j \mid j \geq N\})_{N\in\mathbb{Z}_{>0}}$ in $\overline{\mathbb{R}}$ are monotonically decreasing and monotonically increasing, respectively, with respect to the natural order on $\overline{\mathbb{R}}$. Moreover, note that a monotonically increasing sequence in $\overline{\mathbb{R}}$ is either bounded by some element of $\mathbb{R}$, or it is not. If the sequence is upper bounded by some element of $\mathbb{R}$, then by Theorem 2.3.8 it either converges or is the sequence $(-\infty)_{j\in\mathbb{Z}_{>0}}$. If it is not bounded by some element in $\mathbb{R}$, then either it diverges to ∞, or it is the sequence $(\infty)_{j\in\mathbb{Z}_{>0}}$ (this second case cannot arise in the specific case of the monotonically increasing sequence $(\sup\{x_j \mid j \geq N\})_{N\in\mathbb{Z}_{>0}}$. In all cases, the limit $\lim_{N\to\infty}\big(\sup\{x_j \mid j \geq N\}\big)$ exists or diverges to ∞. A similar argument for holds for $\lim_{N\to\infty}\big(\inf\{x_j \mid j \geq N\}\big)$. ■

2.3.13 Definition (lim sup and lim inf) For a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ denote

$$\begin{aligned}\limsup_{j\to\infty} x_j &= \lim_{N\to\infty}\Big(\sup\{x_j \mid j \geq N\}\Big),\\ \liminf_{j\to\infty} x_j &= \lim_{N\to\infty}\Big(\inf\{x_j \mid j \geq N\}\Big).\end{aligned}$$

•

Before we get to characterising lim sup and lim inf, we give some examples to illustrate all the cases that can arise.

2.3.14 Examples ($\limsup$ and $\liminf$)

1. Consider the sequence $(x_j = (-1)^j)_{j\in\mathbb{Z}_{>0}}$. Here we have $\limsup_{j\to\infty} x_j = 1$ and $\liminf_{j\to\infty} x_j = -1$.
2. Consider the sequence $(x_j = j)_{j\in\mathbb{Z}_{>0}}$. Here $\limsup_{j\to\infty} x_j = \liminf_{j\to\infty} = \infty$.
3. Consider the sequence $(x_j = -j)_{j\in\mathbb{Z}_{>0}}$. Here $\limsup_{j\to\infty} x_j = \liminf_{j\to\infty} = -\infty$.
4. Define
$$x_j = \begin{cases} j, & j \text{ even}, \\ 0, & j \text{ odd}. \end{cases}$$
We then have $\limsup_{j\to\infty} x_j = \infty$ and $\liminf_{j\to\infty} x_j = 0$.
5. Define
$$x_j = \begin{cases} -j, & j \text{ even}, \\ 0, & j \text{ odd}. \end{cases}$$
We then have $\limsup_{j\to\infty} x_j = 0$ and $\liminf_{j\to\infty} = -\infty$.
6. Define
$$x_j = \begin{cases} j, & j \text{ even}, \\ -j, & j \text{ odd}. \end{cases}$$
We then have $\limsup_{j\to\infty} x_j = \infty$ and $\liminf_{j\to\infty} = -\infty$. •

There are many ways to characterise lim sup and lim inf, and we shall indicate but a few of these.

2.3.15 Proposition (Characterisation of $\limsup$) *For a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ and $\alpha \in \mathbb{R}$, the following statements are equivalent:*

(i) $\alpha = \limsup_{j\to\infty} x_j$;

(ii) $\alpha = \inf\{\sup\{x_j \mid j \geq k\} \mid k \in \mathbb{Z}_{>0}\}$;

(iii) for each $\epsilon \in \mathbb{R}_{>0}$ the following statements hold:

(a) there exists $N \in \mathbb{Z}_{>0}$ such that $x_j < \alpha + \epsilon$ for all $j \geq N$;

(b) for an infinite number of $j \in \mathbb{Z}_{>0}$ it holds that $x_j > \alpha - \epsilon$.

Proof (i) $\Longleftrightarrow$ (ii) Let $y_k = \sup\{x_j \mid j \geq k\}$ and note that the sequence $(y_k)_{k\in\mathbb{Z}_{>0}}$ is monotonically decreasing. Therefore, the sequence $(y_k)_{k\in\mathbb{Z}_{>0}}$ converges if and only if it is lower bounded. Moreover, if it converges, it converges to $\inf(y_k)_{k\in\mathbb{Z}_{>0}}$. Putting this all together gives the desired implications.

(i) $\Longrightarrow$ (iii) Let y_k be as in the preceding part of the proof. Since $\lim_{k\to\infty} y_k = \alpha$, for each $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that $|y_k - \alpha| < \epsilon$ for $k \geq N$. In particular, $y_N < \alpha + \epsilon$. Therefore, $x_j < \alpha + \epsilon$ for all $j \geq N$, so (a) holds. We also claim that, for every $\epsilon \in \mathbb{R}_{>0}$ and for every $N \in \mathbb{Z}_{>0}$, there exists $j \geq N$ such that $x_j > y_N - \epsilon$. Indeed, if $x_j \leq y_N - \epsilon$ for every $j \geq N$, then this contradicts the definition of y_N. Since $y_N \geq \alpha$ we have $x_j > y_N - \epsilon \geq \alpha - \epsilon$ for some j. Since N is arbitrary, (b) holds.

(iii) $\implies$ (i) Condition (a) means that there exists $N \in \mathbb{Z}_{>0}$ such that $y_k < \alpha + \epsilon$ for all $k \geq N$. Condition (b) implies that $y_k > \alpha - \epsilon$ for all $k \in \mathbb{Z}_{>0}$. Combining these conclusions shows that $\lim_{k\to\infty} y_k = \alpha$, as desired. ■

The corresponding result for lim inf is the following. The proof follows in the same manner as the result for lim sup.

2.3.16 Proposition (Characterisation of lim inf) *For a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ and $\alpha \in \mathbb{R}$, the following statements are equivalent:*

(i) $\alpha = \liminf_{j\to\infty} x_j$;

(ii) $\alpha = \sup\{\inf\{x_j \mid j \geq k\} \mid k \in \mathbb{Z}_{>0}\}$;

(iii) for each $\epsilon \in \mathbb{R}_{>0}$ the following statements hold:

(a) there exists $N \in \mathbb{Z}_{>0}$ such that $x_j > \alpha - \epsilon$ for all $j \geq N$;

(b) for an infinite number of $j \in \mathbb{Z}_{>0}$ it holds that $x_j < \alpha + \epsilon$.

Finally, we characterise the relationship between lim sup, lim inf, and lim.

2.3.17 Proposition (Relationship between lim sup, lim inf, and lim) *For a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $s_0 \in \mathbb{R}$, the following statements are equivalent:*

(i) $\lim_{j\to\infty} x_j = s_0$;

(ii) $\limsup_{j\to\infty} x_j = \liminf_{j\to\infty} x_j = s_0$.

Proof (i) $\implies$ (ii) Let $\epsilon \in \mathbb{R}_{>0}$ and take $N \in \mathbb{Z}_{>0}$ such that $|x_j - s_0| < \epsilon$ for all $j \geq N$. Then $x_j < s_0 + \epsilon$ and $x_j > s_0 - \epsilon$ for all $j \geq N$. The current implication now follows from Propositions 2.3.15 and 2.3.16.

(ii) $\implies$ (i) Let $\epsilon \in \mathbb{R}_{>0}$. By Propositions 2.3.15 and 2.3.16 there exists $N_1, N_2 \in \mathbb{Z}_{>0}$ such that $x_j - s_0 < \epsilon$ for $j \geq N_1$ and $s_0 - x_j < \epsilon$ for $j \geq N_2$. Thus $|x_j - s_0| < \epsilon$ for $j \geq \max\{N_1, N_2\}$, giving this implication. ■

2.3.5 Multiple sequences

It will be sometimes useful for us to be able to consider sequences indexed, not by a single index, but by multiple indices. We consider the case here of two indices, and extensions to more indices are done by induction.

2.3.18 Definition (Double sequence) A ***double sequence*** in $\mathbb{R}$ is a family of elements of $\mathbb{R}$ indexed by $\mathbb{Z}_{>0} \times \mathbb{Z}_{>0}$. We denote a double sequence by $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$, where x_{jk} is the image of $(j,k) \in \mathbb{Z}_{>0} \times \mathbb{Z}_{>0}$ in $\mathbb{R}$. •

It is not *a priori* obvious what it might mean for a double sequence to converge, so we should carefully say what this means.

2.3.19 Definition (Convergence of double sequences) Let $s_0 \in \mathbb{R}$. A double sequence $(x_{jk})_{j,k \in \mathbb{Z}_{>0}}$:

(i) ***converges to*** $\mathbf{s_0}$, and we write $\lim_{j,k\to\infty} x_{jk} = s_0$, if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|s_0 - x_{jk}| < \epsilon$ for $j, k \geq N$;

(ii) has s_0 as a ***limit*** if it converges to s_0.

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}$;

(iv) ***diverges*** if it does not converge;

(v) ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\lim_{j,k\to\infty} x_{jk} = \infty$ (resp. $\lim_{j,k\to\infty} x_{jk} = -\infty$) if, for each $M \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $x_{jk} > M$ (resp. $x_{jk} < -M$) for $j, k \geq N$;

(vi) has a limit that ***exists*** if $\lim_{j,k\to\infty} x_{jk} \in \mathbb{R}$;

(vii) is ***oscillatory*** if the limit of the sequence does not exist, does not diverge to ∞, or does not diverge to $-\infty$. •

Note that the definition of convergence requires that one check both indices at the same time. Indeed, if one thinks, as it is useful to do, of a double sequence as assigning a real number to each point in an infinite grid defined by the set $\mathbb{Z}_{>0} \times \mathbb{Z}_{>0}$, convergence means that the values on the grid can be made arbitrarily small outside a sufficiently large square (see Figure 2.2). It is useful, however, to have means of computing limits of double sequences by computing limits of sequences in the usual sense. Our next results are devoted to this.

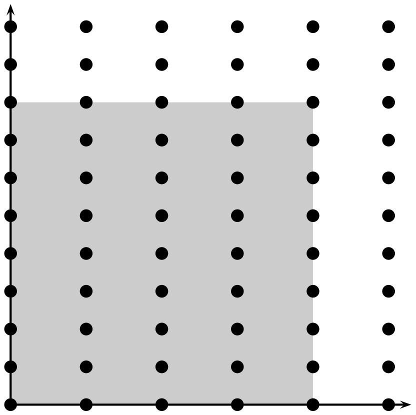


Figure 2.2 Convergence of a double sequence: by choosing the square large enough, the values at the unshaded grid points can be arbitrarily close to the limit

2.3.20 Proposition (Computation of limits of double sequences I) *Suppose that for the double sequence* $(x_{jk})_{j,k \in \mathbb{Z}_{>0}}$ *it holds that*

(i) the double sequence is convergent and

(ii) for each $\mathrm{j} \in \mathbb{Z}_{>0}$*, the limit* $\lim_{\mathrm{k}\to\infty} \mathrm{x}_{\mathrm{jk}}$ *exists.*

Then the limit $\lim_{\mathrm{j}\to\infty}(\lim_{\mathrm{k}\to\infty} \mathrm{x}_{\mathrm{jk}})$ *exists and is equal to* $\lim_{\mathrm{j,k}\to\infty} \mathrm{x}_{\mathrm{jk}}$.

Proof Let $s_0 = \lim_{j,k\to\infty} x_{jk}$ and denote $s_j = \lim_{k\to\infty} x_{jk}$, $j \in \mathbb{Z}_{>0}$. For $\epsilon \in \mathbb{R}_{>0}$ take $N \in \mathbb{Z}_{>0}$ such that $|x_{jk} - s_0| < \frac{\epsilon}{2}$ for $j, k \geq N$. Also take $N_j \in \mathbb{Z}_{>0}$ such that $|x_{jk} - s_j| < \frac{\epsilon}{2}$ for $k \geq N_j$. Next take $j \geq N$ and let $k \geq \max\{N, N_j\}$. We then have

$$|s_j - s_0| = |s_j - x_{jk} + x_{jk} - s_0| \leq |s_j - x_{jk}| + |x_{jk} - s_0| < \epsilon,$$

using the triangle inequality. ■

2.3.21 Proposition (Computation of limits of double sequences II) *Suppose that for the double sequence* $(\mathrm{x}_{\mathrm{jk}})_{\mathrm{j,k}\in\mathbb{Z}_{>0}}$ *it holds that*

(i) the double sequence is convergent,

(ii) for each $\mathrm{j} \in \mathbb{Z}_{>0}$*, the limit* $\lim_{\mathrm{k}\to\infty} \mathrm{x}_{\mathrm{jk}}$ *exists, and*

(iii) for each $\mathrm{k} \in \mathbb{Z}_{>0}$*, the limit* $\lim_{\mathrm{j}\to\infty} \mathrm{x}_{\mathrm{jk}}$ *exists.*

Then the limits $\lim_{\mathrm{j}\to\infty}(\lim_{\mathrm{k}\to\infty} \mathrm{x}_{\mathrm{jk}})$ *and* $\lim_{\mathrm{k}\to\infty}(\lim_{\mathrm{j}\to\infty} \mathrm{x}_{\mathrm{jk}})$ *exist and are equal to* $\lim_{\mathrm{j,k}\to\infty} \mathrm{x}_{\mathrm{jk}}$.

Proof This follows from two applications of Proposition 2.3.20. ■

Let us give some examples that illustrate the idea of convergence of a double sequence.

2.3.22 Examples (Double sequences)

1. It is easy to check that the double sequence $(\frac{1}{j+k})_{j,k\in\mathbb{Z}_{>0}}$ converges to 0. Indeed, for $\epsilon \in \mathbb{R}_{>0}$, if we take $N \in \mathbb{Z}_{>0}$ such that $\frac{1}{2N} < \epsilon$, it follows that $\frac{1}{j+k} < \epsilon$ for $j, k \geq N$.

2. The double sequence $(\frac{j}{j+k})_{j,k\in\mathbb{Z}_{>0}}$ does not converge. To see this we should find $\epsilon \in \mathbb{R}_{>0}$such that, for any $N \in \mathbb{Z}_{>0}$, there exists $j, k \geq N$ for which $\frac{j}{j+k} \geq \epsilon$. Take $\epsilon = \frac{1}{2}$ and let $N \in \mathbb{Z}_{>0}$. Then, if $j, k \geq N$ satisfy $j \geq 2k$, we have $\frac{j}{j+k} \geq \epsilon$.

   Note that for this sequence, the limits $\lim_{j\to\infty} \frac{j}{j+k}$ and $\lim_{k\to\infty} \frac{j}{j+k}$ exist for each fixed k and j, respectively. This cautions about trying to use these limits to infer convergence of the double sequence.

3. The double sequence $(\frac{(-1)^j}{k})_{j,k\in\mathbb{Z}_{>0}}$ is easily seen to converge to 0. However, the limit $\lim_{j\to\infty} \frac{(-1)^j}{k}$ does not exist for any fixed k. Therefore, one needs condition (ii) in Proposition 2.3.20 and conditions (ii) and (iii) in Proposition 2.3.21 in order for the results to be valid. •

2.3.6 Algebraic operations on sequences

It is of frequent interest to add, multiply, or divide sequences and series. In such cases, one would like to ensure that convergence of the sequences or series is sufficient to ensure convergence of the sum, product, or quotient. In this section we address this matter.

2.3.23 Proposition (Algebraic operations on sequences) *Let $(\mathrm{x}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ and $(\mathrm{y}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ be sequences converging to s_0 and t_0, respectively, and let $\alpha \in \mathbb{R}$. Then the following statements hold:*

- *(i) the sequence $(\alpha\mathrm{x}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ converges to $\alpha\mathrm{s}_0$;*
- *(ii) the sequence $(\mathrm{x}_\mathrm{j} + \mathrm{y}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ converges to $\mathrm{s}_0 + \mathrm{t}_0$;*
- *(iii) the sequence $(\mathrm{x}_\mathrm{j}\mathrm{y}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ converges to $\mathrm{s}_0\mathrm{t}_0$;*
- *(iv) if, for all $\mathrm{j} \in \mathbb{Z}_{>0}$, $\mathrm{y}_\mathrm{j} \neq 0$ and if $\mathrm{s}_0 \neq 0$, then the sequence $(\frac{\mathrm{x}_\mathrm{j}}{\mathrm{y}_\mathrm{j}})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ converges to $\frac{\mathrm{s}_0}{\mathrm{t}_0}$.*

Proof (i) The result is trivially true for $a = 0$, so let us suppose that $a \neq 0$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|x_j - s_0| < \frac{\epsilon}{|\alpha|}$. Then, for $j \geq N$,

$$|\alpha x_j - \alpha s_0| = |\alpha||x_j - s_0| < \epsilon.$$

(ii) Let $\epsilon \in \mathbb{R}_{>0}$ and take $N_1, N_2 \in \mathbb{Z}_{>0}$ such that

$$|x_j - s_0| < \tfrac{\epsilon}{2}, \quad j \geq N_1, \qquad |y_j - t_0| < \tfrac{\epsilon}{2}, \quad j \geq N_2.$$

Then, for $j \geq \max\{N_1, N_2\}$,

$$|x_j + y_j - (s_0 + t_0)| \leq |x_j - s_0| + |y_j - t_0| = \epsilon,$$

using the triangle inequality.

(iii) Let $\epsilon \in \mathbb{R}_{>0}$ and define $N_1, N_2, N_3 \in \mathbb{Z}_{>0}$ such that

$$\begin{aligned}
&|x_j - s_0| < 1, \qquad j \geq N_1, \qquad \Longrightarrow \qquad |x_j| < |s_0| + 1, \qquad j \geq N_1,\\
&|x_j - s_0| < \frac{\epsilon}{2(|t_0| + 1)}, \qquad j \geq N_2,\\
&|y_j - t_0| < \frac{\epsilon}{2(|s_0| + 1)}, \qquad j \geq N_2.
\end{aligned}$$

Then, for $j \geq \max\{N_1, N_2, N_3\}$,

$$\begin{aligned}
|x_j y_j - s_0 t_0| &= |x_j y_j - x_j t_0 + x_j t_0 - s_0 t_0|\\
&= |x_j(y_j - t_0) + t_0(x_j - s_0)|\\
&\leq |x_j||y_j - t_0| + |t_0||x_j - s_0|\\
&\leq (|s_0| + 1)\frac{\epsilon}{2(|s_0| + 1)} + (|t_0| + 1)\frac{\epsilon}{2(|t_0| + 1)} = \epsilon.
\end{aligned}$$

(iv) It suffices using part (iii) to consider the case where $x_j = 1$, $j \in \mathbb{Z}_{>0}$. For $\epsilon \in \mathbb{R}_{>0}$ take $N_1.N_2 \in \mathbb{Z}_{>0}$ such that

$$\begin{aligned}
&|y_j - t_0| < \frac{|t_0|}{2}, \qquad j \geq N_1, \qquad \Longrightarrow \qquad |y_j| > \frac{|t_0|}{2}, \qquad j \geq N_1,\\
&|y_j - t_0| < \frac{|t_0|^2\epsilon}{2}, \qquad j \geq N_2.
\end{aligned}$$

Then, for $j \geq \max\{N_1, N_2\}$,

$$\left|\frac{1}{y_j} - \frac{1}{t_0}\right| = \left|\frac{y_j - t_0}{y_j t_0}\right| \leq \frac{|t_0|^2\epsilon}{2}\frac{2}{|t_0|}\frac{1}{|t_0|} = \epsilon,$$

as desired. ■

As we saw in the statement of Proposition 2.2.1, the restriction in part (iv) that $y_j \neq 0$ for all $j \in \mathbb{Z}_{>0}$ is not a real restriction. The salient restriction is that the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ not converge to 0.

2.3.7 Convergence using $\mathbb{R}$-nets

Up to this point in this section we have talked about convergence of sequences. However, in practice it is often useful to take limits of more general objects where the index set is not $\mathbb{Z}_{>0}$, but a subset of $\mathbb{R}$. In Section 1.6.4 we introduced a generalisation of sequences called nets. In this section we consider particular cases of nets, called $\mathbb{R}$-nets, that arise commonly when dealing with real numbers and subsets of real numbers. These will be particularly useful when considering the relationships between limits and functions. As we shall see, this slightly more general notion of convergence can be reduced to standard convergence of sequences. We comment that the notions of convergence in this section can be generalised to general nets, and we refer the reader to Section III-1.5 for details.

Our objective is to understand what is meant by an expression like $\lim_{x\to a} \phi(a)$, where $\phi\colon A \to \mathbb{R}$ is a map from a subset A of $\mathbb{R}$ to $\mathbb{R}$. We will mainly be interested in subsets A of a rather specific form. However, we consider the general case so as to cover all situations that might arise.

2.3.24 Definition ($\mathbb{R}$-directed set) A $\mathbb{R}$-***directed set*** is a pair $D = (A, \preceq)$ where the partial order $\preceq$ is defined by $x \preceq y$ if either

(i) $x \leq y$,

(ii) $x \geq y$, or

(iii) there exists $x_0 \in \mathbb{R}$ such that $|x - x_0| \leq |y - x_0|$ (we abbreviate this relation as $x \leq_{x_0} y$). •

Note that if $D = (A, \preceq)$ is a $\mathbb{R}$-directed set, then it is indeed a directed set because, corresponding to the three cases of the definition,

1. if $x, y \in A$, then $z = \max\{x, y\}$ has the property that $x \preceq z$ and $y \preceq z$ (for the first case in the definition),
2. if $x, y \in A$, then $z = \min\{x, y\}$ has the property that $x \preceq z$ and $y \preceq z$ (for the second case in the definition), or
3. if $x, y \in A$ then, taking z to satisfy $|z - x_0| = \min\{|x - x_0|, |y - x_0|\}$, we have $x \preceq z$ and $y \preceq z$ (for the third case of the definition).

Let us give some examples to illustrate the sort of phenomenon one can see for $\mathbb{R}$-directed sets.

2.3.25 Examples ($\mathbb{R}$-directed sets)

1. Let us take the $\mathbb{R}$-directed set $([0, 1], \leq)$. Here we see that, for any $x, y \in [0, 1]$, we have $x \leq 1$ and $y \leq 1$.

2. Next take the $\mathbb{R}$-directed set $([0,1),\leq)$. Here, there is no element z of $[0,1)$ for which $x \leq z$ and $y \leq z$ for every $x, y \in [0,1)$. However, it obviously holds that $x \leq 1$ and $y \leq 1$ for every $x, y \in [0,1)$.
3. Next we consider the $\mathbb{R}$ directed set $([0,\infty),\geq)$. Here we see that, for any $x, y \in [0,\infty)$, $x \geq 0$ and $y \geq 0$.
4. Next we consider the $\mathbb{R}$ directed set $((0,\infty),\geq)$. Here we see that there is no element $z \in (0,\infty)$ such that, for every $x, y \in (0,\infty)$, $x \geq z$ and $y \geq z$. However, it is true that $x \geq 0$ and $y \geq 0$ for every $x, y \in (0,\infty)$.
5. Now we take the $\mathbb{R}$-directed set $([0,\infty),\leq)$. Here we see that there is no element $z \in [0,\infty)$ such that $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$. Moreover, there is also no element $z \in \mathbb{R}$ for which $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$.
6. Next we take the $\mathbb{R}$-directed set $(\mathbb{Z},\leq)$. As in the preceding example, there is no element $z \in [0,\infty)$ such that $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$. Moreover, there is also no element $z \in \mathbb{R}$ for which $x \leq z$ and $y \leq z$ for every $x, y \in [0,\infty)$.
7. Now consider the $\mathbb{R}$-directed set $(\mathbb{R},\leq_0)$. Note that $0 \in \mathbb{R}$ has the property that, for any $x, y \in \mathbb{R}$, $x \leq_0 0$ and $y \leq_0 0$.
8. Similar to the preceding example, consider the $\mathbb{R}$-directed set $(\mathbb{R}\setminus\{0\},\leq_0)$. Here there is no element $z \in \mathbb{R}\setminus\{0\}$ such that $x \leq_0 z$ and $y \leq_0 z$ for every $x, y \in \mathbb{R}\setminus\{0\}$. However, we clearly have $x \leq_0 0$ and $y \leq_0 0$ for every $x, y \in \mathbb{R}\setminus\{0\}$. •

The examples may seem a little silly, but this is just because the notion of a $\mathbb{R}$-directed set is, in and of itself, not so interesting. What is more interesting is the following notion.

2.3.26 Definition ($\mathbb{R}$-net, convergence in $\mathbb{R}$-nets) If $D = (A,\preceq)$ is a $\mathbb{R}$-directed set, a $\mathbb{R}$-***net*** in D is a map $\phi\colon A \to \mathbb{R}$. A $\mathbb{R}$-net $\phi\colon A \to \mathbb{R}$ in a $\mathbb{R}$-directed set $D = (A,\preceq)$

(i) ***converges*** to $s_0 \in \mathbb{R}$ if, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $x \in A$ such that $|\phi(y) - s_0| < \epsilon$ for any $y \in A$ satisfying $x \preceq y$,

(ii) has s_0 as a ***limit*** if it converges to s_0, and we write $s_0 = \lim_D \phi$,

(iii) ***diverges*** if it does not converge,

(iv) ***diverges to ∞*** ((resp. ***diverges to −∞***, and we write $\lim_D \phi = \infty$ (resp. $\lim_D \phi = -\infty$), if, for each $M \in \mathbb{R}_{>0}$, there exists $x \in A$ such that $\phi(y) > M$ (resp. $\phi(y) < -M$) for every $y \in A$ for which $x \preceq y$,

(v) has a limit that ***exists*** if $\lim_D \phi \in \mathbb{R}$, and

(vi) is ***oscillatory*** if the limit of the $\mathbb{R}$-net does not exist, does not diverge to ∞, and does not diverge to $-\infty$. •

2.3.27 Notation (Limits of $\mathbb{R}$-nets) The importance $\mathbb{R}$-nets can now be illustrated by showing how they give rise to a collection of convergence phenomenon. Let us look at various cases for convergence of a $\mathbb{R}$-net in a $\mathbb{R}$-directed set $D = (A,\preceq)$.

(i) $\preceq = \leq$: Here there are two subcases to consider.

(a) $\sup A = x_0 < \infty$: In this case we write $\lim_D \phi = \lim_{x\uparrow x_0} \phi(x)$.

(b) $\sup A = \infty$: In this case we write $\lim_D \phi = \lim_{x\to\infty} \phi(x)$.

(ii) $\preceq = \geq$: Again we have two subcases.

(a) $\inf A = x_0 > -\infty$: In this case we write $\lim_D \phi = \lim_{x\downarrow x_0} \phi(x)$.

(b) $\inf A = -\infty$: In this case we write $\lim_D \phi = \lim_{x\to -\infty} \phi(x)$.

(iii) $\preceq = \leq_{x_0}$: There are three subcases here that we wish to distinguish.

(a) $\sup A = x_0$: Here we denote $\lim_D \phi = \lim_{x\uparrow x_0} \phi(x)$.

(b) $\inf A = x_0$: Here we denote $\lim_D \phi = \lim_{x\downarrow x_0} \phi(x)$.

(c) $x_0 \notin \{\inf A, \sup A\}$: Here we denote $\lim_D \phi = \lim_{x\to x_0} \phi(x)$. •

In the case when the directed set is an interval, we have the following notation that unifies the various limit notations for this special often encountered case.

2.3.28 Notation (Limit in an interval) Let $I \subseteq \mathbb{R}$ be an interval, let $\phi\colon I \to \mathbb{R}$ be a map, and let $a \in I$. We define $\lim_{x\to_I a} \phi(x)$ by

(i) $\lim_{x\to_I a} \phi(x) = \lim_{x\uparrow a} \phi(x)$ if $a = \sup I$,

(ii) $\lim_{x\to_I a} \phi(x) = \lim_{x\downarrow a} \phi(x)$ if $a = \inf I$, and

(iii) $\lim_{x\to_I a} \phi(x) = \lim_{x\to a} \phi(x)$ otherwise. •

We expect that most readers will be familiar with the idea here, even if the notation is not conventional. Let us also give the notation a precise characterisation in terms of limits of sequences in the case when the point x_0 is in the closure of the set A.

2.3.29 Proposition (Convergence in $\mathbb{R}$-nets in terms of sequences) *Let $(A, \preceq)$ be a $\mathbb{R}$-directed set and let $\phi\colon A \to \mathbb{R}$ be a $\mathbb{R}$-net in $(A, \preceq)$. Then, corresponding to the cases and subcases of Notation 2.3.27, we have the following statements:*

(i) (a) if $x_0 \in \mathrm{cl}(A)$, the following statements are equivalent:

I. $\lim_{x\uparrow x_0} \phi(x) = s_0$;

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ for every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A satisfying $\lim_{j\to\infty} x_j = x_0$;

(b) the following statements are equivalent:

I. $\lim_{x\to\infty} \phi(x) = s_0$;

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ for every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A satisfying $\lim_{j\to\infty} x_j = \infty$;

(ii) (a) if $x_0 \in \mathrm{cl}(A)$, the following statements are equivalent:

I. $\lim_{x\downarrow x_0} \phi(x) = s_0$;

II. $\lim_{j\to\infty} \phi(x_j) = s_0$ for every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A satisfying $\lim_{j\to\infty} x_j = x_0$;

(b) the following statements are equivalent:

I. $\lim_{x\to-\infty}\phi(x)=s_0$;

II. $\lim_{j\to\infty}\phi(x_j)=s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty}x_j=-\infty$;

(iii) (a) *if* $x_0\in\mathrm{cl}(A)$, *the following statements are equivalent:*

I. $\lim_{x\uparrow x_0}\phi(x)=s_0$;

II. $\lim_{j\to\infty}\phi(x_j)=s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty}x_j=x_0$;

(b) *if* $x_0\in\mathrm{cl}(A)$, *the following statements are equivalent:*

I. $\lim_{x\downarrow x_0}\phi(x)=s_0$;

II. $\lim_{j\to\infty}\phi(x_j)=s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty}x_j=x_0$;

(c) *the following statements are equivalent:*

I. $\lim_{x\to\infty}\phi(x)=s_0$;

II. $\lim_{j\to\infty}\phi(x_j)=s_0$ *for every sequence* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *in* A *satisfying* $\lim_{j\to\infty}x_j=\infty$;

Proof These statements are all proved in essentially the same way, so let us prove just, say, part (a).

First suppose that $\lim_{x\uparrow x_0}\phi(x)=s_0$, and let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in A converging to x_0. Let $\epsilon\in\mathbb{R}_{>0}$ and choose $x\in A$ such that $|\phi(y)-s_0|<\epsilon$ whenever $y\in A$ satisfies $x\leq y$. Then, since $\lim_{j\to\infty}x_j=x_0$, there exists $N\in\mathbb{Z}_{>0}$ such that $x\leq x_j$ for all $j\geq N$. Clearly, $|\phi(x_j)-s_0|<\epsilon$, so giving convergence of $(\phi(x_j))_{j\in\mathbb{Z}_{>0}}$ to s_0 for every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A converging to x_0.

For the converse, suppose that $\lim_{x\uparrow x_0}\phi(x)\neq s_0$. Then there exists $\epsilon\in\mathbb{R}_{>0}$ such that, for any $x\in A$, we have a $y\in A$ with $x\leq y$ for which $|\phi(y)-s_0|\geq\epsilon$. Since $x_0\in\mathrm{cl}(A)$ it follows that, for any $j\in\mathbb{Z}_{>0}$, there exists $x_j\in\mathsf{B}(\frac{1}{j},x_0)\cap A$ such that $|\phi(x_j)-s_0|\geq\epsilon$. Thus the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in A converging to x_0 has the property that $(\phi(x_j))_{j\in\mathbb{Z}_{>0}}$ does not converge to s_0. ■

Of course, similar conclusions hold when "convergence to s_0" is replaced with "divergence," "convergence to ∞," "convergence to $-\infty$," or "oscillatory." We leave the precise statements to the reader.

Let us give some examples to illustrate that this is all really nothing new.

2.3.30 Examples (Convergence in $\mathbb{R}$-nets)

1. Consider the $\mathbb{R}$-directed set $([0,\infty),\leq)$ and the corresponding $\mathbb{R}$-net ϕ defined by $\phi(x)=\frac{1}{1+x^2}$. This $\mathbb{R}$-net then converges to 0. Let us verify this using the formal definition of convergence of a $\mathbb{R}$-net. For $\epsilon\in\mathbb{R}_{>0}$ choose $x>0$ such that $x^2=\frac{1}{\epsilon}>\frac{1}{\epsilon}-1$. Then, if $x\leq y$, we have

$$\left|\frac{1}{1+y^2}-0\right|<\frac{1}{1+x^2}<\epsilon,$$

giving convergence to $\lim_{x\to\infty}\phi(x)=0$ as stated.

2. Next consider the $\mathbb{R}$-directed set $((0,1],\geq)$ and the corresponding $\mathbb{R}$-net ϕ defined by $\phi(x)=x\sin\frac{1}{x}$. We claim that this $\mathbb{R}$-net converges to 0. To see this, let $\epsilon\in\mathbb{R}_{>0}$ and let $x\in(0,\epsilon)$. Then we have, for $x\geq y$,
$$\left|y\sin\tfrac{1}{y}-0\right|=y\leq x<\epsilon,$$
giving $\lim_{x\downarrow 0}\phi(x)=0$ as desired.
3. Consider the $\mathbb{R}$-directed set $([0,\infty),\leq)$ and the associated $\mathbb{R}$-net ϕ defined by $\phi(x)=x$. In this case we have $\lim_{x\to\infty}\phi(x)=\infty$.
4. Consider the $\mathbb{R}$-directed set $([0,\infty),\leq)$ and the associated $\mathbb{R}$-net ϕ defined by $\phi(x)=x\sin x$. In this case, due to the oscillatory nature of sin, $\lim_{x\to\infty}\phi(x)$ does not exist, nor does it diverge to either ∞ or $-\infty$.
5. Take the $\mathbb{R}$-directed set $(\mathbb{R}\setminus\{0\},\leq_0)$. Define the $\mathbb{R}$-net ϕ by $\phi(x)=x$. Clearly, $\lim_{x\to 0}\phi(x)=0$. •

There are also generalisations of lim sup and lim inf to $\mathbb{R}$-nets. We let $D=(A,\preceq)$ be a $\mathbb{R}$-directed set and let $\phi\colon A\to\mathbb{R}$ be a $\mathbb{R}$-net in this $\mathbb{R}$-directed set. We denote by $\sup_D\phi,\inf_D\phi\colon A\to\mathbb{R}$ the $\mathbb{R}$-nets in D given by
$$\sup_D\phi(x)=\sup\{\phi(y)\mid x\preceq y\},\quad \inf_D\phi(x)=\inf\{\phi(y)\mid x\preceq y\}.$$
Then we define
$$\limsup_D\phi=\lim_D\sup_D\phi,\quad \liminf_D\phi=\lim_D\inf_D\phi.$$
These allow us to talk of limits in cases where limits in the usual sense to not exist. Let us consider this via an example.

2.3.31 Example (lim sup and lim inf in $\mathbb{R}$-nets) We consider the $\mathbb{R}$-directed set $D=([0,\infty),\leq)$ and let ϕ be the $\mathbb{R}$-net defined by $\phi(x)=\mathrm{e}^{-x}+\sin x$.[5] We claim that $\limsup_D\phi=1$ and that $\liminf_D\phi=-1$. Let us prove the first claim, and leave the second as an exercise. We then have
$$\sup_D\phi(x)=\sup\{\mathrm{e}^{-y}+\sin y\mid x\leq y\}=\mathrm{e}^{-x}+1.$$
First note that $\sup_D\phi(x)\geq 1$ for every $x\in[0,\infty)$, and so $\limsup_D\phi\geq 1$. Now let $\epsilon\in\mathbb{R}_{>0}$ and take $x>\log\epsilon$. Then, for any $y\geq x$,
$$\sup_D\phi(y)=\mathrm{e}^{-y}+1\leq 1+\epsilon.$$
Therefore, $\limsup_D\phi\leq 1$, and so $\limsup_D\phi=1$, as desired. •

[5] We have not yet defined e^{-x} or $\sin x$. The reader who is unable to go on without knowing what these functions really are can skip ahead to Section 3.8.

2.3.8 A first glimpse of Landau symbols

In this section we introduce for the first time the so-called Landau symbols. These provide commonly used notation for when two functions behave "asymptotically" the same. Given our development of $\mathbb{R}$-nets in the preceding section, it is easy for us to be fairly precise here. We also warn the reader that the Landau symbols often get used in an imprecise or vague way. We shall try to avoid such usage.

We begin with the definition.

2.3.32 Definition (Landau symbols "O" and "o") Let $D = (A, \preceq)$ be a $\mathbb{R}$-directed set and let $\phi \colon A \to \mathbb{R}$.

(i) Denote by $O_D(\phi)$ the functions $\psi \colon A \to \mathbb{R}$ for which there exists $x_0 \in A$ and $M \in \mathbb{R}_{>0}$ such that $|\psi(x)| \le M|\phi(x)|$ for $x \in A$ satisfying $x_0 \preceq x$.

(ii) Denote by $o_D(\phi)$ the functions $\psi \colon A \to \mathbb{R}$ such that, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $x_0 \in A$ such that $|\psi(x)| < \epsilon|\phi(x)|$ for $x \in A$ satisfying $x_0 \preceq x$.

If $\psi \in O_D(\phi)$ (resp. $\psi \in o_D(\phi)$) then we say that ψ is ***big oh of*** $\boldsymbol{\phi}$ (resp. ***little oh of*** $\boldsymbol{\phi}$). •

It is very common to see simply $O(\phi)$ and $o(\phi)$ in place of $O_D(\phi)$ and $o_D(\phi)$. This is because the most common situation for using this notation is in the case when $\sup A = \infty$ and $\preceq = \le$. In such cases, the notation indicates means, essentially, that $\psi \in O(\phi)$ if ψ has "size" no larger than ϕ for large values of the argument and that $\psi \in o(\phi)$ if ψ is "small" compared to ϕ for large values of the argument. However, we shall use the Landau symbols in other cases, so we allow the possibility of explicitly including the $\mathbb{R}$-directed set in our notation for the sake of clarity.

It is often the case that the comparison function ϕ is positive on A. In such cases, one can give a somewhat more concrete characterisation of O_D and o_D.

2.3.33 Proposition (Alternative characterisation of Landau symbols) *Let* $\mathrm{D} = (\mathrm{A}, \preceq)$ *be a* $\mathbb{R}$*-directed set, and let* $\phi \colon \mathrm{A} \to \mathbb{R}_{>0}$ *and* $\psi \colon \mathrm{A} \to \mathbb{R}$*. Then*

(i) $\psi \in \mathrm{O}_{\mathrm{D}}(\phi)$ *if and only if* $\limsup_{\mathrm{D}} \frac{\psi}{\phi} < \infty$ *and*

(ii) $\psi \in \mathrm{o}_{\mathrm{D}}(\phi)$ *if and only if* $\lim_{\mathrm{D}} \frac{\psi}{\phi} = 0$*.*

Proof We leave this as Exercise 2.3.5. ■

Let us give some common examples of where the Landau symbols are used. Some examples will make use of ideas we have not yet discussed, but which we imagine are familiar to most readers.

2.3.34 Examples (Landau symbols)

1. Let $I \subseteq \mathbb{R}$ be an interval for which $x_0 \in I$ and let $f \colon I \to \mathbb{R}$. Consider the $\mathbb{R}$-directed set $D = (I \setminus \{x_0\}, \preceq_{x_0})$ and the $\mathbb{R}$-net ϕ in D given by $\phi(x) = 1$. Define $g_{f,x_0} \colon I \to \mathbb{R}$ by $g_{f,x_0}(x) = f(x_0)$. We claim that f is continuous at x_0 if and only if

$f - g_{f,x_0} \in o_D(\phi)$. Indeed, by Theorem 3.1.3 we have that f is continuous at x_0 if and only if

$$\begin{aligned} & \lim_{x \to_I x_0} f(x) = f(x_0) \\ \implies \quad & \lim_{x \to_I x_0} (f(x) - g_{f,x_0}(x)) = 0 \\ \implies \quad & \lim_{x \to_I x_0} \frac{(f(x) - g_{f,x_0}(x))}{\phi(x)} = 0 \\ \implies \quad & f - g_{f,x_0} \in o_D(\phi). \end{aligned}$$

The idea is that f is continuous at x_0 if and only if f is "approximately constant" near x_0.

2. Let $I \subseteq \mathbb{R}$ be an interval for which $x_0 \in I$ and let $f\colon I \to \mathbb{R}$. For $L \in \mathbb{R}$ define $g_{f,x_0,L}\colon I \setminus \{x_0\} \to \mathbb{R}$ by
$$g_{x_0,L}(x) = f(x_0) + L(x - x_0).$$
Consider the $\mathbb{R}$-directed set $D = (I \setminus \{x_0\}, \leq_{x_0})$, and define $\phi\colon I \setminus \{x_0\} \to \mathbb{R}_{>0}$ by $\phi(x) = |x - x_0|$. Then we claim that f is differentiable at x_0 with derivative $f'(x_0) = L$ if and only if $f - g_{f,x_0,L} \in o_D(\phi)$. Indeed, by definition, f is differentiable at x_0 with derivative $f'(x_0) = L$ if and only if, then
$$\begin{aligned} & \lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} = L \\ \iff \quad & \lim_{x \to_I x_0} \frac{1}{x - x_0} \big(f(x) - g_{f,x_0,L}(x)\big) = 0 \\ \iff \quad & \lim_{x \to_I x_0} \frac{1}{|x - x_0|} \big(f(x) - g_{f,x_0,L}(x)\big) = 0 \\ \iff \quad & f(x) - g_{f,x_0,L}(x) \in o_D(\phi), \end{aligned}$$
using Proposition 2.3.33. The idea is that f is differentiable at x_0 if and only if f is "nearly linear" at x_0.

3. We can generalise the preceding two examples. Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in I$, and consider the $\mathbb{R}$-directed set $(I \setminus \{x_0\}, \leq_{x_0})$. For $m \in \mathbb{Z}_{\geq 0}$ define the $\mathbb{R}$-net ϕ_m in D by $\phi_m(x) = |x - x_0|^m$. We shall say that a function $f\colon I \to \mathbb{R}$ ***vanishes to order* $\mathbf{m}$ *at* $\mathbf{x_0}$** if $f \in O_D(\phi_m)$. Moreover, f is m-times differentiable at x_0 with $f^{(j)}(x_0) alpha_j$, $j \in \{0, 1, \dots, m\}$, if and only if $f - g_{f,x_0,\boldsymbol{\alpha}} \in o_D(\phi_m)$, where
$$g_{f,x_0,\boldsymbol{\alpha}}(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_m x^m.$$

4. One of the common places where Landau symbols are used is in the analysis of the complexity of algorithms. An algorithm, loosely speaking, takes some input data, performs operations on the data, and gives an outcome. A very simple example of an algorithm is the multiplication of two square matrices,

and we will use this simple example to illustrate our discussion. It is assumed that the size of the input data is measured by an integer N. For example, for the multiplication of square matrices, this integer is the size of the matrices. The complexity of an algorithm is then determined by the number of steps, denoted by, say, $\psi(N)$, of a certain type in the algorithm. For example, for the multiplication of square matrices, this number is normally taken to be the number of multiplications that are needed, and this is easily seen to be no more than N^2. To describe the complexity of the algorithm, one finds uses Landau symbols in the following way. First of all, we use the $\mathbb{R}$-directed set $D = (\mathbb{Z}_{>0}, \leq)$. If $\phi\colon \mathbb{Z}_{>0} \to \mathbb{R}_{>0}$ is such that $\psi \in O_D(\phi)$, then we say the algorithm ***is* O($\boldsymbol{\phi}$)**. For example, matrix multiplication is $O(N^2)$.

In Theorem IV-7.2.20 we show that the computational complexity of the so-called Cooley–Tukey algorithm for computing the FFT is $O(N \log N)$.

Since we are talking about computational complexity of algorithms, it is a good time to make mention of an important problem in the theory of computational complexity. This discussion is limited to so-called decision algorithms, where the outcome is an affirmative or negative declaration about some problem, e.g., is the determinant of a matrix bounded by some number. For such an algorithm, a ***verification algorithm*** is an algorithm that checks whether given input data does indeed give an affirmative answer. Denote by ***P*** the class of algorithms that are $O(N^m)$ for some $m \in \mathbb{Z}_{>0}$. Such algorithms are known as ***polynomial time*** algorithms. Denote by ***NP*** the class of algorithms for which there exists a verification algorithm that is $O(N^m)$ for some $m \in \mathbb{Z}_{>0}$. An important unresolved question is, "Does P=NP?" •

2.3.9 Notes

Citation for Dedekind cuts.

Exercises

2.3.1 Show that if $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a sequence in $\mathbb{R}$ and if $\lim_{j\to\infty} x_j = x_0$ and $\lim_{j\to\infty} x_j = x_0'$, then $x_0 = x_0'$.

2.3.2 Answer the following questions:

(a) find a subset $S \subseteq \mathbb{Q}$ that possesses an upper bound in $\mathbb{Q}$, but which has no least element;

(b) find a bounded monotonic sequence in $\mathbb{Q}$ that does not converge in $\mathbb{Q}$.

2.3.3 Do the following.

(a) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty} \left|\frac{x_{j+1}}{x_j}\right| = 1$ and which converges in $\mathbb{R}$.

(b) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty} \left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to ∞.

(c) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to $-\infty$.

(d) Find a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which is oscillatory.

In the next exercise you will show that the property that a bounded, monotonically increasing sequence converges implies that Cauchy sequences converge. This completes the argument needed to prove the theorem stated in Aside 2.3.9 concerning characterisations of complete ordered fields.

2.3.4 Assume that every bounded, monotonically increasing sequence in $\mathbb{R}$ converges, and using this show that every Cauchy sequence in $\mathbb{R}$ converges using an argument as follows.

1. Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a Cauchy sequence.
2. Let $I_0 = [a, b]$ be an interval that contains all elements of $(x_j)_{j\in\mathbb{Z}_{>0}}$ (why is this possible?)
3. Split $[a, b]$ into two equal length closed intervals, and argue that in at least one of these there is an infinite number of points from the sequence. Call this interval I_1 and let $x_{k_i} \in (x_j)_{j\in\mathbb{Z}_{>0}} \cap I_1$.
4. Repeat the process for I_1 to find an interval I_2 which contains an infinite number of points from the sequence. Let $x_{k_2} \in (x_j)_{j\in\mathbb{Z}_{>0}} \cap I_2$.
5. Carry on doing this to arrive at a sequence $(x_{k_j})_{j\in\mathbb{Z}_{>0}}$ of points in $\mathbb{R}$ and a sequence $(I_j)_{j\in\mathbb{Z}_{>0}}$.
6. Argue that the sequence of left endpoints of the intervals $(I_j)_{j\in\mathbb{Z}_{>0}}$ is a bounded monotonically increasing sequence, and that the sequence of right endpoints is a bounded monotonically decreasing sequence. and so both converge.
7. Show that they converge to the same number, and that the sequence $(x_{k_j})_{j\in\mathbb{Z}_{>0}}$ also converges to this limit.
8. Show that the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ converges to this limit.

2.3.5 Prove Proposition 2.3.33.

Section 2.4

Series in $\mathbb{R}$

From a sequence $(x_j)_{j\in\mathbb{R}}$ in $\mathbb{R}$, one can consider, in principle, the infinite sum $\sum_{j=1}^{\infty} x_j$. Of course, such a sum *a priori* makes no sense. However, as we shall see in Chapter IV-1, such infinite sums are important for characterising certain discrete-time signal spaces. Moreover, such sums come up frequently in many places in analysis. In this section we outline some of the principle properties of these sums.

Do I need to read this section? Most readers will probably have seen much of the material in this section in their introductory calculus course. What might be new for some readers is the fairly careful discussion in Theorem 2.4.5 of the difference between convergence and absolute convergence of series. Since absolute convergence will be of importance to us, it might be worth understanding in what ways it is different from convergence. The material in Section 2.4.7 can be regarded as optional until it is needed during the course of reading other material in the text. •

2.4.1 Definitions and properties of series

A ***series*** in $\mathbb{R}$ is an expression of the form

$$S = \sum_{j=1}^{\infty} x_j, \tag{2.5}$$

where $x_j \in \mathbb{R}$, $j \in \mathbb{Z}_{>0}$. Of course, the problem with this "definition" is that the expression (2.5) is meaningless as an element of $\mathbb{R}$ unless it possesses additional features. For example, if $x_j = 1$, $j \in \mathbb{Z}_{>0}$, then the sum is infinite. Also, if $x_j = (-1)^j$, $j \in \mathbb{Z}_{>0}$, then it is not clear what the sum is: perhaps it is 0 or perhaps it is 1. Therefore, to be precise, a series is prescribed by the sequence of numbers $(x_j)_{j\in\mathbb{Z}_{>0}}$, and is represented in the form (2.5) in order to distinguish it from the sequence with the same terms.

If the expression (2.5) is to have meaning as a number, we need some sort of condition placed on the terms in the series.

2.4.1 Definition (Convergence and absolute convergence of series) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}$ and consider the series

$$S = \sum_{j=1}^{\infty} x_j.$$

The corresponding sequence of ***partial sums*** is the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ defined by

$$S_k = \sum_{j=1}^{k} x_j.$$

Let $s_0 \in \mathbb{R}$. The series:

(i) ***converges to*** $\mathbf{s_0}$, and we write $\sum_{j=1}^{\infty} x_j = s_0$, if the sequence of partial sums converges to s_0;

(ii) has s_0 as a ***limit*** if it converges to s_0;

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the series

$$\sum_{j=1}^{\infty} |x_j|$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\sum_{j=1}^{\infty} x_j = \infty$ (resp. $\sum_{j=1}^{\infty} x_j = -\infty$), if the sequence of partial sums diverges to ∞ (resp. diverges to $-\infty$);

(viii) has a limit that ***exists*** if $\lim_{j\to\infty} S_j \in \mathbb{R}$;

(ix) is ***oscillatory*** if the sequence of partial sums is oscillatory. •

Let us consider some examples of series in $\mathbb{R}$.

2.4.2 Examples (Series in $\mathbb{R}$)

1. First we consider the ***geometric series*** $\sum_{j=1}^{\infty} x^{j-1}$ for $x \in \mathbb{R}$. We claim that this series converges if and only if $|x| < 1$. To prove this we claim that the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ of partial sums is defined by

$$S_k = \begin{cases} \frac{1-x^{k+1}}{1-x}, & x \neq 1, \\ k, & x = 1. \end{cases}$$

The conclusion is obvious for $x = 1$, so we can suppose that $x \neq 1$. The conclusion is obvious for $k = 1$, so suppose it true for $j \in \{1, \dots, k\}$. Then

$$S_{k+1} = \sum_{j=1}^{k+1} x^j = x^{k+1} + \frac{1-x^{k+1}}{1-x} = \frac{x^{k+1} - x^{k+2} + 1 - x^{k+1}}{1-x} = \frac{1-x^{k+2}}{1-x},$$

as desired. It is clear, then, that if $x = 1$ then the series diverges to ∞. If $x = -1$ then the series is directly checked to be oscillatory; the sequence of partial sums is $\{1, 0, 1, \dots\}$. For $x > 1$ we have

$$\lim_{k\to\infty} S_k = \lim_{k\to\infty} \frac{1 - x^{k+1}}{1 - x} = \infty,$$

showing that the series diverges to ∞ in this case. For $x < -1$ it is easy to see that the sequence of partial sums is oscillatory, but increasing in magnitude. This leaves the case when $|x| < 1$. Here, since the sequence $(x^{k+1})_{k\in\mathbb{Z}_{>0}}$ converges to zero, the sequence of partial sums also converges, and converges to $\frac{1}{1-x}$. (We have used the results concerning the swapping of limits with algebraic operations as described in Section 2.3.6.)

2. We claim that the series $\sum_{j=1}^{\infty} \frac{1}{j}$ diverges to ∞. To show this, we show that the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ is not upper bounded. To show this, we shall show that $S_{2^k} \geq 1 + \frac{1}{2}k$ for all $k \in \mathbb{Z}_{>0}$. This is true directly when $k = 1$. Next suppose that $S_{2^j} \geq 1 + \frac{1}{2}j$ for $j \in \{1, \dots, k\}$. Then

$$\begin{aligned} S_{2^{k+1}} &= S_{2^k} + \frac{1}{2^k + 1} + \frac{1}{2^k + 2} + \dots + \frac{1}{2^{k+1}} \\ &\geq 1 + \frac{1}{2}k + \underbrace{\frac{1}{2^{k+1}} + \dots + \frac{1}{2^{k+1}}}_{2^k \text{ terms}} \\ &= 1 + \frac{1}{2}k + \frac{2^k}{2^{k+1}} = 1 + \frac{1}{2}(k+1). \end{aligned}$$

Thus the sequence of partial sums is indeed unbounded, and since it is monotonically increasing, it diverges to ∞, as we first claimed.

3. We claim that the series $S = \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j}$ converges. To see this, we claim that, for any $m \in \mathbb{Z}_{>0}$, we have

$$S_2 \leq S_4 \leq \dots \leq S_{2m} \leq S_{2m-1} \leq \dots \leq S_3 \leq S_1.$$

That $S_2 \leq S_4 \leq \dots \leq S_{2m}$ follows since $S_{2k} - S_{2k-2} = \frac{1}{2k-1} - \frac{1}{2k} > 0$ for $k \in \mathbb{Z}_{>0}$. That $S_{2m} \leq S_{2m-1}$ follows since $S_{2m-1} - S_{2m} = \frac{1}{2m}$. Finally, $S_{2m-1} \leq \dots \leq S_3 \leq S_1$ since $S_{2k-1} - S_{2k+1} = \frac{1}{2k} - \frac{1}{2k+1} > 0$ for $k \in \mathbb{Z}_{>0}$. Thus the sequences $(S_{2k})_{k\in\mathbb{Z}_{>0}}$ and $(S_{2k-1})_{k\in\mathbb{Z}_{>0}}$ are monotonically increasing and monotonically decreasing, respectively, and their tails are getting closer and closer together since $\lim_{m\to\infty} S_{2m-1} - S_{2m} = \frac{1}{2m} = 0$. By Lemma 2 from the proof of Theorem 2.3.7, it follows that the sequences $(S_{2k})_{k\in\mathbb{Z}_{>0}}$ and $(S_{2k-1})_{k\in\mathbb{Z}_{>0}}$ converge and converge to the same limit. Therefore, the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ converges as well to the same limit. One can moreover show that the limit of the series is $\log 2$, where log denotes the natural logarithm.

Note that we have now shown that the series $\sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j}$ converges, but does not converge absolutely; therefore, it is conditionally convergent.

4. We next consider the ***harmonic series*** $\sum_{j=1}^{\infty} j^{-k}$ for $k \in \mathbb{Z}_{\geq 0}$. For $k = 1$ this agrees with our example of part 2. We claim that this series converges if and only if $k > 1$. We have already considered the case of $k = 1$. For $k < 1$ we have $j^{-k} \geq j^{-1}$ for $j \in \mathbb{Z}_{>0}$. Therefore,
$$\sum_{j=1}^{\infty} j^{-k} \geq \sum_{j=1}^{\infty} j^{-1} = \infty,$$
showing that the series diverges to ∞.
For $k > 1$ we note that the sequence of partial sums is monotonically increasing. Thus, so show convergence of the series it suffices by Theorem 2.3.8 to show that the sequence of partial sums is bounded above. Let $N \in \mathbb{Z}_{>0}$ and take $j \in \mathbb{Z}_{>0}$ such that $N < 2^j - 1$. Then the Nth partial sum satisfies
$$\begin{aligned} S_N \leq S_{2^j-1} &= 1 + \frac{1}{2^k} + \frac{1}{3^k} + \cdots + \frac{1}{(2^j-1)^k} \\ &= 1 + \underbrace{\left(\frac{1}{2^k} + \frac{1}{3^k}\right)}_{2 \text{ terms}} + \underbrace{\left(\frac{1}{4^k} + \cdots + \frac{1}{7^k}\right)}_{4 \text{ terms}} + \cdots + \underbrace{\left(\frac{1}{(2^{j-1})^k} + \cdots + \frac{1}{(2^j-1)^k}\right)}_{2^{j-1} \text{ terms}} \\ &< 1 + \frac{2}{2^k} + \frac{4}{4^k} + \cdots + \frac{2^{j-1}}{(2^{j-1})^k} \\ &= 1 + \frac{1}{2^{k-1}} + \left(\frac{1}{2^{k-1}}\right)^2 + \cdots + \left(\frac{1}{2^{k-1}}\right)^{j-1}. \end{aligned}$$
Now we note that the last expression on the right-hand side is bounded above by the sum $\sum_{j=1}^{\infty}(2^{k-1})^{j-1}$, which is a convergent geometric series as we saw in part 1. This shows that S_N is bounded above by this sum for all N, so showing that the harmonic series converges for $k > 1$.
5. The series $\sum_{j=1}^{\infty}(-1)^{j+1}$ does not converge, and also does not diverge to ∞ or $-\infty$. Therefore, it is oscillatory. •

Let us next explore relationships between the various notions of convergence. First we relate the notions of convergence and absolute convergence in the only possible way, given that the series $\sum_{j=1} \frac{(-1)^{j+1}}{j}$ has been shown to be convergent, but not absolutely convergent.

2.4.3 Proposition (Absolutely convergent series are convergent) *If a series $\sum_{j=1}^{\infty} \mathrm{x_j}$ is absolutely convergent, then it is convergent.*

Proof Denote
$$s_k = \sum_{j=1}^{k} x_j, \quad \sigma_k = \sum_{j=1}^{k} |x_j|,$$
and note that $(\sigma_k)_{k\in\mathbb{Z}_{>0}}$ is a Cauchy sequence since the series $\sum_{j=1}^{\infty} x_j$ is absolutely convergent. Thus let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|\sigma_k - \sigma_l| < \epsilon$ for $k, l \geq N$.

For $m > k$ we then have

$$|s_m - s_k| = \left|\sum_{j=k+1}^{m} x_j\right| \le \sum_{j=k+1}^{m} |x_j| = |\sigma_m - \sigma_k| < \epsilon,$$

where we have used Exercise 2.4.3. Thus, for $m > k \ge N$ we have $|s_m - s_k| < \epsilon$, showing that $(s_k)_{k\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, and so convergent by Theorem 2.3.5. ■

The following result is often useful.

2.4.4 Proposition (Swapping summation and absolute value) *For a sequence* $(\mathrm{x_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$*, if the series* $\mathrm{S} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x_j}$ *is absolutely convergent, then*

$$\left|\sum_{\mathrm{j}=1}^{\infty} \mathrm{x_j}\right| \le \sum_{\mathrm{j}=1}^{\infty} |\mathrm{x_j}|.$$

Proof Define

$$S_m^1 = \left|\sum_{j=1}^{m} x_j\right|, \quad S_m^2 = \sum_{j=1}^{m} |x_j|, \qquad m \in \mathbb{Z}_{>0}.$$

By Exercise 2.4.3 we have $S_m^1 \le S_m^2$ for each $m \in \mathbb{Z}_{>0}$. Moreover, by Proposition 2.4.3 the sequences $(S_m^1)_{m\in\mathbb{Z}_{>0}}$ and $(S_m^2)_{m\in\mathbb{Z}_{>0}}$ converge. It is then clear (why?) that

$$\lim_{m\to\infty} S_m^1 \le \lim_{m\to\infty} S_m^2,$$

which is the result. ■

It is not immediately clear on a first encounter why the notion of absolute convergence is useful. However, as we shall see in Chapter IV-1, it is the notion of absolute convergence that will be of most use to us in our characterisation of discrete signal spaces. The following result indicates why mere convergence of a series is perhaps not as nice a notion as one would like, and that absolute convergence is in some sense better behaved.

2.4.5 Theorem (Convergence and rearrangement of series) *For a series* $\mathrm{S} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x_j}$*, the following statements hold:*

(i) *if* S *is conditionally convergent then, for any* $\mathrm{s_0} \in \mathbb{R}$*, there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the series* $\mathrm{S_\phi} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x}_{\phi(\mathrm{j})}$ *converges to* $\mathrm{s_0}$*;*

(ii) *if* S *is conditionally convergent then there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the series* $\mathrm{S_\phi} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x}_{\phi(\mathrm{j})}$ *diverges to* ∞*;*

(iii) *if* S *is conditionally convergent then there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the series* $\mathrm{S_\phi} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x}_{\phi(\mathrm{j})}$ *diverges to* $-\infty$*;*

(iv) *if* S *is conditionally convergent then there exists a bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$ *such that the limit of the partial sums for the series* $\mathrm{S_\phi} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x}_{\phi(\mathrm{j})}$ *is oscillating;*

(v) *if* S *is absolutely convergent then, for any bijection* $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{>0}$*, the series* $\mathrm{S}_\phi = \sum_{\mathrm{j}=1}^\infty \mathrm{x}_{\phi(\mathrm{j})}$ *converges to the same limit as the series* S*.*

Proof We shall be fairly "descriptive" concerning the first four parts of the proof. More precise arguments can be tediously fabricated from the ideas given. We shall use the fact, given as Exercise 2.4.1, that if a series is conditionally convergent, then the two series formed by the positive terms and the negative terms diverge.

(i) First of all, rearrange the terms in the series so that the positive terms are arranged in decreasing order, and the negative terms are arranged in increasing order. We suppose that $s_0 \geq 0$, as a similar argument can be fabricated when $s_0 < 0$. Take as the first elements of the rearranged sequence the enough of the first few positive terms in the sequence so that their sum exceeds s_0. As the next terms, take enough of the first few negative terms in the series such that their sum, combined with the already chosen positive terms, is less than s_0. Now repeat this process. Because the series was initially rearranged so that the positive and negative terms are in descending and ascending order, respectively, one can show that the construction we have given yields a sequence of partial sums that starts greater than s_0, then monotonically decreases to a value less than s_0, then monotonically increases to a value greater than s_0, and so on. Moreover, at the end of each step, the values get closer to s_0 since the sequence of positive and negative terms both converge to zero. An argument like that used in the proof of Proposition 2.3.10 can then be used to show that the resulting sequence of partial sums converges to s_0.

(ii) To get the suitable rearrangement, proceed as follows. Partition the negative terms in the sequence into disjoint finite sets S_j^-, $j \in \mathbb{Z}_{>0}$. Now partition the positive terms in the sequence as follows. Define S_1^+ to be the first N_1 positive terms in the sequence, where N_1 is sufficiently large that the sum of the elements of S_1^+ exceeds by at least 1 in absolute value the sum of the elements from S_1^-. This is possible since the series of positive terms in the sequence diverges to ∞. Now define S_2^+ by taking the next N_2 positive terms in the sequence so that the sum of the elements of S_2^+ exceeds by at least 1 in absolute value the sum of the elements from S_2^-. Continue in this way, defining $S_3^+, S_4^+, \dots$. The rearrangement of the terms in the series is then made by taking the first collection of terms to be the elements of S_1^+, the second collection to be the elements of S_1^-, the third collection to be the elements of S_2^+, and so on. One can verify that the resulting sequence of partial sums diverges to ∞.

(iii) The argument here is entirely similar to the previous case.

(iv) This result follows from part (i) in the following way. Choose an oscillating sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$. For y_1, by part (i) one can find a finite number of terms from the original series whose sum is as close as desired to y_1. These will form the first terms in the rearranged series. Next, the same argument can be applied to the remaining elements of the series to yield a finite number of terms in the series that are as close as desired to y_2. One carries on in this way, noting that since the sequence $(y_j)_{j\in\mathbb{Z}_{>0}}$ is oscillating, so too will be the sequence of partial sums for the rearranged series.

(v) Let $y_j = x_{\phi(j)}$ for $j \in \mathbb{Z}_{>0}$. Then define sequences $(x_j^+)_{j\in\mathbb{Z}_{>0}}$, $(x_j^-)_{j\in\mathbb{Z}_{>0}}$, $(y_j^+)_{j\in\mathbb{Z}_{>0}}$, and $(y_j^-)_{j\in\mathbb{Z}_{>0}}$ by

$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{-x_j, 0\},$$
$$y_j^+ = \max\{y_j, 0\}, \quad y_j^- = \max\{-y_j, 0\}, \qquad j \in \mathbb{Z}_{>0},$$

noting that $|x_j| = \max\{x_j^-, x_j^+\}$ and $|y_j| = \max\{y_j^-, y_j^+\}$ for $j \in \mathbb{Z}_{>0}$. By Proposition 2.4.8 it follows that the series

$$S^+ = \sum_{j=1}^\infty x_j^+, \quad S^- = \sum_{j=1}^\infty x_j^-, \quad S_\phi^+ = \sum_{j=1}^\infty y_j^+, \quad S_\phi^- = \sum_{j=1}^\infty y_j^-$$

converge. We claim that for each $k \in \mathbb{Z}_{>0}$ we have

$$\sum_{j=1}^k x_j^+ \le \sum_{j=1}^\infty y_j^+.$$

To see this, we need only note that there exists $N \in \mathbb{Z}_{>0}$ such that

$$\{x_1^+, \dots, x_k^+\} \subseteq \{y_1^+, \dots, y_N^+\}.$$

With N having this property,

$$\sum_{j=1}^k x_j^+ \le \sum_{j=1}^N y_j^+ \le \sum_{j=1}^\infty y_j^+,$$

as desired. Therefore,

$$\sum_{j=1}^\infty x_j^+ \le \sum_{j=1}^\infty y_j^+.$$

Reversing the argument gives

$$\sum_{j=1}^\infty y_j^+ \le \sum_{j=1}^\infty x_j^+ \quad \Longrightarrow \quad \sum_{j=1}^\infty x_j^+ = \sum_{j=1}^\infty y_j^+.$$

A similar argument also gives

$$\sum_{j=1}^\infty x_j^- = \sum_{j=1}^\infty y_j^-.$$

This then gives

$$\sum_{j=1}^\infty y_j = \sum_{j=1}^\infty y_j^+ - \sum_{j=1}^\infty y_j^- = \sum_{j=1}^\infty x_j^+ - \sum_{j=1}^\infty x_j^- = \sum_{j=1}^\infty x_j,$$

as desired. ■

The theorem says, roughly, that absolute convergence is necessary and sufficient to ensure that the limit of a series be independent of rearrangement of the terms in the series. Note that the necessity portion of this statement, which is parts (i)–(iv) of the theorem, comes in a rather dramatic form which suggests that conditional convergence behaves maximally poorly with respect to rearrangement.

2.4.2 Tests for convergence of series

In this section we give some of the more popular tests for convergence of a series. It is infeasible to expect an easily checkable general condition for convergence. However, in some cases the tests we give here are sufficient.

First we make a simple general observation that is very often useful; it is merely a reflection that the convergence of a series depends only on the tail of the series. We shall often make use of this result without mention.

2.4.6 Proposition (Convergence is unaffected by changing a finite number of terms) *Let $\sum_{j=1} x_j$ and $\sum_{j=1}^{\infty} y_j$ be series in $\mathbb{R}$ and suppose that there exists $K \in \mathbb{Z}$ and $N \in \mathbb{Z}_{>0}$ such that $x_j = y_{j+K}$ for $j \geq N$. Then the following statements hold:*

(i) the series $\sum_{j=1}^{\infty} x_j$ converges if and only if the series $\sum_{j=1}^{\infty} y_j$ converges;

(ii) the series $\sum_{j=1}^{\infty} x_j$ diverges if and only if the series $\sum_{j=1}^{\infty} y_j$ diverges;

(iii) the series $\sum_{j=1}^{\infty} x_j$ diverges to ∞ if and only if the series $\sum_{j=1}^{\infty} y_j$ diverges to ∞;

(iv) the series $\sum_{j=1}^{\infty} x_j$ diverges to $-\infty$ if and only if the series $\sum_{j=1}^{\infty} y_j$ diverges to $-\infty$.

The next convergence result is also a more or less obvious one.

2.4.7 Proposition (Sufficient condition for a series to diverge) *If the sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ does not converge to zero, then the series $\sum_{j=1}^{\infty} x_j$ diverges.*

Proof Suppose that the series $\sum_{j=1}^{\infty} x_j$ converges to s_0 and let $(S_k)_{k\in\mathbb{Z}_{>0}}$ be the sequence of partial sums. Then $x_k = S_k - S_{k-1}$. Then

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} S_k - \lim_{k\to\infty} S_{k-1} = s_0 - s_0 = 0_{\mathsf{V}},$$

as desired. ■

Note that Example 2.4.2–2 shows that the converse of this result is false. That is to say, for a series to converge, it is not sufficient that the terms in the series go to zero. For this reason, checking the convergence of a series numerically becomes something that must be done carefully, since the blind use of the computer with a prescribed numerical accuracy will suggest the false conclusion that a series converges if and only if the terms in the series go to zero as the index goes to infinity.

Another more or less obvious result is the following.

2.4.8 Proposition (Comparison Test) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ and $(y_j)_{j\in\mathbb{Z}_{>0}}$ be sequences of nonnegative numbers for which there exists $\alpha \in \mathbb{R}_{>0}$ satisfying $y_j \leq \alpha x_j$, $j \in \mathbb{Z}_{>0}$. Then the following statements hold:*

(i) the series $\sum_{j=1}^{\infty} y_j$ converges if the series $\sum_{j=1}^{\infty} x_j$ converges;

(ii) the series $\sum_{j=1}^{\infty} x_j$ diverges if the series $\sum_{j=1}^{\infty} y_j$ diverges.

Proof We shall show that, if the series $\sum_{j=1}^{\infty} x_j$ converges, then the sequence $(T_k)_{k\in\mathbb{Z}_{>0}}$ of partial sums for the series $\sum_{j=1}^{\infty} y_j$ is a Cauchy sequence. Since the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ for $\sum_{j=1}^{\infty} x_j$ is convergent, it is Cauchy. Therefore, for $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that whenever $k, m \geq N$, with $k > m$ without loss of generality,

$$S_k - S_m = \sum_{j=m+1}^{k} x_j < \epsilon\alpha^{-1}.$$

Then, for $k, m \geq N$ with $k > m$ we have

$$T_k - T_m = \sum_{j=m+1}^{k} y_j \leq \alpha \sum_{j=m+1}^{k} x_j < \epsilon,$$

showing that $(T_k)_{k\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, as desired.

The second statement is the contrapositive of the first. ■

Now we can get to some less obvious results for convergence of series. The first result concerns series where the terms alternate sign.

2.4.9 Proposition (Alternating Test) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *be a sequence in* $\mathbb{R}$ *satisfying*

(i) $x_j > 0$ *for* $j \in \mathbb{Z}_{>0}$*,*

(ii) $x_{j+1} \leq x_j$ *for* $j \in \mathbb{Z}_{>0}$*, and*

(iii) $\lim_{j\to\infty} x_j = 0$*.*

Then the series $\sum_{j=1}^{\infty}(-1)^{j+1}x_j$ *converges.*

Proof The proof is a straightforward generalisation of that given for Example 2.4.2–3, and we leave for the reader the simple exercise of verifying that this is so. ■

Our next result is one that is often useful.

2.4.10 Proposition (Ratio Test for series) *Let* $(x_j)_{j\in\mathbb{Z}_{>0}}$ *be a nonzero sequence in* $\mathbb{R}$ *with* $\sum_{j=1}^{\infty} x_j$ *the corresponding series. Then the following statements hold:*

(i) if $\limsup_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| < 1$*, then the series converges absolutely;*

(ii) if there exists $N \in \mathbb{Z}_{>0}$ *such that* $\left|\frac{x_{j+1}}{x_j}\right| > 1$ *for all* $j \geq N$*, then the series diverges.*

Proof (i) By Proposition 2.3.15 there exists $\beta \in (0,1)$ and $N \in \mathbb{Z}_{>0}$ such that $\left|\frac{x_{j+1}}{x_j}\right| < \beta$ for $j \geq N$. Then

$$\left|\frac{x_j}{x_N}\right| = \left|\frac{x_{N+1}}{x_N}\right|\left|\frac{x_{N+2}}{x_{N+1}}\right|\cdots\left|\frac{x_j}{x_{j-1}}\right| < \beta^{j-N}, \qquad j > N,$$

implying that

$$|x_j| < \frac{|x_N|}{\beta^N}\beta^j.$$

Since $\beta < 1$, the geometric series $\sum_{j=1}^{\infty} \beta^j$ converges. The result for $\alpha < 1$ now follows by the Comparison Test.

(ii) The sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ cannot converge to 0 in this case, and so this part of the result follows from Proposition 2.4.7. ■

The following simpler test is often stated as the Ratio Test.

2.4.11 Corollary (Weaker version of the Ratio Test) *If $(x_j)_{j\in\mathbb{Z}_{>0}}$ is a nonzero sequence in $\mathbb{R}$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = \alpha$, then the series $\sum_{j=1}^{\infty} x_j$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.*

2.4.12 Remark (Nonzero assumption in Ratio Test) In the preceding two results we asked that the terms in the series be nonzero. This is not a significant limitation. Indeed, one can enumerate the nonzero terms in the series, and then apply the ratio test to this. •

Our next result has a similar character to the previous one.

2.4.13 Proposition (Root Test) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $\limsup_{j\to\infty}|x_j|^{1/j} = \alpha$. Then the series $\sum_{j=1}^{\infty} x_j$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.*

Proof First take $\alpha < 1$ and define $\beta = \frac{1}{2}(\alpha + 1)$. Then, just as in the proof of Proposition 2.4.10, $\alpha < \beta < 1$. By Proposition 2.3.15 there exists $N \in \mathbb{Z}_{>0}$ such that $|x_j|^{1/j} < \beta$ for $j \geq N$. Thus $|x_j| < \beta^j$ for $j \geq N$. Note that $\sum_{j=N+1}^{\infty} \beta^j$ converges by Example 2.4.2–1. Now $\sum_{j=0}^{\infty}|x_j|$ converges by the Comparison Test.

Next take $\alpha > 1$. In this case we have $\lim_{j\to\infty}|x_j| \neq 0$, and so we conclude divergence from Proposition 2.4.7. ■

The following obvious corollary is often stated as the Root Test.

2.4.14 Corollary (Weaker version of Root Test) *Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $\lim_{j\to\infty}|x_j|^{1/j} = \alpha$. Then the series $\sum_{j=1}^{\infty} x_j$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.*

The Ratio Test and the Root Test are related, as the following result indicates.

2.4.15 Proposition (Root Test implies Ratio Test) *If $(p_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a sequence in $\mathbb{R}_{>0}$ then*

$$\liminf_{j\to\infty} \frac{p_{j+1}}{p_j} \leq \liminf_{j\to\infty} p_j^{1/j}$$

$$\limsup_{j\to\infty} p_j^{1/j} \leq \limsup_{j\to\infty} \frac{p_{j+1}}{p_j}.$$

In particular, for a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$, if $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right|$ exists, then $\lim_{j\to\infty}|x_j|^{1/j} = \lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right|$.

Proof For the first inequality, let $\alpha = \liminf_{j\to\infty} \frac{p_{j+1}}{p_j}$. First consider the case where $\alpha = \infty$. Then, given $M \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $\frac{p_{j+1}}{p_j} > M$ for $j \geq N$. Then we have

$$\left|\frac{p_j}{p_N}\right| = \left|\frac{p_{N+1}}{p_N}\right|\left|\frac{p_{N+1}}{p_{N+1}}\right| \cdots \left|\frac{p_j}{p_{j-1}}\right| > M^{j-N}, \qquad j > N.$$

This gives

$$p_j > \frac{p_N}{M^N} M^j, \qquad j > N.$$

Thus $p_j^{1/j} > (\frac{p_N}{M^N})^{1/j} M$. Since $\lim_{j\to\infty}(p_N\beta^{-N})^{1/j} = 1$ (cf. the definition of P_a in Section 3.8.3), we have $\liminf_{j\to\infty} p_j^{1/j} > M$, giving the desired conclusion in this case, since M is arbitrary. Next consider the case when $\alpha \in \mathbb{R}_{>0}$ and let $\beta < \alpha$. By Proposition 2.3.16 there exists $N \in \mathbb{Z}_{>0}$ such that $\frac{p_{j+1}}{p_j} \geq \beta$ for $j \geq N$. Performing just the same computation as above gives $p_j \geq \beta^{j-N} p_N$ for $j \geq N$. Therefore, $p_j^{1/j} \geq (p_N\beta^{-N})^{1/j}\beta$. Since $\lim_{j\to\infty}(p_N\beta^{-N})^{1/j} = 1$ we have $\liminf_{j\to\infty} p_j^{1/j} \geq \beta$. The first inequality follows since $\beta < \alpha$ is arbitrary.

Now we prove the second inequality. Let $\alpha = \limsup_{j\to\infty} \frac{p_{j+1}}{p_j}$. If $\alpha = \infty$ then the second inequality in the statement of the result is trivial. If $\alpha \in \mathbb{R}_{>0}$ then let $\beta > \alpha$ and note that there exists $N \in \mathbb{Z}_{>0}$ such that $\frac{p_{j+1}}{p_j} \leq \beta$ for $j \geq N$ by Proposition 2.3.15. In particular, just as in the proof of Proposition 2.4.10, $p_j \leq \beta^{j-N} p_N$ for $j \geq N$. Therefore, $p_j^{1/j} \leq (p_N\beta^{-N})^{1/j}\beta$. Since $\lim_{j\to\infty}(p_N\beta^{-N})^{1/j} = 1$ we then have $\liminf_{j\to\infty} p_j^{1/j} \leq \beta$. the second inequality follows since $\beta > \alpha$ is arbitrary.

The final assertion follows immediately from the two inequalities using Proposition 2.3.17. ■

In Exercises 2.4.6 and 2.4.7 the reader can explore the various possibilities for the ratio test and root test when $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and $\lim_{j\to\infty}|x_j|^{1/j} = 1$, respectively.

The final result we state in this section can be thought of as the summation version of integration by parts.

2.4.16 Proposition (Abel's[6] partial summation formula) *For sequences* $(\mathrm{x}_j)_{j\in\mathbb{Z}_{>0}}$ *and* $(\mathrm{y}_j)_{j\in\mathbb{Z}_{>0}}$ *of real numbers, denote* $\mathrm{S}_k = \sum_{j=1}^{k} \mathrm{x}_j$. *Then*

$$\sum_{j=1}^{k} \mathrm{x}_j\mathrm{y}_j = \mathrm{S}_k\mathrm{y}_{k+1} - \sum_{j=1}^{k} \mathrm{S}_j(\mathrm{y}_{j+1} - \mathrm{y}_j) = \mathrm{S}_k\mathrm{y}_1 + \sum_{j=1}^{k}(\mathrm{S}_k - \mathrm{S}_j)(\mathrm{y}_{j+1} - \mathrm{y}_j).$$

[6] Niels Henrik Abel (1802–1829) was a Norwegian mathematician who worked in the area of analysis. An important theorem of Abel, one that is worth knowing for people working in application areas, is a theorem stating that there is no expression for the roots of a quintic polynomial in terms of the coefficients that involves only the operations of addition, subtraction, multiplication, division and taking roots.

Proof Let $S_0 = 0$ by convention. Since $x_j = S_j - S_{j-1}$ we have

$$\sum_{j=1}^{n} x_j y_j = \sum_{j=1}^{n} (S_j - S_{j-1}) y_j = \sum_{j=1}^{n} S_j y_j - \sum_{j=1}^{n} S_{j-1} y_j.$$

Trivially,

$$\sum_{j=1}^{n} S_{j-1} y_j = \sum_{j=1}^{n} S_j y_{j+1} - S_n y_{n+1}.$$

This gives the first equality of the lemma. The second follows from a substitution of

$$y_{n+1} = \sum_{j=1}^{n} (y_{j+1} - y_j) + y_1$$

into the first equality. ■

2.4.3 e and π

In this section we consider two particular convergent series whose limits are among the most important of "physical constants."

2.4.17 Definition (e) $\mathrm{e} = \displaystyle\sum_{j=0}^{\infty} \frac{1}{j!}$. •

Note that the series defining e indeed converges, for example, by the Ratio Test. Another common representation of e as a limit is the following.

2.4.18 Proposition (Alternative representations of e) *We have*

$$\mathrm{e} = \lim_{\mathrm{j}\to\infty} \left(1 + \tfrac{1}{\mathrm{j}}\right)^{\mathrm{j}} = \lim_{\mathrm{j}\to\infty} \left(1 + \tfrac{1}{\mathrm{j}}\right)^{\mathrm{j}+1}.$$

Proof First note that if the limit $\lim_{j\to\infty} \left(1 + \frac{1}{j}\right)^j$ exists, then, by Proposition 2.3.23,

$$\lim_{j\to\infty} \left(1 + \tfrac{1}{j}\right)^{j+1} = \lim_{j\to\infty} \left(1 + \tfrac{1}{j}\right)\left(1 + \tfrac{1}{j}\right)^{j} = \lim_{j\to\infty} \left(1 + \tfrac{1}{j}\right)^{j}.$$

Thus we will only prove that $\mathrm{e} = \lim_{j\to\infty} \left(1 + \frac{1}{j}\right)^j$.

Let

$$S_k = \sum_{j=0}^{k} \frac{1}{k!}, \quad A_k = \left(1 + \tfrac{1}{k}\right)^k, \quad B_k = \left(1 + \tfrac{1}{k}\right)^{k+1},$$

be the kth partial sum of the series for e and the kth term in the proposed sequence for e. By the Binomial Theorem (Exercise 2.2.1) we have

$$A_k = \left(1 + \tfrac{1}{k}\right)^k = \sum_{j=0}^{k} \binom{k}{j} \frac{1}{k^j}.$$

Moreover, the exact form for the binomial coefficients can directly be seen to give

$$A_k = \sum_{j=0}^{k} \frac{1}{j!}\Big(1 - \frac{1}{k}\Big)\Big(1 - \frac{2}{k}\Big)\dots\Big(1 - \frac{j-1}{k}\Big).$$

Each coefficient of $\frac{1}{j!}$, $j \in \{0, 1, \dots, k\}$ is then less than 1. Thus $A_k \leq S_k$ for each $k \in \mathbb{Z}_{\geq 0}$. Therefore, $\limsup_{k\to\infty} A_k \leq \limsup_{k\to\infty} S_k$. For $m \leq k$ the same computation gives

$$A_k \geq \sum_{j=0}^{m} \frac{1}{j!}\Big(1 - \frac{1}{k}\Big)\Big(1 - \frac{2}{k}\Big)\dots\Big(1 - \frac{j-1}{k}\Big).$$

Fixing m and letting $k \to \infty$ gives

$$\liminf_{k\to\infty} A_k \geq \sum_{j=0}^{m} \frac{1}{j!} = S_m.$$

Thus $\liminf_{k\to\infty} A_k \geq \liminf_{m\to\infty} S_m$, which gives the result when combined with our previous estimate $\limsup_{k\to\infty} A_k \leq \limsup_{k\to\infty} S_k$. ■

It is interesting to note that the series representation of e allows us to conclude that e is irrational.

2.4.19 Proposition (Irrationality of e) $\mathrm{e} \in \mathbb{R} \setminus \mathbb{Q}$.

Proof Suppose that $\mathrm{e} = \frac{l}{m}$ for $l, m \in \mathbb{Z}_{>0}$. We compute

$$(m-1)!l = m!\mathrm{e} = m!\sum_{j=0}^{\infty} \frac{1}{j!} = \sum_{j=0}^{m} \frac{m!}{j!} + \sum_{j=m+1}^{\infty} \frac{m!}{j!},$$

which then gives

$$\sum_{j=m+1}^{\infty} \frac{m!}{j!} = (m-1)!l - \sum_{j=0}^{m} \frac{m!}{j!},$$

which implies that $\sum_{j=m+1}^{\infty} \frac{m!}{j!} \in \mathbb{Z}_{>0}$. We then compute, using Example 2.4.2–1,

$$0 < \sum_{j=m+1}^{\infty} \frac{m!}{j!} < \sum_{j=m+1}^{\infty} \frac{1}{(m+1)^{j-m}} = \sum_{j=1}^{\infty} \frac{1}{(m+1)^j} = \frac{\frac{1}{m+1}}{1 - \frac{1}{m+1}} = \frac{1}{m} \leq 1.$$

Thus $\sum_{j=m+1}^{\infty} \frac{m!}{j!} \in \mathbb{Z}_{>0}$, being an integer, must equal 1, and, moreover, $m = 1$. Thus we have

$$\sum_{j=2}^{\infty} \frac{1}{j!} = \mathrm{e} - 2 = 1 \quad \Longrightarrow \quad \mathrm{e} = 3.$$

Next let

$$\alpha = \sum_{j=1}^{\infty} \Big(\frac{1}{2^{j-1}} - \frac{1}{j!}\Big),$$

noting that this series for α converges, and converges to a positive number since each term in the series is positive. Then, using Example 2.4.2–1,

$$\alpha = (2 - (e - 1)) \qquad \Longrightarrow \qquad e = 3 - \alpha.$$

Thus e < 3, and we have arrived at a contradiction. ■

Next we turn to the number π. Perhaps the best description of π is that it is the ratio of the circumference of a circle with the diameter of the circle. Indeed, the use of the Greek letter "p" (i.e., π) has its origins in the word "perimeter." However, to make sense of this definition, one must be able to talk effectively about circles, what the circumference means, etc. This is more trouble than it is worth for us at this point. Therefore, we give a more analytic description of π, albeit one that, at this point, is not very revealing of what the reader probably already knows about it.

2.4.20 Definition (π) $\pi = 4\sum_{j=0}^{\infty} \frac{(-1)^j}{2j+1}$. •

By the Alternating Test, this series representation for π converges.

We can also fairly easily show that π is irrational, although our proof uses some facts about functions on $\mathbb{R}$ that we will not discuss until Chapter 3.

2.4.21 Proposition (Irrationality of π) $\pi \in \mathbb{R} \setminus \mathbb{Q}$.

Proof In Section 3.8.4 we will give a definition of the trigonometric functions, sin and cos, and prove that, on $(0, \pi)$, sin is positive, and that $\sin 0 = \sin \pi = 0$. We will also prove the rules of differentiation for trigonometric functions necessary for the proof we now present.

Note that if π is rational, then π^2 is also rational. Therefore, it suffices to show that π^2 is irrational.

Let us suppose that $\pi^2 = \frac{l}{m}$ for $l, m \in \mathbb{Z}_{>0}$. For $k \in \mathbb{Z}_{>0}$ define $f_k\colon [0,1] \to \mathbb{R}$ by

$$f_k(x) = \frac{x^k(1-x)^k}{k!},$$

noting that $\operatorname{image}(f) \subseteq [0, \frac{1}{k!}]$. It is also useful to write

$$f_k(x) = \frac{1}{k!}\sum_{j=k}^{2k} c_j x^j,$$

where we observe that c_j, $j \in \{k, k+1, \dots, 2k\}$ are integers. Define $g_j\colon [0,1] \to \mathbb{R}$ by

$$g_k(x) = k^j \sum_{j=0}^{k} (-1)^j \pi^{2(k-j)} f^{(2j)}(x).$$

A direct computation shows that

$$f_k^{(j)}(0) = 0, \qquad j < k,\ j > 2k,$$

and that

$$f_k^{(j)}(0) = \frac{j!}{k!}c_j, \qquad j \in \{k, k+1, \dots, 2k\},$$

is an integer. Thus f and all of its derivatives take integer values at $x = 0$, and therefore also at $x = 1$ since $f_k(x) = f_k(1-x)$. One also verifies directly that $g_k(0)$ and $g_k(1)$ are integers.

Now we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}(g_k'(x)\sin\pi x - \pi g_k(x)\cos\pi x) &= (g_k''(x) + \pi^2 g_k(x))\sin\pi x \\ &= m^k\pi^{2k+2}f(x)\sin\pi x = \pi^2 l^k f(x)\sin\pi x,\end{aligned}$$

using the definition of g_k and the fact that $\pi^2 = \frac{l}{m}$. By the Fundamental Theorem of Calculus we then have, after a calculation,

$$\pi l^k \int_0^1 f(x)\sin\pi x\,\mathrm{d}x = g_k(0) + g_k(1) \in \mathbb{Z}_{>0}.$$

But we then have, since the integrand in the above integral is nonnegative,

$$0 < \pi l^k \int_0^1 f(x)\sin\pi x\,\mathrm{d}x < \frac{\pi l^k}{k!}$$

given the bounds on f_k. Note that $\lim_{k\to\infty}\frac{l^k}{k!} = 0$. Since the above computations hold for any k, if we take k sufficiently large that $\frac{\pi l^k}{k!} < 1$, we arrive at a contradiction. ■

2.4.4 Doubly infinite series

We shall frequently encounter series whose summation index runs not from 1 to ∞, but from $-\infty$ to ∞. Thus we call a family $(x_j)_{j\in\mathbb{Z}}$ of elements of $\mathbb{R}$ a ***doubly infinite sequence*** in $\mathbb{R}$, and a sum of the form $\sum_{j=-\infty}^{\infty} x_j$ a ***doubly infinite series***. A little care need to be shown when defining convergence for such series, and here we give the appropriate definitions.

2.4.22 Definition (Convergence and absolute convergence of doubly infinite series) Let $(x_j)_{j\in\mathbb{Z}}$ be a doubly infinite sequence and let $S = \sum_{j=-\infty}^{\infty} x_j$ be the corresponding doubly infinite series. The sequence of ***single partial sums*** is the sequence $(S_k)_{k\in\mathbb{Z}_{>0}}$ where

$$S_k = \sum_{j=-k}^{k} x_j,$$

and the sequence of ***double partial sums*** is the double sequence $(S_{k,l})_{k,l\in\mathbb{Z}_{>0}}$ defined by

$$S_{k,l} = \sum_{j=-k}^{l} x_j.$$

Let $s_0 \in \mathbb{R}$. The doubly infinite series:

(i) ***converges to*** $\mathbf{s_0}$ if the double sequence of partial sums converges to s_0;

(ii) has s_0 as a ***limit*** if it converges to s_0;

(iii) is ***convergent*** if it converges to some element of $\mathbb{R}$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the doubly infinite series

$$\sum_{j=-\infty}^{\infty} |x_j|$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\sum_{j=-\infty}^{\infty} x_j = \infty$ (resp. $\sum_{j=-\infty}^{\infty} x_j = -\infty$), if the sequence of double partial sums diverges to ∞ (resp. diverges to $-\infty$);

(viii) has a limit that ***exists*** if $\sum_{j=-\infty}^{\infty} x_j \in \mathbb{R}$;

(ix) is ***oscillatory*** if the limit of the double sequence of partial sums is oscillatory. •

2.4.23 Remark (Partial sums versus double partial sums) Note that the convergence of the sequence of partial sums is not a very helpful notion, in general. For example, the series $\sum_{j=-\infty}^{\infty} j$ possesses a sequence of partial sums that is identically zero, and so the sequence of partial sums obviously converges to zero. However, it is not likely that one would wish this doubly infinite series to qualify as convergent. Thus partial sums are not a particularly good measure of convergence. However, there are situations—for example, the convergence of Fourier series (see Chapter IV-5)—where the standard notion of convergence of a doubly infinite series is made using the partial sums. However, in these cases, there is additional structure on the setup that makes this a reasonable thing to do. •

The convergence of a doubly infinite series has the following useful, intuitive characterisation.

2.4.24 Proposition (Characterisation of convergence of doubly infinite series) *For a doubly infinite series* $S = \sum_{j=-\infty}^{\infty} x_j$*, the following statements are equivalent:*

(i) S *converges;*

(ii) the two series $\sum_{j=0}^{\infty} x_j$ *and* $\sum_{j=1}^{\infty} x_{-j}$ *converge.*

Proof For $k, l \in \mathbb{Z}_{>0}$, denote

$$S_{k,l} = \sum_{-k}^{l} x_j, \quad S_k^+ = \sum_{j=0}^{k} x_j, \quad S_k^- = \sum_{-k}^{-1} x_j,$$

so that $S_{k,l} = S^-_k + S^+_l$.

(i) $\Longrightarrow$ (ii) Let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $|S_{j,k} - s_0| < \frac{\epsilon}{2}$ for $j,k \geq N$. Now let $j,k \geq N$, choose some $l \geq N$, and compute

$$|S^+_j - S^+_k| \leq |S^+_j + S^-_l - s_0| + |S^+_k + S^-_l - s_0| < \epsilon.$$

Thus $(S^+_j)_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence, and so is convergent. A similar argument shows that $(S^-_j)_{j\in\mathbb{Z}_{>0}}$ is also a Cauchy sequence.

(ii) $\Longrightarrow$ (i) Let s^+ be the limit of $\sum_{j=0}^{\infty} x_j$ and let s^- be the limit of $\sum_{j=1}^{\infty} x_{-j}$. For $\epsilon \in \mathbb{R}_{>0}$ define $N^+, N^- \in \mathbb{Z}_{>0}$ such that $|S^+_j - s^+| < \frac{\epsilon}{2}$, $j \geq N^+$, and $|S^-_j - s^-| < \frac{\epsilon}{2}$, $j \leq -N^-$. Then, for $j,k \geq \max\{N^-, N^+\}$,

$$|S_{j,k} - (s^+ + s^-)| = |S^+_k - s_+ + S^-_j - s_-| \leq |S^+_k - s_+| + |S^-_j - s_-| < \epsilon,$$

thus showing that S converges. ■

Thus convergent doubly infinite series are really just combinations of convergent series in the sense that we have studied in the preceding sections. Thus, for example, one can use the tests of Section 2.4.2 to check for convergence of a doubly infinite series by applying them to both "halves" of the series. Also, the relationships between convergence and absolute convergence for series also hold for doubly infinite series. And a suitable version of Theorem 2.4.5 also holds for doubly infinite series. These facts are so straightforward that we will assume them in the sequel without explicit mention; they all follow directly from Proposition 2.4.24.

2.4.5 Multiple series

Just as we considered multiple sequences in Section 2.3.5, we can consider multiple series. As we did with sequences, we content ourselves with double series.

2.4.25 Definition (Double series) A ***double series*** in $\mathbb{R}$ is a sum of the form $\sum_{j,k=1}^{\infty} x_{jk}$ where $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$ is a double sequence in $\mathbb{R}$. •

While our definition of a series was not entirely sensible since it was not really identifiable as anything unless it had certain convergence properties, for double series, things are even worse. In particular, it is not clear what $\sum_{j,k=1}^{\infty} x_{jk}$ means. Does it mean $\sum_{j=1}^{\infty}\left(\sum_{k=1}^{\infty} x_{jk}\right)$? Does it mean $\sum_{k=1}^{\infty}\left(\sum_{j=1}^{\infty} x_{jk}\right)$? Or does it mean something different from both of these? The only way to rectify our poor mathematical manners is to define convergence for double series as quickly as possible.

2.4.26 Definition (Convergence and absolute convergence of double series) Let $(x_{jk})_{j,k\in\mathbb{Z}_{>0}}$ be a double sequence in $\mathbb{R}$ and consider the double series

$$S = \sum_{j,k=1}^{\infty} x_{jk}.$$

The corresponding sequence of ***partial sums*** is the double sequence $(S_{jk})_{j,k\in\mathbb{Z}_{>0}}$ defined by

$$S_{jk} = \sum_{l=1}^{j}\sum_{m=1}^{k} x_{lm}.$$

Let $s_0 \in \mathbb{R}$. The double series:

(i) ***converges to*** $\mathbf{s_0}$, and we write $\sum_{j,k=1}^{\infty} x_{jk} = s_0$, if the double sequence of partial sums converges to s_0;

(ii) has s_0 as a ***limit*** if it converges to s_0;

(iii) is ***convergent*** if it converges to some member of $\mathbb{R}$;

(iv) ***converges absolutely***, or is ***absolutely convergent***, if the series

$$\sum_{j,k=1}^{\infty} |x_{jk}|$$

converges;

(v) ***converges conditionally***, or is ***conditionally convergent***, if it is convergent, but not absolutely convergent;

(vi) ***diverges*** if it does not converge;

(vii) ***diverges to*** $\boldsymbol{\infty}$ (resp. ***diverges to*** $\boldsymbol{-\infty}$), and we write $\sum_{j,k=1}^{\infty} x_{jk} = \infty$ (resp. $\sum_{j,k=1}^{\infty} x_{jk} = -\infty$), if the double sequence of partial sums diverges to ∞ (resp. diverges to $-\infty$);

(viii) has a limit that ***exists*** if $\sum_{j,k=1}^{\infty} x_{jk} \in \mathbb{R}$;

(ix) is ***oscillatory*** if the sequence of partial sums is oscillatory. •

Note that the definition of the partial sums, S_{jk}, $j,k \in \mathbb{Z}_{>0}$, for a double series is unambiguous since

$$\sum_{l=1}^{j}\sum_{m=1}^{k} x_{lm} = \sum_{m=1}^{k}\sum_{l=1}^{j} x_{lm},$$

this being valid for finite sums. The idea behind convergence of double series, then, has an interpretation that can be gleaned from that in Figure 2.2 for double sequences.

Let us state a result, derived from similar results for double sequences, that allows the computation of limits of double series by computing one limit at a time.

2.4.27 Proposition (Computation of limits of double series I) *Suppose that for the double series* $\sum_{j,k=1}^{\infty} x_{jk}$ *it holds that*

(i) the double series is convergent and

(ii) for each $j \in \mathbb{Z}_{>0}$, *the series* $\sum_{k=1}^{\infty} x_{jk}$ *converges.*

Then the series $\sum_{j=1}^{\infty}(\sum_{k=1}^{\infty} x_{jk})$ *converges and its limit is equal to* $\sum_{j,k=1}^{\infty} x_{jk}$.

Proof This follows directly from Proposition 2.3.20. ■

2.4.28 Proposition (Computation of limits of double series II) *Suppose that for the double series $\sum_{\mathrm{j,k}=1}^{\infty} \mathrm{x_{jk}}$ it holds that*

(i) the double series is convergent,

(ii) for each $\mathrm{j} \in \mathbb{Z}_{>0}$, the series $\sum_{\mathrm{k}=1}^{\infty} \mathrm{x_{jk}}$ converges, and

(iii) for each $\mathrm{k} \in \mathbb{Z}_{>0}$, the limit $\sum_{\mathrm{j}=1}^{\infty} \mathrm{x_{jk}}$ converges.

Then the series $\sum_{\mathrm{j}=1}^{\infty}(\sum_{\mathrm{k}=1}^{\infty} \mathrm{x_{jk}})$ and $\sum_{\mathrm{k}=1}^{\infty}(\sum_{\mathrm{j}=1}^{\infty} \mathrm{x_{jk}})$ converge and their limits are both equal to $\sum_{\mathrm{j,k}=1}^{\infty} \mathrm{x_{jk}}$.

Proof This follows directly from Proposition 2.3.21. ■

2.4.6 Algebraic operations on series

In this section we consider the manner in which series interact with algebraic operations. The results here mirror, to some extent, the results for sequences in Section 2.3.6. However, the series structure allows for different ways of thinking about the product of sequences. Let us first give these definitions. For notational convenience, we use sums that begin at 0 rather than 1. This clearly has no affect on the definition of a series, or on any of its properties.

2.4.29 Definition (Products of series) Let $S = \sum_{j=0}^{\infty} x_j$ and $T = \sum_{j=0}^{\infty} y_j$ be series in $\mathbb{R}$.

(i) The ***product*** of S and T is the double series $\sum_{j,k=0}^{\infty} x_j y_k$.

(ii) The ***Cauchy product*** of S and T is the series $\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} x_j y_{k-j}\right)$. •

Now we can state the basic results on algebraic manipulation of series.

2.4.30 Proposition (Algebraic operations on series) *Let $\mathrm{S} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{x_j}$ and $\mathrm{T} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{y_j}$ be series in $\mathbb{R}$ that converges to $\mathrm{s_0}$ and $\mathrm{t_0}$, respectively, and let $\alpha \in \mathbb{R}$. Then the following statements hold:*

(i) the series $\sum_{\mathrm{j}=0}^{\infty} \alpha \mathrm{x_j}$ converges to $\alpha \mathrm{s_0}$;

(ii) the series $\sum_{\mathrm{j}=0}^{\infty}(\mathrm{x_j} + \mathrm{y_j})$ converges to $\mathrm{s_0} + \mathrm{t_0}$;

(iii) if S and T are absolutely convergent, then the product of S and T is absolutely convergent and converges to $\mathrm{s_0 t_0}$;

(iv) if S and T are absolutely convergent, then the Cauchy product of S and T is absolutely convergent and converges to $\mathrm{s_0 t_0}$;

(v) if S or T are absolutely convergent, then the Cauchy product of S and T is convergent and converges to $\mathrm{s_0 t_0}$;

(vi) if S and T are convergent, and if the Cauchy product of S and T is convergent, then the Cauchy product of S and T converges to $\mathrm{s_0 t_0}$.

Proof (i) Since $\sum_{j=0}^{k} \alpha x_j = \alpha \sum_{j=0}^{k} x_j$, this follows from part (i) of Proposition 2.3.23.

(ii) Since $\sum_{j=0}^{\infty}(x_j + y_j) = \sum_{j=0}^{k} x_j + \sum_{j=0}^{k} y_j$, this follows from part (ii) of Proposition 2.3.23.

(iii) and (iv) To prove these parts of the result, we first make a general argument. We note that $\mathbb{Z}_{\geq 0} \times \mathbb{Z}_{\geq 0}$ is a countable set (e.g., by Proposition 1.7.16), and so there exists a bijection, in fact many bijections, $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0}\times\mathbb{Z}_{\geq 0}$. For such a bijection ϕ, suppose that we are given a double sequence $(x_{jk})_{j,k\in\mathbb{Z}_{\geq 0}}$ and define a sequence $(x_j^\phi)_{j\in\mathbb{Z}_{>0}}$ by $x_j^\phi = x_{kl}$ where $(k,l) = \phi(j)$. We then claim that, for any bijection $\phi\colon \mathbb{Z}_{>0} \to \mathbb{Z}_{\geq 0}\times\mathbb{Z}_{\geq 0}$, the double series $A = \sum_{k,l=1}^\infty x_{kl}$ converges absolutely if and only if the series $A^\phi = \sum_{j=1}^\infty x_j^\phi$ converges absolutely.

Indeed, suppose that the double series $|A| = \sum_{k,l=1}^\infty |x_{kl}|$ converges to $\beta\in\mathbb{R}$. For $\epsilon\in\mathbb{R}_{>0}$ the set

$$\{(k,l)\in\mathbb{Z}_{\geq 0}\times\mathbb{Z}_{\geq 0} \mid \||A|_{kl}-\beta| \geq \epsilon\}$$

is then finite. Therefore, there exists $N\in\mathbb{Z}_{>0}$ such that, if $(k,l)=\phi(j)$ for $j\geq N$, then $\||A|_{kl}-\beta| < \epsilon$. It therefore follows that $\||A^\phi|_j - \beta| < \epsilon$ for $j\geq N$, where $|A^\phi|$ denotes the series $\sum_{j=1}^\infty |x_j^\phi|$. This shows that the series $|A^\phi|$ converges to β.

For the converse, suppose that the series $|A^\phi|$ converges to β. Then, for $\epsilon\in\mathbb{R}_{>0}$ the set

$$\{j\in\mathbb{Z}_{>0} \mid \||A^\phi|_j-\beta|\geq\epsilon\}$$

is finite. Therefore, there exists $N\in\mathbb{Z}_{>0}$ such that

$$\{(k,l)\in\mathbb{Z}_{\geq 0} \mid k,l\geq N\}\cap\{(k,l)\in\mathbb{Z}_{\geq 0} \mid \||A^\phi|_{\phi^{-1}(k,l)}-\beta|\geq\epsilon\}=\varnothing.$$

It then follows that for $k,l\geq N$ we have $\||A|_{kl}-\beta|<\epsilon$, showing that $|A|$ converges to β.

Thus we have shown that A is absolutely convergent if and only if A^ϕ is absolutely convergent for any bijection $\phi\colon \mathbb{Z}_{>0}\to\mathbb{Z}_{\geq 0}\times\mathbb{Z}_{\geq 0}$. From part (v) of Theorem 2.4.5, and its generalisation to double series, we know that the limit of an absolutely convergent series or double series is independent of the manner in which the terms in the series are arranged.

Consider now a term in the product of S and T. It is easy to see that this term appears exactly once in the Cauchy product of S and T. Conversely, each term in the Cauchy product appears exactly one in the product. Thus the product and Cauchy product are simply rearrangements of one another. Moreover, each term in the product and the Cauchy product appears exactly once in the expression

$$\left(\sum_{j=0}^N x_j\right)\left(\sum_{k=0}^N y_k\right)$$

as we allow N to go to ∞. That is to say,

$$\sum_{j,k=0}^\infty x_j y_k = \sum_{k=0}^\infty\left(\sum_{j=k}^k x_j y_{k-j}\right) = \lim_{N\to\infty}\left(\sum_{j=0}^N x_j\right)\left(\sum_{k=0}^N y_k\right).$$

However, this last limit is exactly s_0t_0, using part (iii) of Proposition 2.3.23.

(v) Without loss of generality, suppose that S converges absolutely. Let $(S_k)_{k\in\mathbb{Z}_{>0}}$, $(T_k)_{k\in\mathbb{Z}_{>0}}$, and $((ST)_k)_{k\in\mathbb{Z}_{>0}}$ be the sequences of partial sums for S, T, and the Cauchy product, respectively. Also define $\tau_k = T_k - t_0$, $k \in \mathbb{Z}_{\geq 0}$. Then

$$\begin{aligned}(ST)_k &= x_0 y_0 + (x_0 y_1 + x_1 y_0) + \cdots + (x_0 y_k + \cdots + x_k y_0)\\ &= x_0 T_k + x_1 T_{k-1} + \cdots + x_k T_0\\ &= x_0(t_0 + \tau_k) + x_1(t_0 + \tau_{k-1}) + \cdots + x_k(t_0 + \tau_0)\\ &= S_k t_0 + x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0.\end{aligned}$$

Since $\lim_{k\to\infty} S_k t_0 = s_0 t_0$ by part (i), this part of the result will follow if we can show that

$$\lim_{k\to\infty}(x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0) = 0. \tag{2.6}$$

Denote

$$\sigma = \sum_{j=0}^{\infty} |x_j|,$$

and for $\epsilon \in \mathbb{R}_{>0}$ choose $N_1 \in \mathbb{Z}_{>0}$ such that $|\tau_j| \leq \frac{\epsilon}{2\sigma}$ for $j \geq N_1$, this being possible since $(\tau_j)_{j\in\mathbb{Z}_{>0}}$ clearly converges to zero. Then, for $k \geq N_1$,

$$\begin{aligned}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| &\leq |x_0\tau_k + \cdots + x_{k-N_1-1}\tau_{N_1-1}| + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0|\\ &\leq \tfrac{\epsilon}{2} + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0|.\end{aligned}$$

Since $\lim_{k\to\infty} x_k = 0$, choose $N_2 \in \mathbb{Z}_{>0}$ such that

$$|x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0| < \tfrac{\epsilon}{2}$$

for $k \geq N_2$. Then

$$\begin{aligned}\limsup_{k\to\infty}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| &= \lim_{k\to\infty}\sup\{|x_0\tau_j + x_1\tau_{j-1} + \cdots + x_j\tau_0| \mid j \geq k\}\\ &\leq \lim_{k\to\infty}\sup\{\tfrac{\epsilon}{2} + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0| \mid j \geq k\}\\ &\leq \sup\{\tfrac{\epsilon}{2} + |x_{k-N_1}\tau_{N_1} + \cdots + x_k\tau_0| \mid j \geq N_2\} \leq \epsilon.\end{aligned}$$

Thus

$$\limsup_{k\to\infty}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| \leq 0,$$

and since clearly

$$\liminf_{k\to\infty}|x_0\tau_k + x_1\tau_{k-1} + \cdots + x_k\tau_0| \geq 0,$$

we infer that (2.6) holds by Proposition 2.3.17.

(vi) The reader can prove this as Exercise 3.7.3. ■

The reader is recommended to remember the Cauchy product when we talk about convolution of discrete-time signals in Section IV-4.1.4.

2.4.7 Series with arbitrary index sets

It will be helpful on a few occasions to be able to sum series whose index set is not necessarily countable, and here we indicate how this can be done. This material should be considered optional until one comes to that point in the text where it is needed.

2.4.31 Definition (Sum of series for arbitrary index sets) Let A be a set and let $(x_a)_{a\in A}$ be a family of elements of $\overline{\mathbb{R}}$. Let $A_+ = \{a \in A \mid x_a \in [0,\infty]\}$ and $A_- = \{a \in A \mid x_a \in [-\infty, 0]\}$.

(i) If $x_a \in [0,\infty]$ for $a \in A$, then $\sum_{a\in A} x_a = \sup\{\sum_{a\in A'} x_a \mid A' \subseteq A \text{ is finite}\}$.

(ii) For a general family, $\sum_{a\in A} x_a = \sum_{a_+\in A_+} x_{a_+} - \sum_{a_-\in A_-}(-x_{a_-})$, provided that at least one of $\sum_{a_+\in A_+} x_{a_+}$ or $\sum_{a_-\in A_-}(-x_{a_-})$ is finite.

(iii) If both $\sum_{a_+\in A_+} x_{a_+}$ are $\sum_{a_-\in A_-}(-x_{a_-})$ are finite, then $(x_a)_{a\in A}$ is ***summable***. •

We should understand the relationship between this sort of summation and our existing notion of the sum of a series in the case where the index set is $\mathbb{Z}_{>0}$.

2.4.32 Proposition (A summable series with index set $\mathbb{Z}_{>0}$ is absolutely convergent) *A sequence $(\mathrm{x_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ is summable if and only if the series $\mathrm{S} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{x_j}$ is absolutely convergent.*

Proof Consider the sequences $(x_j^+)_{j\in\mathbb{Z}_{>0}}$ and $(x_j^-)_{j\in\mathbb{Z}_{>0}}$ defined by

$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{-x_j, 0\}, \qquad j \in \mathbb{Z}_{>0}.$$

Then $(x_j)_{j\in\mathbb{Z}_{>0}}$ is summable if and only if both of the expressions

$$\sup\left\{\sum_{j\in A'} x_j^+ \;\middle|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\right\}, \quad \sup\left\{\sum_{j\in A'} x_j^- \;\middle|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\right\} \tag{2.7}$$

are finite.

First suppose that $(x_j)_{j\in\mathbb{Z}_{>0}}$ is summable. Therefore, if $(S_k^+)_{k\in\mathbb{Z}_{>0}}$ and $(S_k^-)_{k\in\mathbb{Z}_{>0}}$ are the sequences of partial sums

$$S_k^+ = \sum_{j=1}^{k} x_j^+, \quad S_k^- = \sum_{j=1}^{k} x_j^-,$$

then these sequences are increasing and so convergent by (2.7). Then, by Proposition 2.3.23,

$$\sum_{j=1}^{\infty} |x_j| = \sum_{j=1}^{\infty} x_j^+ + \sum_{j=1}^{\infty} x_j^-$$

giving absolute convergence of S.

Now suppose that S is absolutely convergent. Then the subsets $\{S_k^+ \mid k \in \mathbb{Z}_{>0}\}$ and $\{S_k^- \mid k \in \mathbb{Z}_{>0}\}$ are bounded above (as well as being bounded below by zero) so that both expressions

$$\sup\{S_k^+ \mid k \in \mathbb{Z}_{>0}\}, \quad \sup\{S_k^- \mid k \in \mathbb{Z}_{>0}\}$$

are finite. Then for any finite set $A' \subseteq \mathbb{Z}_{>0}$ we have

$$\sum_{j\in A'} x_j^+ \leq S^+_{\sup A'}, \quad \sum_{j\in A'} x_j^- \leq S^-_{\sup A'}.$$

From this we deduce that

$$\sup\left\{\sum_{j\in A'} x_j^+ \;\middle|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\right\} \leq \sup\{S_k^+ \mid k \in \mathbb{Z}_{>0}\},$$
$$\sup\left\{\sum_{j\in A'} x_j^- \;\middle|\; A' \subseteq \mathbb{Z}_{>0} \text{ is finite}\right\} \leq \sup\{S_k^- \mid k \in \mathbb{Z}_{>0}\},$$

which implies that $(x_j)_{j\in\mathbb{Z}_{>0}}$ is summable. ■

Now we can actually show that, for a summable family of real numbers, only countably many of them can be nonzero.

2.4.33 Proposition (A summable family has countably many nonzero members) *If* $(\mathrm{x}_\mathrm{a})_{\mathrm{a}\in\mathrm{A}}$ *is summable, then the set* $\{\mathrm{a} \in \mathrm{A} \mid \mathrm{x}_\mathrm{a} \neq 0\}$ *is countable.*

Proof Note that for any $k \in \mathbb{Z}_{>0}$, the set $\{a \in A \mid |x_a| \geq \frac{1}{k}\}$ must be finite if $(x_a)_{a\in A}$ is summable (why?). Thus, since

$$\{a \in A \mid |x_a| \neq 0\} = \cup_{k\in\mathbb{Z}_{>0}}\{a \in A \mid |x_a| \geq \tfrac{1}{k}\},$$

the set $\{a \in A \mid x_a \neq 0\}$ is a countable union of finite sets, and so is countable by Proposition 1.7.16. ■

A legitimate question is, since a summable family reduces to essentially being countable, why should we bother with the idea at all? The reason is simply that it will be notationally convenient in Section 3.3.4.

2.4.8 Notes

The numbers e and π are not only irrational, but have the much stronger property of being ***transcendental***. This means that they are not the roots of any polynomial having rational coefficients (see Definition 4.6.12). That e is transcendental was proved by Hermite[7] in 1873, and the that π is transcendental was proved by Lindemann[8] in 1882.

The proof we give for the irrationality of π is essentially that of Niven [1947]; this is the most commonly encountered proof, and is simpler than the original proof of Lambert[9] presented to the Berlin Academy in 1768.

[7] Charles Hermite (1822–1901) was a French mathematician who made contributions to the fields of number theory, algebra, differential equations, and analysis.

[8] Carl Louis Ferdinand von Lindemann (1852–1939) was born in what is now Germany. His mathematical contributions were in the areas of analysis and geometry. He also was interested in physics.

[9] Johann Heinrich Lambert (1728–1777) was born in France. His mathematical work included contributions to analysis, geometry, and probability. He also made contributions to astronomical theory.

Exercises

2.4.1 Let $S = \sum_{j=1}^{\infty} x_j$ be a series in $\mathbb{R}$, and, for $j \in \mathbb{Z}_{>0}$, define

$$x_j^+ = \max\{x_j, 0\}, \quad x_j^- = \max\{0, -x_j\}.$$

Show that, if S is conditionally convergent, then the series $S^+ = \sum_{j=1}^{\infty} x_j^+$ and $S^- = \sum_{j=1}^{\infty} x_j^-$ diverge to ∞.

2.4.2 In this exercise we consider more carefully the paradox of Zeno given in Exercise 1.9.2. Let us attach some symbols to the relevant data, so that we can say useful things. Suppose that the tortoise travels with constant velocity v_t and that Achilles travels with constant velocity v_a. Suppose that the tortoise gets a head start of t_0 seconds.

(a) Compute directly using elementary physics (i.e., time/distance/velocity relations) the time at which Achilles will overtake the tortoise, and the distance both will have travelled during that time.

(b) Consider the sequences $(d_j)_{j\in\mathbb{Z}_{>0}}$ and $(t_j)_{j\in\mathbb{Z}_{>0}}$ defined so that

1. d_1 is the distance travelled by the tortoise during the head start time t_0,
2. t_j, $j \in \mathbb{Z}_{>0}$, is the time it takes Achilles to cover the distance d_j,
3. d_j, $j \geq 2$, is the distance travelled by the tortoise in time t_{j-1}.

Find explicit expressions for these sequences in terms of t_0, v_t, and v_a.

(c) Show that the series $\sum_{j=1}^{\infty} d_j$ and $\sum_{j=1}^{\infty} t_j$ converge, and compute their limits.

(d) What is the relationship between the limits of the series in part (c) and the answers to part (a).

(e) Does this shed some light on how to resolve Zeno's paradox?

2.4.3 Show that

$$\left|\sum_{j=1}^{m} x_j\right| \leq \sum_{j=1}^{m} |x_j|$$

for any finite family $(x_1, \dots, x_m) \subseteq \mathbb{R}$.

2.4.4 State the correct version of Proposition 2.4.4 in the case that $S = \sum_{j=1}^{\infty} x_j$ is not absolutely convergent, and indicate why it is not a very interesting result.

2.4.5 For a sum

$$S = \sum_{j=1}^{\infty} s_j,$$

answer the following questions.

(a) Show that if S converges then the sequence $(s_j)_{j\in\mathbb{Z}_{>0}}$ converges to 0.

(b) Is the converse of part (a) true? That is to say, if the sequence $(s_j)_{j\in\mathbb{Z}_{>0}}$ converges to zero, does S converge? If this is true, prove it. If it is not true, give a counterexample.

2.4.6 Do the following.

(a) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which converges in $\mathbb{R}$.

(b) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to ∞.

(c) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which diverges to $-\infty$.

(d) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}\left|\frac{x_{j+1}}{x_j}\right| = 1$ and which is oscillatory.

2.4.7 Do the following.

(a) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}|x_j|^{1/j} = 1$ and which converges in $\mathbb{R}$.

(b) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}|x_j|^{1/j} = 1$ and which diverges to ∞.

(c) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}|x_j|^{1/j} = 1$ and which diverges to $-\infty$.

(d) Find a series $\sum_{j=1}^{\infty} x_j$ for which $\lim_{j\to\infty}|x_j|^{1/j} = 1$ and which is oscillatory.

The next exercise introduces the notion of the decimal expansion of a real number. An ***infinite decimal expansion*** is a series in $\mathbb{Q}$ of the form

$$\sum_{j=0}^{\infty} \frac{a_j}{10^j}$$

where $a_0 \in \mathbb{Z}$ and where $a_j \in \{0, 1, \dots, 9\}$, $j \in \mathbb{Z}_{>0}$. An infinite decimal expansion is ***eventually periodic*** if there exists $k, m \in \mathbb{Z}_{>0}$ such that $a_{j+k} = a_j$ for all $j \geq m$.

2.4.8 (a) Show that the sequence of partial sums for an infinite decimal expansion is a Cauchy sequence.

(b) Show that, for every Cauchy sequence $(q_j)_{j\in\mathbb{Z}_{>0}}$, there exists a sequence $(d_j)_{j\in\mathbb{Z}_{>0}}$ of partial sums for a decimal expansion having the property that $[(q_j)_{j\in\mathbb{Z}_{>0}}] = [(d_j)_{j\in\mathbb{Z}_{>0}}]$ (the equivalence relation is that in the Cauchy sequences in $\mathbb{Q}$ as defined in Definition 2.1.16).

(c) Give an example that shows that two distinct infinite decimal expansions can be equivalent.

(d) Show that if two distinct infinite decimal expansions are equivalent, and if one of them is eventually periodic, then the other is also eventually periodic.

The previous exercises show that every real number is the limit of a (not necessarily unique) infinite decimal expansion. The next exercises characterise the infinite decimal expansions that correspond to rational numbers.

First you will show that an eventually periodic decimal expansion corresponds to a rational number. Let $\sum_{j=0}^{\infty} \frac{a_j}{10^j}$ be an eventually periodic infinite decimal expansion and let $k, m \in \mathbb{Z}_{>0}$ have the property that $a_{j+k} = a_j$ for $j \geq m$. Denote by $x \in \mathbb{R}$ the number to which the infinite decimal expansion converges.

(e) Show that
$$10^{m+k}x = \sum_{j=0}^{\infty} \frac{b_j}{10^j}, \quad 10^m x = \sum_{j=0}^{\infty} \frac{c_j}{10^j}$$
are decimal expansions, and give expressions for b_j and c_j, $j \in \mathbb{Z}_{>0}$, in terms of a_j, $j \in \mathbb{Z}_{>0}$. In particular, show that $b_j = c_j$ for $j \geq 1$.

(f) Conclude that $(10^{m+k} - 10^m)x$ is an integer, and so x is therefore rational.

Next you will show that the infinite decimal expansion of a rational number is eventually periodic. Thus let $q \in \mathbb{Q}$.

(g) Let $q = \frac{a}{b}$ for $a, b \in \mathbb{Z}$ and with $b > 0$. For $j \in \{0, 1, \dots, b\}$, let $r_j \in \{0, 1, \dots, b-1\}$ satisfy $\frac{10^j}{b} = s_j + \frac{r_j}{b}$ for $s_j \in \mathbb{Z}$, i.e., r_j is the remainder after dividing 10^j by b. Show that at least two of the numbers $\{r_0, r_1, \dots, r_b\}$ must agree, i.e., conclude that $r_m = r_{m+k}$ for $k, m \in \mathbb{Z}_{\geq 0}$ satisfying $0 \leq m < m + k \leq b$.

Hint: *There are only* b *possible values for these* b + 1 *numbers.*

(h) Show that b exactly divides $10^{m+k} - 10^k$ with k and m as above. Thus $bc = 10^{m+k} - 10^k$ for some $c \in \mathbb{Z}$.

(i) Show that
$$\frac{a}{b} = 10^{-m} \frac{ac}{10^k - 1},$$
and so write
$$q = 10^{-m}\left(s + \frac{r}{10^k - 1}\right)$$
for $s \in \mathbb{Z}$ and $r \in \{0, 1, \dots, 10^k - 1\}$, i.e., r is the remainder after dividing ac by $10^k - 1$.

(j) Argue that we can write
$$b = \sum_{j=1}^{k} b_j 10^j,$$
for $b_j \in \{0, 1, \dots, 9\}$, $j \in \{1, \dots, k\}$.

(k) With b_j, $j \in \{1, \dots, k\}$ as above, define an infinite decimal expansion $\sum_{j=0}^{\infty} \frac{a_j}{10^j}$ by asking that $a_0 = 0$, that $a_j = b_j$, $j \in \{1, \dots, k\}$, and that $a_{j+km} = a_j$ for $j, m \in \mathbb{Z}_{>0}$. Let $d \in \mathbb{R}$ be the number to which this decimal expansion converges. Show that $(10^k - 1)d = b$, so $d \in \mathbb{Q}$.

(l) Show that $10^m q = s + d$, and so conclude that $10^m q$ has the eventually periodic infinite decimal expansion $s + \sum_{j=1}^{\infty} \frac{a_j}{10^j}$.

(m) Conclude that q has an eventually periodic infinite decimal expansion, and then conclude from (d) that any infinite decimal expansion for q is eventually periodic.

Section 2.5

Subsets of $\mathbb{R}$

In this section we study in some detail the nature of various sorts of subsets of $\mathbb{R}$. The character of these subsets will be of some importance when we consider the properties of functions defined on $\mathbb{R}$, and/or taking values in $\mathbb{R}$. Our presentation also gives us an opportunity to introduce, in a fairly simple setting, some concepts that will appear later in more abstract settings, e.g., open sets, closed sets, compactness.

Do I need to read this section? Unless you know the material here, it is indeed a good idea to read this section. Many of the ideas are basic, but some are not (e.g., the Heine–Borel Theorem). Moreover, many of the not-so-basic ideas will appear again later, particularly in Chapter III-1, and if a reader does not understand the ideas in the simple case of $\mathbb{R}$, things will only get more difficult. Also, the ideas expressed here will be essential in understanding even basic things about signals as presented in Chapter IV-1. •

2.5.1 Open sets, closed sets, and intervals

One of the basic building blocks in the understanding of the real numbers is the idea of an open set. In this section we define open sets and some related notions, and provide some simple properties associated to these ideas.

First, it is convenient to introduce the following ideas.

2.5.1 Definition (Open ball, closed ball) For $r \in \mathbb{R}_{>0}$ and $x_0 \in \mathbb{R}$,

(i) the ***open ball*** in $\mathbb{R}$ of radius r about x_0 is the set

$$\mathsf{B}(r, x_0) = \{x \in \mathbb{R} \mid |x - x_0| < r\},$$

and

(ii) the ***closed ball*** of radius r about x_0 is the set

$$\overline{\mathsf{B}}(r, x_0) = \{x \in \mathbb{R} \mid |x - x_0| \leq r\}. \qquad \bullet$$

These sets are simple to understand, and we depict them in Figure 2.3. With

Figure 2.3 An open ball (left) and a closed ball (right) in $\mathbb{R}$

the notion of an open ball, it is easy to give some preliminary definitions.

2.5.2 Definition (Open and closed sets in $\mathbb{R}$) A set $A \subseteq \mathbb{R}$ is:

(i) ***open*** if, for every $x \in A$, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\epsilon, x) \subseteq A$ (the empty set is also open, by declaration);

(ii) ***closed*** if $\mathbb{R} \setminus A$ is open. •

A trivial piece of language associated with an open set is the notion of a neighbourhood.

2.5.3 Definition (Neighbourhood in $\mathbb{R}$) A ***neighbourhood*** of an element $x \in \mathbb{R}$ is an open set U for which $x \in U$. •

Some authors allow a "neighbourhood" to be a set A which contains a neighbourhood in our sense. Such authors will then frequently call what we call a neighbourhood an "open neighbourhood."

Let us give some examples of sets that are open, closed, or neither. The examples we consider here are important ones, since they are all examples of ***intervals***, which will be of interest at various times, and for various reasons, throughout these volumes. In particular, the notation we introduce here for intervals will be used a great deal.

2.5.4 Examples (Intervals)

1. For $a, b \in \mathbb{R}$ with $a < b$ the set

$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$$

is open. Indeed, let $x \in (a, b)$ and let $\epsilon = \frac{1}{2}\min\{b - x, x - a\}$. It is then easy to see that $\mathsf{B}(\epsilon, x) \subseteq (a, b)$. If $a \geq b$ we take the convention that $(a, b) = \emptyset$.

2. For $a \in \mathbb{R}$ the set

$$(a, \infty) = \{x \in \mathbb{R} \mid a < x\}$$

is open. For example, if $x \in (a, \infty)$ then, if we define $\epsilon = \frac{1}{2}(x - a)$, we have $\mathsf{B}(\epsilon, x) \subseteq (a, \infty)$.

3. For $b \in \mathbb{R}$ the set

$$(-\infty, b) = \{x \in \mathbb{R} \mid x < b\}$$

is open.

4. For $a, b \in \mathbb{R}$ with $a \leq b$ the set

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

is closed. Indeed, $\mathbb{R} \setminus [a, b] = (-\infty, a) \cup (b, \infty)$. The sets $(-\infty, a)$ and (b, ∞) are both open, as we have already seen. Moreover, it is easy to see, directly from the definition, that the union of open sets is also an open set. Therefore, $\mathbb{R} \setminus [a, b]$ is open, and so $[a, b]$ is closed.

5. For $a \in \mathbb{R}$ the set
$$[a, \infty) = \{x \in \mathbb{R} \mid a \leq x\}$$
is closed since it complement in $\mathbb{R}$ is $(-\infty, a)$ which is open.
6. For $b \in \mathbb{R}$ the set
$$(-\infty, b] = \{x \in \mathbb{R} \mid x \leq b\}$$
is closed.
7. For $a, b \in \mathbb{R}$ with $a < b$ the set
$$(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}$$
is neither open nor closed. To see that it is not open, note that $b \in (a, b]$, but that any open ball about b will contain points not in $(a, b]$. To see that $(a, b]$ is not closed, note that $a \in \mathbb{R} \setminus (a, b]$, and that any open ball about a will contain points not in $\mathbb{R} \setminus (a, b]$.
8. For $a, b \in \mathbb{R}$ with $a < b$ the set
$$[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\}$$
is neither open nor closed.
9. The set $\mathbb{R}$ is both open and closed. That it is open is clear. That it is closed follows since $\mathbb{R} \setminus \mathbb{R} = \varnothing$, and $\varnothing$ is, by convention, open. We will sometimes, although not often, write $\mathbb{R} = (-\infty, \infty)$. •

We shall frequently denote typical interval by I, and the set of intervals we denote by $\mathscr{I}$. If I and J are intervals with $J \subseteq I$, we will say that J is a ***subinterval*** of I. The expressions "open interval" and "closed interval" have their natural meanings as intervals that are, as subsets of $\mathbb{R}$, open and closed, respectively. An interval that is neither open nor closed will be called ***half-open*** or ***half-closed***. A ***left endpoint*** (resp. ***right endpoint***) for an interval I is a number $x \in \mathbb{R}$ such that $\inf I = x$ (resp. $\sup I = x$). An endpoint x, be it left or right, is ***open*** if $x \notin I$ and is ***closed*** if $x \in I$. If $\inf I = -\infty$ (resp. $\sup I = \infty$), then we saw that I is ***unbounded on the left*** (resp. ***unbounded on the right***). We will also use the interval notation to denote subsets of the extended real numbers $\overline{\mathbb{R}}$. Thus, we may write

1. $(a, \infty] = (a, \infty) \cup \{\infty\}$,
2. $[a, \infty] = [a, \infty) \cup \{\infty\}$,
3. $[-\infty, b) = (-\infty, b) \cup \{-\infty\}$,
4. $[-\infty, b] = (-\infty, b] \cup \{-\infty\}$, and
5. $[-\infty, \infty] = (-\infty, \infty) \cup \{-\infty, \infty\} = \overline{\mathbb{R}}$.

The following characterisation of intervals is useful.

2.5.5 Proposition (Characterisation of intervals) *A subset* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval if and only if, for each* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$ *with* $\mathrm{a} < \mathrm{b}$*,* $[\mathrm{a}, \mathrm{b}] \subseteq \mathrm{I}$*.*

Proof It is clear from the definition that, if I is an interval, then, for each $a, b \in I$ with $a < b$, $[a, b] \subseteq I$. So suppose that, for each $a, b \in I$ with $a < b$, $[a, b] \subseteq I$. Let $A = \inf I$ and let $B = \sup I$. We have the following cases to consider.

1. $A = B$: Trivially I is an interval.
2. $A, B \in \mathbb{R}$ and $A \neq B$: Choose $a_1, b_1 \in I$ such that $a_1 < b_1$. Define $a_{j+1}, b_{j+1} \in I$, $j \in \mathbb{Z}_{>0}$, inductively as follows. Let a_{j+1} be a point in I to the left of $\frac{1}{2}(A + a_j)$ and let b_{j+1} be a point in I to the right of $\frac{1}{2}(b_j + B)$. These constructions make sense by definition of A and B. Note that $(a_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically decreasing sequence converging to A and that $(b_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically increasing sequence converging to B. Also,
$$\bigcup_{j\in\mathbb{Z}_{>0}} [a_j, b_j] \subseteq I.$$
We also have either $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (A, B)$, $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = [A, B)$, $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (A, B]$, or $\cup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = [A, B]$. Therefore we conclude that I is an interval with endpoints A and B.
3. $A = -\infty$ and $B \in \mathbb{R}$. Choose $a_1, b_1 \in I$ with $a_a < b_1 < B$. Define $a_{j+1}, b_{j+1} \in I$, $j \in \mathbb{Z}_{>0}$, inductively by asking that a_{j+1} be a point in I to the left of $a_j - 1$ and that b_{j+1} be a point in I to the right of $\frac{1}{2}(b_j + B)$. These constructions make sense by definition of A and B. Thus $(a_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically decreasing sequence in I diverging to $-\infty$ and $(b_j)_{j\in\mathbb{Z}_{>0}}$ is a monotonically increasing sequence in I converging to B. Thus
$$\bigcup_{j\in\mathbb{Z}_{>0}} [a_j, b_j] =\subseteq I.$$
Note that either $\bigcup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (-\infty, B)$ or $\bigcup_{j\in\mathbb{Z}_{>0}}[a_j, b_j] = (-\infty, B]$. This means that either $I = (-\infty, B)$ or $I = (-\infty, B]$.
4. $A \in \mathbb{R}$ and $B = \infty$: A construction entirely like the preceding one shows that either $I = (A, \infty)$ or $I = [A, \infty)$.
5. $A = -\infty$ and $B = \infty$: Choose $a_1, b_1 \in I$ with $a_1 < b_1$. Inductively define $a_{j+1}, b_{j+1} \in I$, $j \in \mathbb{Z}_{>0}$, by asking that a_{j+1} be a point in I to the left of a_j and that b_{j+1} be a point in I to the right of b_j. We then conclude that
$$\bigcup_{j\in\mathbb{Z}_{>0}} [a_j, b_j] = \mathbb{R} =\subseteq I,$$
and so $I = \mathbb{R}$.

In all cases we have concluded that I is an interval. ■

The following property of open sets will be useful for us, and tells us a little about the character of open sets.

2.5.6 Proposition (Open sets in $\mathbb{R}$ are unions of open intervals) *If* $\mathrm{U} \subseteq \mathbb{R}$ *is a nonempty open set then* U *is a countable union of disjoint open intervals.*

Proof Let $x \in U$ and let I_x be the largest open interval containing x and contained in U. This definition of I_x makes sense since the union of open intervals containing x is also an open interval containing x. Now to each interval can be associated a rational number within the interval. Therefore, the number of intervals to cover U can be associated with a subset of $\mathbb{Q}$, and is therefore countable. This shows that U is indeed a countable union of open intervals. ■

2.5.2 Partitions of intervals

In this section we consider the idea of partitioning an interval of the form $[a, b]$. This is a construction that will be useful in a variety of places, but since we dealt with intervals in the previous section, this is an appropriate time to make the definition and the associated constructions.

2.5.7 Definition (Partition of an interval) A ***partition*** of an interval $[a, b]$ is a family $(I_1, \dots, I_k)$ of intervals such that

(i) $\mathrm{int}(I_j) \neq \emptyset$ for $j \in \{1, \dots, k\}$,

(ii) $[a, b] = \cup_{j=1}^{k} I_j$, and

(iii) $I_j \cap I_l = \emptyset$ for $j \neq l$.

We denote by $\mathrm{Part}([a, b])$ the set of partitions of $[a, b]$. •

We shall always suppose that a partition $(I_1, \dots, I_k)$ is totally ordered so that the left endpoint of I_{j+1} agrees with the right endpoint of I_j for each $j \in \{1, \dots, k-1\}$. That is to say, when we write a partition, we shall list the elements of the set according to this total order. Note that associated to a partition $(I_1, \dots, I_k)$ are the endpoints of the intervals. Thus there exists a family $(x_0, x_1, \dots, x_k)$ of $[a, b]$, ordered with respect to the natural total order on $\mathbb{R}$, such that, for each $j \in \{1, \dots, k\}$, x_{j-1} is the left endpoint of I_j and x_j is the right endpoint of I_j. Note that necessarily we have $x_0 = a$ and $x_k = b$. The set of endpoints of the intervals in a partition $P = (I_1, \dots, I_k)$ we denote by $\mathrm{EP}(P)$. In Figure 2.4 we show a partition with all ingredients labelled. For a partition P with $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$, denote

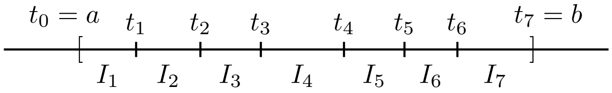


Figure 2.4 A partition

$$|P| = \max\{|x_j - x_l| \mid j, l \in \{1, \dots, k\}\},$$

which is the ***mesh*** of P. Thus $|P|$ is the length of the largest interval of the partition.

It is often useful to be able to say one partition is finer than another, and the following definition makes this precise.

2.5.8 Definition (Refinement of a partition) If P_1 and P_2 are partitions of an interval $[a, b]$, then P_2 is a ***refinement*** of P_1 if $\mathrm{EP}(P_1) \subseteq \mathrm{EP}(P_2)$. •

Next we turn to a sometimes useful construction involving the addition of certain structure onto a partition. This construction is rarely used in the text, so may be skipped until it is encountered.

2.5.9 Definition (Tagged partition, δ-fine tagged partition) Let $[a, b]$ be an interval and let $\delta\colon [a, b] \to \mathbb{R}_{>0}$.

(i) A ***tagged partition*** of $[a, b]$ is a finite family of pairs $((c_1, I_1), \dots, (c_k, I_k))$ where $(I_1, \dots, I_k)$ is a partition and where c_j is contained in the union of I_j with its endpoints.

(ii) A tagged partition $((c_1, I_1), \dots, (c_k, I_k))$ is ***δ-fine*** if the interval I_j, along with its endpoints, is a subset of $\mathsf{B}(\delta(c_j), c_j)$. •

The following result asserts that δ-fine tagged partitions always exist.

2.5.10 Proposition (δ-fine tagged partitions exist) *For any positive function $\delta\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}_{>0}$, there exists a δ-fine tagged partition.*

Proof Let Δ be the set of all points $x \in (a, b]$ such that there exists a δ-fine tagged partition of $[a, x]$. Note that $(a, a + \delta(a)) \subseteq \Delta$ since, for each $x \in (a, a + \delta(a))$, $((a, [a, x]))$ is a δ-fine tagged partition of $[a, x]$. Let $b' = \sup \Delta$. We will show that $b' = b$ and that $b' \in \Delta$.

Since $b' = \sup \Delta$ there exists $b'' \in \Delta$ such that $b' - \delta(b') < b'' < b'$. Then there exists a δ-fine partition P' of $[a, b']$. Now $P' \cup ((b', (b'', b']))$ is δ-fine tagged partition of $[a, b']$. Thus $b' \in \Delta$.

Now suppose that $b' < b$ and choose $b'' < b$ such that $b' < b'' < b' + \delta(b')$. If P is a tagged partition of $[a, b']$ (this exists since $b' \in \Delta$), then $P \cup ((b', (b', b'']))$ is a δ-fine tagged partition of $[a, b'']$. This contradicts the fact that $b' = \sup \Delta$. Thus we conclude that $b' = b$. ■

2.5.3 Interior, closure, boundary, and related notions

Associated with the concepts of open and closed are a collection of useful concepts.

2.5.11 Definition (Accumulation point, cluster point, limit point in $\mathbb{R}$) Let $A \subseteq \mathbb{R}$. A point $x \in \mathbb{R}$ is:

(i) an ***accumulation point*** for A if, for every neighbourhood U of x, the set $A \cap (U \setminus \{x\})$ is nonempty;

(ii) a ***cluster point*** for A if, for every neighbourhood U of x, the set $A \cap U$ is infinite;

(iii) a ***limit point*** of A if there exists a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ in A converging to x.

The set of accumulation points of A is called the ***derived set*** of A, and is denoted by $\mathrm{der}(A)$. •

2.5.12 Remark (Conventions concerning "accumulation point," "cluster point," and "limit point") There seems to be no agreed upon convention about what is meant by the three concepts of accumulation point, cluster point, and limit point. Some authors make no distinction between the three concepts at all. Some authors lump two together, but give the third a different meaning. As we shall see in Proposition 2.5.13 below, sometimes there is no need to distinguish between two of the concepts. However, in order to keep as clear as possible the transition to the more abstract presentation of Chapter III-1, we have gone with the most pedantic interpretation possible for the concepts of "accumulation point," "cluster point," and "limit point." •

The three concepts of accumulation point, cluster point, and limit point are actually excessive for $\mathbb{R}$ since, as the next result shall indicate, two of them are exactly the same. However, in the more general setup of Chapter III-1, the concepts are no longer equivalent.

2.5.13 Proposition ("Accumulation point" equals "cluster point" in $\mathbb{R}$) *For a set* $A \subseteq \mathbb{R}$, $x \in \mathbb{R}$ *is an accumulation point for* A *if and only if it is a cluster point for* A.

Proof It is clear that a cluster point for A is an accumulation point for A. Suppose that x is not a cluster point. Then there exists a neighbourhood U of x for which the set $A \cap U$ is finite. If $A \cap U = \{x\}$, then clearly x is not an accumulation point. If $A \cap U \neq \{x\}$, then $A \cap (U \setminus \{x\}) \supseteq \{x_1, \dots, x_k\}$ where the points $x_1, \dots, x_k$ are distinct from x. Now let

$$\epsilon = \tfrac{1}{2}\min\{|x_1 - x|, \dots, |x_k - x|\}.$$

Clearly $A \cap (\mathsf{B}(\epsilon, x) \setminus \{x\})$ is then empty, and so x is not an accumulation point for A. ■

Now let us give some examples that illustrate the differences between accumulation points (or equivalently cluster points) and limit points.

2.5.14 Examples (Accumulation points and limit points)

1. For any subset $A \subseteq \mathbb{R}$ and for every $x \in A$, x is a limit point for A. Indeed, the constant sequence $(x_j = x)_{j \in \mathbb{Z}_{>0}}$ is a sequence in A converging to x. However, as we shall see in the examples to follow, it is not the case that all points in A are accumulation points.
2. Let $A = (0, 1)$. The set of accumulation points of A is then easily seen to be $[0, 1]$. The set of limit points is also $[0, 1]$.
3. Let $A = [0, 1)$. Then, as in the preceding example, both the set of accumulation points and the set of limit points are the set $[0, 1]$.
4. Let $A = [0, 1] \cup \{2\}$. Then the set of accumulation points is $[0, 1]$ whereas the set of limit points is A.
5. Let $A = \mathbb{Q}$. One can readily check that the set of accumulation points of A is $\mathbb{R}$ and the set of limit points of A is also $\mathbb{R}$. •

The following result gives some properties of the derived set.

2.5.15 Proposition (Properties of the derived set in $\mathbb{R}$) *For* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for a family of subsets* $(\mathrm{A}_{\mathrm{i}})_{\mathrm{i}\in\mathrm{I}}$ *of* $\mathbb{R}$*, the following statements hold:*

(i) $\mathrm{der}(\varnothing) = \varnothing$*;*

(ii) $\mathrm{der}(\mathbb{R}) = \mathbb{R}$*;*

(iii) if $\mathrm{A} \subseteq \mathrm{B}$ *then* $\mathrm{der}(\mathrm{A}) \subseteq \mathrm{der}(\mathrm{B})$*;*

(iv) $\mathrm{der}(\mathrm{A} \cup \mathrm{B}) = \mathrm{der}(\mathrm{A}) \cup \mathrm{der}(\mathrm{B})$*;*

(v) $\mathrm{der}(\mathrm{A} \cap \mathrm{B}) \subseteq \mathrm{der}(\mathrm{A}) \cap \mathrm{der}(\mathrm{B})$.

Proof Parts (i) and (ii) follow directly from the definition of the derived set.

(iii) Let $x \in \mathrm{der}(A)$ and let U be a neighbourhood of x. Then the set $A \cap (U \setminus \{x\})$ is nonempty, implying that the set $B \cap (U \setminus \{x\})$ is also nonempty. Thus $x \in \mathrm{der}(B)$.

(iv) Let $x \in \mathrm{der}(A \cup B)$ and let U be a neighbourhood of x. Then the set $U \cap ((A \cup B) \setminus \{x\})$ is nonempty. But

$$\begin{aligned} U \cap ((A \cup B) \setminus \{x\}) &= U \cap ((A \setminus \{x\}) \cup (B \setminus \{x\})) \\ &= (U \cap (A \setminus \{x\})) \cup (U \cap (B \setminus \{x\})). \end{aligned} \tag{2.8}$$

Thus it cannot be that both $U \cap (A \setminus \{x\})$ and $U \cap (B \setminus \{x\})$ are empty. Thus x is an element of either $\mathrm{der}(A)$ or $\mathrm{der}(B)$.

Now let $x \in \mathrm{der}(A) \cup \mathrm{der}(A)$. Then, using (2.8), $U \cap ((A \cup B) \setminus \{x\})$ is nonempty, and so $x \in \mathrm{der}(A \cup B)$.

(v) Let $x \in \mathrm{der}(A \cap B)$ and let U be a neighbourhood of x. Then $U \cap ((A \cap B) \setminus \{x\}) \neq \varnothing$. We have

$$U \cap ((A \cap B) \setminus \{x\}) = U \cap ((A \setminus \{x\}) \cap (B \setminus \{x\}))$$

Thus the sets $U \cap (A \setminus \{x\})$ and $U \cap (B \setminus \{x\})$ are both nonempty, showing that $x \in \mathrm{der}(A) \cap \mathrm{der}(B)$. ■

Next we turn to characterising distinguished subsets of subsets of $\mathbb{R}$.

2.5.16 Definition (Interior, closure, and boundary in $\mathbb{R}$) Let $A \subseteq \mathbb{R}$.

(i) The ***interior*** of A is the set

$$\mathrm{int}(A) = \cup\{U \mid U \subseteq A,\ U \text{ open}\}.$$

(ii) The ***closure*** of A is the set

$$\mathrm{cl}(A) = \cap\{C \mid A \subseteq C,\ C \text{ closed}\}.$$

(iii) The ***boundary*** of A is the set $\mathrm{bd}(A) = \mathrm{cl}(A) \cap \mathrm{cl}(\mathbb{R} \setminus A)$. •

In other words, the interior of A is the largest open set contained in A. Note that this definition makes sense since a union of open sets is open (Exercise 2.5.1). In like manner, the closure of A is the smallest closed set containing A, and this definition makes sense since an intersection of closed sets is closed (Exercise 2.5.1 again). Note that $\mathrm{int}(A)$ is open and $\mathrm{cl}(A)$ is closed. Moreover, since $\mathrm{bd}(A)$ is the intersection of two closed sets, it too is closed (Exercise 2.5.1 yet again).

Let us give some examples of interiors, closures, and boundaries.

2.5.17 Examples (Interior, closure, and boundary)

1. Let $A = \operatorname{int}(0,1)$. Then $\operatorname{int}(A) = (0,1)$ since A is open. We claim that $\operatorname{cl}(A) = [0,1]$. Clearly $[0,1] \subseteq \operatorname{cl}(A)$ since $[0,1]$ is closed and contains A. Moreover, the only smaller subsets contained in $[0,1]$ and containing A are $[0,1)$, $(0,1]$, and $(0,1)$, none of which are closed. We may then conclude that $\operatorname{cl}(A) = [0,1]$. Finally we claim that $\operatorname{bd}(A) = \{0,1\}$. To see this, note that we have $\operatorname{cl}(A) = [0,1]$ and $\operatorname{cl}(\mathbb{R} \setminus A) = (-\infty,0] \cup [1,\infty)$ (by an argument like that used to show that $\operatorname{cl}(A) = [0,1]$). Therefore, $\operatorname{bd}(A) = \operatorname{cl}(A) \cap \operatorname{cl}(\mathbb{R} \setminus A) = \{0,1\}$, as desired.
2. Let $A = [0,1]$. Then $\operatorname{int}(A) = (0,1)$. To see this, we note that $(0,1) \subseteq \operatorname{int}(A)$ since $(0,1)$ is open and contained in A. Moreover, the only larger sets contained in A are $[0,1)$, $(0,1]$, and $[0,1]$, none of which are open. Thus $\operatorname{int}(A) = (0,1)$, just as claimed. Since A is closed, $\operatorname{cl}(A) = A$. Finally we claim that $\operatorname{bd}(A) = \{0,1\}$. Indeed, $\operatorname{cl}(A) = [0,1]$ and $\operatorname{cl}(\mathbb{R} \setminus A) = (-\infty,0] \cup [1,\infty)$. Therefore, $\operatorname{bd}(A) = \operatorname{cl}(A) \cap \operatorname{cl}(\mathbb{R} \setminus A) = \{0,1\}$, as claimed.
3. Let $A = (0,1) \cup \{2\}$. We have $\operatorname{int}(A) = (0,1)$, $\operatorname{cl}(A) = [0,1] \cup \{2\}$, and $\operatorname{bd}(A) = \{0,1,2\}$. We leave the simple details of these assertions to the reader.
4. Let $A = \mathbb{Q}$. One readily ascertains that $\operatorname{int}(A) = \varnothing$, $\operatorname{cl}(A) = \mathbb{R}$, and $\operatorname{bd}(A) = \mathbb{R}$. •

Now let us give a characterisation of interior, closure, and boundary that are often useful in practice. Indeed, we shall often use these characterisations without explicitly mentioning that we are doing so.

2.5.18 Proposition (Characterisation of interior, closure, and boundary in $\mathbb{R}$) *For* $\mathrm{A} \subseteq \mathbb{R}$, *the following statements hold:*

(i) $\mathrm{x} \in \operatorname{int}(\mathrm{A})$ *if and only if there exists a neighbourhood* U *of* x *such that* $\mathrm{U} \subseteq \mathrm{A}$*;*

(ii) $\mathrm{x} \in \operatorname{cl}(\mathrm{A})$ *if and only if, for each neighbourhood* U *of* x*, the set* $\mathrm{U} \cap \mathrm{A}$ *is nonempty;*

(iii) $\mathrm{x} \in \operatorname{bd}(\mathrm{A})$ *if and only if, for each neighbourhood* U *of* x*, the sets* $\mathrm{U} \cap \mathrm{A}$ *and* $\mathrm{U} \cap (\mathbb{R} \setminus \mathrm{A})$ *are nonempty.*

Proof (i) Suppose that $x \in \operatorname{int}(A)$. Since $\operatorname{int}(A)$ is open, there exists a neighbourhood U of x contained in $\operatorname{int}(A)$. Since $\operatorname{int}(A) \subseteq A$, $U \subseteq A$.

Next suppose that $x \notin \operatorname{int}(A)$. Then, by definition of interior, for any open set U for which $U \subseteq A$, $x \notin U$.

(ii) Suppose that there exists a neighbourhood U of x such that $U \cap A = \varnothing$. Then $\mathbb{R} \setminus U$ is a closed set containing A. Thus $\operatorname{cl}(A) \subseteq \mathbb{R} \setminus U$. Since $x \notin \mathbb{R} \setminus U$, it follows that $x \notin \operatorname{cl}(A)$.

Suppose that $x \notin \operatorname{cl}(A)$. Then x is an element of the open set $\mathbb{R} \setminus \operatorname{cl}(A)$. Thus there exists a neighbourhood U of x such that $U \subseteq \mathbb{R} \setminus \operatorname{cl}(A)$. In particular, $U \cap A = \varnothing$.

(iii) This follows directly from part (ii) and the definition of boundary. ■

Now let us state some useful properties of the interior of a set.

2.5.19 Proposition (Properties of interior in $\mathbb{R}$) *For* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for a family of subsets* $(\mathrm{A}_\mathrm{i})_{\mathrm{i}\in\mathrm{I}}$ *of* $\mathbb{R}$*, the following statements hold:*

(i) $\mathrm{int}(\varnothing) = \varnothing$*;*

(ii) $\mathrm{int}(\mathbb{R}) = \mathbb{R}$*;*

(iii) $\mathrm{int}(\mathrm{int}(\mathrm{A})) = \mathrm{int}(\mathrm{A})$*;*

(iv) if $\mathrm{A} \subseteq \mathrm{B}$ *then* $\mathrm{int}(\mathrm{A}) \subseteq \mathrm{int}(\mathrm{B})$*;*

(v) $\mathrm{int}(\mathrm{A} \cup \mathrm{B}) \supseteq \mathrm{int}(\mathrm{A}) \cup \mathrm{int}(\mathrm{B})$*;*

(vi) $\mathrm{int}(\mathrm{A} \cap \mathrm{B}) = \mathrm{int}(\mathrm{A}) \cap \mathrm{int}(\mathrm{B})$*;*

(vii) $\mathrm{int}(\cup_{\mathrm{i}\in\mathrm{I}}\mathrm{A}_\mathrm{i}) \supseteq \cup_{\mathrm{i}\in\mathrm{I}}\mathrm{int}(\mathrm{A}_\mathrm{i})$*;*

(viii) $\mathrm{int}(\cap_{\mathrm{i}\in\mathrm{I}}\mathrm{A}_\mathrm{i}) \subseteq \cap_{\mathrm{i}\in\mathrm{I}}\mathrm{int}(\mathrm{A}_\mathrm{i})$.

Moreover, a set $\mathrm{A} \subseteq \mathbb{R}$ *is open if and only if* $\mathrm{int}(\mathrm{A}) = \mathrm{A}$*.*

Proof Parts (i) and (ii) are clear by definition of interior. Part (v) follows from part (vii), so we will only prove the latter.

(iii) This follows since the interior of an open set is the set itself.

(iv) Let $x \in \mathrm{int}(A)$. Then there exists a neighbourhood U of x such that $U \subseteq A$. Thus $U \subseteq B$, and the result follows from Proposition 2.5.18.

(vi) Let $x \in \mathrm{int}(A) \cap \mathrm{int}(B)$. Since $\mathrm{int}(A) \cap \mathrm{int}(B)$ is open by Exercise 2.5.1, there exists a neighbourhood U of x such that $U \subseteq \mathrm{int}(A) \cap \mathrm{int}(B)$. Thus $U \subseteq A \cap B$. This shows that $x \in \mathrm{int}(A \cap B)$. This part of the result follows from part (viii).

(vii) Let $x \in \cup_{i\in I}\mathrm{int}(A_i)$. By Exercise 2.5.1 the set $\cup_{i\in I}\mathrm{int}(A_i)$ is open. Thus there exists a neighbourhood U of x such that $U \subseteq \cup_{i\in I}\mathrm{int}(A_i)$. Thus $U \subseteq \cup_{i\in I}A_i$, from which we conclude that $x \in \mathrm{int}(\cup_{i\in I}A_i)$.

(viii) Let $x \in \mathrm{int}(\cap_{i\in I}A_i)$. Then there exists a neighbourhood U of x such that $U \subseteq \cap_{i\in I}A_i$. It therefore follows that $U \subseteq A_i$ for each $i \in I$, and so that $x \in \mathrm{int}(A_i)$ for each $i \in I$.

The final assertion follows directly from Proposition 2.5.18. ■

Next we give analogous results for the closure of a set.

2.5.20 Proposition (Properties of closure in $\mathbb{R}$) *For* $\mathrm{A}, \mathrm{B} \subseteq \mathbb{R}$ *and for a family of subsets* $(\mathrm{A}_\mathrm{i})_{\mathrm{i}\in\mathrm{I}}$ *of* $\mathbb{R}$*, the following statements hold:*

(i) $\mathrm{cl}(\varnothing) = \varnothing$*;*

(ii) $\mathrm{cl}(\mathbb{R}) = \mathbb{R}$*;*

(iii) $\mathrm{cl}(\mathrm{cl}(\mathrm{A})) = \mathrm{cl}(\mathrm{A})$*;*

(iv) if $\mathrm{A} \subseteq \mathrm{B}$ *then* $\mathrm{cl}(\mathrm{A}) \subseteq \mathrm{cl}(\mathrm{B})$*;*

(v) $\mathrm{cl}(\mathrm{A} \cup \mathrm{B}) = \mathrm{cl}(\mathrm{A}) \cup \mathrm{cl}(\mathrm{B})$*;*

(vi) $\mathrm{cl}(\mathrm{A} \cap \mathrm{B}) \subseteq \mathrm{cl}(\mathrm{A}) \cap \mathrm{cl}(\mathrm{B})$*;*

(vii) $\mathrm{cl}(\cup_{\mathrm{i}\in\mathrm{I}}\mathrm{A}_\mathrm{i}) \supseteq \cup_{\mathrm{i}\in\mathrm{I}}\mathrm{cl}(\mathrm{A}_\mathrm{i})$*;*

(viii) $\mathrm{cl}(\cap_{\mathrm{i}\in\mathrm{I}}\mathrm{A}_\mathrm{i}) \subseteq \cap_{\mathrm{i}\in\mathrm{I}}\mathrm{cl}(\mathrm{A}_\mathrm{i})$.

Moreover, a set $\mathrm{A} \subseteq \mathbb{R}$ *is closed if and only if* $\mathrm{cl}(\mathrm{A}) = \mathrm{A}$*.*

Proof Parts (i) and (ii) follow immediately from the definition of closure. Part (vi) follows from part (viii), so we will only prove the latter.

(iii) This follows since the closure of a closed set is the set itself.

(iv) Suppose that $x \in \mathrm{cl}(A)$. Then, for any neighbourhood U of x, the set $U \cap A$ is nonempty, by Proposition 2.5.18. Since $A \subseteq B$, it follows that $U \cap B$ is also nonempty, and so $x \in \mathrm{cl}(B)$.

(v) Let $x \in \mathrm{cl}(A \cup B)$. Then, for any neighbourhood U of x, the set $U \cap (A \cup B)$ is nonempty by Proposition 2.5.18. By Proposition 1.1.4, $U \cap (A \cup B) = (U \cap A) \cup (U \cap B)$. Thus the sets $U \cap A$ and $U \cap B$ are not both nonempty, and so $x \in \mathrm{cl}(A) \cup \mathrm{cl}(B)$. That $\mathrm{cl}(A) \cup \mathrm{cl}(B) \subseteq \mathrm{cl}(A \cup B)$ follows from part (vii).

(vi) Let $x \in \mathrm{cl}(A \cap B)$. Then, for any neighbourhood U of x, the set $U \cap (A \cap B)$ is nonempty. Thus the sets $U \cap A$ and $U \cap B$ are nonempty, and so $x \in \mathrm{cl}(A) \cap \mathrm{cl}(B)$.

(vii) Let $x \in \cup_{i \in I} \mathrm{cl}(A_i)$ and let U be a neighbourhood of x. Then, for each $i \in I$, $U \cap A_i \neq \emptyset$. Therefore, $\cup_{i \in I}(U \cap A_i) \neq \emptyset$. By Proposition 1.1.7, $\cup_{i \in I}(U \cap A_i) = U \cap (\cup_{i \in I} A_i)$, showing that $U \cap (\cup_{i \in I} A_i) \neq \emptyset$. Thus $x \in \mathrm{cl}(\cup_{i \in I} A_i)$.

(viii) Let $x \in \mathrm{cl}(\cap_{i \in I} A_i)$ and let U be a neighbourhood of x. Then the set $U \cap (\cap_{i \in I} A_i)$ is nonempty. This means that, for each $i \in I$, the set $U \cap A_i$ is nonempty. Thus $x \in \mathrm{cl}(A_i)$ for each $i \in I$, giving the result. ■

Note that there is a sort of "duality" between int and cl as concerns their interactions with union and intersection. This is reflective of the fact that open and closed sets themselves have such a "duality," as can be seen from Exercise 2.5.1. We refer the reader to Exercise 2.5.6 to construct counterexamples to any missing opposite inclusions in Propositions 2.5.19 and 2.5.20.

Let us state some relationships between certain of the concepts we have thus far introduced.

2.5.21 Proposition (Joint properties of interior, closure, boundary, and derived set in $\mathbb{R}$) *For $A \subseteq \mathbb{R}$, the following statements hold:*

(i) $\mathbb{R} \setminus \mathrm{int}(A) = \mathrm{cl}(\mathbb{R} \setminus A)$*;*

(ii) $\mathbb{R} \setminus \mathrm{cl}(A) = \mathrm{int}(\mathbb{R} \setminus A)$*.*

(iii) $\mathrm{cl}(A) = A \cup \mathrm{bd}(A)$*;*

(iv) $\mathrm{int}(A) = A - \mathrm{bd}(A)$*;*

(v) $\mathrm{cl}(A) = \mathrm{int}(A) \cup \mathrm{bd}(A)$*;*

(vi) $\mathrm{cl}(A) = A \cup \mathrm{der}(A)$*;*

(vii) $\mathbb{R} = \mathrm{int}(A) \cup \mathrm{bd}(A) \cup \mathrm{int}(\mathbb{R} \setminus A)$*.*

Proof (i) Let $x \in \mathbb{R} \setminus \mathrm{int}(A)$. Since $x \notin \mathrm{int}(A)$, for every neighbourhood U of x it holds that $U \not\subset A$. Thus, for any neighbourhood U of x, we have $U \cap (\mathbb{R} \setminus A) \neq \emptyset$, showing that $x \in \mathrm{cl}(\mathbb{R} \setminus A)$.

Now let $x \in \mathrm{cl}(\mathbb{R} \setminus A)$. Then for any neighbourhood U of x we have $U \cap (\mathbb{R} \setminus A) \neq \emptyset$. Thus $x \notin \mathrm{int}(A)$, so $x \in \mathbb{R} \setminus A$.

(ii) The proof here strongly resembles that for part (i), and we encourage the reader to provide the explicit arguments.

(iii) This follows from part (v).

(iv) Clearly $\mathrm{int}(A) \subseteq A$. Suppose that $x \in A \cap \mathrm{bd}(A)$. Then, for any neighbourhood U of x, the set $U \cap (\mathbb{R} \setminus A)$ is nonempty. Therefore, no neighbourhood of x is a subset of A, and so $x \notin \mathrm{int}(A)$. Conversely, if $x \in \mathrm{int}(A)$ then there is a neighbourhood U of x such that $U \subseteq A$. The precludes the set $U \cap (\mathbb{R} \setminus A)$ from being nonempty, and so we must have $x \notin \mathrm{bd}(A)$.

(v) Let $x \in \mathrm{cl}(A)$. For a neighbourhood U of x it then holds that $U \cap A \neq \emptyset$. If there exists a neighbourhood V of x such that $V \subseteq A$, then $x \in \mathrm{int}(A)$. If there exists *no* neighbourhood V of x such that $V \subseteq A$, then for every neighbourhood V of x we have $V \cap (\mathbb{R} \setminus A) \neq \emptyset$, and so $x \in \mathrm{bd}(A)$.

Now let $x \in \mathrm{int}(A) \cup \mathrm{bd}(A)$. If $x \in \mathrm{int}(A)$ then $x \in A$ and so $x \in\subseteq \mathrm{cl}(A)$. If $x \in \mathrm{bd}(A)$ then it follows immediately from Proposition 2.5.18 that $x \in \mathrm{cl}(A)$.

(vi) Let $x \in \mathrm{cl}(A)$. If $x \notin A$ then, for every neighbourhood U of x, $U \cap A = U \cap (A \setminus \{x\}) \neq \emptyset$, and so $x \in \mathrm{der}(A)$.

If $x \in A \cup \mathrm{der}(A)$ then either $x \in A \subseteq \mathrm{cl}(A)$, or $x \notin A$. In this latter case, $x \in \mathrm{der}(A)$ and so the set $U \cap (A \setminus \{x\})$ is nonempty for each neighbourhood U of x, and we again conclude that $x \in \mathrm{cl}(A)$.

(vii) Clearly $\mathrm{int}(A) \cap \mathrm{int}(\mathbb{R} \setminus A) = \emptyset$ since $A \cap (\mathbb{R} \setminus A) = \emptyset$. Now let $x \in \mathbb{R} \setminus (\mathrm{int}(A) \cup \mathrm{int}(\mathbb{R} \setminus A))$. Then, for any neighbourhood U of x, we have $U \not\subseteq A$ and $U \not\subseteq (\mathbb{R} \setminus A)$. Thus the sets $U \cap (\mathbb{R} \setminus A)$ and $U \cap A$ must both be nonempty, from which we conclude that $x \in \mathrm{bd}(A)$. ■

An interesting class of subset of $\mathbb{R}$ is the following.

2.5.22 Definition (Discrete subset of $\mathbb{R}$) A subset $A \subseteq \mathbb{R}$ is ***discrete*** if there exists $\epsilon \in \mathbb{R}_{>0}$ such that, for each $x, y \in A$, $|x - y| \geq \epsilon$. •

Let us give a characterisation of discrete sets.

2.5.23 Proposition (Characterisation of discrete sets in $\mathbb{R}$) *A discrete subset* A $\subseteq \mathbb{R}$ *is countable and has no accumulation points.*

Proof It is easy to show (Exercise 2.5.8) that if A is discrete and if $N \in \mathbb{Z}_{>0}$, then the set $A \cap [-N, N]$ is finite. Therefore

$$A = \cup_{N \in \mathbb{Z}_{>0}} A \cap [-N, N],$$

which gives A as a countable union of finite sets, implying that A is countable by Proposition 1.7.16. Now let $\epsilon \in \mathbb{R}_{>0}$ satisfy $|x - y| \geq \epsilon$ for $x, y \in A$. Then, if $x \in A$ then the set $A \cap \mathsf{B}(\frac{\epsilon}{2}, x)$ is empty, implying that x is not an accumulation point. If $x \notin A$ then $\mathsf{B}(\frac{\epsilon}{2}, x)$ can contain at most one point from A, which again prohibits x from being an accumulation point. ■

The notion of a discrete set is actually a more general one having to do with what is known as the discrete topology (cf. Example III-1.2.3–6). The reader can explore some facts about discrete subsets of $\mathbb{R}$ in Exercise 2.5.8.

2.5.4 Compactness

The idea of compactness is absolutely fundamental in much of mathematics. The reasons for this are not at all clear to a newcomer to analysis. Indeed, the definition we give for compactness comes across as extremely unmotivated. This might be particularly since for $\mathbb{R}$ (or more generally, in $\mathbb{R}^n$) compact sets have a fairly banal characterisation as sets that are closed and bounded (Theorem 2.5.27). However, the original definition we give for a compact set is the most useful one. The main reason it is useful is that it allows for certain pointwise properties to be automatically extended to the entire set. A good example of this is Theorem 3.1.24, where continuity of a function on a compact set is extended to uniform continuity on the set. This idea of uniformity is an important one, and accounts for much of the value of the notion of compactness. But we are getting ahead of ourselves.

As indicated in the above paragraph, we shall give a rather strange seeming definition of compactness. Readers looking for a quick and dirty definition of compactness, valid for subsets of $\mathbb{R}$, can refer ahead to Theorem 2.5.27. Our construction relies on the following idea.

2.5.24 Definition (Open cover of a subset of $\mathbb{R}$) Let $A \subseteq \mathbb{R}$.

(i) An ***open cover*** for A is a family $(U_i)_{i\in I}$ of open subsets of $\mathbb{R}$ having the property that $A \subseteq \cup_{i\in I} U_i$.

(ii) A ***subcover*** of an open cover $(U_i)_{i\in I}$ of A is an open cover $(V_j)_{j\in J}$ of A having the property that $(V_j)_{j\in J} \subseteq (U_i)_{i\in I}$. •

The following property of open covers of subsets of $\mathbb{R}$ is useful.

2.5.25 Lemma (Lindelöf[10] Lemma for $\mathbb{R}$) *If* $(\mathrm{U_i})_{\mathrm{i}\in\mathrm{I}}$ *is an open cover of* $\mathrm{A} \subseteq \mathbb{R}$*, then there exists a countable subcover of* A.

Proof Let $\mathscr{B} = \{\mathsf{B}(r,x) \mid x, r \in \mathbb{Q}\}$. Note that $\mathscr{B}$ is a countable union of countable sets, and so is countable by Proposition 1.7.16. Therefore, we can write $\mathscr{B} = (\mathsf{B}(r_j, x_j))_{j\in\mathbb{Z}_{>0}}$. Now define

$$\mathscr{B}' = \{\mathsf{B}(r_j, x_j) \mid \mathsf{B}(r_j, x_j) \subseteq U_i \text{ for some } i \in I\}.$$

Let us write $\mathscr{B}' = (\mathsf{B}(r_{j_k}, x_{j_k}))_{k\in\mathbb{Z}_{>0}}$. We claim that $\mathscr{B}'$ covers A. Indeed, if $x \in A$ then $x \in U_i$ for some $i \in I$. Since U_i is open there then exists $k \in \mathbb{Z}_{>0}$ such that $x \in \mathsf{B}(r_{j_k}, x_{j_k}) \subseteq U_i$. Now, for each $k \in \mathbb{Z}_{>0}$, let $i_k \in I$ satisfy $\mathsf{B}(r_{j_k}, x_{j_k}) \subseteq U_{i_k}$. Then the countable collection of open sets $(U_{i_k})_{k\in\mathbb{Z}_{>0}}$ clearly covers A since $\mathscr{B}'$ covers A. ■

Now we define the important notion of compactness, along with some other related useful concepts.

2.5.26 Definition (Bounded, compact, and totally bounded in $\mathbb{R}$) A subset $A \subseteq \mathbb{R}$ is:

(i) ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $A \subseteq \overline{\mathsf{B}}(M, 0)$;

(ii) ***compact*** if every open cover $(U_i)_{i\in I}$ of A possesses a finite subcover;

(iii) ***precompact***[11] if $\mathrm{cl}(A)$ is compact;

[10] Ernst Leonard Lindelöf (1870–1946) was a Finnish mathematician who worked in the areas of differential equations and complex analysis.

[11] What we call "precompact" is very often called "relatively compact." However, we shall use the term "relatively compact" for something different.

(iv) ***totally bounded*** if, for every $\epsilon \in \mathbb{R}_{>0}$ there exists $x_1, \dots, x_k \in \mathbb{R}$ such that $A \subseteq \cup_{j=1}^{k} \mathsf{B}(\epsilon, x_j)$. •

The simplest characterisation of compact subsets of $\mathbb{R}$ is the following. We shall freely interchange our use of the word compact between the definition given in Definition 2.5.26 and the conclusions of the following theorem.

2.5.27 Theorem (Heine–Borel[12] Theorem in $\mathbb{R}$) *A subset* $\mathrm{K} \subseteq \mathbb{R}$ *is compact if and only if it is closed and bounded.*

Proof Suppose that K is closed and bounded. We first consider the case when $K = [a, b]$. Let $\mathscr{O} = (U_i)_{i \in I}$ be an open cover for $[a, b]$ and let

$$S_{[a,b]} = \{x \in \mathbb{R} \mid \ x \leq b \text{ and } [a, x] \text{ has a finite subcover in } \mathscr{O}\}.$$

Note that $S_{[a,b]} \neq \emptyset$ since $a \in S_{[a,b]}$. Let $c = \sup S_{[a,b]}$. We claim that $c = b$. Suppose that $c < b$. Since $c \in [a, b]$ there is some $\bar{i} \in I$ such that $c \in U_{\bar{i}}$. As $U_{\bar{i}}$ is open, there is some $\epsilon \in \mathbb{R}_{>0}$ sufficiently small that $\mathsf{B}(\epsilon, c) \subseteq U_{\bar{i}}$. By definition of c, there exists some $x \in (c - \epsilon, c)$ for which $x \in S_{[a,b]}$. By definition of $S_{[a,b]}$ there is a finite collection of open sets $U_{i_1}, \dots, U_{i_m}$ from $\mathscr{O}$ which cover $[a, x]$. Therefore, the finite collection $U_{i_1}, \dots, U_{i_m}, U_{\bar{i}}$ of open sets covers $[a, c+\epsilon)$. This then contradicts the fact that $c = \sup S_{[a,b]}$, so showing that $b = \sup S_{[a,b]}$. The result follows by definition of $S_{[a,b]}$.

Now suppose that K is a general closed and bounded set. Then $K \subseteq [a, b]$ for some suitable $a, b \in \mathbb{R}$. Suppose that $\mathscr{O} = (U_i)_{i \in I}$ is an open cover of K, and define a new open cover $\tilde{\mathscr{O}} = \mathscr{O} \cup (\mathbb{R} \setminus K)$. Note that $\cup_{i \in I} U_i \cup (\mathbb{R} \setminus K) = \mathbb{R}$ showing that $\tilde{\mathscr{O}}$ is an open cover for $\mathbb{R}$, and therefore also is an open cover for $[a, b]$. By the first part of the proof, there exists a finite subset of $\tilde{\mathscr{O}}$ which covers $[a, b]$, and therefore also covers K. We must show that this finite cover can be chosen so as not to include the set $\mathbb{R} \setminus K$ as this set is not necessarily in $\mathscr{O}$. However, if $[a, b]$ is covered by $U_{i_1}, \dots, U_{i_k}, \mathbb{R} \setminus K$, then one sees that K is covered by $U_{i_1}, \dots, U_{i_k}$, since $K \cap (\mathbb{R} \setminus K) = \emptyset$. Thus we have arrived at a finite subset of $\mathscr{O}$ covering K, as desired.

Now suppose that K is compact. Consider the following collection of open subsets: $\mathscr{O}_K = (\mathsf{B}(\epsilon, x))_{x \in K}$. Clearly this is an open cover of K. Thus there exists a finite collection of point $x_1, \dots, x_k \in K$ such that $(\mathsf{B}(\epsilon, x_j))_{j \in \{1,\dots,k\}}$ covers K. If we take

$$M = \max\{|x_1|, \dots, |x_k|\} + 2$$

then we easily see that $K \subseteq \overline{\mathsf{B}}(M, 0)$, so that K is bounded. Now suppose that K is not closed. Then $K \subset \mathrm{cl}(K)$. By part (vi) of Proposition 2.5.21 there exists an accumulation point x_0 of K that is not in K. Then, for any $j \in \mathbb{Z}_{>0}$ there exists a point $x_j \in K$ such that $|x_0 - x_j| < \frac{1}{j}$. Define

$$U_j = (-\infty, x_0 - \tfrac{1}{j}) \cup (x_0 + \tfrac{1}{j}, \infty),$$

noting that U_j is open, since it is the union of open sets (see Exercise 2.5.1). We claim that $(U_j)_{j \in \mathbb{Z}_{>0}}$ is an open cover of K. Indeed, we will show that $\cup_{j \in \mathbb{Z}_{>0}} U_j = \mathbb{R} \setminus \{x_0\}$.

[12] Heinrich Eduard Heine (1821–1881) was a German mathematician who worked mainly with special functions. Félix Edouard Justin Emile Borel (1871–1956) was a French mathematician, and he worked mainly in the area of analysis.

To see this, let $x \in \mathbb{R} \setminus \{x_0\}$ and choose $k \in \mathbb{Z}_{>0}$ such that $\frac{1}{k} < |x - x_0|$. Then it follows by definition of U_k that $x \in U_k$. Since $x_0 \notin K$, we then have $K \subseteq \cup_{j \in \mathbb{Z}_{>0}} U_j$. Next we show that there is no finite subset of $(U_j)_{j \in \mathbb{Z}_{>0}}$ that covers K. Indeed, consider a finite set $j_1, \dots, j_k \in \mathbb{Z}_{>0}$, and suppose without loss of generality that $j_1 < \cdots < j_k$. Then the point x_{j_k+1} satisfies $|x_0 - x_{j_k+1}| < \frac{1}{j_k+1} < \frac{1}{j_k}$, implying that $x_{j_k+1} \notin U_{j_k} \supseteq \cdots \supseteq U_{j_1}$. Thus, if K is not closed, we have constructed an open cover of K having no finite subcover. From this we conclude that if K is compact, then it is closed. ■

The Heine–Borel Theorem has the following useful corollary.

2.5.28 Corollary (Closed subsets of compact sets in $\mathbb{R}$ are compact) *If* $A \subseteq \mathbb{R}$ *is compact and if* $B \subseteq A$ *is closed, then* B *is compact.*

Proof Since A is bounded by the Heine–Borel Theorem, B is also bounded. Thus B is also compact, again by the Heine–Borel Theorem. ■

In Chapter III-1 we shall encounter many of the ideas in this section in the more general setting of topological spaces. Many of the ideas for $\mathbb{R}$ transfer directly to this more general setting. However, with compactness, some care must be exercised. In particular, it is *not* true that, in a general topological space, a subset is compact if and only if it is closed and bounded. Indeed, in a general topological space, the notion of bounded is not defined. It is not an uncommon error for newcomers to confuse "compact" with "closed and bounded" in situations where this is not the case.

The following result is another equivalent characterisation of compact subsets of $\mathbb{R}$, and is often useful.

2.5.29 Theorem (Bolzano–Weierstrass[13] Theorem in $\mathbb{R}$) *A subset* $K \subseteq \mathbb{R}$ *is compact if and only if every sequence in* K *has a subsequence which converges in* K*.*

Proof First suppose that K is compact. Let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in K. Since K is bounded by Theorem 2.5.27, the sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ is bounded. We next show that there exists either a monotonically increasing, or a monotonically decreasing, subsequence of $(x_j)_{j \in \mathbb{Z}_{>0}}$. Define

$$D = \{j \in \mathbb{Z}_{>0} \mid x_k > x_j, \ k > j\}$$

If the set D is infinite, then we can write $D = (j_k)_{k \in \mathbb{Z}_{>0}}$. By definition of D, it follows that $x_{j_{k+1}} > x_{j_k}$ for each $k \in \mathbb{Z}_{>0}$. Thus the subsequence $(x_{j_k})_{k \in \mathbb{Z}_{>0}}$ is monotonically increasing. If the set D is finite choose $j_1 > \sup D$. Then there exists $j_2 > j_1$ such that $x_{j_2} \leq x_{j_1}$. Since $j_2 > \sup D$, there then exists $j_3 > j_2$ such that $x_{j_3} \leq x_{j_2}$. By definition of D, this process can be repeated inductively to yield a monotonically decreasing subsequence $(x_{j_k})_{k \in \mathbb{Z}_{>0}}$. It now follows from Theorem 2.3.8 that the sequence $(x_{j_k})_{k \in \mathbb{Z}_{>0}}$, be it monotonically increasing or monotonically decreasing, converges.

[13] Bernard Placidus Johann Nepomuk Bolzano (1781–1848) was a Czechoslovakian philosopher, mathematician, and theologian who made mathematical contributions to the field of analysis. Karl Theodor Wilhelm Weierstrass (1815–1897) is one of the greatest of all mathematicians. He made significant contributions to the fields of analysis, complex function theory, and the calculus of variations.

Next suppose that every sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in K possesses a convergent subsequence. Let $(U_i)_{i\in I}$ be an open cover of K, and by Lemma 2.5.25 choose a countable subcover which we denote by $(U_j)_{j\in\mathbb{Z}_{>0}}$. Now suppose that every finite subcover of $(U_j)_{j\in\mathbb{Z}_{>0}}$ does not cover K. This means that, for every $k\in\mathbb{Z}_{>0}$, the set $C_k = K\setminus\left(\cup_{j=1}^{k}U_j\right)$ is nonempty. Thus we may define a sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ in $\mathbb{R}$ such that $x_k\in C_k$. Since the sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ is in K, it possesses a convergent subsequence $(x_{k_m})_{m\in\mathbb{Z}_{>0}}$, by hypotheses. Let x be the limit of this subsequence. Since $x\in K$ and since $K=\cup_{j\in\mathbb{Z}_{>0}}U_j$, $x\in U_l$ for some $l\in\mathbb{Z}_{>0}$. Since the sequence $(x_{k_m})_{m\in\mathbb{Z}_{>0}}$ converges to x, it follows that there exists $N\in\mathbb{Z}_{>0}$ such that $x_{k_m}\in U_l$ for $m\geq N$. But this contradicts the definition of the sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$, forcing us to conclude that our assumption is wrong that there is no finite subcover of K from the collection $(U_j)_{j\in\mathbb{Z}_{>0}}$. ■

The following property of compact intervals of $\mathbb{R}$ is useful.

2.5.30 Theorem (Lebesgue[14] number for compact intervals) *Let* $\mathrm{I}=[\mathrm{a},\mathrm{b}]$ *be a compact interval. Then for any open cover* $(\mathrm{U}_\alpha)_{\alpha\in\mathrm{A}}$ *of* $[\mathrm{a},\mathrm{b}]$*, there exists* $\delta\in\mathbb{R}_{>0}$*, called the* ***Lebesgue number*** *of* I*, such that, for each* $\mathrm{x}\in[\mathrm{a},\mathrm{b}]$*, there exists* $\alpha\in\mathrm{A}$ *such that* $\mathsf{B}(\delta,\mathrm{x})\cap\mathrm{I}\subseteq\mathrm{U}_\alpha$*.*

Proof Suppose there exists an open cover $(U_\alpha)_{\alpha\in A}$ such that, for all $\delta\in\mathbb{R}_{>0}$, there exists $x\in[a,b]$ such that none of the sets U_α, $\alpha\in A$, contains $\mathsf{B}(\delta,x)\cap I$. Then there exists a sequence $(x_j)_{j\in\mathbb{Z}_{>0}}$ in I such that

$$\{\alpha\in A \mid \mathsf{B}(\tfrac{1}{j},x_j)\subseteq U_\alpha\}=\emptyset$$

for each $j\in\mathbb{Z}_{>0}$. By the Bolzano–Weierstrass Theorem there exists a subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ that converges to a point, say x, in $[a,b]$. Then there exists $\epsilon\in\mathbb{R}_{>0}$ and $\alpha\in A$ such that $\mathsf{B}(\epsilon,x)\subseteq U_\alpha$. Now let $N\in\mathbb{Z}_{>0}$ be sufficiently large that $|x_{j_k}-x|<\frac{\epsilon}{2}$ for $k\geq N$ and such that $\frac{1}{j_N}<\frac{\epsilon}{2}$. Now let $k\geq N$. Then, if $y\in\mathsf{B}(\frac{1}{j_k},x_{j_k})$ we have

$$|y-x|=|y-x_{j_k}+x_{j_k}-x|\leq|y-x_{j_k}|+|x-x_{j_k}|<\epsilon.$$

Thus we arrive at the contradiction that $\mathsf{B}(\frac{1}{j_k},x_{j_k})\subseteq U_\alpha$. ■

The following result is sometimes useful.

2.5.31 Proposition (Countable intersections of nested compact sets are nonempty) *Let* $(\mathrm{K}_j)_{j\in\mathbb{Z}_{>0}}$ *be a collection of compact subsets of* $\mathbb{R}$ *satisfying* $\mathrm{K}_{j+1}\subseteq\mathrm{K}_j$*. Then* $\cap_{j\in\mathbb{Z}_{>0}}\mathrm{K}_j$ *is nonempty.*

Proof It is clear that $K=\cap_{j\in\mathbb{Z}_{>0}}K_j$ is bounded, and moreover it is closed by Exercise 2.5.1. Thus K is compact by the Heine–Borel Theorem. Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence for which $x_j\in K_j$ for $j\in\mathbb{Z}_{>0}$. This sequence is thus a sequence in K_1 and so, by the Bolzano–Weierstrass Theorem, has a subsequence $(x_{j_k})_{k\in\mathbb{Z}_{>0}}$ converging to $x\in K_1$. The sequence $(x_{j_{k+1}})_{k\in\mathbb{Z}_{>0}}$ is then a sequence in K_2 which is convergent, so showing that $x\in K_2$. Similarly, one shows that $x\in K_j$ for all $j\in\mathbb{Z}_{>0}$, giving the result. ■

[14] Henri Léon Lebesgue (1875–1941) was a French mathematician. His work was in the area of analysis. The Lebesgue integral is considered to be one of the most significant contributions to mathematics in the past century or so.

Finally, let us indicate the relationship between the notions of relative compactness and total boundedness. We see that for $\mathbb{R}$ these concepts are the same. This may not be true in general; see Exercise III-1.1.19.

2.5.32 Proposition ("Precompact" equals "totally bounded" in $\mathbb{R}$) *A subset of $\mathbb{R}$ is precompact if and only if it is totally bounded.*

Proof Let $A \subseteq \mathbb{R}$.

First suppose that A is precompact. Since $A \subseteq \mathrm{cl}(A)$ and since $\mathrm{cl}(A)$ is bounded by the Heine–Borel Theorem, it follows that A is bounded. It is then easy to see that A is totally bounded.

Now suppose that A is totally bounded. For $\epsilon \in \mathbb{R}_{>0}$ let $x_1, \dots, x_k \in \mathbb{R}$ have the property that $A \subseteq \cup_{j=1}^{k} \mathsf{B}(\epsilon, x_j)$. If

$$M_0 = \max\{|x_j - x_l| \mid j, l \in \{i, \dots, k\}\} + 2\epsilon,$$

then it is easy to see that $A \subseteq \mathsf{B}(M, 0)$ for any $M > M_0$. Then $\mathrm{cl}(A) \subseteq \overline{\mathsf{B}}(M, 0)$ by part (iv) of Proposition 2.5.20, and so $\mathrm{cl}(A)$ is bounded. Since $\mathrm{cl}(A)$ is closed, it follows from the Heine–Borel Theorem that A is precompact. ■

2.5.5 Connectedness

The idea of a connected set will come up occasionally in these volumes. Intuitively, a set is connected if it cannot be "broken in two." We will study it more systematically in Section III-1.7, and here we only give enough detail to effectively characterise connected subsets of $\mathbb{R}$.

2.5.33 Definition (Connected subset of $\mathbb{R}$) Subsets $A, B \subseteq \mathbb{R}$ are ***separated*** if $A \cap \mathrm{cl}(B) = \emptyset$ and $\mathrm{cl}(A) \cap B = \emptyset$. A subset $S \subseteq \mathbb{R}$ is ***disconnected*** if $S = A \cup B$ for nonempty separated subsets A and B. A subset $S \subseteq \mathbb{R}$ is ***connected*** if it is not disconnected. •

Rather than give examples, let us simply immediately characterise the connected subsets of $\mathbb{R}$, since this renders all examples trivial to understand.

2.5.34 Theorem (Connected subsets of $\mathbb{R}$ are intervals and vice versa) *A subset $\mathsf{S} \subseteq \mathbb{R}$ is connected if and only if S is an interval.*

Proof Suppose that S is not an interval. Then, by Proposition 2.5.5, there exists $a, b \in S$ with $a < b$ and $c \in (a, b)$ such that $c \notin S$. Let $A_c = S \cap (-\infty, c)$ and $B_c = S \cap (c, \infty)$, and note that both A_c and B_c are nonempty. Also, since $c \notin S$, $S = A_c \cup B_c$. Since $(-\infty, c) \cap [c, \infty) = \emptyset$ and $(-\infty, c] \cap (c, \infty) = \emptyset$, A_c and B_c are separated. That S is not connected follows.

Now suppose that S is not connected, and write $S = A \cup B$ for nonempty separated sets A and B. Without loss of generality, let $a \in A$ and $b \in B$ have the property that $a < b$. Note that $A \cap [a, b]$ is bounded so that $c = \sup A \cap [a, b]$ exists in $\mathbb{R}$. Then $c \in \mathrm{cl}(A \cap [a, b]) \subseteq \mathrm{cl}(A) \cap [a, b]$. In other words, $c \in \mathrm{cl}(A)$. Since $\mathrm{cl}(A) \cap B = \emptyset$, $c \notin B$. If $c \notin A$ then $c \notin S$, and so S is not connected by Proposition 2.5.5. If $c \in A$ then, since $A \cap \mathrm{cl}(B) = \emptyset$, $c \notin \mathrm{cl}(B)$. In this case there exists an open interval containing c that does not intersect $\mathrm{cl}(B)$. In particular, there exists $d > c$ such that $d \notin B$. Since $d > c$

we also have $d \notin A$, and so $d \notin S$. Again we conclude that S is not an interval by Proposition 2.5.5. ■

Let us consider a few examples.

2.5.35 Examples (Connected subsets of sets)

1. If $D \subseteq \mathbb{R}$ is a discrete set as given in Definition 2.5.22. From Theorem 2.5.34 we see that the only subsets of D that are connected are singletons.
2. Note that it also follows from Theorem 2.5.34 that the only connected subsets of $\mathbb{Q} \subseteq \mathbb{R}$ are singletons. However, $\mathbb{Q}$ is not discrete. •

2.5.6 Sets of measure zero

The topic of this section will receive a full treatment in the context of measure theory as presented in Chapter III-2. However, it is convenient here to talk about a simple concepts from measure theory, one which formalises the idea of a set being "small." We shall only give here the definition and a few examples. The reader should look ahead to Chapter III-2 for more detail.

2.5.36 Definition (Set of measure zero in $\mathbb{R}$) A subset $A \subseteq \mathbb{R}$ has ***measure zero***, or is ***of measure zero***, if

$$\inf\left\{\sum_{j=1}^{\infty}|b_j - a_j| \;\middle|\; A \subseteq \bigcup_{j\in\mathbb{Z}_{>0}}(a_j, b_j)\right\} = 0. \qquad \bullet$$

The idea, then, is that one can cover a set A with open intervals, each of which have some length. One can add all of these lengths to get a total length for the intervals used to cover A. Now, if one can make this total length arbitrarily small, then the set has measure zero.

2.5.37 Notation ("Almost everywhere" and "a.e.") We give here an important piece of notation associated to the notion of a set of measure zero. Let $A \subseteq \mathbb{R}$ and let $P\colon A \to \{\text{true}, \text{false}\}$ be a property defined on A (see the prelude to the Principle of Transfinite Induction, Theorem 1.5.14). The property P holds ***almost everywhere***, ***a.e.***, or ***for almost every* x ∈ A** if the set $\{x \in A \mid P(x) = \text{false}\}$ has measure zero. •

This is best illustrated with some examples.

2.5.38 Examples (Sets of measure zero)

1. Let $A = \{x_1, \dots, x_k\}$ for some distinct $x_1, \dots, x_k \in \mathbb{R}$. We claim that this set has measure zero. Note that for any $\epsilon \in \mathbb{R}_{>0}$ the intervals $(x_j - \frac{\epsilon}{4k}, x_j + \frac{\epsilon}{4k})$, $j \in \{1, \dots, k\}$, clearly cover A. Now consider the countable collection of open intervals

$$((x_j - \tfrac{\epsilon}{4k}, x_j + \tfrac{\epsilon}{4k}))_{j\in\{1,\dots,k\}} \cup ((0, \tfrac{\epsilon}{2^{j+1}}))_{j\in\mathbb{Z}_{>0}}$$

obtained by adding to the intervals covering A a collection of intervals around zero. The total length of these intervals is

$$\sum_{j=1}^{k} |(x_j + \tfrac{\epsilon}{4k}) - (x_j - \tfrac{\epsilon}{4k})| + \frac{\epsilon}{2}\sum_{j=1}^{\infty} \frac{1}{2^j} = \frac{\epsilon}{2} + \frac{\epsilon}{2},$$

using the fact that $\sum_{j=1}^{\infty} \frac{\epsilon}{2^j} = 1$ (by Example 2.4.2–1). Since $\inf\{2k\epsilon \mid \epsilon \in \mathbb{R}_{>0}\} = 0$, our claim that A has zero measure is validated.

2. Now let $A = \mathbb{Q}$ be the set of rational numbers. To show that A has measure zero, note that from Exercise 2.1.3 that A is countable. Thus we can write the elements of A as $(q_j)_{j\in\mathbb{Z}_{>0}}$. Now let $\epsilon \in \mathbb{R}_{>0}$ and for $j \in \mathbb{Z}_{>0}$ define $a_j = q_j - \frac{\epsilon}{2^j}$ and $b_j = q_j + \frac{\epsilon}{2^j}$. Then the collection (a_j, b_j), $j \in \mathbb{Z}_{>0}$, covers A. Moreover,

$$\sum_{j=1}^{\infty} |b_j - a_j| = \sum_{j=1}^{\infty} \frac{2\epsilon}{2^j} = 2\epsilon,$$

using the fact, shown in Example 2.4.2–1, that the series $\sum_{j=1}^{\infty} \frac{1}{2^j}$ converges to 1. Now, since $\inf\{2\epsilon \mid \epsilon \in \mathbb{R}_{>0}\} = 0$, it follows that A indeed has measure zero.

3. Let $A = \mathbb{R} \setminus \mathbb{Q}$ be the set of irrational numbers. We claim that this set does not have measure zero. To see this, let $k \in \mathbb{Z}_{>0}$ and consider the set $A_k = A \cap [-k, k]$. Now let $\epsilon \in \mathbb{R}_{>0}$. We claim that if $((a_j, b_j))_{j\in\mathbb{Z}_{>0}}$, is a collection of open intervals for which $A_k \subseteq \cup_{j\in\mathbb{Z}_{>0}}(a_j, b_j)$, then

$$\sum_{j=1}^{\infty} |b_j - a_j| \geq 2k - \epsilon. \tag{2.9}$$

To see this, let $((c_l, d_l))_{l\in\mathbb{Z}_{>0}}$ be a collection of intervals such that $\mathbb{Q} \cap [-k, k] \subseteq \cup_{l\in\mathbb{Z}_{>0}}(c_l, d_l)$ and such that

$$\sum_{l=1}^{\infty} |d_l - c_l| < \epsilon.$$

Such a collection of intervals exists since we have already shown that $\mathbb{Q}$, and therefore $\mathbb{Q} \cap [-k, k]$, has measure zero (see Exercise 2.5.9). Now note that

$$[-k, k] \subseteq \left(\bigcup_{j\in\mathbb{Z}_{>0}} (a_j, b_j)\right) \cup \left(\bigcup_{l\in\mathbb{Z}_{>0}} (c_l, d_l)\right),$$

so that

$$\left(\sum_{j=1}^{\infty} |b_j - a_j|\right) + \left(\sum_{l=1}^{\infty} |d_l - c_l|\right) \geq 2k.$$

From this we immediately conclude that (2.9) does indeed hold. Moreover, (2.9) holds for every $k \in \mathbb{Z}_{>0}$, for every $\epsilon \in \mathbb{R}_{>0}$, and for every open cover $((a_j, b_j))_{j\in\mathbb{Z}_{>0}}$ of A_k. Thus,

$$\inf\left\{\sum_{l=1}^{\infty}|\tilde{b}_l - \tilde{a}_l| \;\middle|\; A \subseteq \bigcup_{l\in\mathbb{Z}_{>0}}(\tilde{a}_l, \tilde{b}_l)\right\} \\ \geq \inf\left\{\sum_{j=1}^{\infty}|b_j - a_j| \;\middle|\; A_k \subseteq \bigcup_{j\in\mathbb{Z}_{>0}}(a_j, b_j)\right\} \geq 2k - \epsilon$$

for every $k \in \mathbb{Z}_{>0}$ and for every $\epsilon \in \mathbb{R}_{>0}$. This precludes A from having measure zero. •

The preceding examples suggest sets of measure zero are countable. This is not so, and the next famous example gives an example of an uncountable set with measure zero.

2.5.39 Example (An uncountable set of measure zero: the middle-thirds Cantor set) In this example we construct one of the standard "strange" sets used in real analysis to exhibit some of the characteristics that can possibly be attributed to subsets of $\mathbb{R}$. We shall also use this set in a construction in Example 3.2.27 to give an example of a continuous monotonically increasing function whose derivative is zero almost everywhere.

Let $C_0 = [0, 1]$. Then define

$$\begin{aligned} C_1 &= [0, \tfrac{1}{3}] \cup [\tfrac{2}{3}, 1], \\ C_2 &= [0, \tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{1}{3}] \cup [\tfrac{2}{3}, \tfrac{7}{9}] \cup [\tfrac{8}{9}, 1], \\ &\vdots \end{aligned}$$

so that C_k is a collection of 2^k disjoint closed intervals each of length 3^{-k} (see Figure 2.5). We define $C = \cap_{k\in\mathbb{Z}_{>0}} C_k$, which we call the ***middle-thirds Cantor set***.

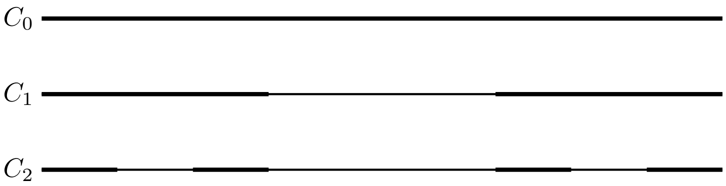


Figure 2.5 The first few sets used in the construction of the middle-thirds Cantor set

Let us give some of the properties of C.

1 **Lemma** *C has the same cardinality as* $[0,1]$.

Proof Note that each of the sets C_k, $k \in \mathbb{Z}_{\geq 0}$, is a collection of disjoint closed intervals. Let us write $C_k = \cup_{j=1}^{2^k} I_{k,j}$, supposing that the intervals $I_{k,j}$ are enumerated such that the right endpoint of $I_{k,j}$ lies to the left of the left endpoint of $I_{k,j+1}$ for each $k \in \mathbb{Z}_{\geq 0}$ and $j \in \{1,\dots,2^k\}$. Now note that each interval $I_{k+1,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1,\dots,2^{k+1}\}$ comes from assigning two intervals to each of the intervals $I_{k,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1,\dots,2^k\}$. Assign to an interval $I_{k+1,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1,\dots,2^k\}$, the number 0 (resp. 1) if it the left (resp. right) interval coming from an interval $I_{k,j'}$ of C_k. In this way, each interval in C_k, $k \in \mathbb{Z}_{\geq 0}$, is assigned a 0 or a 1 in a unique manner. Since, for each point in $x \in C$, there is exactly one $j \in \{1,\dots,2^k\}$ such that $x \in I_{k,j}$. Therefore, for each point in C there is a unique decimal expansion $0.n_1n_2n_3\dots$ where $n_k \in \{0,1\}$. Moreover, for every such decimal expansion, there is a corresponding point in C. However, such decimal expansions are exactly binary decimal expansions for points in $[0,1]$. In other words, there is a bijection from C to $[0,1]$. ▼

2 **Lemma** *C is a set of measure zero.*

Proof Let $\epsilon \in \mathbb{R}_{>0}$. Note that each of the sets C_k can be covered by a finite number of closed intervals whose lengths sum to $\left(\frac{2}{3}\right)^k$. Therefore, each of the sets C_k can be covered by open intervals whose lengths sum to $\left(\frac{2}{3}\right)^k + \frac{\epsilon}{2}$. Choosing k sufficiently large that $\left(\frac{2}{3}\right)^k < \frac{\epsilon}{2}$ we see that C is contained in the union of a finite collection of open intervals whose lengths sum to ϵ. Since ϵ is arbitrary, it follows that C has measure zero. ▼

This example thus shows that sets of measure zero, while "small" in some sense, can be "large" in terms of the number of elements they possess. Indeed, in terms of cardinality, C has the same size as $[0,1]$, although their measures differ by as much as possible. •

2.5.7 Cantor sets

The remainder of this section is devoted to a characterisation of certain sorts of exotic sets, perhaps the simplest example of which is the middle-thirds Cantor set of Example 2.5.39. This material is only used occasionally, and so can be omitted until the reader feels they need/want to understand it.

The qualifier "middle-thirds" in Example 2.5.39 makes one believe that there might be a general notion of a "Cantor set." This is indeed the case.

2.5.40 **Definition (Cantor set)** Let $I \subseteq \mathbb{R}$ be a closed interval. A subset $A \subseteq I$ is a ***Cantor set*** if

(i) A is closed,

(ii) $\operatorname{int}(A) = \emptyset$, and

(iii) every point of A is an accumulation point of A. •

We leave it to the reader to verify in Exercise 2.5.12 that the middle-thirds Cantor set is a Cantor set, according to the previous definition.

One might wonder whether all Cantor sets have the properties of having the cardinality of an interval and of having measure zero. To address this, we give a result and an example. The result shows that all Cantor sets are uncountable.

2.5.41 Proposition (Cantor sets are uncountable) *If* $A \subseteq \mathbb{R}$ *is a nonempty set having the property that each of its points is an accumulation point, then* A *is uncountable. In particular, Cantor sets are uncountable.*

Proof Any finite set has no accumulation points by Proposition 2.5.13. Therefore A must be either enumerable or uncountable. Suppose that A is enumerable and write $A = (x_j)_{j\in\mathbb{Z}_{>0}}$. Let $y_1 \in A \setminus \{x_1\}$. For $r_1 < |x_1 - y_1|$ we have $x_1 \notin \overline{\mathsf{B}}(r_1, y_1)$. We note that y_1 is an accumulation point for $A \setminus \{x_1, x_2\}$; this follows immediately from Proposition 2.5.13. Thus there exists $y_2 \in A \setminus \{x_1, x_2\}$ such that $y_2 \in \mathsf{B}(r_1, y_1)$ and such that $y_2 \neq y_1$. If $r_2 < \min\{|x_2 - y_2|, r_1 - |y_2 - y_2|\}$ then $x_2 \notin \overline{\mathsf{B}}(r_2, y_2)$ and $\overline{\mathsf{B}}(r_2, y_2) \subseteq \mathsf{B}(r_1, y_1)$ by a simple application of the triangle inequality. Continuing in this way we define a sequence $(\overline{\mathsf{B}}(r_j, y_j))_{j\in\mathbb{Z}_{>0}}$ of closed balls having the following properties:

1. $\overline{\mathsf{B}}(r_{j+1}, y_{j+1}) \subseteq \overline{\mathsf{B}}(r_j, y_j)$ for each $j \in \mathbb{Z}_{>0}$;
2. $x_j \notin \overline{\mathsf{B}}(r_j, y_j)$ for each $j \in \mathbb{Z}_{>0}$.

Note that $(\overline{\mathsf{B}}(r_j, y_j) \cap A)_{j\in\mathbb{Z}_{>0}}$ is a nested sequence of compact subsets of A, and so by Proposition 2.5.31, $\cap_{j\in\mathbb{Z}_{>0}}(\overline{\mathsf{B}}(r_j, y_j) \cap A)$ is a nonempty subset of A. However, for any $j \in \mathbb{Z}_{>0}$, $x_j \notin \cap_{j\in\mathbb{Z}_{>0}}(\overline{\mathsf{B}}(r_j, y_j) \cap A)$, and so we arrive, by contradiction, to the conclusion that A is not enumerable. ■

The following example shows that Cantor sets may not have measure zero.

2.5.42 Example (A Cantor set not having zero measure) We will define a subset of $[0, 1]$ that is a Cantor set, but does not have measure zero. The construction mirrors closely that of Example 2.5.39.

We let $\epsilon \in (0, 1)$. Let $C_{\epsilon,0} = [0, 1]$ and define $C_{\epsilon,1}$ by deleting from $C_{\epsilon,0}$ an open interval of length $\frac{\epsilon}{2}$ centered at the midpoint of $C_{\epsilon,0}$. Note that $C_{\epsilon,1}$ consists of two disjoint closed intervals whose lengths sum to $1 - \frac{\epsilon}{2}$. Next define $C_{\epsilon,2}$ by deleting from $C_{\epsilon,1}$ two open intervals, each of length $\frac{\epsilon}{8}$, centered at the midpoints of each of the intervals comprising $C_{\epsilon,1}$. Note that $C_{\epsilon,2}$ consists of four disjoint closed intervals whose lengths sum to $1 - \frac{\epsilon}{4}$. Proceed in this way, defining a sequence of sets $(C_{\epsilon,k})_{k\in\mathbb{Z}_{>0}}$, where $C_{\epsilon,k}$ consists of 2^k disjoint closed intervals whose lengths sum to $1 - \sum_{j=1}^{k} \frac{\epsilon}{2^j} = 1 - \epsilon$. Take $C_\epsilon = \cap_{k\in\mathbb{Z}_{>0}} C_{\epsilon,k}$.

Let us give the properties of C_ϵ in a series of lemmata.

1 Lemma C_ϵ *is a Cantor set.*

Proof That C_ϵ is closed follows from Exercise 2.5.1 and the fact that it is the intersection of a collection of closed sets. To see that $\operatorname{int}(C_\epsilon) = \emptyset$, let $I \subseteq [0, 1]$ be an open interval and suppose that $I \subseteq C_\epsilon$. This means that $I \subseteq C_{\epsilon,k}$ for each $k \in \mathbb{Z}_{>0}$. Note that the sets $C_{\epsilon,k}$, $k \in \mathbb{Z}_{>0}$, are unions of closed intervals, and that for any $\delta \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that the lengths of the intervals comprising $C_{\epsilon,k}$ are less than δ for $k \geq N$. Thus the length of I must be zero, and so $I = \emptyset$. Thus C_ϵ contains no nonempty open intervals, and so must have an empty interior. To see that every point of C_ϵ is an accumulation point of C_ϵ, we note that all points in C_ϵ are endpoints for one of the closed intervals comprising $C_{\epsilon,k}$ for some $k \in \mathbb{Z}_{>0}$. Moreover, it is clear that every neighbourhood of a point in C_ϵ must contain another endpoint from one of the closed intervals comprising $C_{\epsilon,k}$ for some $k \in \mathbb{Z}_{>0}$. Indeed, were this not the case, this would imply the existence of a nonempty open interval contained in C_ϵ, and we have seen that there can be no such interval. ▼

2 Lemma *C_ϵ is uncountable.*

Proof This can be proved in exactly the same manner as the middle-thirds Cantor set was shown to be uncountable. ▼

3 Lemma *C_ϵ does not have measure zero.*

Proof Once one knows the basic properties of Lebesgue measure, it follows immediately that C_ϵ has, in fact, measure $1 - \epsilon$. However, since we have not yet defined measure, let us prove that C_ϵ does not have measure zero, using only the definition of a set of measure zero. Let $((a_j, b_j))_{j\in\mathbb{Z}_{>0}}$ be a countable collection of open intervals having the property that

$$C_\epsilon \subseteq \bigcup_{j\in\mathbb{Z}_{>0}} (a_j, b_j).$$

Since C_ϵ is closed, it is compact by Corollary 2.5.28. Therefore, there exists a finite collection $((a_{j_l}, b_{j_l}))_{l\in\{1,\dots,m\}}$ of intervals having the property that

$$C_\epsilon \subseteq \bigcup_{l=1}^{m} (a_{j_l}, b_{j_l}). \tag{2.10}$$

We claim that there exists $k \in \mathbb{Z}_{>0}$ such that

$$C_{\epsilon,k} \subseteq \bigcup_{l=1}^{m} (a_{j_l}, b_{j_l}). \tag{2.11}$$

Indeed, suppose that, for each $k \in \mathbb{Z}_{>0}$ there exists $x_k \in C_{\epsilon,k}$ such that $x_k \notin \cup_{l=1}^{m}(a_{j_l}, b_{j_l})$. The sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ is then a sequence in the compact set $C_{\epsilon,1}$, and so by the Bolzano–Weierstrass Theorem, possesses a subsequence $(x_{k_r})_{r\in\mathbb{Z}_{>0}}$ converging to $x \in C_{\epsilon,1}$. But the sequence $(x_{k_{r+1}})_{r\in\mathbb{Z}_{>0}}$ is then a convergent sequence in $C_{\epsilon,2}$, so $x \in C_{\epsilon,2}$. Continuing in this way, $x \in \cap_{k\in\mathbb{Z}_{>0}} C_{\epsilon,k}$. Moreover, the sequence $(x_k)_{k\in\mathbb{Z}_{>0}}$ is also a sequence in the closed set $[0,1] - \cup_{l=1}^{m}(a_{j_l}, b_{j_l})$, and so we conclude that $x \in [0,1] - \cup_{l=1}^{m}(a_{j_l}, b_{j_l})$. Thus we contradict the condition (2.10), and so there indeed must be a $k \in \mathbb{Z}_{>0}$ such that (2.11) holds. However, this implies that any collection of open intervals covering C_ϵ must have lengths which sum to at least $1 - \epsilon$. Thus C_ϵ cannot have measure zero. ▼

Cantor sets such as C_ϵ are sometimes called ***fat Cantor sets***, reflecting the fact that they do not have measure zero. Note, however, that they are not *that* fat, since they have an empty interior! •

2.5.8 Notes

Some uses of δ-fine tagged partitions in real analysis can be found in the paper of Gordon [1998].

Exercises

2.5.1 For an arbitrary collection $(U_a)_{a\in A}$ of open sets and an arbitrary collection $(C_b)_{b\in B}$ of closed sets, do the following:

(a) show that $\cup_{a\in A} U_a$ is open;

(b) show that $\cap_{b\in B} C_b$ is closed;

For open sets U_1 and U_2 and closed sets C_1 and C_2, do the following:

(c) show that $U_1 \cap U_2$ is open;

(d) show that $C_1 \cup C_2$ is closed.

2.5.2 Show that a set $A \subseteq \mathbb{R}$ is closed if and only if it contains all of its limit points.

2.5.3 Let $A \subseteq \mathbb{R}$. A point $x \in A$ is an ***isolated point*** if there exists a neighbourhood U of x such that $A \cap U = \{x\}$. Answer the following questions.

(a) Show that the set of isolated points of A is closed.

(b) Show that $\mathrm{cl}(A)$ is the disjoint union of $\mathrm{der}(A)$ and the set of accumulation points.

2.5.4 Answer the following questions.

(a) Give an example of a subset A of $\mathbb{R}$ for which $\mathrm{der}(\mathrm{der}(A)) = \emptyset$ and $\mathrm{der}(A) \neq \emptyset$.

(b) Show that, if A is a nonempty subset of $\mathbb{R}^d$, then $\mathrm{der}(\mathrm{der}(A)) \subseteq \mathrm{der}(A)$.

(c) Give an example of a subset A of $\mathbb{R}$ for which $\mathrm{der}(\mathrm{der}(A)) = \mathrm{der}(A)$.

(d) Give an example of a subset A of $\mathbb{R}$ for which $\emptyset \neq \mathrm{der}(\mathrm{der}(A)) \neq \mathrm{der}(A)$.

2.5.5 For $A \subseteq \mathbb{R}$, show that $\mathrm{bd}(A) = \mathrm{bd}(\mathbb{R} \setminus A)$.

2.5.6 Find counterexamples to the following statements (cf. Propositions 2.5.15, 2.5.19, and 2.5.20):

(a) $\mathrm{int}(A \cup B) \subseteq \mathrm{int}(A) \cup \mathrm{int}(B)$;

(b) $\mathrm{int}(\cup_{i\in I} A_i) \subseteq \cup_{i\in I} \mathrm{int}(A_i)$;

(c) $\mathrm{int}(\cap_{i\in I} A_i) \supseteq \cap_{i\in I} \mathrm{int}(A_i)$;

(d) $\mathrm{cl}(A \cap B) \supseteq \mathrm{cl}(A) \cap \mathrm{cl}(B)$;

(e) $\mathrm{cl}(\cup_{i\in I}A_i) \subseteq \cup_{i\in I}\mathrm{cl}(A_i)$;

(f) $\mathrm{cl}(\cap_{i\in I}A_i) \supseteq \cap_{i\in I}\mathrm{cl}(A_i)$.

Hint: *No fancy sets are required. Intervals will suffice in all cases.*

2.5.7 For each of the following statements, prove the statement if it is true, and give a counterexample if it is not:

(a) $\mathrm{int}(A_1 \cup A_2) = \mathrm{int}(A_1) \cup \mathrm{int}(A_2)$;

(b) $\mathrm{int}(A_1 \cap A_2) = \mathrm{int}(A_1) \cap \mathrm{int}(A_2)$;

(c) $\mathrm{cl}(A_1 \cup A_2) = \mathrm{cl}(A_1) \cup \mathrm{cl}(A_2)$;

(d) $\mathrm{cl}(A_1 \cap A_2) = \mathrm{cl}(A_1) \cap \mathrm{cl}(A_2)$;

(e) $\mathrm{bd}(A_1 \cup A_2) = \mathrm{bd}(A_1) \cup \mathrm{bd}(A_2)$;

(f) $\mathrm{bd}(A_1 \cap A_2) = \mathrm{bd}(A_1) \cap \mathrm{bd}(A_2)$.

2.5.8 Do the following:

(a) show that any finite subset of $\mathbb{R}$ is discrete;

(b) show that a discrete bounded set is finite;

(c) find a set $A \subseteq \mathbb{R}$ that is countable and has no accumulation points, but that is not discrete.

2.5.9 Show that if $A \subseteq \mathbb{R}$ has measure zero and if $B \subseteq A$, then B has measure zero.

2.5.10 Show that any countable subset of $\mathbb{R}$ has measure zero.

2.5.11 Let $(Z_j)_{j\in\mathbb{Z}_{>0}}$ be a family of subsets of $\mathbb{R}$ that each have measure zero. Show that $\cup_{j\in\mathbb{Z}_{>0}}Z_j$ also has measure zero.

2.5.12 Show that the set C constructed in Example 2.5.39 is a Cantor set.

Chapter 3

Functions of a single real variable

In the preceding chapter we endowed the set $\mathbb{R}$ with a great deal of structure. In this chapter we employ this structure to endow functions whose domain and range is $\mathbb{R}$ with some useful properties. These properties include the usual notions of continuity and differentiability given in first-year courses on calculus. The theory of the Riemann integral is also covered here, and it can be expected that students will have at least a functional familiarity with this. However, students who have had the standard engineering course (at least in North American universities) dealing with these topics will find the treatment here a little different than what they are used to. Moreover, there are also topics covered that are simply not part of the standard undergraduate curriculum, but which still fit under the umbrella of "functions of a real variable." These include a detailed discussion of functions of bounded variation, an introductory treatment of absolutely continuous functions, and a generalisation of the Riemann integral called the Riemann–Stieltjes integral.

Do I need to read this chapter? For readers having had a good course in analysis, this chapter can easily be bypassed completely. It can be expected that all other readers will have some familiarity with the material in this chapter, although not perhaps with the level of mathematical rigour we undertake. This level of mathematical rigour is not necessarily needed, if all one wishes to do is deal with $\mathbb{R}$-valued functions defined on $\mathbb{R}$ (as is done in most engineering undergraduate programs). However, we will wish to use the ideas introduced in this chapter, particularly those from Section 3.1, in contexts far more general than the simple one of $\mathbb{R}$-valued functions. Therefore, it will be helpful, at least, to understand the simple material in this chapter in the rigorous manner in which it is presented.

As for the more advanced material, such as is contained in Sections 3.3 and 3.5, it is probably best left aside on a first reading. The reader will be warned when this material is needed in the presentation.

Some of what we cover in this chapter, particularly notions of continuity, differentiability, and Riemann integrability, will be covered in more generality in Chapter II-1. Aggressive readers may want to skip this material here and proceed directly to the more general case. •

Contents

Section 3.1

Continuous $\mathbb{R}$-valued functions on $\mathbb{R}$

The notion of continuity is one of the most important in all of mathematics. Here we present this important idea in its simplest form: continuity for functions whose domain and range are subsets of $\mathbb{R}$.

Do I need to read this section? Unless you are familiar with this material, it is probably a good idea to read this section fairly carefully. It builds on the structure of $\mathbb{R}$ built up in Chapter 2 and uses this structure in an essential way. It is essential to understand this if one is to understand the more general ideas of continuity that will arise in Chapter III-1. This section also provides an opportunity to improve one's facility with the $\epsilon - \delta$ formalism. •

3.1.1 Definition and properties of continuous functions

In this section we will deal with functions defined on an interval $I \subseteq \mathbb{R}$. This interval might be open, closed, or neither, and bounded, unbounded, or neither. In this section, we shall reserve the letter I to denote such a general interval. It will also be convenient to say that a subset $A \subseteq I$ is ***open*** if $A = U \cap I$ for an open subset U of $\mathbb{R}$.[1] For example, if $I = [0, 1]$, then the subset $[0, \frac{1}{2})$ is an open subset of I, but not an open subset of $\mathbb{R}$. We will be careful to explicitly say that a subset is open *in* I if this is what we mean. *There is a chance for confusion here, so the reader is advised to be alert!*

Let us give the standard definition of continuity.

3.1.1 Definition (Continuous function) Let $I \subseteq \mathbb{R}$ be an interval. A map $f: I \to \mathbb{R}$ is:

(i) ***continuous at*** $\mathbf{x_0 \in I}$ if, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ whenever $x \in I$ satisfies $|x - x_0| < \delta$;

(ii) ***continuous*** if it is continuous at each $x_0 \in I$;

(iii) ***discontinuous at*** $\mathbf{x_0 \in I}$ if it is not continuous at x_0;

(iv) ***discontinuous*** if it is not continuous. •

The idea behind the definition of continuity is this: one can make the values of a continuous function as close as desired by making the points at which the function is evaluated sufficiently close. Readers not familiar with the definition should be prepared to spend some time embracing it. An often encountered oversimplification of continuity is illustrated in Figure 3.1. The idea is supposed to be that the function whose graph is shown on the left is continuous because its

[1] This is entirely related to the notion of relative topology which we will discuss in Section II-1.2.8 for sets of multiple real variables and in Definition III-1.4.1 within the general context of topological spaces.

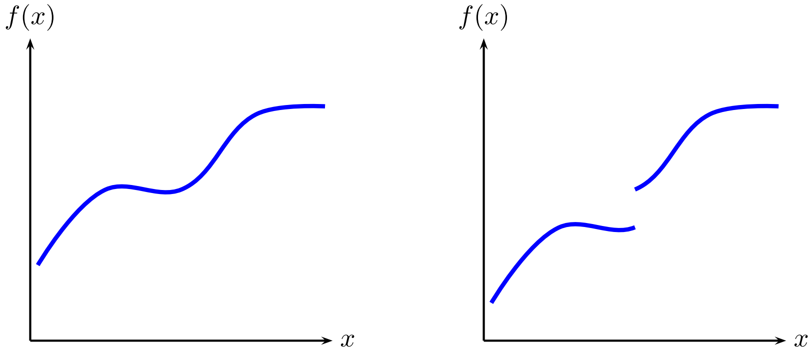


Figure 3.1 Probably not always the best way to envision continuity versus discontinuity

graph has no "gaps," whereas the function on the right is discontinuous because its graph does have a "gap." As we shall see in Example 3.1.2–4 below, it is possible for a function continuous at a point to have a graph with lots of "gaps" in a neighbourhood of that point. Thus the "graph gap" characterisation of continuity is a little misleading.

Let us give some examples of functions that are continuous or not. More examples of discontinuous functions are given in Example 3.1.9 below. We suppose the reader to be familiar with the usual collection of "standard functions," at least for the moment. We shall consider some such functions in detail in Section 3.8.

3.1.2 Examples (Continuous and discontinuous functions)

1. For $\alpha \in \mathbb{R}$, define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = \alpha$. Since $|f(x) - f(x_0)| = 0$ for all $x, x_0 \in \mathbb{R}$, it follows immediately that f is continuous.
2. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = x$. For $x_0 \in \mathbb{R}$ and $\epsilon \in \mathbb{R}_{>0}$ take $\delta = \epsilon$. It then follows that if $|x - x_0| < \delta$ then $|f(x) - f(x_0)| < \epsilon$, giving continuity of f.
3. Define $f\colon \mathbb{R} \to \mathbb{R}$ by
$$f(x) = \begin{cases} x \sin\frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$
We claim that f is continuous. We first note that the functions $f_1, f_2\colon \mathbb{R} \to \mathbb{R}$ defined by
$$f_1(x) = x, \quad f_2(x) = \sin x$$
are continuous. Indeed, f_1 is continuous from part 2 and in Section 3.8 we will prove that f_2 is continuous. The function $f_3\colon \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by $f_3(x) = \frac{1}{x}$ is continuous on any interval not containing 0 by Proposition 3.1.15 below. It then follows from Propositions 3.1.15 and 3.1.16 below that f is continuous at x_0, provided that $x_0 \neq 0$. To show continuity at $x = 0$, let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \epsilon$. Then, provided that $|x| < \delta$,
$$|f(x) - f(0)| = \left|x \sin\tfrac{1}{x}\right| \leq |x| < \epsilon,$$

using the fact that image(sin) $\subseteq [-1,1]$. This shows that f is continuous at 0, and so is continuous.

4. Define $f\colon \mathbb{R} \to \mathbb{R}$ by
$$f(x) = \begin{cases} x, & x \in \mathbb{Q}, \\ 0, & \text{otherwise}. \end{cases}$$
We claim that f is continuous at $x_0 = 0$ and discontinuous everywhere else.
To see that f is continuous at $x_0 = 0$, let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta = \epsilon$. Then, for $|x - x_0| < \delta$ we have either $f(x) = x$ or $f(x) = 0$. In either case, $|f(x) - f(x_0)| < \epsilon$, showing that f is indeed continuous at $x_0 = 0$. Note that this is a function whose continuity at $x_0 = 0$ is not subject to an interpretation like that of Figure 3.1 since the graph of f has an uncountable number of "gaps" near 0.
Next we show that f is discontinuous at x_0 for $x_0 \neq 0$. We have two possibilities.
   (a) $x_0 \in \mathbb{Q}$: Let $\epsilon < \frac{1}{2}|x_0|$. For any $\delta \in \mathbb{R}_{>0}$ the set $\mathsf{B}(\delta, x_0)$ will contain points $x \in \mathbb{R}$ for which $f(x) = 0$. Thus for any $\delta \in \mathbb{R}_{>0}$ the set $\mathsf{B}(\delta, x_0)$ will contain points x such that $|f(x) - f(x_0)| = |x_0| > \epsilon$. This shows that f is discontinuous at nonzero rational numbers.
   (b) $x_0 \in \mathbb{R} \setminus \mathbb{Q}$: Let $\epsilon = \frac{1}{2}|x_0|$. For any $\delta \in \mathbb{R}_{>0}$ we claim that the set $\mathsf{B}(\delta, x_0)$ will contain points $x \in \mathbb{R}$ for which $|f(x)| > \epsilon$ (why?). It then follows that for any $\delta \in \mathbb{R}_{>0}$ the set $\mathsf{B}(\delta, x_0)$ will contain points x such that $|f(x) - f(x_0)| = |f(x)| > \epsilon$, so showing that f is discontinuous at all irrational numbers.

5. Let $I = (0, \infty)$ and on I define the function $f\colon I \to \mathbb{R}$ by $f(x) = \frac{1}{x}$. It follows from Proposition 3.1.15 below that f is continuous on I.

6. Next take $I = [0, \infty)$ and define $f\colon I \to \mathbb{R}$ by
$$f(x) = \begin{cases} \frac{1}{x}, & x \in \mathbb{R}_{>0}, \\ 0, & x = 0. \end{cases}$$
In the previous example we saw that f is continuous at all points in $(0, \infty)$. However, at $x = 0$ the function is discontinuous, as is easily verified. •

The following alternative characterisations of continuity are sometimes useful. The first of these, part (ii) in the theorem, will also be helpful in motivating the general definition of continuity given for topological spaces in Section III-1.3. The reader will wish to recall from Notation 2.3.28 the notation $\lim_{x \to_I x_0} f(x)$ for taking limits in intervals.

3.1.3 Theorem (Alternative characterisations of continuity) *For a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *defined on an interval* I *and for* $\mathrm{x}_0 \in \mathrm{I}$, *the following statements are equivalent:*

(i) f *is continuous at* x_0*;*

(ii) for every neighbourhood V *of* $\mathrm{f}(\mathrm{x}_0)$ *there exists a neighbourhood* U *of* x_0 *in* I *such that* $\mathrm{f}(\mathrm{U}) \subseteq \mathrm{V}$*;*

(iii) $\lim_{x\to_I x_0} f(x) = f(x_0)$.

Proof (i) $\Longrightarrow$ (ii) Let $V \subseteq \mathbb{R}$ be a neighbourhood of $f(x_0)$. Let $\epsilon \in \mathbb{R}_{>0}$ be defined such that $\mathsf{B}(\epsilon, f(x_0)) \subseteq V$, this being possible since V is open. Since f is continuous at x_0, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $x \in \mathsf{B}(\delta, x_0) \cap I$, then we have $f(x) \in \mathsf{B}(\epsilon, f(x_0))$. This shows that, around the point x_0, we can find an open set in I whose image lies in V.

(ii) $\Longrightarrow$ (iii) Let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in I converging to x_0 and let $\epsilon \in \mathbb{R}_{>0}$. By hypothesis there exists a neighbourhood U of x_0 in I such that $f(U) \subseteq \mathsf{B}(\epsilon, f(x_0))$. Thus there exists $\delta \in \mathbb{R}_{>0}$ such that $f(\mathsf{B}(\delta, x_0) \cap I) \subseteq \mathsf{B}(\epsilon, f(x_0))$ since U is open in I. Now choose $N \in \mathbb{Z}_{>0}$ sufficiently large that $|x_j - x_0| < \delta$ for $j \geq N$. It then follows that $|f(x_j) - f(x_0)| < \epsilon$ for $j \geq N$, so giving convergence of $(f(x_j))_{j\in\mathbb{Z}_{>0}}$ to $f(x_0)$, as desired, after an application of Proposition 2.3.29.

(iii) $\Longrightarrow$ (i) Let $\epsilon \in \mathbb{R}_{>0}$. Then, by definition of $\lim_{x\to_I x_0} f(x) = f(x_0)$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{B}(\delta, x_0) \cap I$, $|f(x) - f(x_0)| < \epsilon$, which is exactly the definition of continuity of f at x_0. ■

3.1.4 Corollary *For an interval* $I \subseteq \mathbb{R}$*, a function* $f\colon I \to \mathbb{R}$ *is continuous if and only if* $f^{-1}(V)$ *is open in* I *for every open subset* V *of* $\mathbb{R}$*.*

Proof Suppose that f is continuous. If $V \cap \mathrm{image}(f) = \emptyset$ then clearly $f^{-1}(V) = \emptyset$ which is open. So assume that $V \cap \mathrm{image}(f) \neq \emptyset$ and let $x \in f^{-1}(V)$. Since f is continuous at x and since V is a neighbourhood of $f(x)$, there exists a neighbourhood U of x such that $f(U) \subseteq V$. Thus $U \subseteq f^{-1}(V)$, showing that $f^{-1}(V)$ is open.

Now suppose that $f^{-1}(V)$ is open for each open set V and let $x \in \mathbb{R}$. If V is a neighbourhood of $f(x)$ then $f^{-1}(V)$ is open. Then there exists a neighbourhood U of x such that $U \subseteq f^{-1}(V)$. By Proposition 1.3.5 we have $f(U) \subseteq f(f^{-1}(V)) \subseteq V$, thus showing that f is continuous. ■

The reader can explore these alternative representations of continuity in Exercise 3.1.9.

A stronger notion of continuity is sometimes useful. As well, the following definition introduces for the first time the important notion of "uniform."

3.1.5 Definition (Uniform continuity) Let $I \subseteq \mathbb{R}$ be an interval. A map $f\colon I \to \mathbb{R}$ is ***uniformly continuous*** if, for every $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x_1) - f(x_2)| < \epsilon$ whenever $x_1, x_2 \in I$ satisfy $|x_1 - x_2| < \delta$. •

3.1.6 Remark (On the idea of "uniformly") In the preceding definition we have encountered for the first time the idea of a property holding "uniformly." This is an important idea that comes up often in mathematics. Moreover, it is an idea that is often useful in applications of mathematics, since the absence of a property holding "uniformly" can have undesirable consequences. Therefore, we shall say some things about this here.

In fact, the comparison of continuity versus uniform continuity is a good one for making clear the character of something holding "uniformly." Let us compare the definitions.

1. One defines continuity of a function at a point x_0 by asking that, for each $\epsilon \in \mathbb{R}_{>0}$, one can find $\delta \in \mathbb{R}_{>0}$ such that if x is within δ of x_0, then $f(x)$ is within ϵ of $f(x_0)$. Note that δ will generally depend on ϵ, and most importantly for our discussion here, on x_0. Often authors explicitly write $\delta(\epsilon, x_0)$ to denote this dependence of δ on both ϵ and x_0.
2. One defines uniform continuity of a function on the interval I by asking that, for each $\epsilon \in \mathbb{R}_{>0}$, one can find $\delta \in \mathbb{R}_{>0}$ such that if x_1 and x_2 are within δ of one another, then $f(x_1)$ and $f(x_2)$ are within ϵ of one another. Here, the number δ depends *only* on ϵ. Again, to reflect this, some authors explicitly write $\delta(\epsilon)$, or state explicitly that δ is independent of x.

The idea of "uniform" then is that a property, in this case the existence of $\delta \in \mathbb{R}_{>0}$ with a certain property, holds for the entire set I, and not just for a single point. •

Let us give an example to show that uniformly continuous is not the same as continuous.

3.1.7 Example (Uniform continuity versus continuity) Let us give an example of a function that is continuous, but not uniformly continuous. Define $f\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = x^2$. We first show that f is continuous at each point $x_0 \in \mathbb{R}$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose δ such that $2|x_0|\delta + \delta^2 < \epsilon$ (why is this possible?). Then, provided that $|x - x_0| < \delta$, we have

$$\begin{aligned}|f(x) - f(x_0)| &= |x^2 - x_0^2| = |x - x_0||x + x_0|\\ &\le |x - x_0|(|x| + |x_0|) \le |x - x_0|(2|x_0| + |x - x_0|)\\ &\le \delta(2|x_0| + \delta) < \epsilon.\end{aligned}$$

Thus f is continuous.

Now let us show that f is not uniformly continuous. We will show that there exists $\epsilon \in \mathbb{R}_{>0}$ such that there is no $\delta \in \mathbb{R}_{>0}$ for which $|x - x_0| < \delta$ ensures that $|f(x) - f(x_0)| < \epsilon$ *for all* x_0. Let us take $\epsilon = 1$ and let $\delta \in \mathbb{R}_{>0}$. Then define $x_0 \in \mathbb{R}$ such that $\frac{\delta}{2}\left|2x_0 + \frac{\delta}{2}\right| > 1$ (why is this possible?). We then note that $x = x_0 + \frac{\delta}{2}$ satisfies $|x - x_0| < \delta$, but that

$$|f(x) - f(x_0)| = |x^2 - x_0^2| = |x - x_0||x + x_0| = \tfrac{\delta}{2}\left|2x_0 + \tfrac{\delta}{2}\right| > 1 = \epsilon.$$

This shows that f is not uniformly continuous. •

3.1.2 Discontinuous functions[2]

It is often useful to be specific about the nature of a discontinuity of a function that is not continuous. The following definition gives names to all possibilities. The reader may wish to recall from Section 2.3.7 the discussion concerning taking limits using an index set that is a subset of $\mathbb{R}$.

[2] This section is rather specialised and technical and so can be omitted until needed. However, the material is needed at certain points in the text.

3.1.8 Definition (Types of discontinuity) Let $I \subseteq \mathbb{R}$ be an interval and suppose that $f\colon I \to \mathbb{R}$ is discontinuous at $x_0 \in I$. The point x_0 is:

(i) a ***removable discontinuity*** if $\lim_{x\to_I x_0} f(x)$ exists;

(ii) a ***discontinuity of the first kind***, or a ***jump discontinuity***, if the limits $\lim_{x\downarrow x_0} f(x)$ and $\lim_{x\uparrow x_0} f(x)$ exist;

(iii) a ***discontinuity of the second kind***, or an ***essential discontinuity***, if at least one of the limits $\lim_{x\downarrow x_0} f(x)$ and $\lim_{x\uparrow x_0} f(x)$ does not exist.

The set of all discontinuities of f is denoted by D_f. •

In Figure 3.2 we depict the various sorts of discontinuity. We can also illustrate these with explicit examples.

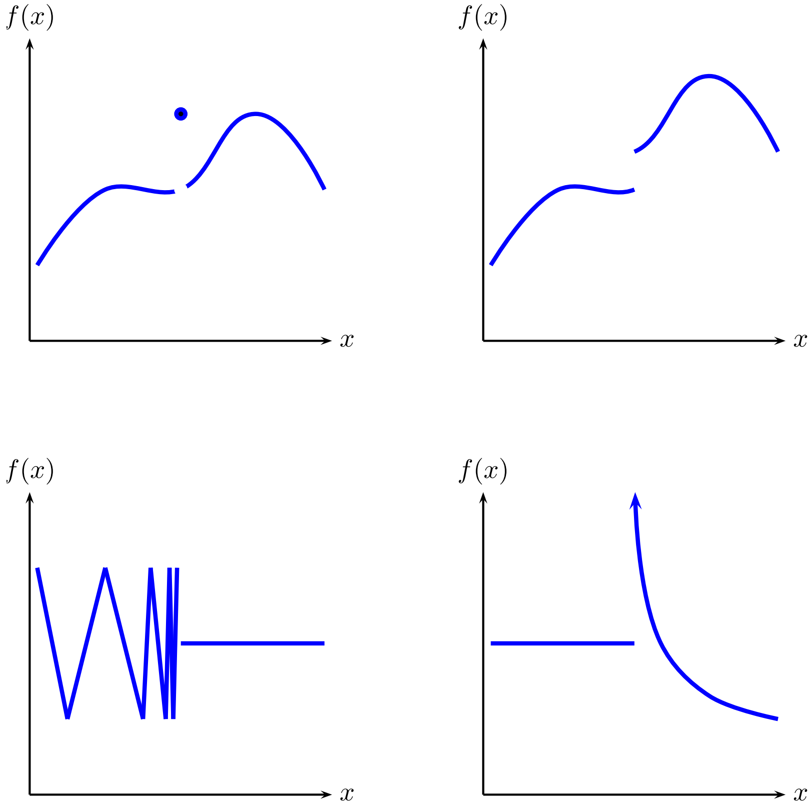


Figure 3.2 A removable discontinuity (top left), a jump discontinuity (top right), and two essential discontinuities (bottom)

3.1.9 Examples (Types of discontinuities)

1. Let $I = [0,1]$ and let $f\colon I \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x, & x \in (0,1], \\ 1, & x = 0. \end{cases}$$

It is clear that f is continuous for all $x \in (0,1]$, and is discontinuous at $x = 0$. However, since we have $\lim_{x \to_I 0} f(x) = 0$ (note that the requirement that this limit be taken in I amounts to the fact that the limit is given by $\lim_{x \downarrow 0} f(x) = 0$), it follows that the discontinuity is removable.

Note that one might be tempted to also say that the discontinuity is a jump discontinuity since the limit $\lim_{x \downarrow 0} f(x)$ exists and since the limit $\lim_{x \uparrow 0} f(x)$ cannot be defined here since 0 is a left endpoint for I. However, we do require that both limits exist at a jump discontinuity, which has as a consequence the fact that jump discontinuities can only occur at interior points of an interval.

2. Let $I = [-1,1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = \operatorname{sign}(x)$. We may easily see that f is continuous at $x \in [-1,1] \setminus \{0\}$, and is discontinuous at $x = 0$. Then, since we have $\lim_{x \downarrow 0} f(x) = 1$ and $\lim_{x \uparrow 0} f(x) = -1$, it follows that the discontinuity at 0 is a jump discontinuity.
3. Let $I = [-1,1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} \sin\frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Then, by Proposition 3.1.15 (and accepting continuity of sin), f is continuous at $x \in [-1,1] \setminus \{0\}$. At $x = 0$ we claim that we have an essential discontinuity. To see this we note that, for any $\epsilon \in \mathbb{R}_{>0}$, the function f restricted to $[0,\epsilon)$ and $(-\epsilon,0]$ takes all possible values in set $[-1,1]$. This is easily seen to preclude existence of the limits $\lim_{x \downarrow 0} f(x)$ and $\lim_{x \uparrow 0} f(x)$.
4. Let $I = [-1,1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} \frac{1}{x}, & x \in (0,1], \\ 0, & x \in [-1,0]. \end{cases}$$

Then f is continuous at $x \in [-1,1] \setminus \{0\}$ by Proposition 3.1.15. At $x = 0$ we claim that f has an essential discontinuity. Indeed, we have $\lim_{x \downarrow} f(x) = \infty$, which precludes f having a removable or jump discontinuity at $x = 0$. •

The following definition gives a useful quantitative means of measuring the discontinuity of a function.

3.1.10 Definition (Oscillation) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function. The ***oscillation*** of f is the function $\omega_f\colon I \to \mathbb{R}$ defined by

$$\omega_f(x) = \inf\{\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} \mid \delta \in \mathbb{R}_{>0}\}. \qquad \bullet$$

Note that the definition makes sense since the function

$$\delta \mapsto \sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\}$$

is monotonically increasing (see Definition 3.1.27 for a definition of monotonically increasing in this context). In particular, if f is bounded (see Definition 3.1.20 below) then ω_f is also bounded. The following result indicates in what way ω_f measures the continuity of f.

3.1.11 Proposition (Oscillation measures discontinuity) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$, f *is continuous at* $\mathrm{x} \in \mathrm{I}$ *if and only if* $\omega_{\mathrm{f}}(\mathrm{x}) = 0$.

Proof Suppose that f is continuous at x and let $\epsilon \in \mathbb{R}_{>0}$. Choose $\delta \in \mathbb{R}_{>0}$ such that if $y \in \mathsf{B}(\delta, x) \cap I$ then $|f(y) - f(x)| < \frac{\epsilon}{2}$. Then, for $x_1, x_2 \in \mathsf{B}(\delta, x)$ we have

$$|f(x_1) - f(x_2)| \leq |f(x_1) - f(x)| + |f(x) - f(x_2)| < \epsilon.$$

Therefore,

$$\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} < \epsilon.$$

Since ϵ is arbitrary this gives

$$\inf\{\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} \mid \delta \in \mathbb{R}_{>0}\} = 0,$$

meaning that $\omega_f(x) = 0$.

Now suppose that $\omega_f(x) = 0$. For $\epsilon \in \mathbb{R}_{>0}$ let $\delta \in \mathbb{R}_{>0}$ be chosen such that

$$\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\} < \epsilon.$$

In particular, $|f(y) - f(x)| < \epsilon$ for all $y \in \mathsf{B}(\delta, x) \cap I$, giving continuity of f at x. ■

Let us consider a simple example.

3.1.12 Example (Oscillation for a discontinuous function) We let $I = [-1, 1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = \mathrm{sign}(x)$. It is then easy to see that

$$\omega_f(x) = \begin{cases} 0, & x \neq 0, \\ 2, & x = 0. \end{cases} \quad \bullet$$

We close this section with a technical property of the oscillation of a function. This property will be useful during the course of some proofs in the text.

3.1.13 Proposition (Closed preimages of the oscillation of a function) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function. Then, for every* $\alpha \in \mathbb{R}_{\geq 0}$, *the set*

$$\mathrm{A}_\alpha = \{\mathrm{x} \in \mathrm{I} \mid \omega_{\mathrm{f}}(\mathrm{x}) \geq \alpha\}$$

is closed in I.

Proof The result where $\alpha = 0$ is clear, so we assume that $\alpha \in \mathbb{R}_{>0}$. For $\delta \in \mathbb{R}_{>0}$ define

$$\omega_f(x, \delta) = \sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\}$$

so that $\omega_f(x) = \lim_{\delta \to 0} \omega_f(x, \delta)$. Let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in A_α converging to $x \in \mathbb{R}$ and let $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $(0, \alpha)$ converging to zero. Let $j \in \mathbb{Z}_{>0}$. We claim that there exists points $y_j, z_j \in \mathsf{B}(\epsilon_j, x_j) \cap I$ such that $|f(y_j) - f(z_j)| \geq \alpha - \epsilon_j$. Suppose otherwise so that for every $y, z \in \mathsf{B}(\epsilon_j, x_j) \cap I$ we have $|f(y) - f(z)| < \alpha - \epsilon_j$. It then follows that $\lim_{\delta \to 0} \omega_f(x_j, \delta) \leq \alpha - \epsilon_j < \alpha$, contradicting the fact that $x_j \in A_\alpha$. We claim that $(y_j)_{j \in \mathbb{Z}_{>0}}$ and $(z_j)_{j \in \mathbb{Z}_{>0}}$ converge to x. Indeed, let $\epsilon \in \mathbb{R}_{>0}$ and choose $N_1 \in \mathbb{Z}_{>0}$

sufficiently large that $\epsilon_j < \frac{\epsilon}{2}$ for $j \geq N_1$ and choose $N_2 \in \mathbb{Z}_{>0}$ such that $|x_j - x| < \frac{\epsilon}{2}$ for $j \geq N_2$. Then, for $j \geq \max\{N_1, N_2\}$ we have

$$|y_j - x| \leq |y_j - x_j| + |x_j - x| < \epsilon.$$

Thus $(y_j)_{j \in \mathbb{Z}_{>0}}$ converges to x, and the same argument, and therefore the same conclusion, also applies to $(z_j)_{j \in \mathbb{Z}_{>0}}$.

Thus we have sequences of points $(y_j)_{j \in \mathbb{Z}_{>0}}$ and $(z_j)_{j \in \mathbb{Z}_{>0}}$ in I converging to x and a sequence $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ in $(0, \alpha)$ converging to zero for which $|f(y_j) - f(z_j)| \geq \alpha - \epsilon_j$. We claim that this implies that $\omega_f(x) \geq \alpha$. Indeed, suppose that $\omega_f(x) < \alpha$. There exists $N \in \mathbb{Z}_{>0}$ such that $\alpha - \epsilon_j > \alpha - \omega_f(x)$ for every $j \geq N$. Therefore,

$$|f(y_j) - f(z_j)| \geq \alpha - \epsilon_j > \alpha - \omega_f(x)$$

for every $j \geq N$. This contradicts the definition of $\omega_f(x)$ since the sequences $(y_j)_{j \in \mathbb{Z}_{>0}}$ and $(z_j)_{j \in \mathbb{Z}_{>0}}$ converge to x.

Now we claim that the sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ converges to x. Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1 \in \mathbb{Z}_{>0}$ be large enough that $|x - y_j| < \frac{\epsilon}{2}$ for $j \geq N_1$ and let $N_2 \in \mathbb{Z}_{>0}$ be large enough that $\epsilon_j < \frac{\epsilon}{2}$ for $j \geq N_2$. Then, for $j \geq \max\{N_1, N_2\}$ we have

$$|x - x_j| \leq |x - y_j| + |y_j - x_j| < \epsilon,$$

as desired.

This shows that every sequence in A_α converges to a point in A_α. It follows from Exercise 2.5.2 that A_α is closed. ■

The following corollary is somewhat remarkable, in that it shows that the set of discontinuities of a function cannot be arbitrary.

3.1.14 Corollary (Discontinuities are the countable union of closed sets) *Let* $I \subseteq \mathbb{R}$ *be an interval and let* $f \colon I \to \mathbb{R}$ *be a function. Then the set*

$$D_f = \{x \in I \mid f \textit{ is not continuous at } x\}$$

is the countable union of closed sets.

Proof This follows immediately from Proposition 3.1.13 after we note that

$$D_f = \cup_{k \in \mathbb{Z}_{>0}} \{x \in I \mid \omega_f(x) \geq \tfrac{1}{k}\}.$$ ■

3.1.3 Continuity and operations on functions

Let us consider how continuity behaves relative to simple operations on functions. To do so, we first note that, given an interval I and two functions $f, g \colon I \to \mathbb{R}$, one can define two functions $f + g, fg \colon I \to \mathbb{R}$ by

$$(f+g)(x) = f(x) + g(x), \qquad (fg)(x) = f(x)g(x),$$

respectively. Moreover, if $g(x) \neq 0$ for all $x \in I$, then we define

$$\left(\frac{f}{g}\right)(x) = \frac{f(x)}{g(x)}.$$

Thus one can add and multiply $\mathbb{R}$-valued functions using the operations of addition and multiplication in $\mathbb{R}$.

3.1.15 Proposition (Continuity, and addition and multiplication) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$, *if* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *are continuous at* $\mathrm{x}_0 \in \mathrm{I}$, *then both* $\mathrm{f} + \mathrm{g}$ *and* fg *are continuous at* x_0. *If additionally* $\mathrm{g}(\mathrm{x}) \neq 0$ *for all* $\mathrm{x} \in \mathrm{I}$, *then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is continuous at* x_0.

Proof To show that $f + g$ and fg are continuous at x_0 if f and g are continuous at x_0, let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in I converging to x_0. Then, by Theorem 3.1.3 the sequences $(f(x_j))_{j\in\mathbb{Z}_{>0}}$ and $(g(x_j))_{j\in\mathbb{Z}_{>0}}$ converge to $f(x_0)$ and $g(x_0)$, respectively. Then, by Proposition 2.3.23, the sequences $(f(x_j) + g(x_j))_{j\in\mathbb{Z}_{>0}}$ and $(f(x_j)g(x_j))_{j\in\mathbb{Z}_{>0}}$ converge to $f(x_0) + g(x_0)$ and $f(x_0)g(x_0)$, respectively. Then $\lim_{j\to\infty}(f + g)(x_j) = (f + g)(x_0)$ and $\lim_{j\to\infty}(fg)(x_j) = (fg)(x_0)$, and the result follows by Proposition 2.3.29 and Theorem 3.1.3.

Now suppose that $g(x) \neq 0$ for every $x \in I$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that $|g(x_0)| > 2\epsilon$. By Theorem 3.1.3 take $\delta \in \mathbb{R}_{>0}$ such that $g(\mathsf{B}(\delta, x_0)) \subseteq \mathsf{B}(\epsilon, g(x_0))$. Thus g is nonzero on the ball $\mathsf{B}(\delta, x_0)$. Now let $(x_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathsf{B}(\delta, x_0)$ converging to x_0. Then, as above, the sequences $(f(x_j))_{j\in\mathbb{Z}_{>0}}$ and $(g(x_j))_{j\in\mathbb{Z}_{>0}}$ converge to $f(x_0)$ and $g(x_0)$, respectively. We can now employ Proposition 2.3.23 to conclude that the sequence $\left(\frac{f(x_j)}{g(x_j)}\right)_{j\in\mathbb{Z}_{>0}}$ converges to $\frac{f(x_0)}{g(x_0)}$, and the last part of the result follows by Proposition 2.3.29 and Theorem 3.1.3. ■

3.1.16 Proposition (Continuity and composition) *Let* $\mathrm{I}, \mathrm{J} \subseteq \mathbb{R}$ *be intervals and let* $\mathrm{f}\colon \mathrm{I} \to \mathrm{J}$ *and* $\mathrm{f}\colon \mathrm{J} \to \mathbb{R}$ *be continuous at* $\mathrm{x}_0 \in \mathrm{I}$ *and* $\mathrm{f}(\mathrm{x}_0) \in \mathrm{J}$, *respectively. Then* $\mathrm{g} \circ \mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is continuous at* x_0.

Proof Let W be a neighbourhood of $g \circ f(x_0)$. Since g is continuous at $f(x_0)$ there exists a neighbourhood V of $f(x_0)$ such that $g(V) \subseteq W$. Since f is continuous at x_0 there exists a neighbourhood U of x_0 such that $f(U) \subseteq V$. Clearly $g \circ f(U) \subseteq W$, and the result follows from Theorem 3.1.3. ■

3.1.17 Proposition (Continuity and restriction) *If* $\mathrm{I}, \mathrm{J} \subseteq \mathbb{R}$ *are intervals for which* $\mathrm{J} \subseteq \mathrm{I}$, *and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is continuous at* $\mathrm{x}_0 \in \mathrm{J} \subseteq \mathrm{I}$, *then* $\mathrm{f}|\mathrm{J}$ *is continuous at* x_0.

Proof This follows immediately from Theorem 3.1.3, also using Proposition 1.3.5, after one notes that open subsets of J are of the form $U \cap I$ where U is an open subset of I. ■

Note that none of the proofs of the preceding results use the definition of continuity, but actually use the alternative characterisations of Theorem 3.1.3. Thus these alternative characterisations, while less intuitive initially (particularly the one involving open sets), they are in fact quite useful.

Let us finally consider the behaviour of continuity with respect to the operations of selection of maximums and minimums.

3.1.18 Proposition (Continuity and min and max) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *are continuous functions, then the functions*

$$\mathrm{I} \ni \mathrm{x} \mapsto \min\{\mathrm{f}(\mathrm{x}), \mathrm{g}(\mathrm{x})\} \in \mathbb{R}, \qquad \mathrm{I} \ni \mathrm{x} \mapsto \max\{\mathrm{f}(\mathrm{x}), \mathrm{g}(\mathrm{x})\} \in \mathbb{R}$$

are continuous.

Proof Let $x_0 \in I$ and let $\epsilon \in \mathbb{R}_{>0}$. Let us first assume that $f(x_0) > g(x_0)$. That is to say, assume that $(f-g)(x_0) \in \mathbb{R}_{>0}$. Continuity of f and g ensures that there exists $\delta_1 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}(\delta_1, x_0) \cap I$ then $(f-g)(x) \in \mathbb{R}_{>0}$. That is, if $x \in \mathsf{B}(\delta_1, x_0) \cap I$ then

$$\min\{f(x), g(x)\} = g(x), \quad \max\{f(x), g(x)\} = f(x).$$

Continuity of f ensures that there exists $\delta_2 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}(\delta_2, x_0) \cap I$ then $|f(x) - f(x_0)| < \epsilon$. Similarly, continuity of f ensures that there exists $\delta_3 \in \mathbb{R}_{>0}$ such that if $x \in \mathsf{B}(\delta_3, x_0) \cap I$ then $|g(x) - g(x_0)| < \epsilon$. Let $\delta_4 = \min\{\delta_1, \delta_2\}$. If $x \in \mathsf{B}(\delta_4, x_0) \cap I$ then

$$|\min\{f(x), g(x)\} - \min\{f(x_0), g(x_0)\}| = |g(x) - g(x_0)| < \epsilon$$

and

$$|\max\{f(x), g(x)\} - \max\{f(x_0), g(x_0)\}| = |f(x) - f(x_0)| < \epsilon.$$

This gives continuity of the two functions in this case. Similarly, swapping the rôle of f and g, if $f(x_0) < g(x_0)$ one can arrive at the same conclusion. Thus we need only consider the case when $f(x_0) = g(x_0)$. In this case, by continuity of f and g, choose $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ and $|g(x) - g(x_0)| < \epsilon$ for $x \in \mathsf{B}(\delta, x_0) \cap I$. Then let $x \in \mathsf{B}(\delta, x_0) \cap I$. If $f(x) \geq g(x)$ then we have

$$|\min\{f(x), g(x)\} - \min\{f(x_0), g(x_0)\}| = |g(x) - g(x_0)| < \epsilon$$

and

$$|\max\{f(x), g(x)\} - \max\{f(x_0), g(x_0)\}| = |f(x) - f(x_0)| < \epsilon.$$

This gives the result in this case, and one similarly gets the result when $f(x) < g(x)$. ■

3.1.4 Continuity, and compactness and connectedness

In this section we will consider some of the relationships that exist between continuity, and compactness and connectedness. We see here for the first time some of the benefits that can be drawn from the notion of continuity. Moreover, if one studies the proofs of the results in this section, one can see that we use the actual definition of compactness (rather than the simpler alternative characterisation of compact sets as being closed and bounded) to great advantage.

The first result is a simple and occasionally useful one.

3.1.19 Proposition (The continuous image of a compact set is compact) *If* $I \subseteq \mathbb{R}$ *is a compact interval and if* $f \colon I \to \mathbb{R}$ *is continuous, then* $\mathrm{image}(f)$ *is compact.*

Proof Let $(U_a)_{a\in A}$ be an open cover of image(f). Then $(f^{-1}(U_a))_{a\in A}$ is an open cover of I, and so there exists a finite subset $(a_1, \dots, a_k) \subseteq A$ such that $\cup_{j=1}^k f^{-1}(U_{a_k}) = I$. It is then clear that $(f(f^{-1}(U_{a_1})), \dots, f(f^{-1}(U_{a_k})))$ covers image(f). Moreover, by Proposition 1.3.5, $f(f^{-1}(U_{a_j})) \subseteq U_{a_j}$, $j \in \{1, \dots, k\}$. Thus $(U_{a_1}, \dots, U_{a_k})$ is a finite subcover of $(U_a)_{a\in A}$. ■

A useful feature that a function might possess is that of having bounded values.

3.1.20 Definition (Bounded function) For an interval I, a function $f\colon I \to \mathbb{R}$ is:

(i) ***bounded*** if there exists $M \in \mathbb{R}_{>0}$ such that $\operatorname{image}(f) \subseteq \overline{\mathsf{B}}(M, 0)$;

(ii) ***locally bounded*** if $f|J$ is bounded for every compact interval $J \subseteq I$;

(iii) ***unbounded*** if it is not bounded. •

3.1.21 Remark (On "locally") This is our first encounter with the qualifier "locally" assigned to a property, in this case, of a function. This concept will appear frequently, as for example in this chapter with the notion of "locally bounded variation" (Definition 3.3.6) and "locally absolutely continuous" (Definition III-2.9.23). The idea in all cases is the same; that a property holds "locally" if it holds on every compact subset. •

For continuous functions it is sometimes possible to immediately assert boundedness simply from the property of the domain.

3.1.22 Theorem (Continuous functions on compact intervals are bounded) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval, then a continuous function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is bounded.*

Proof Let $x \in I$. As f is continuous, there exists $\delta \in \mathbb{R}_{>0}$ so that $|f(y) - f(x)| < 1$ provided that $|y - x| < \delta$. In particular, if $x \in I$, there is an open interval I_x in I with $x \in I_x$ such that $|f(y)| \leq |f(x)| + 1$ for all $x \in I_x$. Thus f is bounded on I_x. This can be done for each $x \in I$, so defining a family of open sets $(I_x)_{x \in I}$. Clearly $I \subseteq \cup_{x \in I} I_x$, and so, by Theorem 2.5.27, there exists a finite collection of points $x_1, \dots, x_k \in I$ such that $I \subseteq \cup_{j=1}^{k} I_{x_j}$. Obviously for any $x \in I$,

$$|f(x)| \leq 1 + \max\{f(x_1), \dots, f(x_k)\},$$

thus showing that f is bounded. ■

In Exercise 3.1.7 the reader can explore cases where the theorem does not hold. Related to the preceding result is the following.

3.1.23 Theorem (Continuous functions on compact intervals achieve their extreme values) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval and if* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is continuous, then there exist points* $\mathrm{x}_{\min}, \mathrm{x}_{\max} \in [\mathrm{a}, \mathrm{b}]$ *such that*

$$\mathrm{f}(\mathrm{x}_{\min}) = \inf\{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in [\mathrm{a}, \mathrm{b}]\}, \quad \mathrm{f}(\mathrm{x}_{\max}) = \sup\{\mathrm{f}(\mathrm{x}) \mid \mathrm{x} \in [\mathrm{a}, \mathrm{b}]\}.$$

Proof It suffices to show that f achieves its maximum on I since if f achieves its maximum, then $-f$ will achieve its minimum. So let $M = \sup\{f(x) \mid x \in I\}$, and suppose that there is no point $x_{\max} \in I$ for which $f(x_{\max}) = M$. Then $f(x) < M$ for each $x \in I$. For a given $x \in I$ we have

$$f(x) = \tfrac{1}{2}(f(x) + f(x)) < \tfrac{1}{2}(f(x) + M).$$

Continuity of f ensures that there is an open interval I_x containing x such that, for each $y \in I_x \cap I$, $f(y) < \frac{1}{2}(f(x) + M)$. Since $I \subseteq \cup_{x \in I} I_x$, by the Heine–Borel theorem, there exists a finite number of points $x_1, \dots, x_k$ such that $I \subseteq \cup_{j=1}^k I_{x_j}$. Let $m = \max\{f(x_1), \dots, f(x_k)\}$ so that, for each $y \in I_{x_j}$, and for each $j \in \{1, \dots, k\}$, we have

$$f(y) < \tfrac{1}{2}(f(x_j) + M) < \tfrac{1}{2}(m + M),$$

which shows that $\frac{1}{2}(m+M)$ is an upper bound for f. However, since f attains the value m on I, we have $m < M$ and so $\frac{1}{2}(m + M) < M$, contradicting the fact that M is the least upper bound. Thus our assumption that f cannot attain the value M on I is false. ■

The theorem tells us that a continuous function on a bounded interval actually *attains* its maximum and minimum value *on the interval*. You should understand that this is not the case if I is neither closed nor bounded (see Exercise 3.1.8).

Our next result gives our first connection between the concepts of uniformity and compactness. This is something of a theme in analysis where continuity is involved. A good place to begin to understand the relationship between compactness and uniformity is the proof of the following theorem, since it is one of the simplest instances of the phenomenon.

3.1.24 Theorem (Heine–Cantor Theorem) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval. If* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is continuous, then it is uniformly continuous.*

Proof Let $x \in [a, b]$ and let $\epsilon \in \mathbb{R}_{>0}$. Since f is continuous, then there exists $\delta_x \in \mathbb{R}_{>0}$ such that, if $|y - x| < \delta_x$, then $|f(y) - f(x)| < \frac{\epsilon}{2}$. Now define an open interval $I_x = (x - \frac{1}{2}\delta_x, x + \frac{1}{2}\delta_x)$. Note that $[a, b] \subseteq \cup_{x\in[a,b]} I_x$, so that the open sets $(I_x)_{x\in[a,b]}$ cover $[a, b]$. By definition of compactness, there then exists a finite number of open sets from $(I_x)_{x\in[a,b]}$ that cover $[a, b]$. Denote this finite family by $(I_{x_1}, \dots, I_{x_k})$ for some $x_1, \dots, x_k \in [a, b]$. Take $\delta = \frac{1}{2}\min\{\delta_{x_1}, \dots, \delta_{x_k}\}$. Now let $x, y \in [a, b]$ satisfy $|x - y| < \delta$. Then there exists $j \in \{1, \dots, k\}$ such that $x \in I_{x_j}$ since the sets $I_{x_1}, \dots, I_{x_k}$ cover $[a, b]$. We also have

$$|y - x_j| = |y - x + x - x_j| \le |y - x| + |x - x_j| < \tfrac{1}{2}\delta_{x_j} + \tfrac{1}{2}\delta_{x_j} = \delta_{x_j},$$

using the triangle inequality. Therefore,

$$\begin{aligned}|f(y) - f(x)| &= |f(y) - f(x_j) + f(x_j) - f(x)| \\ &\le |f(y) - f(x_j)| + |f(x_j) - f(x)| < \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon,\end{aligned}$$

again using the triangle inequality. Since this holds for *any* $x \in [a, b]$, it follows that f is uniformly continuous. ■

Next we give a standard result from calculus that is frequently useful.

3.1.25 Theorem (Intermediate Value Theorem) *Let* I *be an interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be continuous. If* $\mathrm{x}_1, \mathrm{x}_2 \in \mathrm{I}$ *then, for any* $\mathrm{y} \in [\mathrm{f}(\mathrm{x}_1), \mathrm{f}(\mathrm{x}_2)]$, *there exists* $\mathrm{x} \in \mathrm{I}$ *such that* $\mathrm{f}(\mathrm{x}) = \mathrm{y}$.

Proof Since otherwise the result is obviously true, we may suppose that $y \in (f(x_1), f(x_2))$. Also, since we may otherwise replace f with $-f$, we may without loss of generality suppose that $x_1 < x_2$. Now define $S = \{x \in [x_1, x_2] \mid f(x) \le y\}$ and let

$x_0 = \sup S$. We claim that $f(x_0) = y$. Suppose not. Then first consider the case where $f(x_0) > y$, and define $\epsilon = f(x_0) - y$. Then there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ for $|x - x_0| < \delta$. In particular, $f(x_0 - \delta) > y$, contradicting the fact that $x_0 = \sup S$. Next suppose that $f(x_0) < y$. Let $\epsilon = y - f(x_0)$ so that there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(x_0)| < \epsilon$ for $|x - x_0| < \delta$. In particular, $f(x_0 + \delta) < y$, contradicting again the fact that $x_0 = \sup S$. ■

In Figure 3.3 we give the idea of the proof of the Intermediate Value Theorem.

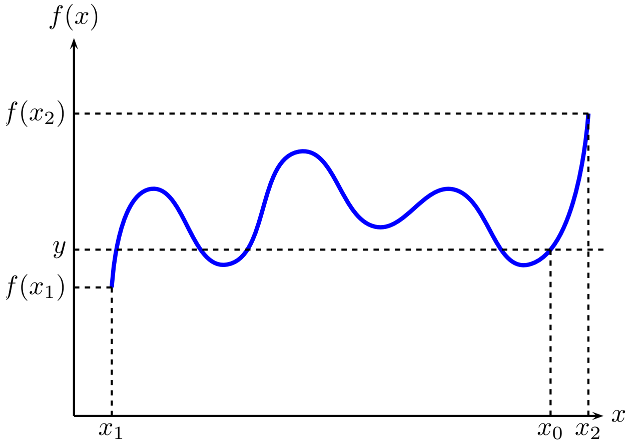


Figure 3.3 Illustration of the Intermediate Value Theorem

There is also a useful relationship between continuity and connected sets.

3.1.26 Proposition (The continuous image of a connected set is connected) *If* $I \subseteq \mathbb{R}$ *is an interval, if* $S \subseteq I$ *is connected, and if* $f\colon I \to \mathbb{R}$ *is continuous, then* $f(S)$ *is connected.*

Proof Suppose that $f(S)$ is not connected. Then there exist nonempty separated sets A and B such that $f(S) = A \cup B$. Let $C = S \cap f^{-1}(A)$ and $D = S \cap f^{-1}(B)$. By Propositions 1.1.4 and 1.3.5 we have

$$\begin{aligned} C \cup D &= (S \cap f^{-1}(A)) \cup (S \cap f^{-1}(B)) \\ &= S \cap (f^{-1}(A) \cup f^{-1}(B)) = S \cap f^{-1}(A \cup B) = S. \end{aligned}$$

By Propositions 2.5.20 and 1.3.5, and since $f^{-1}(\mathrm{cl}(A))$ is closed, we have

$$\mathrm{cl}(C) = \mathrm{cl}(f^{-1}(A)) \subseteq \mathrm{cl}(f^{-1}(\mathrm{cl}(A)) = f^{-1}(\mathrm{cl}(A)).$$

We also clearly have $D \subseteq f^{-1}(B)$. Therefore, by Proposition 1.3.5,

$$\mathrm{cl}(C) \cap D \subseteq f^{-1}(\mathrm{cl}(A)) \cap f^{-1}(B) = f^{-1}(\mathrm{cl}(A) \cap B) = \emptyset.$$

We also similarly have $C \cap \mathrm{cl}(D) = \emptyset$. Thus S is not connected, which gives the result. ■

3.1.5 Monotonic functions and continuity

In this section we consider a special class of functions, namely those that are "increasing" or "decreasing."

3.1.27 Definition (Monotonic function) For $I \subseteq \mathbb{R}$ an interval, a function $f\colon I \to \mathbb{R}$ is:

(i) ***monotonically increasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \le f(x_2)$;

(ii) ***strictly monotonically increasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) < f(x_2)$;

(iii) ***monotonically decreasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) \ge f(x_2)$;

(iv) ***strictly monotonically decreasing*** if, for every $x_1, x_2 \in I$ with $x_1 < x_2$, $f(x_1) > f(x_2)$;

(v) ***constant*** if there exists $\alpha \in \mathbb{R}$ such that $f(x) = \alpha$ for every $x \in I$. •

Let us see how monotonicity can be used to make some implications about the continuity of a function. In Theorem 3.2.26 below we will explore some further properties of monotonic functions.

3.1.28 Theorem (Characterisation of monotonic functions I) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is either monotonically increasing or monotonically decreasing, then the following statements hold:*

(i) the limits $\lim_{\mathrm{x}\downarrow\mathrm{x}_0} \mathrm{f}(\mathrm{x})$ *and* $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \mathrm{f}(\mathrm{x})$ *exist whenever they make sense as limits in* I*;*

(ii) the set on which f *is discontinuous is countable.*

Proof We can assume without loss of generality (why?), we assume that $I = [a, b]$ and that f is monotonically increasing.

(i) First let us consider limits from the left. Thus let $x_0 > a$ and consider $\lim_{x\uparrow x_0} f(x)$. For any increasing sequence $(x_j)_{j\in\mathbb{Z}_{>0}} \subseteq [a, x_0)$ converging to x_0 the sequence $(f(x_j))_{j\in\mathbb{Z}_{>0}}$ is bounded and increasing. Therefore it has a limit by Theorem 2.3.8. In a like manner, one shows that right limits also exist.

(ii) Define

$$j(x_0) = \lim_{x\downarrow x_0} f(x) - \lim_{x\uparrow x_0} f(x)$$

as the jump at x_0. This is nonzero if and only if x_0 is a point of discontinuity of f. Let A_f be the set of points of discontinuity of f. Since f is monotonically increasing and defined on a compact interval, it is bounded and we have

$$\sum_{x\in A_f} j(x) \le f(b) - f(a). \tag{3.1}$$

Now let $n \in \mathbb{Z}_{>0}$ and denote

$$A_n = \left\{x \in [a, b] \;\middle|\; j(x) > \tfrac{1}{n}\right\}.$$

The set A_n must be finite by (3.1). We also have

$$A_f = \bigcup_{n\in\mathbb{Z}_{>0}} A_n,$$

meaning that A_f is a countable union of finite sets. Thus A_f is itself countable. ■

Sometimes the following "local" characterisation of monotonicity is useful.

3.1.29 Proposition (Monotonicity is "local") *A function* $f\colon I \to \mathbb{R}$ *defined on an interval* I *is*

(i) monotonically increasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is monotonically increasing;*

(ii) strictly monotonically increasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is strictly monotonically increasing;*

(iii) monotonically decreasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is monotonically decreasing;*

(iv) strictly monotonically decreasing if and only if, for every $x \in I$*, there exists a neighbourhood* U *of* x *such that* $f|U \cap I$ *is strictly monotonically decreasing.*

Proof We shall only prove the first assertion as the other follow from an identical sort of argument. Also, the "only if" assertion is clear, so we need only prove the "if" assertion.

Let $x_1, x_2 \in I$ with $x_1 < x_2$. By hypothesis, for $x \in [x_1, x_2]$, there exists $\epsilon_x \in \mathbb{R}_{>0}$ such that, if we define $U_x = (x - \epsilon, x + \epsilon)$, then $f|U_x \cap I$ is monotonically increasing. Note that $(U_x)_{x\in[x_1,x_2]}$ covers $[x_1, x_2]$ and so, by the Heine–Borel Theorem, there exists $\xi_1, \dots, \xi_k \in [x_1, x_2]$ such that $[x_1, x_2] \subseteq \cup_{j=1}^k U_{\xi_j}$. We can assume that $\xi_1, \dots, \xi_k$ are ordered so that $x_1 \in U_{\xi_1}$, that $U_{\xi_{j+1}} \cap U_{\xi_j} \neq \emptyset$, and such that $x_2 \in U_{\xi_k}$. We have that $f|U_{\xi_1} \cap I$ is monotonically increasing. Since $f|U_{\xi_2} \cap I$ is monotonically increasing and since $U_{\xi_1} \cap U_{\xi_2} \neq \emptyset$, we deduce that $f|(U_{\xi_1} \cup U_{\xi_2}) \cap I$ is monotonically increasing. We can continue this process to show that

$$f|(U_{\xi_1} \cup \dots \cup U_{\xi_k}) \cap I$$

is monotonically increasing, which is the result. ■

In thinking about the graph of a continuous monotonically increasing function, it will not be surprising that there might be a relationship between monotonicity and invertibility. In the next result we explore the precise nature of this relationship.

3.1.30 Theorem (Strict monotonicity and continuity implies invertibility) *Let* $I \subseteq \mathbb{R}$ *be an interval, let* $f\colon I \to \mathbb{R}$ *be continuous and strictly monotonically increasing (resp. strictly monotonically decreasing). If* $J = \mathrm{image}(f)$ *then the following statements hold:*

(i) J *is an interval;*

(ii) there exists a continuous, strictly monotonically increasing (resp. strictly monotonically decreasing) inverse $g\colon J \to I$ *for* f*.*

Proof We suppose f to be strictly monotonically increasing; the case where it is strictly monotonically decreasing is handled similarly (or follows by considering $-f$, which is strictly monotonically increasing if f is strictly monotonically decreasing).

(i) This follows from Theorem 2.5.34 and Proposition 3.1.26, where it is shown that intervals are the only connected sets, and that continuous images of connected sets are connected.

(ii) Since f is strictly monotonically increasing, if $f(x_1) = f(x_2)$, then $x_1 = x_2$. Thus f is injective as a map from I to J. Clearly $f\colon I \to J$ is also surjective, and so is invertible. Let $y_1, y_2 \in J$ and suppose that $y_1 < y_2$. Then $f(g(y_1)) < f(g(y_2))$, implying that $g(y_1) < g(y_2)$. Thus g is strictly monotonically increasing. It remains to show that the inverse g is continuous. Let $y_0 \in J$ and let $\epsilon \in \mathbb{R}_{>0}$. First suppose that $y_0 \in \operatorname{int}(J)$. Let $x_0 = g(y_0)$ and, supposing ϵ sufficiently small, define $y_1 = f(x_0 - \epsilon)$ and $y_2 = f(x_0 + \epsilon)$. Then let $\delta = \min\{y_0 - y_1, y_2 - y_0\}$. If $y \in \mathsf{B}(\delta, y_0)$ then $y \in (y_1, y_2)$, and since g is strictly monotonically increasing

$$x_0 - \epsilon = g(y_1) < g(y) < g(y_2) = x_0 + \epsilon.$$

Thus $g(y) \in \mathsf{B}(\epsilon, y_0)$, giving continuity of g at x_0. An entirely similar argument can be given if y_0 is an endpoint of J. ■

3.1.6 Convex functions and continuity

In this section we see for the first time the important notion of convexity, here in a fairly simple setting.

Let us first define what we mean by a convex function.

3.1.31 Definition (Convex function) For an interval $I \subseteq \mathbb{R}$, a function $f\colon I \to \mathbb{R}$ is:

(i) ***convex*** if
$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2)$$
for every $x_1, x_2 \in I$ and $s \in [0,1]$;

(ii) ***strictly convex*** if
$$f((1-s)x_1 + sx_2) < (1-s)f(x_1) + sf(x_2)$$
for every distinct $x_1, x_2 \in I$ and for every $s \in (0,1)$;

(iii) ***concave*** if $-f$ is convex;

(iv) ***strictly concave*** if $-f$ is strictly convex. •

Let us give some examples of convex functions.

3.1.32 Examples (Convex functions)

1. A constant function $x \mapsto c$, defined on any interval, is both convex and concave in a trivial way. It is neither strictly convex nor strictly concave.
2. A linear function $x \mapsto ax + b$, defined on any interval, is both convex and concave. It is neither strictly convex nor strictly concave.
3. The function $x \mapsto x^2$, defined on any interval, is strictly convex. Let us verify this. For $s \in (0,1)$ and for $x, y \in \mathbb{R}$ we have, using the triangle inequality,

$$((1-s)x + sy)^2 \leq |(1-s)x + sy|^2 < (1-s)^2x^2 + s^2y^2 \leq (1-s)x^2 + sy^2.$$

4. We refer to Section 3.8.1 for the definition of exponential function $\exp\colon \mathbb{R} \to \mathbb{R}$. We claim that exp is strictly convex. This can be verified explicitly with some effort. However, it follows easily from the fact, proved as Proposition 3.2.30 below, that a function like exp that is twice continuously differentiable with a positive second-derivative is strictly convex. (Note that $\exp'' = \exp$.)
5. We claim that the function log defined in Section 3.8.2 is strictly concave as a function on $\mathbb{R}_{>0}$. Here we compute $\log''(x) = -\frac{1}{x^2}$, which gives strict convexity of $-\log$ (and hence strict concavity of log) by Proposition 3.2.30 below.
6. For $x_0 \in \mathbb{R}$, the function $n_{x_0}\colon \mathbb{R} \to \mathbb{R}$ defined by $n_{x_0} = |x - x_0|$ is convex. Indeed, if $x_1, x_2 \in \mathbb{R}$ and $s \in [0, 1]$ then

$$\begin{aligned} n_{x_0}((1-s)x_1 + sx_2) = |(1-s)x_1 + sx_2 - x_0| &= |(1-s)(x_1 - x_0) + s(x_2 - x_0)| \\ &\leq (1-s)|x_1 - x_0| + s|x_2 - x_0| = (1-s)n_{x_0}(x_1) + sn_{x_0}(x_2), \end{aligned}$$

using the triangle inequality. •

Let us give an alternative and insightful characterisation of convex functions. For an interval $I \subseteq \mathbb{R}$ define

$$E_I = \{(x, y) \in I^2 \mid s < t\}$$

and, for $a, b \in I$, denote

$$L_b = \{a \in I \mid (a, b) \in E_I\}, \quad R_a = \{b \in I \mid (a, b) \in E_I\}.$$

Now, for $f\colon I \to \mathbb{R}$ define $s_f\colon E_I \to \mathbb{R}$ by

$$s_f(a, b) = \frac{f(b) - f(a)}{b - a}.$$

With this notation at hand, we have the following result.

3.1.33 Lemma (Alternative characterisation of convexity) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$*, a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is (strictly) convex if and only if, for every* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$*, the functions*

$$\mathrm{L_b} \ni \mathrm{a} \mapsto \mathrm{s_f}(\mathrm{a}, \mathrm{b}) \in \mathbb{R}, \quad \mathrm{R_a} \ni \mathrm{b} \mapsto \mathrm{s_f}(\mathrm{a}, \mathrm{b}) \in \mathbb{R} \tag{3.2}$$

are (strictly) monotonically increasing.

Proof First suppose that f is convex. Let $a, b, c \in I$ satisfy $a < b < c$. Define $s \in (0, 1)$ by $s = \frac{b-a}{c-a}$ and note that the definition of convexity using this value of s gives

$$f(b) \leq \frac{c-b}{c-a} f(a) + \frac{b-a}{c-a} f(c).$$

Simple rearrangement gives

$$\frac{c-b}{c-a} f(a) + \frac{b-a}{c-a} f(c) = f(a) + \frac{f(c) - f(a)}{c-a}(b-a) = f(c) - \frac{f(c) - f(a)}{c-a}(c-b),$$

and so we have

$$\frac{f(b)-f(a)}{b-a} \leq \frac{f(c)-f(a)}{c-a}, \quad \frac{f(c)-f(a)}{c-a} \leq \frac{f(c)-f(b)}{c-b}.$$

In other words, $s_f(a,b) \leq s_f(a,c)$ and $s_f(a,c) \leq s_f(b,c)$. Since this holds for every $a,b,c \in I$ with $a < b < c$, we conclude that the functions (3.2) are monotonically increasing, as stated. If f is strictly convex, then the inequalities in the above computation are strict, and one concludes that the functions (3.2) are strictly monotonically increasing.

Next suppose that the functions (3.2) are monotonically increasing and let $a,c \in I$ with $a < c$ and let $s \in (0,1)$. Define $b = (1-s)a + sc$. A rearrangement of the inequality $s_f(a,b) \leq s_f(a,c)$ gives

$$\begin{aligned} & f(b) \leq \frac{c-b}{c-a}f(a) + \frac{b-a}{c-a}f(c) \\ \implies \quad & f((1-s)a+sc) \leq (1-s)f(a) + sf(c), \end{aligned}$$

showing that f is convex since $a,c \in I$ with $a < c$ and $s \in (0,1)$ are arbitrary in the above computation. If the functions (3.2) are strictly monotonically increasing, then the inequalities in the preceding computations are strict, and so one deduces that f is strictly convex. ■

In Figure 3.4 we depict what the lemma is telling us about convex functions.

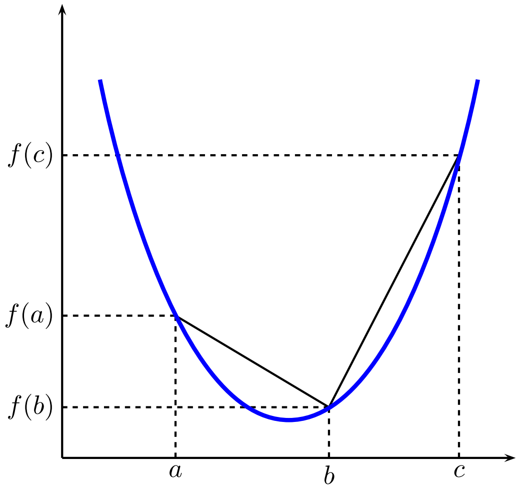


Figure 3.4 A characterisation of a convex function

The idea is that the slope of the line connecting the points $(a, f(a))$ and $(b, f(b))$ in the plane is nondecreasing in a and b.

The following inequality for convex functions is very often useful.

3.1.34 Theorem (Jensen's inequality) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$, *for a convex function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$, *for* $\mathrm{x}_1, \dots, \mathrm{x}_k \in \mathrm{I}$, *and for* $\lambda_1, \dots, \lambda_k \in \mathbb{R}_{\geq 0}$, *we have*

$$\mathrm{f}\left(\frac{\lambda_1}{\sum_{j=1}^{k}\lambda_j}\mathrm{x}_1 + \dots + \frac{\lambda_k}{\sum_{j=1}^{k}\lambda_j}\mathrm{x}_k\right) \leq \frac{\lambda_1}{\sum_{j=1}^{k}\lambda_j}\mathrm{f}(\mathrm{x}_1) + \dots + \frac{\lambda_k}{\sum_{j=1}^{k}\lambda_j}\mathrm{f}(\mathrm{x}_k).$$

Moreover, if f *is strictly convex and if* $\lambda_1, \dots, \lambda_\mathrm{k} \in \mathbb{R}_{>0}$*, than we have equality in the preceding expression if and only if* $\mathrm{x}_1 = \cdots = \mathrm{x}_\mathrm{k}$.

Proof We first comment that, with $\lambda_1, \dots, \lambda_k$ and $x_1, \dots, x_k$ as stated,

$$\frac{\lambda_1}{\sum_{j=1}^k \lambda_j} x_1 + \cdots + \frac{\lambda_k}{\sum_{j=1}^k \lambda_j} x_k \in I.$$

This is because intervals are convex, something that will become clear in Section II-1.9.2.

It is clear that we can without loss of generality, by replacing λ_j with

$$\lambda'_m = \frac{\lambda_m}{\sum_{j=1}^k \lambda_j}, \qquad m \in \{1, \dots, k\},$$

if necessary, that we can assume that $\sum_{j=1}^k \lambda_j = 1$.

We first note that if $x_1 = \cdots = x_k$ then the inequality in the statement of the theorem is an equality, no matter what the character of f.

The proof is by induction on k, the result being obvious when $k = 1$. So suppose the result is true when $k = m$ and let $x_1, \dots, x_{m+1} \in I$ and let $\lambda_1, \dots, \lambda_{m+1} \in \mathbb{R}_{\geq 0}$ satisfy $\sum_{j=1}^{m+1} \lambda_j = 1$. Without loss of generality (by reindexing if necessary), suppose that $\lambda_{m+1} \in [0, 1)$. Note that

$$\frac{\lambda_1}{1 - \lambda_{m+1}} + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}} = 1$$

so that, by the induction hypothesis,

$$f\left(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\right) \leq \frac{\lambda_1}{1 - \lambda_{m+1}} f(x_1) + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}} f(x_m).$$

Now, by convexity of f,

$$\begin{aligned} f\Bigg((1 - \lambda_{m+1})\left(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\right) + \lambda_{m+1} x_{m+1}\Bigg) \\ \leq (1 - \lambda_{m+1}) f\left(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \cdots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\right) + \lambda_{m+1} f(x_{m+1}). \end{aligned}$$

The desired inequality follows by combining the previous two equations.

To prove the final assertion of the theorem, suppose that f is strictly convex, that $\lambda_1, \dots, \lambda_k \in \mathbb{R}_{>0}$ satisfy $\sum_{j=1}^k \lambda_j = 1$, and that the inequality in the theorem is equality. We prove by induction that $x_1 = \cdots = x_k$. For $k = 1$ the assertion is obvious. Let us prove the assertion for $k = 2$. Thus suppose that

$$f((1 - \lambda)x_1 + \lambda x_2) = (1 - \lambda) f(x_1) + \lambda f(x_2)$$

for $x_1, x_2 \in I$ and for $\lambda \in (0, 1)$. If $x_1 \neq x_2$ then we have, by definition of strict convexity,

$$f((1 - \lambda)x_1 + \lambda x_2) < (1 - \lambda) f(x_1) + \lambda f(x_2),$$

contradicting our hypotheses. Thus we must have $x_1 = x_2$. Now suppose the assertion is true for $k = m$ and let $x_1, \dots, x_{m+1} \in I$, let $\lambda_1, \dots, \lambda_{m+1} \in \mathbb{R}_{>0}$ satisfy $\sum_{j=1}^{m+1} \lambda_j = 1$, and suppose that

$$f(\lambda_1 x_1 + \dots + \lambda_{m+1} x_{m+1}) = \lambda_1 f(x_1) + \dots + \lambda_{m+1} f(x_{m+1}).$$

Since none of $\lambda_1, \dots, \lambda_{m+1}$ are zero we must have $\lambda_{m+1} \in (0, 1)$. Now note that

$$f(\lambda_1 x_1 + \dots + \lambda_{m+1} x_{m+1}) = f\Big((1 - \lambda_{m+1})\Big(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\Big) + \lambda_{m+1} x_{m+1}\Big) \tag{3.3}$$

and that

$$\begin{aligned} &\lambda_1 f(x_1) + \dots + \lambda_{m+1} f(x_{m+1}) \\ &\qquad = (1 - \lambda_{m+1}) f\Big(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\Big) + \lambda_{m+1} f(x_{m+1}). \end{aligned}$$

Therefore, by assumption,

$$\begin{aligned} &f\Big((1 - \lambda_{m+1})\Big(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\Big) + \lambda_{m+1} x_{m+1}\Big) \\ &\qquad = (1 - \lambda_{m+1}) f\Big(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\Big) + \lambda_{m+1} f(x_{m+1}). \end{aligned} \tag{3.4}$$

Since the assertion we are proving holds for $k = 2$ this implies that

$$x_{m+1} = \frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m. \tag{3.5}$$

Now suppose that the numbers $x_1, \dots, x_m$ are not all equal. Then, by the induction hypothesis,

$$f\Big(\frac{\lambda_1}{1 - \lambda_{m+1}} x_1 + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} x_m\Big) < \frac{\lambda_1}{1 - \lambda_{m+1}} f(x_1) + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} f(x_m)$$

since

$$\frac{\lambda_1}{1 - \lambda_{m+1}} + \dots + \frac{\lambda_m}{1 - \lambda_{m+1}} = 1.$$

Therefore, combining (3.3) and (3.4)

$$f(\lambda_1 x_1 + \dots + \lambda_{m+1} x_{m+1}) < \lambda_1 f(x_1) + \dots + \lambda_{m+1} f(x_{m+1}),$$

contradicting our hypotheses. Thus we must have $x_1 = \dots = x_m$. From (3.5) we then conclude that $x_1 = \dots = x_{m+1}$, as desired. ■

An interesting application of Jensen's inequality is the derivation of the so-called arithmetic/geometric mean inequalities. If $x_1, \dots, x_k \in \mathbb{R}_{>0}$, their ***arithmetic mean*** is

$$\frac{1}{k}(x_1 + \dots + x_k)$$

and their ***geometric mean*** is

$$(x_1 \cdots x_k)^{1/k}.$$

We first state a result which relates generalisations of the arithmetic and geometric means.

3.1.35 Corollary (Weighted arithmetic/geometric mean inequality) *Let* $x_1, \dots, x_k \in \mathbb{R}_{\geq 0}$ *and suppose that* $\lambda_1, \dots, \lambda_k \in \mathbb{R}_{>0}$ *satisfy* $\sum_{j=1}^k \lambda_j = 1$*. Then*

$$x_1^{\lambda_1} \cdots x_k^{\lambda_k} \leq \lambda_1 x_1 + \cdots + \lambda_k x_k,$$

and equality holds if and only if $x_1 = \cdots = x_k$*.*

Proof Since the inequality obviously holds if any of $x_1, \dots, x_k$ are zero, let us suppose that these numbers are all positive. By Example 3.1.32–5, $-\log$ is convex. Thus Jensen's inequality gives

$$-\log(\lambda_1 x_1 + \cdots + \lambda_k x_k) \leq -\lambda_1 \log(x_1) - \cdots - \lambda_k \log(x_k) = -\log(x_1^{\lambda_1} \cdots x_k^{\lambda_k}).$$

Since $-\log$ is strictly monotonically decreasing by Proposition 3.8.6(ii), the result follows. Moreover, since $-\log$ is strictly convex by Proposition 3.2.30, the final assertion of the corollary follows from the final assertion of Theorem 3.1.34. ■

The corollary gives the following inequality as a special case.

3.1.36 Corollary (Arithmetic/geometric mean inequality) *If* $x_1, \dots, x_k \in \mathbb{R}_{\geq 0}$ *then*

$$(x_1 \cdots x_k)^{1/k} \leq \frac{x_1 + \cdots + x_k}{k},$$

and equality holds if and only if $x_1 = \cdots = x_k$*.*

Let us give some properties of convex functions. Further properties of convex function are give in Proposition 3.2.29

3.1.37 Proposition (Properties of convex functions I) *For an interval* $I \subseteq \mathbb{R}$ *and for a convex function* $f\colon I \to \mathbb{R}$*, the following statements hold:*

(i) if I *is open, then* f *is continuous;*

(ii) for any compact interval $K \subseteq \operatorname{int}(I)$*, there exists* $L \in \mathbb{R}_{>0}$ *such that*

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|, \qquad x_1, x_2 \in K.$$

Proof (ii) Let $K = [a, b] \subseteq \operatorname{int}(I)$ and let $a', b' \in I$ satisfy $a' < a$ and $b' > b$, this being possible since $K \subseteq \operatorname{int}(I)$. Now let $x_1, x_2 \in K$ and note that, by Lemma 3.1.33,

$$s_f(a', a) \leq s_f(x_1, x_2) \leq s_f(b, b')$$

since $a' < x_1$, $a \leq x_2$, $x_1 \leq b$, and $x_2 < b'$. Thus, taking $L = \max\{s_f(a', a), s_f(b, b')\}$, we have

$$-L \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq L,$$

which gives the result.

(i) This follows from part (ii) easily. Indeed let $x \in I$ and let K be a compact subinterval of I such that $x \in \operatorname{int}(K)$, this being possible since I is open. If $\epsilon \in \mathbb{R}_{>0}$, let $\delta = \frac{\epsilon}{L}$. It then immediately follows that if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$. ■

Let us give some an example that illustrates that openness is necessary in the first part of the preceding result.

3.1.38 Example (A convex discontinuous function) Let $I = [0,1]$ and define $f\colon [0,1] \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x = 1, \\ 0, & x \in [0,1). \end{cases}$$

If $x_1, x_2 \in [0,1)$ and if $s \in [0,1]$ then

$$0 = f((1-s)x_1 + sx_2) = (1-s)f(x_1) + sf(x_2).$$

If $x_1 \in [0,1)$, if $x_2 = 1$, and if $s \in (0,1)$ then

$$0 = f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2) = s,$$

showing that f is convex as desired. Note that f is not continuous, but that its discontinuity is on the boundary, as must be the case since convex functions on open sets are continuous. •

Let us also present some operations that preserve convexity.

3.1.39 Proposition (Convexity and operations on functions) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and for convex functions* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$*, the following statements hold:*

(i) the function $\mathrm{I} \ni \mathrm{x} \mapsto \max\{\mathrm{f}(\mathrm{x}), \mathrm{g}(\mathrm{x})\}$ *is convex;*

(ii) the function af *is convex if* $\mathrm{a} \in \mathbb{R}_{\geq 0}$*;*

(iii) the function $\mathrm{f} + \mathrm{g}$ *is convex;*

(iv) if $\mathrm{J} \subseteq \mathbb{R}$ *is an interval, if* f *takes values in* J*, and if* $\phi\colon \mathrm{J} \to \mathbb{R}$ *is convex and monotonically increasing, then* $\phi \circ \mathrm{f}$ *is convex;*

(v) if $\mathrm{x}_0 \in \mathrm{I}$ *is a local minimum for* f *(see Definition 3.2.15). then* x_0 *is a minimum for* f.

Proof (i) Let $x_1, x_2 \in I$ and let $s \in [0,1]$. Then, by directly applying the definition of convexity to f and g, we have

$$\begin{aligned} &\max\{f((1-s)x_1 + sx_2), g((1-s)x_1 + sx_2)\} \\ &\qquad\qquad \leq (1-s)\max\{f(x_1), g(x_1)\} + s\max\{f(x_2), g(x_2)\}. \end{aligned}$$

(ii) This follows immediately from the definition of convexity.

(iii) For $x_1, x_2 \in I$ and for $s \in [0,1]$ we have

$$\begin{aligned} f((1-s)x_1 + sx_2) + g((1-s)x_1 + sx_2) &\leq (1-s)f(x_1) + sf(x_2) + (1-s)g(x_1) + sg(x_2) \\ &= (1-s)(f(x_1) + g(x_1)) + s(f(x_2 + g(x_2)), \end{aligned}$$

by applying the definition of convexity to f and g.

(iv) For $x_1, x_2 \in I$ and for $s \in [0,1]$, convexity of f gives

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2)$$

and so monotonicity of ϕ gives

$$\phi\circ f((1-s)x_1+sx_2)\le\phi((1-s)f(x_1)+sf(x_2)).$$

Now convexity of ϕ gives

$$\phi\circ f((1-s)x_1+sx_2)\le(1-s)\phi\circ f(x_1)+s\phi\circ f(x_2),$$

as desired.

(v) Suppose that x_0 is a local minimum for f, i.e., there is a neighbourhood $U\subseteq I$ of x_0 such that $f(x)\ge f(x_0)$ for all $x\in U$. Now let $x\in I$ and note that

$$s\mapsto(1-s)x_0+sx$$

is continuous and $\lim_{s\to 0}(1-s)x_0+sx=x_0$. Therefore, there exists $s_0\in(0,1]$ such that $(1-s)x_0+sx\in U$ for all $s\in(0,s_0)$. Thus

$$f(x_0)\le f((1-s)x_0+sx)\le(1-s)f(x_0)+sf(x)$$

for $s\in(0,s_0)$. Simplification gives $f(x_0)\le f(x)$ and so x_0 is a minimum for f. ■

3.1.7 Piecewise continuous functions

It is often of interest to consider functions that are not continuous, but which possess only jump discontinuities, and only “few” of these. In order to do so, it is convenient to introduce some notation. For and interval $I\subseteq\mathbb{R}$, a function $f\colon I\to\mathbb{R}$, and $x\in I$ define

$$f(x-)=\lim_{\epsilon\downarrow 0}f(x-\epsilon),\quad f(x+)=\lim_{\epsilon\downarrow 0}f(x+\epsilon),$$

allowing that these limits may not be defined (or even make sense if $x\in\mathrm{bd}(I)$).

We then have the following definition, recalling our notation concerning partitions of intervals given in and around Definition 2.5.7.

3.1.40 Definition (Piecewise continuous function) A function $f\colon[a,b]\to\mathbb{R}$ is ***piecewise continuous*** if there exists a partition $P=(I_1,\dots,I_k)$, with $\mathrm{EP}(P)=(x_0,x_1,\dots,x_k)$, of $[a,b]$ with the following properties:

(i) $f|\,\mathrm{int}(I_j)$ is continuous for each $j\in\{1,\dots,k\}$;

(ii) for $j\in\{1,\dots,k-1\}$, the limits $f(x_j+)$ and $f(x_j-)$ exist;

(iii) the limits $f(a+)$ and $f(b-)$ exist. •

Let us give a couple of examples to illustrate some of the things that can happen with piecewise continuous functions.

3.1.41 Examples (Piecewise continuous functions)

1. Let $I = [-1, 1]$ and define $f_1, f_2, f_3 \colon I \to \mathbb{R}$ by
$$f_1(x) = \operatorname{sign}(x),$$
$$f_2(x) = \begin{cases} \operatorname{sign}(x), & x \neq 0, \\ 1, & x = 0, \end{cases}$$
$$f_2(x) = \begin{cases} \operatorname{sign}(x), & x \neq 0, \\ -1, & x = 0. \end{cases}$$
One readily verifies that all of these functions are piecewise continuous with a single discontinuity at $x = 0$. Note that the functions do not have the same value at the discontinuity. Indeed, the definition of piecewise continuity is unconcerned with the value of the function at discontinuities.

2. Let $I = [-1, 1]$ and define $f \colon I \to \mathbb{R}$ by
$$f(x) = \begin{cases} 1, & x \neq 0, \\ 0, & x = 0. \end{cases}$$
This function is, by definition, piecewise continuous with a single discontinuity at $x = 0$. This shows that the definition of piecewise continuity includes functions, not just with jump discontinuities, but with removable discontinuities. •

Exercises

3.1.1

Oftentimes, a continuity novice will think that the definition of continuity at x_0 of a function $f \colon I \to \mathbb{R}$ is as follows: for every $\epsilon \in \mathbb{R}_{>0}$ there exists $\delta \in \mathbb{R}_{>0}$ such that if $|f(x) - f(x_0)| < \epsilon$ then $|x - x_0| < \delta$. Motivated by this, let us call a function ***fresh-from-high-school continuous*** if it has the preceding property at each point $x \in I$.

3.1.2 Answer the following two questions.

(a) Find an interval $I \subseteq \mathbb{R}$ and a function $f \colon I \to \mathbb{R}$ such that f is continuous but not fresh-from-high-school continuous.

(b) Find an interval $I \subseteq \mathbb{R}$ and a function $f \colon I \to \mathbb{R}$ such that f is fresh-from-high-school continuous but not continuous.

3.1.3 Let $I \subseteq \mathbb{R}$ be an interval and let $f, g \colon I \to \mathbb{R}$ be functions.

(a) Show that $D_{fg} \subseteq D_f \cup D_g$.

(b) Show that it is not generally true that $D_f \cap D_g \subseteq D_{fg}$.

(c) Suppose that f is bounded. Show that if $x \in (D_f \cap (I \setminus D_g)) \cap (I \setminus D_{fg})$, then $g(x) = 0$.

3.1.4 Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function. For $x \in I$ and $\delta \in \mathbb{R}_{>0}$ define

$$\omega_f(x, \delta) = \sup\{|f(x_1), f(x_2)| \mid x_1, x_2 \in \mathsf{B}(\delta, x) \cap I\}.$$

Show that, if $y \in \mathsf{B}(\delta, x)$, then $\omega_f(y, \frac{\delta}{2}) \leq \omega_f(x, \delta)$.

3.1.5 Recall from Theorem 3.1.24 that a continuous function defined on a compact interval is uniformly continuous. Show that this assertion is generally false if the interval is not compact.

3.1.6 Give an example of an interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ that is locally bounded but not bounded.

3.1.7 Answer the following three questions.

(a) Find a bounded interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is continuous but not bounded.

(b) Find a compact interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is bounded but not continuous.

(c) Find a closed but unbounded interval $I \subseteq \mathbb{R}$ and a function $f\colon I \to \mathbb{R}$ such that f is continuous but not bounded.

3.1.8 Answer the following two questions.

(a) For $I = [0, 1)$ find a bounded, continuous function $f\colon I \to \mathbb{R}$ that does not attain its maximum on I.

(b) For $I = [0, \infty)$ find a bounded, continuous function $f\colon I \to \mathbb{R}$ that does not attain its maximum on I.

3.1.9 Explore your understanding of Theorem 3.1.3 and its Corollary 3.1.4 by doing the following.

(a) For the continuous function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$, verify Theorem 3.1.3 by (1) determining $f^{-1}(I)$ for a general open interval I and (2) showing that this is sufficient to ensure continuity.
Hint: *For the last part, consider using Proposition 2.5.6 and part (iv) of Proposition 1.3.5.*

(b) For the discontinuous function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(x) = \operatorname{sign}(x)$, verify Theorem 3.1.3 by (1) finding an open subset $U \subseteq \mathbb{R}$ for which $f^{-1}(U)$ is not open and (2) finding a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ converging to $x_0 \in \mathbb{R}$ for which $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ does not converge to $f(x_0)$.

3.1.10 Find a continuous function $f\colon I \to \mathbb{R}$ defined on some interval I and a sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ such that the sequence $(x_j)_{j \in \mathbb{Z}_{>0}}$ does not converge but the sequence $(f(x_j))_{j \in \mathbb{Z}_{>0}}$ does converge.

3.1.11 Let $I \subseteq \mathbb{R}$ be an interval and let $f, g\colon I \to \mathbb{R}$ be convex.

(a) Is it true that $x \mapsto \min\{f(x), g(x)\}$ is convex?

(b) Is it true that $f - g$ is convex?

3.1.12 Let $U \subseteq \mathbb{R}$ be open and suppose that $f\colon U \to \mathbb{R}$ is continuous and has the property that

$$\{x \in U \mid f(x) \neq 0\}$$

has measure zero. Show that $f(x) = 0$ for *all* $x \in U$.

Section 3.2

Differentiable $\mathbb{R}$-valued functions on $\mathbb{R}$

In this section we deal systematically with another topic with which most readers are at least somewhat familiar: differentiation. However, as with everything we do, we do this here is a manner that is likely more thorough and systematic than that seen by some readers. We do suppose that the reader has had that sort of course where one learns the derivatives of the standard functions, and learns to apply some of the standard rules of differentiation, such as we give in Section 3.2.3.

Do I need to read this section? If you are familiar with, or perhaps even if you only think you are familiar with, the meaning of "continuously differentiable," then you can probably forgo the details of this section. However, if you have not had the benefit of a rigorous calculus course, then the material here might at least be interesting. •

3.2.1 Definition of the derivative

The definition we give of the derivative is as usual, with the exception that, as we did when we talked about continuity, we allow functions to be defined on general intervals. In order to do this, we recall from Section 2.3.7 the notation $\lim_{x\to_I x_0} f(x)$.

3.2.1 Definition (Derivative and differentiable function) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function.

(i) The function f is ***differentiable at*** $\mathbf{x_0 \in I}$ if the limit

$$\lim_{x\to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} \tag{3.6}$$

exists.

(ii) If the limit (3.6) exists, then it is denoted by $f'(x_0)$ and called the ***derivative*** of f at x_0.

(iii) If f is differentiable at each point $x \in I$, then f is ***differentiable***.

(iv) If f is differentiable and if the function $x \mapsto f'(x)$ is continuous, then f is ***continuously differentiable***, or of ***class*** $\mathbf{C^1}$. •

3.2.2 Notation (Alternative notation for derivative) In applications where $\mathbb{R}$-valued functions are clearly to be thought of as functions of "time," we shall sometimes write $\dot{f}$ rather than f' for the derivative.

Sometimes it is convenient to write the derivative using the convention $f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}$. This notation for derivative suffers from the same problems as the notation

"$f(x)$" to denote a function as discussed in Notation 1.3.2. That is to say, one cannot really use $\frac{\mathrm{d}f}{\mathrm{d}x}$ as a substitute for f', but only for $f'(x)$. Sometimes one can kludge one's way around this with something like $\frac{\mathrm{d}f}{\mathrm{d}x}\big|_{x=x_0}$ to specify the derivative at x_0. But this still leaves unresolved the matter of what is the rôle of "x" in the expression $\frac{\mathrm{d}f}{\mathrm{d}x}\big|_{x=x_0}$. For this reason, we will generally (but not exclusively) stick to f', or sometimes $\dot{f}$. For notation for the derivative for multivariable functions, we refer to Definition II-1.4.2. •

Let us consider some examples that illustrate the definition.

3.2.3 Examples (Derivative)

1. Take $I = \mathbb{R}$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x^k$ for $k \in \mathbb{Z}_{>0}$. We claim that f is continuously differentiable, and that $f'(x) = kx^{k-1}$. To prove this we first note that
$$(x - x_0)(x^{k-1} + x^{k-1}x_0 + \cdots + xx_0^{k-2} + x_0^{k-1}) = x^k - x_0^k,$$
as can be directly verified. Then we compute
$$\begin{aligned}\lim_{x\to x_0}\frac{f(x)-f(x_0)}{x-x_0} &= \lim_{x\to x_0}\frac{x^k - x_0^k}{x-x_0}\\ &= \lim_{x\to x_0}(x^{k-1} + x^{k-1}x_0 + \cdots + xx_0^{k-2} + x_0^{k-1}) = kx_0^{k-1},\end{aligned}$$
as desired. Since f' is obviously continuous, we obtain that f is continuously differentiable, as desired.
2. Let $I = [0,1]$ and define $f\colon I \to \mathbb{R}$ by
$$f(x) = \begin{cases} x, & x \neq 0,\\ 1, & x = 0.\end{cases}$$
From Example 1 we know that f is continuously differentiable at points in $(0,1]$. We claim that f is not differentiable at $x = 0$. This will follow from Proposition 3.2.7 below, but let us show this here directly. We have
$$\lim_{x\to_I 0}\frac{f(x)-f(0)}{x-0} = \lim_{x\downarrow 0}\frac{x-1}{x} = -\infty.$$
Thus the limit does not exist, and so f is not differentiable at $x = 0$, albeit in a fairly stupid way.
3. Let $I = [0,1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = \sqrt{x(1-x)}$. We claim that f is differentiable at points in $(0,1)$, but is not differentiable at $x = 0$ or $x = 1$. Providing that one believes that the function $x \mapsto \sqrt{x}$ is differentiable on $\mathbb{R}_{>0}$ (see Section 3.8.3, then the continuous differentiability of f on $(0,1)$ follows

from the results of Section 3.2.3. Moreover, the derivative of f at $x \in (0,1)$ can be explicitly computed as

$$f'(x) = \frac{1-2x}{2\sqrt{x(1-x)}}.$$

To show that f is not differentiable at $x = 0$ we compute

$$\lim_{x \to_I 0} \frac{f(x) - f(0)}{x - 0} = \lim_{x \downarrow 0} \frac{\sqrt{1-x}}{\sqrt{x}} = \infty.$$

Similarly, at $x = 1$ we compute

$$\lim_{x \to_I 1} \frac{f(x) - f(1)}{x - 1} = \lim_{x \uparrow 1} \frac{-\sqrt{x}}{\sqrt{x-1}} = -\infty.$$

Since neither of these limits are elements of $\mathbb{R}$, it follows that f is not differentiable at $x = 0$ or $x = 1$.

4. Let $I = \mathbb{R}$ and define $f \colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

We first claim that f is differentiable. The differentiability of f at points $x \in \mathbb{R}\setminus\{0\}$ will follow from our results in Section 3.2.3 concerning differentiability, and algebraic operations along with composition. Indeed, using these rules for differentiation we compute that for $x \neq 0$ we have

$$f'(x) = 2x \sin \tfrac{1}{x} - \cos \tfrac{1}{x}.$$

Next let us prove that f is differentiable at $x = 0$ and that $f'(0) = 0$. We have

$$\lim_{x \to 0} \frac{f(x) - f(x)}{x - 0} = \lim_{x \to 0} x \sin \tfrac{1}{x}.$$

Now let $\epsilon \in \mathbb{R}_{>0}$. Then, for $\delta = \epsilon$ we have

$$\left| x \sin \tfrac{1}{x} - 0 \right| < \epsilon$$

since $\left|\sin \frac{1}{x}\right| \leq 1$. This shows that $f'(0) = 0$, as claimed. This shows that f is differentiable.

However, we claim that f is not *continuously* differentiable. Clearly there are no problems away from $x = 0$, again by the results of Section 3.2.3. But we note that f' is discontinuous at $x = 0$. Indeed, we note that f is the sum of two functions, one ($x \sin \frac{1}{x}$) of which goes to zero as x goes to zero, and the other ($-\cos\frac{1}{x}$) of which, when evaluated in any neighbourhood of $x = 0$, takes all possible values in the interval $[-1, 1]$. This means that in any sufficiently small neighbourhood of $x = 0$, the function f' will take all possible values in the interval $[-\frac{1}{2}, \frac{1}{2}]$. This precludes the limit $\lim_{x\to 0} f'(x)$ from existing, and so precludes f' from being continuous at $x = 0$ by Theorem 3.1.3. •

Let us give some intuition about the derivative. Given an interval I and functions $f, g\colon I \to \mathbb{R}$, we say that f and g are ***tangent*** at $x_0 \in \mathbb{R}$ if

$$\lim_{x\to_I x_0} \frac{f(x) - g(x)}{x - x_0} = 0.$$

In Figure 3.5 we depict the idea of two functions being tangent. Using this idea,

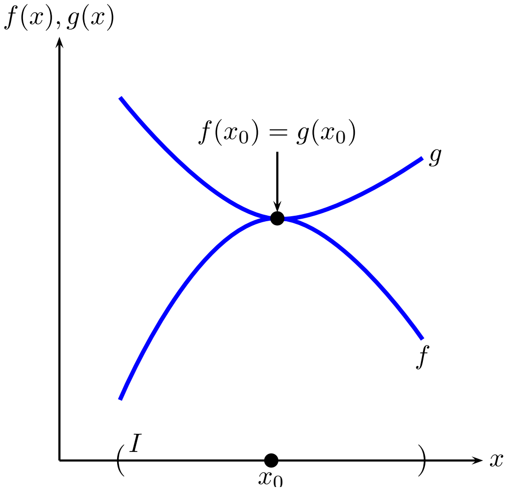


Figure 3.5 Functions that are tangent

we can give the following interpretation of the derivative.

3.2.4 Proposition (Derivative and linear approximation) *Let* $\mathrm{I} \subseteq \mathbb{R}$, *let* $\mathrm{x}_0 \in \mathrm{I}$, *and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function. Then there exists at most one number* $\alpha \in \mathbb{R}$ *such that* f *is tangent at* x_0 *with the function* $\mathrm{x} \mapsto \mathrm{f}(\mathrm{x}_0) + \alpha(\mathrm{x} - \mathrm{x}_0)$. *Moreover, such a number* α *exists if and only if* f *is differentiable at* x_0, *in which case* $\alpha = \mathrm{f}'(\mathrm{x}_0)$.

Proof Suppose there are two such numbers α_1 and α_2. Thus

$$\lim_{x\to_I x_0} \frac{f(x) - (f(x_0) + \alpha_j(x - x_0))}{x - x_0} = 0, \qquad j \in \{1, 2\}, \tag{3.7}$$

We compute

$$\begin{aligned}|\alpha_1 - \alpha_2| &= \frac{|\alpha_1(x - x_0) - \alpha_2(x - x_0)|}{|x - x_0|} \\ &= \frac{|-f(x) + f(x_0) + \alpha_1(x - x_0) + f(x) - f(x_0) - \alpha_2(x - x_0)|}{|x - x_0|} \\ &\le \frac{|f(x) - f(x_0) - \alpha_1(x - x_0)|}{|x - x_0|} + \frac{|f(x) - f(x_0) - \alpha_2(x - x_0)|}{|x - x_0|}.\end{aligned}$$

Since α_1 and α_2 satisfy (3.7), as we let $x \to x_0$ the right-hand side goes to zero showing that $|\alpha_1 - \alpha_2| = 0$. This proves the first part of the result.

Next suppose that there exists $\alpha \in \mathbb{R}$ such that

$$\lim_{x \to_I x_0} \frac{f(x) - (f(x_0) + \alpha(x - x_0))}{x - x_0} = 0.$$

It then immediately follows that

$$\lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} = \alpha.$$

Thus f is differentiable at x_0 with derivative equal to α. Conversely, if f is differentiable at x_0 then we have

$$\begin{aligned} & f'(x_0) = \lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0}, \\ \implies \quad & \lim_{x \to_I x_0} \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} = 0,\end{aligned}$$

which completes the proof. ■

The idea, then, is that the derivative serves, as we are taught in first-year calculus, as the best linear approximation to the function, since the function $x \mapsto f(x_0) + \alpha(x - x_0)$ is a linear function with slope α passing through $f(x_0)$.

We may also define derivatives of higher-order. Suppose that $f\colon I \to \mathbb{R}$ is differentiable, so that the function $f'\colon I \to \mathbb{R}$ can be defined. If the limit

$$\lim_{x \to_I x_0} \frac{f'(x) - f'(x_0)}{x - x_0}$$

exists, then we say that f is ***twice differentiable at*** $\mathbf{x_0}$. We denote the limit by $f''(x_0)$, and call it the ***second derivative*** of f at x_0. If f is differentiable at each point $x \in I$ then f is ***twice differentiable***. If additionally the map $x \mapsto f''(x)$ is continuous, then f is ***twice continuously differentiable***, or of ***class*** $\mathbf{C^2}$. Clearly this process can be continued inductively. Let us record the language coming from this iteration.

3.2.5 Definition (Higher-order derivatives) Let $I \subseteq \mathbb{R}$ be an interval, let $f\colon I \to \mathbb{R}$ be a function, let $r \in \mathbb{Z}_{>0}$, and suppose that f is $(r-1)$ times differentiable with g the corresponding $(r-1)$st derivative.

(i) The function f is ***r times differentiable at $\mathbf{x_0 \in I}$*** if the limit

$$\lim_{x \to_I x_0} \frac{g(x) - g(x_0)}{x - x_0} \tag{3.8}$$

exists.

(ii) If the limit (3.8) exists, then it is denoted by $f^{(r)}(x_0)$ and called the ***rth derivative*** of f at x_0.

(iii) If f is r times differentiable at each point $x \in I$, then f is ***r times differentiable***.

(iv) If f is r times differentiable and if the function $x \mapsto f^{(r)}(x)$ is continuous, then f is ***r times continuously differentiable***, or of ***class $\mathbf{C^r}$***.

If f is of class C^r for each $r \in \mathbb{Z}_{>0}$, then f is ***infinitely differentiable***, or of ***class $\mathbf{C^\infty}$***. •

3.2.6 Notation (Class $\mathbf{C^0}$) A continuous function will sometimes be said to be of ***class $\mathbf{C^0}$***, in keeping with the language used for functions that are differentiable to some order. •

3.2.2 The derivative and continuity

In this section we simply do two things. We show that differentiable functions are continuous (Proposition 3.2.7), and we (dramatically) show that the converse of this is not true (Example 3.2.9).

3.2.7 Proposition (Differentiable functions are continuous) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is a function differentiable at* $\mathrm{x_0} \in \mathrm{I}$*, then* f *is continuous at* $\mathrm{x_0}$*.*

Proof Using Propositions 2.3.23 and 2.3.29 the limit

$$\lim_{x \to_I x_0} \left(\frac{f(x) - f(x_0)}{x - x_0}\right)(x - x_0)$$

exists, and is equal to the product of the limits

$$\lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0}, \qquad \lim_{x \to_I x_0} (x - x_0),$$

i.e., is equal to zero. We therefore can conclude that

$$\lim_{x \to_I x_0} (f(x) - f(x_0)) = 0,$$

and the result now follows from Theorem 3.1.3. ■

If the derivative is bounded, then there is more that one can say.

3.2.8 Proposition (Functions with bounded derivative are uniformly continuous) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is differentiable with* $\mathrm{f}'\colon \mathrm{I} \to \mathbb{R}$ *bounded, then* f *is uniformly continuous.*

Proof Let

$$M = \sup\{f'(t) \mid t \in I\}.$$

Then, for every $x, y \in I$, by the Mean Value Theorem, Theorem 3.2.19 below, there exists $z \in [x, y]$ such that

$$f(x) - f(y) = f'(z)(x - y) \quad \implies \quad |f(x) - f(y)| \leq M||x - y||.$$

Now let $\epsilon \in \mathbb{R}_{>0}$ and let $x \in I$. Define $\delta = \frac{\epsilon}{M}$ and note that if $y \in I$ satisfies $|x - y| < \delta$ then we have

$$|f(x) - f(y)| \leq M||x - y|| \leq \epsilon,$$

giving the desired uniform continuity. ■

Of course, it is not true that a continuous function is differentiable; we have an example of this as Example 3.2.3–3. However, things are much worse than that, as the following example indicates.

3.2.9 Example (A continuous but nowhere differentiable function) For $k \in \mathbb{Z}_{>0}$ define $g_k\colon \mathbb{R} \to \mathbb{R}$ as shown in Figure 3.6. Thus g_k is periodic with period $4 \cdot 2^{-2^k}$.[3] We

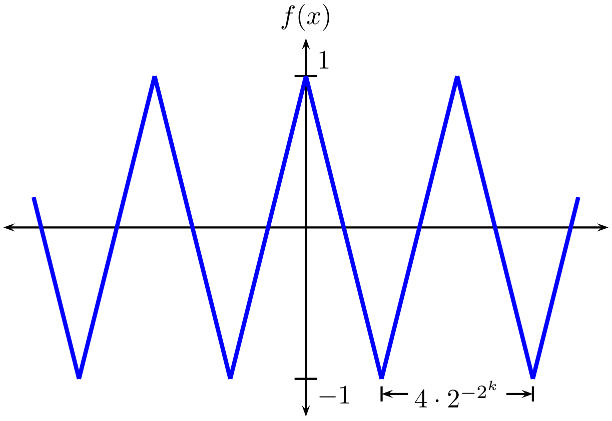


Figure 3.6 The function g_k

then define

$$f(x) = \sum_{k=1}^{\infty} 2^{-k} g_k(x).$$

[3] We have not yet defined what is meant by a periodic function, although this is likely clear. In case it is not, a function $f\colon \mathbb{R} \to \mathbb{R}$ is ***periodic*** with period $T \in \mathbb{R}_{>0}$ if $f(x + T) = f(x)$ for every $x \in \mathbb{R}$. Periodic functions will be discussed in some detail in Section IV-1.1.6.

Since g_k is bounded in magnitude by 1, and since the sum $\sum_{k=1}^{\infty} 2^{-k}$ is absolutely convergent (Example 2.4.2–4), for each x the series defining f converges, and so f is well-defined. We claim that f is continuous but is nowhere differentiable.

It is easily shown by the Weierstrass M-test (see Theorem 3.6.15 below) that the series converges uniformly, and so defines a continuous function in the limit by Theorem 3.6.8. Thus f is continuous.

Now let us show that f is nowhere differentiable. Let $x \in \mathbb{R}$, $k \in \mathbb{Z}_{>0}$, and choose $h_k \in \mathbb{R}$ such that $|h| = 2^{-2^k}$ and such that x and $x + h_k$ lie on the line segment in the graph of g_k (this is possible since h_k is small enough, as is easily checked). Let us prove a few lemmata for this choice of x and h_k.

1 **Lemma** $\mathrm{g_l(x + h_k) = g(x)}$ *for* $\mathrm{l > k}$.

Proof This follows since g_l is periodic with period $4 \cdot 2^{-2^l}$, and so is therefore also periodic with period 2^{-2^k} since

$$\frac{4 \cdot 2^{-2^l}}{2^{-2^k}} = 4 \cdot 2^{-2^l - 2^k} \in \mathbb{Z}$$

for $l > k$. ▼

2 **Lemma** $\mathrm{|g_k(x + h_k) - g_k(x)| = 1}$.

Proof This follows from the fact that we have chosen h_k such that x and $x+h_k$ lie on the same line segment in the graph of g_k, and from the fact that $|h_k|$ is one-quarter the period of g_k (cf. Figure 3.6). ▼

3 **Lemma** $\mathrm{\left|\sum_{j=1}^{k-1} 2^{-j} g_j(x + h_k) - \sum_{j=1}^{k-1} 2^{-j} g_j(x)\right| \leq 2^k 2^{-2^{k-1}}}$.

Proof We note that if x and $x + h_k$ are on the same line segment in the graph of g_k, then they are also on the same line segment of the graph of g_j for $j \in \{1, \dots, k\}$. Using this fact, along with the fact that the slope of the line segments of the function g_j have magnitude 2^{2^j}, we compute

$$\begin{aligned}
&\left|\sum_{j=1}^{k-1} 2^{-j} g_j(x + h_k) - \sum_{j=1}^{k-1} 2^{-j} g_j(x)\right| \\
&\qquad\qquad \leq (k-1)\max\{|2^{-j} g_j(x + h_k) - 2^{-j} g_j(x)| \mid j \in \{1, \dots, k\}\} \\
&\qquad\qquad = (k-1) 2^{2^{k-1}} 2^{-2^k} < 2^k 2^{-2^{k-1}}.
\end{aligned}$$

The final inequality follows since $k - 1 < 2^k$ for $k \geq 1$ and since $2^{2^{k-1}} 2^{-2^k} = 2^{-2^{k-1}}$. ▼

Now we can assemble these lemmata to give the conclusion that f is not differentiable at x. Let $x \in \mathbb{R}$, let $\epsilon \in \mathbb{R}_{>0}$, choose $k \in \mathbb{Z}_{>0}$ such that $2^{-2^k} < \epsilon$, and choose

h_k as above. We then have

$$\begin{aligned}\Big|\frac{f(x+h_k)-f(x)}{h_k}\Big| &= \Big|\frac{\sum_{j=1}^{\infty}2^{-j}g_j(x+h_k)-\sum_{j=1}^{\infty}2^{-j}g_j(x)}{h_k}\Big|\\ &= \Big|\frac{\sum_{j=1}^{k-1}2^{-j}g_j(x+h_k)-\sum_{j=1}^{k-1}2^{-j}g_j(x)}{h_k}+\frac{2^{-k}(g_k(x+h_k)-g_k(x))}{h_k}\Big|\\ &\geq 2^{-k}2^{2^k}-2^k2^{-2^{k-1}}.\end{aligned}$$

Since $\lim_{k\to\infty}(2^{-k}2^{2^k}-2^k2^{-2^{k-1}})=\infty$, it follows that any neighbourhood of x will contain a point y for which $\frac{f(y)-f(x)}{y-x}$ will be as large in magnitude as desired. This precludes f from being differentiable at x. Now, since x was arbitrary in our construction, we have shown that f is nowhere differentiable as claimed.

In Figure 3.7 we plot the function

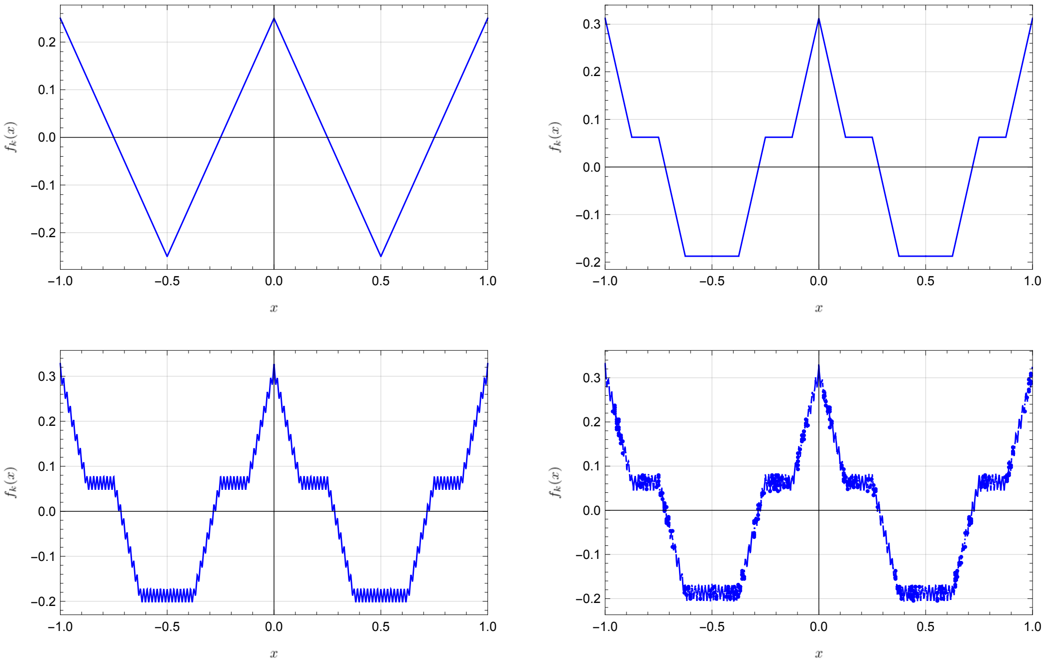


Figure 3.7 The first four partial sums for f

$$f_k(x)=\sum_{j=1}^{k}2^{-j}g_j(x)$$

for $j\in\{1,2,3,4\}$. Note that, to the resolution discernible by the eye, there is no difference between f_3 and f_4. However, if we were to magnify the scale, we would see the effects that lead to the limit function not being differentiable. •

3.2.3 The derivative and operations on functions

In this section we provide the rules for using the derivative in conjunction with the natural algebraic operations on functions as described at the beginning of Section 3.1.3. Most readers will probably be familiar with these ideas, at least inasmuch as how to use them in practice.

3.2.10 Proposition (The derivative, and addition and multiplication) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be functions differentiable at* $\mathrm{x}_0 \in \mathrm{I}$*. Then the following statements hold:*

(i) $\mathrm{f} + \mathrm{g}$ *is differentiable at* x_0 *and* $(\mathrm{f} + \mathrm{g})'(\mathrm{x}_0) = \mathrm{f}'(\mathrm{x}_0) + \mathrm{g}'(\mathrm{x}_0)$*;*

(ii) fg *is differentiable at* x_0 *and* $(\mathrm{fg})'(\mathrm{x}_0) = \mathrm{f}'(\mathrm{x}_0)\mathrm{g}(\mathrm{x}_0) + \mathrm{f}(\mathrm{x}_0)\mathrm{g}'(\mathrm{x}_0)$ *(**product rule** or **Leibniz'**[4] **rule**);*

(iii) if additionally $\mathrm{g}(\mathrm{x}_0) \neq 0$*, then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is differentiable at* x_0 *and*

$$\left(\frac{\mathrm{f}}{\mathrm{g}}\right)'(\mathrm{x}_0) = \frac{\mathrm{f}'(\mathrm{x}_0)\mathrm{g}(\mathrm{x}_0) - \mathrm{f}(\mathrm{x}_0)\mathrm{g}'(\mathrm{x}_0)}{\mathrm{g}(\mathrm{x}_0)^2} \quad \textit{(\textbf{quotient rule})}.$$

Proof (i) We have

$$\frac{(f+g)(x) - (f+g)(x_0)}{x - x_0} = \frac{f(x) - f(x_0)}{x - x_0} + \frac{g(x) - g(x_0)}{x - x_0}.$$

Now we may apply Propositions 2.3.23 and 2.3.29 to deduce that

$$\lim_{x \to_I x_0} \frac{(f+g)(x) - (f+g)(x_0)}{x - x_0} = \lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0} + \lim_{x \to_I x_0} \frac{g(x) - g(x_0)}{x - x_0} = f'(x_0) + g'(x_0),$$

as desired.

(ii) Here we note that

$$\begin{aligned}\frac{(fg)(x) - (fg)(x_0)}{x - x_0} &= \frac{f(x)g(x) - f(x)g(x_0) + f(x)g(x_0) - f(x_0)g(x_0)}{x - x_0} \\ &= f(x)\frac{g(x) - g(x_0)}{x - x_0} + g(x_0)\frac{f(x) - f(x_0)}{x - x_0}.\end{aligned}$$

Since f is continuous at x_0 by Proposition 3.2.7, we may apply Propositions 2.3.23 and 2.3.29 to conclude that

$$\lim_{x \to_I x_0} \frac{(fg)(x) - (fg)(x_0)}{x - x_0} = f'(x_0)g(x_0) + f(x_0)g'(x_0),$$

[4] Gottfried Wilhelm von Leibniz (1646–1716) was born in Leipzig (then a part of Saxony), and was a lawyer, philosopher, and mathematician. His main mathematical contributions were to the development of calculus, where he had a well-publicised feud over priority with Newton, and algebra. His philosophical contributions, mainly in the area of logic, were also of some note.

just as claimed.

(iii) By using part (ii), it suffices to consider the case where f is defined by $f(x) = 1$ (why?). Note that if $g(x_0) \neq 0$, then there is a neighbourhood of x_0 to which the restriction of g is nowhere zero. Thus, without loss of generality, we suppose that $g(x) \neq 0$ for all $x \in I$. But in this case we have

$$\lim_{x\to_I x_0} \frac{\frac{1}{g(x)} - \frac{1}{g(x_0)}}{x - x_0} = \lim_{x\to_I x_0} \frac{1}{g(x)g(x_0)} \frac{g(x_0)}{x - x_0} = -\frac{g'(x_0)}{g(x_0)^2},$$

giving the result in this case. We have used Propositions 2.3.23 and 2.3.29 as usual. ■

The following generalisation of the product rule will be occasionally useful.

3.2.11 Proposition (Higher-order product rule) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{x}_0 \in \mathrm{I}$*, let* $\mathrm{r} \in \mathbb{Z}_{>0}$*, and suppose that* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *are of class* $\mathrm{C}^{\mathrm{r}-1}$ *and are* r*-times differentiable at* x_0*. Then* fg *is* r*-times differentiable at* x_0*, and*

$$(\mathrm{fg})^{(\mathrm{r})}(\mathrm{x}_0) = \sum_{\mathrm{j}=0}^{\mathrm{r}} \binom{\mathrm{r}}{\mathrm{j}} \mathrm{f}^{(\mathrm{j})}(\mathrm{x}_0)\mathrm{g}^{(\mathrm{r}-\mathrm{j})}(\mathrm{x}_0),$$

where

$$\binom{\mathrm{r}}{\mathrm{j}} = \frac{\mathrm{r}!}{\mathrm{j}!(\mathrm{r}-\mathrm{j})!}.$$

Proof The result is true for $r = 1$ by Proposition 3.2.10. So suppose the result true for $k \in \{1, \dots, r\}$. We then have

$$\begin{aligned}\frac{(fg)^{(r)}(x) - (fg)^{(r)}(x_0)}{x - x_0} &= \frac{\sum_{j=0}^{r} \binom{r}{j} f^{(j)}(x)g^{(r-j)}(x) - \sum_{j=0}^{r} \binom{r}{j} f^{(j)}(x_0)g^{(r-j)}(x_0)}{x - x_0} \\ &= \sum_{j=0}^{r} \binom{r}{j} \frac{f^{(j)}(x)g^{(r-j)}(x) - f^{(j)}(x_0)g^{(r-j)}(x_0)}{x - x_0}.\end{aligned}$$

Now we note that

$$\lim_{x\to_I x_0} \frac{f^{(j)}(x)g^{(r-j)}(x) - f^{(j)}(x_0)g^{(r-j)}(x_0)}{x - x_0} = f^{(j+1)}(x_0)g^{(r-j)}(x_0) + f^{(j)}(x_0)g^{(r-j+1)}(x_0).$$

Therefore,

$$
\begin{aligned}
\lim_{x \to_I x_0} & \frac{(fg)^{(r)}(x) - (fg)^{(r)}(x_0)}{x - x_0} \\
&= \sum_{j=0}^{r} \binom{r}{j} \big(f^{(j+1)}(x_0) g^{(r-j)}(x_0) + f^{(j)}(x_0) g^{(r-j+1)}(x_0) \big) \\
&= f(x_0) g^{(r+1)}(x_0) + \sum_{j=0}^{r} \binom{r}{j} f^{(j+1)}(x_0) g^{(r-j)}(x_0) + \sum_{j=1}^{r} \binom{r}{j} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= f(x_0) g^{(r+1)}(x_0) + \sum_{j=1}^{r+1} \binom{r}{j-1} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&\quad + \sum_{j=1}^{r} \binom{r}{j} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= f^{(r+1)}(x_0) g(x_0) + f(x_0) g^{(r+1)}(x_0) \\
&\quad + \sum_{j=1}^{r} \left(\binom{r}{j} + \binom{r}{j-1} \right) f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= f^{(r+1)}(x_0) g(x_0) + f(x_0) g^{(r+1)}(x_0) + \sum_{j=1}^{r} \binom{r+1}{j} f^{(j)}(x_0) g^{(r-j+1)}(x_0) \\
&= \sum_{j=0}^{r+1} \binom{r+1}{j} f^{(j)}(x_0) g^{(r-j)}(x_0).
\end{aligned}
$$

In the penultimate step we have used ***Pascal's***[5] ***Rule*** which states that

$$\binom{r}{j} + \binom{r}{j-1} = \binom{r+1}{j}.$$

We leave the direct proof of this fact to the reader. ■

The preceding two results had to do with differentiability at a point. For convenience, let us record the corresponding results when we consider the derivative, not just at a point, but on the entire interval.

3.2.12 Proposition (Class C^r, and addition and multiplication) *Let* $I \subseteq \mathbb{R}$ *be an interval and let* $f, g \colon I \to \mathbb{R}$ *be functions of class* C^r*. Then the following statements hold:*

(i) $f + g$ *is of class* C^r*;*

(ii) fg *is of class* C^r*;*

(iii) if additionally $g(x) \neq 0$ *for all* $x \in I$*, then* $\frac{f}{g}$ *is of class* C^r*.*

[5] Blaise Pascal (1623–1662) was a French mathematician and philosopher. Much of his mathematical work was on analytic geometry and probability theory.

Proof This follows directly from Propositions 3.2.10 and 3.2.11, along with the fact, following from Proposition 3.1.15, that the expressions for the derivatives of sums, products, and quotients are continuous, as they are themselves sums, products, and quotients. ■

The following rule for differentiating the composition of functions is one of the more useful of the rules concerning the behaviour of the derivative.

3.2.13 Theorem (Chain Rule) *Let* $I, J \subseteq \mathbb{R}$ *be intervals and let* $f\colon I \to J$ *and* $g\colon J \to \mathbb{R}$ *be functions for which* f *is differentiable at* $x_0 \in I$ *and* g *is differentiable at* $f(x_0) \in J$*. Then* $g \circ f$ *is differentiable at* x_0*, and* $(g \circ f)'(x_0) = g'(f(x_0))f'(x_0)$*.*

Proof Let us define $h\colon J \to \mathbb{R}$ by

$$h(y) = \begin{cases} \frac{g(y)-g(f(x_0))}{y-f(x_0)}, & g(y) \neq g(f(x_0)), \\ g'(f(x_0)), & g(y) = g(f(x_0)). \end{cases}$$

We have

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = \frac{(g \circ f)(x) - (g \circ f)(x_0)}{f(x) - f(x_0)} \frac{f(x) - f(x_0)}{x - x_0} = h(f(x))\frac{f(x) - f(x_0)}{x - x_0},$$

provided that $f(x) \neq f(x_0)$. On the other hand, if $f(x) = f(x_0)$, we immediately have

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = h(f(x))\frac{f(x) - f(x_0)}{x - x_0}$$

since both sides of this equation are zero. Thus we simply have

$$\frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = h(f(x))\frac{f(x) - f(x_0)}{x - x_0}$$

for all $x \in I$. Note that h is continuous at $f(x_0)$ by Theorem 3.1.3 since

$$\lim_{y \to_I f(x_0)} h(y) = g'(x_0) = h(x_0),$$

using the fact that g is differentiable at x_0. Now we can use Propositions 2.3.23 and 2.3.29 to ascertain that

$$\lim_{x \to_I x_0} \frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = \lim_{x \to_I x_0} h(f(x))\frac{f(x) - f(x_0)}{x - x_0} = g'(f(x_0))f'(x_0),$$

as desired. ■

The derivative behaves as one would expect when restricting a differentiable function.

3.2.14 Proposition (The derivative and restriction) *If* $\mathrm{I}, \mathrm{J} \subseteq \mathbb{R}$ *are intervals for which* $\mathrm{J} \subseteq \mathrm{I}$*, and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is differentiable at* $\mathrm{x}_0 \in \mathrm{J} \subseteq \mathrm{I}$*, then* $\mathrm{f}|\mathrm{J}$ *is differentiable at* x_0*.*

Proof This follows since if the limit

$$\lim_{x \to_I x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

exists, then so too does the limit

$$\lim_{x \to_J x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

provided that $J \subseteq I$. ■

3.2.4 The derivative and function behaviour

From the behaviour of the derivative of a function, it is often possible to deduce some important features of the function itself. One of the most important of these concerns maxima and minima of a function. Let us define these concepts precisely.

3.2.15 Definition (Local maximum and local minimum) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function. A point $x_0 \in I$ is a:

(i) ***local maximum*** if there exists a neighbourhood U of x_0 such that $f(x) \leq f(x_0)$ for every $x \in U$;

(ii) ***strict local maximum*** if there exists a neighbourhood U of x_0 such that $f(x) < f(x_0)$ for every $x \in U \setminus \{x_0\}$;

(iii) ***local minimum*** if there exists a neighbourhood U of x_0 such that $f(x) \geq f(x_0)$ for every $x \in U$;

(iv) ***strict local minimum*** if there exists a neighbourhood U of x_0 such that $f(x) > f(x_0)$ for every $x \in U \setminus \{x_0\}$. •

Now we have the standard result that relates derivatives to maxima and minima.

3.2.16 Theorem (Derivatives, and maxima and minima) *For* $\mathrm{I} \subseteq \mathbb{R}$ *an interval,* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *a function, and* $\mathrm{x}_0 \in \mathrm{int}(\mathrm{I})$*, the following statements hold:*

(i) if f *is differentiable at* x_0 *and if* x_0 *is a local maximum or a local minimum for* f*, then* $\mathrm{f}'(\mathrm{x}_0) = 0$*;*

(ii) if f *is twice differentiable at* x_0*, and if* x_0 *is a local maximum (resp. local minimum) for* f*, then* $\mathrm{f}''(\mathrm{x}_0) \leq 0$ *(resp.* $\mathrm{f}''(\mathrm{x}_0) \geq 0$*);*

(iii) if f *is twice differentiable at* x_0*, and if* $\mathrm{f}'(\mathrm{x}_0) = 0$ *and* $\mathrm{f}''(\mathrm{x}_0) \in \mathbb{R}_{<0}$ *(resp.* $\mathrm{f}''(\mathrm{x}_0) \in \mathbb{R}_{>0}$*), then* x_0 *is a strict local maximum (resp. strict local minimum) for* f*.*

Proof (i) We will prove the case where x_0 is a local minimum, since the case of a local maximum is similar. If x_0 is a local minimum, then there exists $\epsilon \in \mathbb{R}_{>0}$ such that

$f(x) \geq f(x_0)$ for all $x \in \mathsf{B}(\epsilon, x_0)$. Therefore, $\frac{f(x)-f(x_0)}{x-x_0} \geq 0$ for $x \geq x_0$ and $\frac{f(x)-f(x_0)}{x-x_0} \leq 0$ for $x \leq x_0$. Since the limit $\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$ exists, it must be equal to both limits

$$\lim_{x\downarrow x_0} \frac{f(x)-f(x_0)}{x-x_0}, \quad \lim_{x\uparrow x_0} \frac{f(x)-f(x_0)}{x-x_0}.$$

However, since the left limit is nonnegative and the right limit is nonpositive, we conclude that $f'(x_0) = 0$.

(ii) We shall show that if f is twice differentiable at x_0 and $f''(x_0)$ is not less than or equal to zero, then x_0 is not a local maximum. The statement concerning the local minimum is argued in the same way. Now, if f is twice differentiable at x_0, and if $f''(x_0) \in \mathbb{R}_{>0}$, then x_0 is a local minimum by part (iii), which prohibits it from being a local maximum.

(iii) We consider the case where $f''(x_0) \in \mathbb{R}_{>0}$, since the other case follows in the same manner. Choose $\epsilon \in \mathbb{R}_{>0}$ such that, for $x \in \mathsf{B}(\epsilon, x_0)$,

$$\left|\frac{f'(x)-f'(x_0)}{x-x_0} - f''(x_0)\right| < \tfrac{1}{2} f''(x_0),$$

this being possible since $f''(x_0) > 0$ and since f is twice differentiable at x_0. Since $f''(x_0) > 0$ it follows that, for $x \in \mathsf{B}(\epsilon, x_0)$,

$$\frac{f'(x)-f'(x_0)}{x-x_0} > 0,$$

from which we conclude that $f'(x) > 0$ for $x \in (x_0, x_0+\epsilon)$ and that $f'(x) < 0$ for $x \in (x_0-\epsilon, x_0)$. Now we prove a technical lemma.

1 **Lemma** *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an open interval, let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a continuous function that is differentiable, except possibly at* $\mathrm{x}_0 \in \mathrm{I}$*. If* $\mathrm{f}'(\mathrm{x}) > 0$ *for every* $\mathrm{x} > \mathrm{x}_0$ *and if* $\mathrm{f}'(\mathrm{x}) < 0$ *for every* $\mathrm{x} < \mathrm{x}_0$*, then* x_0 *is a strict local minimum for* f.

Proof We will use the Mean Value Theorem (Theorem 3.2.19) which we prove below. Note that our proof of the Mean Value Theorem depends on part (i) of the present theorem, but not on part that we are now proving. Let $x \in I \setminus \{x_0\}$. We have two cases.

1. $x > x_0$: By the Mean Value Theorem there exists $a \in (x, x_0)$ such that $f(x) - f(x_0) = (x - x_0)f'(a)$. Since $f'(a) > 0$ it then follows that $f(x) > f(x_0)$.
2. $x < x_0$: A similar argument as in the previous case again gives $f(x) > f(x_0)$.

Combining these conclusions, we see that $f(x) > f(x_0)$ for all $x \in I$, and so x_0 is a strict local maximum for f. ▼

The lemma now immediately applies to the restriction of f to $\mathsf{B}(\epsilon, x_0)$, and so gives the result. ■

Let us give some examples that illustrate the value and limitations of the preceding result.

3.2.17 Examples (Derivatives, and maxima and minima)

1. Let $I = \mathbb{R}$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x^2$. Note that f is infinitely differentiable, so Theorem 3.2.16 can be applied freely. We compute $f'(x) = 2x$, and so $f'(x) = 0$ if and only if $x = 0$. Therefore, the only local maxima and local minima must occur at $x = 0$. To check whether a local maxima, a local minima, or neither exists at $x = 0$, we compute the second derivative which is $f''(x) = 2$. This is positive at $x = 0$ (and indeed everywhere), so we may conclude that $x = 0$ is a strict local maximum for f from part (iii) of the theorem.

   Applying the same computations to $g(x) = -x^2$ shows that $x = 0$ is a strict local maximum for g.
2. Let $I = \mathbb{R}$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x^3$. We compute $f'(x) = 3x^2$, from which we ascertain that all maxima and minima must occur, if at all, at $x = 0$. However, since $f''(x) = 6x$, $f''(0) = 0$, and we cannot conclude from Theorem 3.2.16 whether there is a local maximum, a local minimum, or neither at $x = 0$. In fact, one can see "by hand" that $x = 0$ is neither a local maximum nor a local minimum for f.

   The same arguments apply to the functions $g(x) = x^4$ and $h(x) = -x^4$ to show that when the second derivative vanishes, it is possible to have all possibilities—a local maximum, a local minimum, or neither—at a point where both f' and f'' are zero.
3. Let $I = [-1, 1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1 - x, & x \in [0,1], \\ 1 + x, & x \in [-1, 0). \end{cases}$$

   "By hand," one can check that f has a strict local maximum at $x = 0$, and strict local minima at $x = -1$ and $x = 1$. However, we can detect none of these using Theorem 3.2.16. Indeed, the local minima at $x = -1$ and $x = 1$ occur at the boundary of I, and so the hypotheses of the theorem do not apply. This, indeed, is why we demand that x_0 lie in $\mathrm{int}(I)$ in the theorem statement. For the local maximum at $x = 0$, the theorem does not apply since f is not differentiable at $x = 0$. However, we do note that Lemma 1 (with modifications to the signs of the derivative in the hypotheses, and changing "minimum" to "maximum" in the conclusions) in the proof of the theorem *does* apply, since f is differentiable at points in $(-1, 0)$ and $(0, 1)$, and for $x > 0$ we have $f'(x) < 0$ and for $x < 0$ we have $f'(x) > 0$. The lemma then allows us to conclude that f has a strict local maximum at $x = 0$. •

Next let us prove a simple result that, while not always of great value itself, leads to the important Mean Value Theorem below.

3.2.18 Theorem (Rolle's[6] Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be continuous,*

[6] Michel Rolle (1652–1719) was a French mathematician whose primary contributions were to algebra.

and suppose that for $a, b \in I$ *it holds that* $f|(a, b)$ *is differentiable and that* $f(a) = f(b)$*. Then there exists* $c \in (a, b)$ *such that* $f'(c) = 0$*.*

Proof Since $f|[a, b]$ is continuous, by Theorem 3.1.23 there exists $x_1, x_2 \in [a, b]$ such that $\operatorname{image}(f|[a, b]) = [f(x_1), f(x_2)]$. We have three cases to consider.

1. $x_1, x_2 \in \operatorname{bd}([a, b])$: In this case it holds that f is constant since $f(a) = f(b)$. Thus the conclusions of the theorem hold for any $c \in (a, b)$.
2. $x_1 \in \operatorname{int}([a, b])$: In this case, f has a local minimum at x_1, and so by Theorem 3.2.16(i) we conclude that $f'(x_1) = 0$.
3. $x_2 \in \operatorname{int}([a, b])$: In this case, f has a local maximum at x_2, and so by Theorem 3.2.16(i) we conclude that $f'(x_2) = 0$. ■

Rolle's Theorem has the following generalisation, which is often quite useful, since it establishes links between the values of a function and the values of its derivative.

3.2.19 Theorem (Mean Value Theorem) *Let* $I \subseteq \mathbb{R}$ *be an interval, let* $f\colon I \to \mathbb{R}$ *be continuous, and suppose that for* $a, b \in I$ *it holds that* $f|(a, b)$ *is differentiable. Then there exists* $c \in (a, b)$ *such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Proof Define $g\colon I \to \mathbb{R}$ by

$$g(x) = f(x) - \frac{f(b) - f(a)}{b - a}(x - a).$$

Using the results of Section 3.2.3 we conclude that g is continuous and differentiable on (a, b). Moreover, direct substitution shows that $g(b) = g(a)$. Thus Rolle's Theorem allows us to conclude that there exists $c \in (a, b)$ such that $g'(c) = 0$. However, another direct substitution shows that $g'(c) = f'(c) - \frac{f(b)-f(a)}{b-a}$. ■

In Figure 3.8 we give the intuition for Rolle's Theorem, the Mean Value Theo-

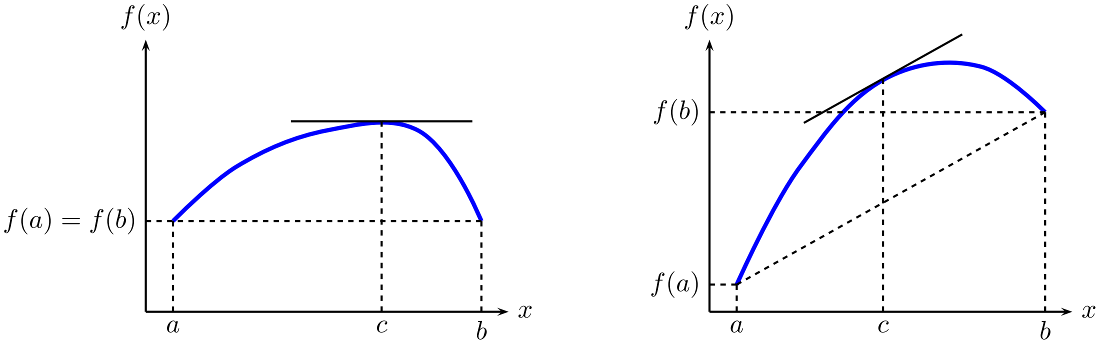


Figure 3.8 Illustration of Rolle's Theorem (left) and the Mean Value Theorem (right)

rem, and the relationship between the two results.

Another version of the Mean Value Theorem relates the values of two functions with the values of their derivatives.

3.2.20 Theorem (Cauchy's Mean Value Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be continuous, and suppose that for* $\mathrm{a}, \mathrm{b} \in \mathrm{I}$ *it holds that* $\mathrm{f}|(\mathrm{a},\mathrm{b})$ *and* $\mathrm{g}|(\mathrm{a},\mathrm{b})$ *are differentiable, and that* $\mathrm{g}'(\mathrm{x}) \neq 0$ *for each* $\mathrm{x} \in (\mathrm{a},\mathrm{b})$*. Then there exists* $\mathrm{c} \in (\mathrm{a},\mathrm{b})$ *such that*

$$\frac{\mathrm{f}'(\mathrm{c})}{\mathrm{g}'(\mathrm{c})} = \frac{\mathrm{f}(\mathrm{b}) - \mathrm{f}(\mathrm{a})}{\mathrm{g}(\mathrm{b}) - \mathrm{g}(\mathrm{a})}.$$

Proof Note that $g(b) \neq g(a)$ by Rolle's Theorem, since $g'(x) \neq 0$ for $x \in \mathrm{int}(a,b)$. Let

$$\alpha = \frac{f(b) - f(a)}{g(b) - g(a)}$$

and define $h\colon I \to \mathbb{R}$ by $h(x) = f(x) - \alpha g(x)$. Using the results of Section 3.2.3, one verifies that h is continuous on I and differentiable on (a,b). Moreover, one can also verify that $h(a) = h(b)$. Thus Rolle's Theorem implies the existence of $c \in (a,b)$ for which $h'(c) = 0$. A simple computation verifies that $h'(c) = 0$ is equivalent to the conclusion of the theorem. ■

We conclude this section with the useful L'Hôpital's Rule. This rule for finding limits is sufficiently useful that we state and prove it here in an unusual level of generality.

3.2.21 Theorem (L'Hôpital's[7] Rule) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{x}_0 \in \mathbb{R}$*, and let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be differentiable functions with* $\mathrm{g}'(\mathrm{x}) \neq 0$ *for all* $\mathrm{x} \in \mathrm{I} - \{\mathrm{x}_0\}$*. Then the following statements hold.*

(i) Suppose that x_0 *is an open right endpoint for* I *and suppose that either*

(a) $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \mathrm{f}(\mathrm{x}) = 0$ *and* $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \mathrm{g}(\mathrm{x}) = 0$ *or*

(b) $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \mathrm{f}(\mathrm{x}) = \infty$ *and* $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \mathrm{g}(\mathrm{x}) = \infty$*,*

and suppose that $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \frac{\mathrm{f}'(\mathrm{x})}{\mathrm{g}'(\mathrm{x})} = \mathrm{s}_0 \in \overline{\mathbb{R}}$*. Then* $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \frac{\mathrm{f}(\mathrm{x})}{\mathrm{g}(\mathrm{x})} = \mathrm{s}_0$*.*

(ii) Suppose that x_0 *is an left right endpoint for* I *and suppose that either*

(a) $\lim_{\mathrm{x}\downarrow\mathrm{x}_0} \mathrm{f}(\mathrm{x}) = 0$ *and* $\lim_{\mathrm{x}\downarrow\mathrm{x}_0} \mathrm{g}(\mathrm{x}) = 0$ *or*

(b) $\lim_{\mathrm{x}\uparrow\mathrm{x}_0} \mathrm{f}(\mathrm{x}) = \infty$ *and* $\lim_{\mathrm{x}\downarrow\mathrm{x}_0} \mathrm{g}(\mathrm{x}) = \infty$*,*

and suppose that $\lim_{\mathrm{x}\downarrow\mathrm{x}_0} \frac{\mathrm{f}'(\mathrm{x})}{\mathrm{g}'(\mathrm{x})} = \mathrm{s}_0 \in \overline{\mathbb{R}}$*. Then* $\lim_{\mathrm{x}\downarrow\mathrm{x}_0} \frac{\mathrm{f}(\mathrm{x})}{\mathrm{g}(\mathrm{x})} = \mathrm{s}_0$*.*

(iii) Suppose that $\mathrm{x}_0 \in \mathrm{int}(\mathrm{I})$ *and suppose that either*

(a) $\lim_{\mathrm{x}\to\mathrm{x}_0} \mathrm{f}(\mathrm{x}) = 0$ *and* $\lim_{\mathrm{x}\to\mathrm{x}_0} \mathrm{g}(\mathrm{x}) = 0$ *or*

(b) $\lim_{\mathrm{x}\to\mathrm{x}_0} \mathrm{f}(\mathrm{x}) = \infty$ *and* $\lim_{\mathrm{x}\to\mathrm{x}_0} \mathrm{g}(\mathrm{x}) = \infty$*,*

[7] Guillaume François Antoine Marquis de L'Hôpital (1661–1704) was one of the early developers of calculus.

and suppose that $\lim_{x\to x_0}\frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. *Then* $\lim_{x\to x_0}\frac{f(x)}{g(x)} = s_0$.

The following two statements which are independent of x_0 *(thus we ask that* $g'(x) \neq 0$ *for all* $x \in I$*) also hold.*

(iv) *Suppose that* I *is unbounded on the right and suppose that either*

(a) $\lim_{x\to\infty} f(x) = 0$ *and* $\lim_{x\to\infty} g(x) = 0$ *or*

(b) $\lim_{x\to\infty} f(x) = \infty$ *and* $\lim_{x\to\infty} g(x) = \infty$,

and suppose that $\lim_{x\to\infty}\frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. *Then* $\lim_{x\to\infty}\frac{f(x)}{g(x)} = s_0$.

(v) *Suppose that* I *is unbounded on the left and suppose that either*

(a) $\lim_{x\to-\infty} f(x) = 0$ *and* $\lim_{x\to-\infty} g(x) = 0$ *or*

(b) $\lim_{x\to-\infty} f(x) = \infty$ *and* $\lim_{x\to-\infty} g(x) = \infty$,

and suppose that $\lim_{x\to-\infty}\frac{f'(x)}{g'(x)} = s_0 \in \overline{\mathbb{R}}$. *Then* $\lim_{x\to-\infty}\frac{f(x)}{g(x)} = s_0$.

Proof (i) First suppose that $\lim_{x\uparrow x_0} f(x) = 0$ and $\lim_{x\uparrow x_0} g(x) = 0$ and that $s_0 \in \mathbb{R}$. We may then extend f and g to be defined at x_0 by taking their values at x_0 to be zero, and the resulting function will be continuous by Theorem 3.1.3. We may now apply Cauchy's Mean Value Theorem to assert that for $x \in I$ there exists $c_x \in (x, x_0)$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x_0) - f(x)}{g(x_0) - g(x)} = \frac{f(x)}{g(x)}.$$

Now let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta \in \mathbb{R}_{>0}$ such that $\left|\frac{f'(x)}{g'(x)} - s_0\right| < \epsilon$ for $x \in \mathsf{B}(\delta, x_0) \cap I$. Then, for $x \in \mathsf{B}(\delta, x_0) \cap I$ we have

$$\left|\frac{f(x)}{g(x)} - s_0\right| = \left|\frac{f'(c_x)}{g'(c_x)} - s_0\right| < \epsilon$$

since $c_x \in \mathsf{B}(\delta, x_0) \cap I$. This shows that $\lim_{x\uparrow x_0}\frac{f(x)}{g(x)} = s_0$, as claimed.

Now suppose that $\lim_{x\uparrow x_0} f(x) = \infty$ and $\lim_{x\uparrow x_0} g(x) = \infty$ and that $s_0 \in \mathbb{R}$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta_1 \in \mathbb{R}_{>0}$ such that $\left|\frac{f'(x)}{g'(x)} - s_0\right| < \frac{\epsilon}{2(1+|s_0|)}$ for $x \in \mathsf{B}(\delta_1, x_0) \cap I$. For $x \in \mathsf{B}(\delta_1, x_0) \cap I$, by Cauchy's Mean Value Theorem there exists $c_x \in \mathsf{B}(\delta_1, x_0) \cap I$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x) - f(x-\delta_1)}{g(x) - g(x-\delta_1)}.$$

Therefore,

$$\left|\frac{f(x) - f(x-\delta_1)}{g(x) - g(x-\delta_1)} - s_0\right| < \frac{\epsilon}{2(1+|s_0|)}$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$. Now define

$$h(x) = \frac{1 - \frac{f(x-\delta_1)}{f(x)}}{1 - \frac{g(x-\delta_1)}{g(x)}}$$

and note that

$$\frac{f(x) - f(x-\delta_1)}{g(x) - g(x-\delta_1)} = h(x)\frac{f(x)}{g(x)}.$$

Therefore we have

$$\left|h(x)\frac{f(x)}{g(x)} - s_0\right| < \frac{\epsilon}{2(1+|s_0|)}$$

for $x \in \mathsf{B}(\delta_1, x_0) \cap I$. Note also that $\lim_{x\uparrow x_0} h(x) = 1$. Thus we can choose $\delta_2 \in \mathbb{R}_{>0}$ such that $|h(x) - 1| < \frac{\epsilon}{2(1+|s_0|)}$ and $h(x) > \frac{1}{2}$ for $x \in \mathsf{B}(\delta_2, x_0) \cap I$. Then define $\delta = \min\{\delta_1, \delta_2\}$. For $x \in \mathsf{B}(\delta, x_0) \cap I$ we then have

$$\begin{aligned}\left|h(x)\left(\frac{f(x)}{g(x)} - s_0\right)\right| &= \left|h(x)\frac{f(x)}{g(x)} - h(x)s_0\right| \\ &\le \left|h(x)\frac{f(x)}{g(x)} - s_0\right| + |(1-h(x))s_0| \\ &< \frac{\epsilon}{2(1+|s_0|)} + \frac{\epsilon}{2(1+|s_0|)}|s_0| = \frac{\epsilon}{2}.\end{aligned}$$

Then, finally,

$$\left|\frac{f(x)}{g(x)} - s_0\right| < \frac{\epsilon}{2h(x)} < \epsilon,$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$.

Now we consider the situation when $s_0 \in \{-\infty, \infty\}$. We shall take only the case of $s_0 = \infty$ since the other follows in a similar manner. We first take the case where $\lim_{x\uparrow x_0} f(x) = 0$ and $\lim_{x\uparrow x_0} g(x) = 0$. In this case, for $x \in I$, from the Cauchy Mean Value Theorem we can find $c_x \in (x, x_0)$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x)}{g(x)}.$$

Now for $M \in \mathbb{R}_{>0}$ we choose $\delta \in \mathbb{R}_{>0}$ such that for $x \in \mathsf{B}(\delta, x_0) \cap I$ we have $\frac{f'(x)}{g'(x)} > M$. Then we immediately have

$$\frac{f(x)}{g(x)} = \frac{f'(c_x)}{g'(c_x)} > M$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$ since $c_x \in \mathsf{B}(\delta, x_0)$, which gives the desired conclusion.

The final case we consider in this part of the proof is that where $s_0 = \infty$ and $\lim_{x\uparrow x_0} f(x) = \infty$ and $\lim_{x\uparrow x_0} g(x) = \infty$. For $M \in \mathbb{R}_{>0}$ choose $\delta_1 \in \mathbb{R}_{>0}$ such that $\frac{f'(x)}{g'(x)} > 2M$ provided that $x \in \mathsf{B}(\delta_1, x_0) \cap I$. Then, using Cauchy's Mean Value Theorem, for $x \in \mathsf{B}(\delta_1, x_0) \cap I$ there exists $c_x \in \mathsf{B}(\delta_1, x_0)$ such that

$$\frac{f'(c_x)}{g'(c_x)} = \frac{f(x) - f(x-\delta_1)}{g(x) - g(x-\delta_1)}.$$

Therefore,

$$\frac{f(x) - f(x-\delta_1)}{g(x) - g(x-\delta_1)} > 2M$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$. As above, define

$$h(x) = \frac{1 - \frac{f(x-\delta_1)}{f(x)}}{1 - \frac{g(x-\delta_1)}{g(x)}}$$

and note that

$$\frac{f(x)-f(x-\delta_1)}{g(x)-g(x-\delta_1)} = h(x)\frac{f(x)}{g(x)}.$$

Therefore

$$h(x)\frac{f(x)}{g(x)} > 2M$$

for $x \in \mathsf{B}(\delta_1, x_0)$. Now take $\delta_2 \in \mathbb{R}_{>0}$ such that, if $x \in \mathsf{B}(\delta_2, x_0) \cap I$, then $h(x) \in [\frac{1}{2}, 2]$, this being possible since $\lim_{x\uparrow x_0} h(x) = 1$. It then follows that

$$\frac{f(x)}{g(x)} > \frac{2M}{h(x)} > M$$

for $x \in \mathsf{B}(\delta, x_0) \cap I$ where $\delta = \min\{\delta_1, \delta_2\}$.

(ii) This follows in the same manner as part (i).

(iii) This follows from parts (i) and (ii).

(iv) Let us define $\phi\colon (0,\infty) \to (0,\infty)$ by $\phi(x) = \frac{1}{x}$. Then define $\tilde{I} = \phi(I)$, noting that $\tilde{I}$ is an interval having 0 as an open left endpoint. Now define $\tilde{f}, \tilde{g}\colon \tilde{I} \to \mathbb{R}$ by $\tilde{f} = f \circ \phi$ and $\tilde{g} = g \circ \phi$. Using the Chain Rule (Theorem 3.2.13 below) we compute

$$\tilde{f}'(\tilde{x}) = f'(\phi(\tilde{x}))\phi'(\tilde{x}) = -\frac{f'(\frac{1}{\tilde{x}})}{\tilde{x}^2}$$

and similarly $\tilde{g}'(\tilde{x}) = -\frac{f'(\frac{1}{\tilde{x}})}{\tilde{x}^2}$. Therefore, for $\tilde{x} \in \tilde{I}$,

$$\frac{f'(\frac{1}{\tilde{x}})}{g'(\frac{1}{\tilde{x}})} = \frac{\tilde{f}'(\tilde{x})}{\tilde{g}'(\tilde{x})}.$$

and so, using part (ii) (it is easy to see that the hypotheses are verified),

$$\begin{aligned} & \lim_{\tilde{x}\downarrow 0}\frac{f'(\frac{1}{\tilde{x}})}{g'(\frac{1}{\tilde{x}})} = \lim_{\tilde{x}\downarrow 0}\frac{\tilde{f}'(\tilde{x})}{\tilde{g}'(\tilde{x})} \\ \implies \quad & \lim_{x\to\infty}\frac{f'(x)}{g'(x)} = \lim_{\tilde{x}\downarrow 0}\frac{\tilde{f}(\tilde{x})}{\tilde{g}(\tilde{x})} \\ \implies \quad & \lim_{x\to\infty}\frac{f'(x)}{g'(x)} = \lim_{x\to\infty}\frac{f(x)}{g(x)}, \end{aligned}$$

which is the desired conclusion.

(v) This follows in the same manner as part (iv). ■

3.2.22 Examples (Uses of L'Hôpital's Rule)

1. Let $I = \mathbb{R}$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = \sin x$ and $g(x) = x$. Note that f and g satisfy the hypotheses of Theorem 3.2.21 with $x_0 = 0$. Therefore we may compute

$$\lim_{x\to 0}\frac{f(x)}{g(x)} = \lim_{x\to 0}\frac{f'(x)}{g'(x)} = \frac{\cos 0}{1} = 1.$$

2. Let $I = [0,1]$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = \sin x$ and $g(x) = x^2$. We can verify that f and g satisfy the hypotheses of L'Hôpital's Rule with $x_0 = 0$. Therefore we compute
$$\lim_{x\downarrow 0} \frac{f(x)}{g(x)} = \lim_{x\downarrow 0} \frac{f'(x)}{g'(x)} = \lim_{x\downarrow 0} \frac{\cos x}{2x} = \infty.$$
3. Let $I = \mathbb{R}_{>0}$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = \mathrm{e}^x$ and $g(x) = -x$. Note that $\lim_{x\to\infty} f(x) = \infty$ and that $\lim_{x\to\infty} g(x) = -\infty$. Thus f and g do not quite satisfy the hypotheses of part (iv) of Theorem 3.2.21 since $\lim_{x\to\infty} g(x) \neq \infty$. However, the problem is a superficial one, as we now illustrate. Define $\tilde{g}(x) = -g(x) = x$. Then f and $\tilde{g}$ *do* satisfy the hypotheses of Theorem 3.2.21(iv). Therefore,
$$\lim_{x\to\infty} \frac{f(x)}{\tilde{g}(x)} = \lim_{x\to\infty} \frac{f'(x)}{\tilde{g}'(x)} = \lim_{x\to\infty} \frac{\mathrm{e}^x}{1} = \infty,$$
and so
$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} -\frac{f(x)}{\tilde{g}(x)} = -\infty.$$
4. Consider the function $h\colon \mathbb{R} \to \mathbb{R}$ defined by $h(x) = \frac{x}{\sqrt{1+x^2}}$. We wish to determine $\lim_{x\to\infty} h(x)$, if this limit indeed exists. We will try to use L'Hôpital's Rule with $f(x) = x$ and $g(x) = \sqrt{1+x^2}$. First, one should check that f and g satisfy the hypotheses of the theorem taking $x_0 = 0$. One can check that f and g are differentiable on I and that $g'(x)$ is nonzero for $x \in I \setminus \{x_0\}$. Moreover, $\lim_{x\to 0} f(x) = 0$ and $\lim_{x\to 0} g(x) = 0$. Thus it only remains to check that $\lim_{x\to 0} \frac{f'(x)}{g'(x)} \in \overline{\mathbb{R}}$. To this end, one can easily compute that
$$\frac{f'(x)}{g'(x)} = \frac{g(x)}{f(x)},$$
which immediately implies that an application of L'Hôpital's Rule is destined to fail. However, the actual limit $\lim_{x\to\infty} h(x)$ does exist, however, and is readily computed, using the definition of limit, to be 1. Thus the converse of L'Hôpital's Rule does not hold. •

3.2.5 Monotonic functions and differentiability

In Section 3.1.5 we considered the notion of monotonicity, and its relationship with continuity. In this section we see how monotonicity is related to differentiability.

For functions that are differentiable, the matter of deciding on their monotonicity properties is straightforward.

3.2.23 Proposition (Monotonicity for differentiable functions) *For* $\mathrm{I} \subseteq \mathbb{R}$ *an interval and* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *a differentiable function, the following statements hold:*

(i) f *is constant if and only if* $\mathrm{f}'(\mathrm{x}) = 0$ *for all* $\mathrm{x} \in \mathrm{I}$*;*

(ii) f *is monotonically increasing if and only* $\mathrm{f}'(\mathrm{x}) \geq 0$ *for all* $\mathrm{x} \in \mathrm{I}$*;*

(iii) f *is strictly monotonically increasing if and only* $\mathrm{f}'(\mathrm{x}) > 0$ *for all* $\mathrm{x} \in \mathrm{I}$*;*

(iv) f *is monotonically decreasing if and only if* $\mathrm{f}'(\mathrm{x}) \leq 0$ *for all* $\mathrm{x} \in \mathrm{I}$*.*

(v) f *is strictly monotonically decreasing if and only if* $\mathrm{f}'(\mathrm{x}) < 0$ *for all* $\mathrm{x} \in \mathrm{I}$*.*

Proof In each case the "only if" assertions follow immediately from the definition of the derivative. To prove the "if" assertions, let $x_1, x_2 \in I$ with $x_1 < x_2$. By the Mean Value Theorem there exists $c \in [x_1, x_2]$ such that $f(x_1) - f(x_2) = f'(c)(x_1 - x_2)$. The result follows by considering the three cases of $f'(c) = 0$, $f'(c) \leq 0$, $f'(c) > 0$, $f'(c) \leq 0$, and $f'(c) < 0$, respectively. ■

The previous result gives the relationship between the derivative and monotonicity. Combining this with Theorem 3.1.30 which relates monotonicity with invertibility, we obtain the following characterisations of the derivative of the inverse function.

3.2.24 Theorem (Inverse Function Theorem for $\mathbb{R}$) *Let* $\mathrm{I} \subseteq \mathrm{J}$ *be an interval, let* $\mathrm{x}_0 \in \mathrm{I}$*, and let* $\mathrm{f}\colon \mathrm{I} \to \mathrm{J} = \mathrm{image}(\mathrm{f})$ *be a continuous, strictly monotonically increasing function that is differentiable at* x_0 *and for which* $\mathrm{f}'(\mathrm{x}_0) \neq 0$*. Then* $\mathrm{f}^{-1}\colon \mathrm{J} \to \mathrm{I}$ *is differentiable at* $\mathrm{f}(\mathrm{x}_0)$ *and the derivative is given by*

$$(\mathrm{f}^{-1})'(\mathrm{f}(\mathrm{x}_0)) = \frac{1}{\mathrm{f}'(\mathrm{x}_0)}.$$

Proof From Theorem 3.1.30 we know that f is invertible. Let $y_0 = f(x_0)$, let $y_1 \in J$, and define $x_1 \in I$ by $f(x_1) = y_1$. Then, if $x_1 \neq x_0$,

$$\frac{f^{-1}(y_1) - f^{-1}(y_0)}{y_1 - y_0} = \frac{x_1 - x_0}{f(x_1) - f(x_0)}.$$

Therefore,

$$(f^{-1})'(y_0) = \lim_{y_1 \to_J y_0} \frac{f^{-1}(y_1) - f^{-1}(y_0)}{y_1 - y_0} = \lim_{x_1 \to_I x_0} \frac{x_1 - x_0}{f(x_1) - f(x_0)} = \frac{1}{f'(x_0)},$$

as desired. ■

3.2.25 Corollary (Alternate version of Inverse Function Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{x}_0 \in \mathrm{I}$*, and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function of class* C^1 *such that* $\mathrm{f}'(\mathrm{x}_0) \neq 0$*. Then there exists a neighbourhood* U *of* x_0 *in* I *and a neighbourhood* V *of* $\mathrm{f}(\mathrm{x}_0)$ *such that* $\mathrm{f}|\mathrm{U}$ *is invertible, and such that* $(\mathrm{f}|\mathrm{U})^{-1}$ *is differentiable, and the derivative is given by*

$$((\mathrm{f}|\mathrm{U})^{-1})'(\mathrm{y}) = \frac{1}{\mathrm{f}'(\mathrm{f}^{-1}(\mathrm{y}))}$$

for each $\mathrm{y} \in \mathrm{V}$*.*

Proof Since f' is continuous and is nonzero at x_0, there exists a neighbourhood U of x_0 such that $f'(x)$ has the same sign as $f'(x_0)$ for all $x \in U$. Thus, by Proposition 3.2.23, $f|U$ is either strictly monotonically increasing (if $f'(x_0) > 0$) or strictly monotonically decreasing (if $f'(x_0) < 0$). The result now follows from Theorem 3.2.24. ■

For general monotonic functions, Proposition 3.2.23 turns out to be "almost" enough to characterise them. To understand this, we recall from Section 2.5.6 the notion of a subset of $\mathbb{R}$ of measure zero. With this recollection having been made, we have the following characterisation of general monotonic functions.

3.2.26 Theorem (Characterisation of monotonic functions II) *If* $I \subseteq \mathbb{R}$ *is an interval and if* $f\colon I \to \mathbb{R}$ *is either monotonically increasing (resp. monotonically decreasing), then* f *is differentiable almost everywhere, and* $f'(x) \geq 0$ *(resp.* $f'(x) \leq 0$*) at all points* $x \in I$ *where* f *is differentiable.*

Proof We first prove a technical lemma.

1 **Lemma** *If* $g\colon [a,b] \to \mathbb{R}$ *has the property that, for each* $x \in [a,b]$*, the limits* $g(x+)$ *and* $g(x-)$ *exist whenever they are defined as limits in* $[a,b]$*. If we define*

$$S = \{x \in [a,b] \mid \textit{there exists } x' > x \textit{ such that } g(x') > \max\{g(x-), g(x), g(x+)\}\},$$

then S *is a disjoint union of a countable collection* $\{I_\alpha \mid \alpha \in A\}$ *of intervals that are open as subsets of* $[a,b]$ *(cf. the beginning of Section 3.1.1).*

Proof Let $x \in S$. We have three cases.

1. There exists $x' > x$ such that $g(x') > g(x-)$, and $g(x-) \geq g(x)$ and $g(x-) \geq g(x+)$: Define $g_{x,-}, g_{x,+}\colon [a,b] \to \mathbb{R}$ by

$$g_{x,-}(y) = \begin{cases} g(y), & y \neq 1, \\ g(x-), & y = x, \end{cases} \qquad g_{x,+}(y) = \begin{cases} g(y), & y \neq 1, \\ g(x+), & y = x. \end{cases}$$

Since the limit $g(x-)$ exists, $g_{x,-}|[a,x]$ is continuous at x by Theorem 3.1.3. Since $g(x') > g_{x,-}(x)$, there exists $\epsilon_1 \in \mathbb{R}_{>0}$ such that $g(x') > g_{x,-}(y) = g(y)$ for all $y \in (x-\epsilon_1, x)$. Now note that $g(x') > g(x-) \geq g_{x,+}(x)$. Arguing similarly to what we have done, there exists $\epsilon_2 \in \mathbb{R}_{>0}$ such that $g(x') > g_{x,+}(y) = g(y)$ for all $y \in (x, x+\epsilon_2)$. Let $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. Since $g(x') > g(x-) \geq g(x)$, it follows that $g(x') > g(y)$ for all $y \in (x-\epsilon, x+\epsilon)$, so we can conclude that S is open.

2. There exists $x' > x$ such that $g(x') > g(x)$, and $g(x) \geq g(x-)$ and $g(x) \geq g(x+)$: Define $g_{x,-}$ and $g_{x,+}$ as above. Then, since $g(x') > g(x) \geq g(x-)$ and $g(x') > g(x) \geq g(x+)$, we can argue as in the previous case that there exists $\epsilon \in \mathbb{R}_{>0}$ such that $g(x') > g(y)$ for all $y \in (x-\epsilon, x+\epsilon)$. Thus S is open.

3. There exists $x' > x$ such that $g(x') > g(x+)$, and $g(x+) \geq g(x)$ and $g(x+) \geq g(x-)$: Here we can argue in a manner entirely similar to the first case that S is open.

The preceding arguments show that S is open, and so by Proposition 2.5.6 it is a countable union of open intervals. ▼

Now define

$$\Lambda_l(x) = \limsup_{h\downarrow 0} \frac{f(x-h)-f(x)}{-h} \qquad \lambda_l(x) = \liminf_{h\downarrow 0} \frac{f(x-h)-f(x)}{-h}$$
$$\Lambda_r(x) = \limsup_{h\downarrow 0} \frac{f(x+h)-f(x)}{h} \qquad \lambda_r(x) = \liminf_{h\downarrow 0} \frac{f(x+h)-f(x)}{h}.$$

If f is differentiable at x then these four numbers will be finite and equal. We shall show that

1. $\Lambda_r(x) < \infty$ and
2. $\Lambda_r(x) \le \lambda_l(x)$

for almost every $x \in [a, b]$. Since the relations

$$\lambda_l \le \Lambda_l \le \lambda_r \le \Lambda_r$$

hold due to monotonicity of f, the differentiability of f for almost all x will then follow.

For 1, if $M \in \mathbb{R}_{>0}$ denote

$$S_M = \{x \in [a, b] \mid \Lambda_r(x) > M\}.$$

Thus, for $x_0 \in S_M$, there exists $x > x_0$ such that

$$\frac{f(x)-f(x_0)}{x-x_0} > M.$$

Defining $g_M(x) = f(x) - Mx$ this asserts that $g_M(x) > g_M(x_0)$. The function g_M satisfies the hypotheses of Lemma 1 by part (i). This means that S_M is contained in a countable disjoint union of intervals $\{I_\alpha \mid \alpha \in A\}$, open in $[a, b]$, for which

$$g_M(a_\alpha) \le \max\{g_M(b_\alpha-), g_M(b_\alpha), g_M(b_\alpha+)\}, \qquad \alpha \in A,$$

where a_α and b_α are the left and right endpoints, respectively, for I_α, $\alpha \in A$. In particular, $g_M(a_\alpha) \le g_M(b_\alpha)$. A trivial manipulation then gives

$$M(b_\alpha - a_\alpha) \le f(b_\alpha) - f(a_\alpha), \qquad \alpha \in A.$$

We have

$$M\sum_{\alpha\in A}|b_\alpha - a_\alpha| \le \sum_{\alpha\in A}|f(b_\alpha) - f(a_\alpha)| \le f(b) - f(a)$$

since f is monotonically increasing. Since f is bounded, this shows that as $M \to \infty$ the length of the open intervals $\{(a_\alpha, b_\alpha) \mid \alpha \in A\}$ covering S_M must go to zero. This shows that the set of points where 1 holds has zero measure.

Now we turn to 2. Let $0 < m < M$, define $g_m(x) = -f(x) + mx$ and $g_M(x) = f(x) - Mx$. Also define

$$S_m = \{x \in [a, b] \mid \lambda_l(x) < m\}.$$

For $x_0 \in S_m$ there exists $x < x_0$ such that

$$\frac{f(x)-f(x_0)}{x-x_0} < m,$$

which is equivalent to $g_m(x) > g_m(x_0)$. Therefore, by Lemma 1, note that S_m is contained in a countable disjoint union of intervals $\{I_\alpha \mid \alpha \in A\}$, open in $[a, b]$. Denote by a_α and b_α the left and right endpoints, respectively, for I_α for $\alpha \in A$. For $\alpha \in A$ denote

$$S_{\alpha,M} = \{x \in [a_\alpha, b_\alpha] \mid \Lambda_r(x) > M\},$$

and arguing as we did in the proof that 1 holds almost everywhere, denote by $\{I_{\alpha,\beta} \mid \beta \in B_\alpha\}$ the countable collection of subintervals, open in $[a, b]$, of (a_α, b_α) that contain $S_{\alpha,M}$. Denote by $a_{\alpha,\beta}$ and $b_{\alpha,\beta}$ the left and right endpoints, respectively, of $I_{\alpha,\beta}$ for $\alpha \in A$ and $\beta \in B_\alpha$. Note that the relations

$$\begin{aligned} &g_m(a_\alpha) \le \max\{g_m(b_\alpha-), g_m(b_\alpha), g_m(b_\alpha+)\}, \qquad \alpha \in A, \\ &g_M(a_{\alpha,\beta}) \le \max\{g_M(b_{\alpha,\beta}-), g_M(b_{\alpha,\beta}), g_M(b_{\alpha,\beta}+)\}, \qquad \alpha \in A,\ \beta \in B_\alpha \end{aligned}$$

hold. We then may easily compute

$$\begin{aligned} &f(b_\alpha) - f(a_\alpha) \le m(b_\alpha - a_\alpha), \qquad \alpha \in A, \\ &f(b_{\alpha,\beta}) - f(a_{\alpha,\beta}) \ge M(b_{\alpha,\beta} - b_{\alpha,\beta}), \qquad \alpha \in A,\ \beta \in A_\alpha. \end{aligned}$$

Therefore, for each $\alpha \in A$,

$$M \sum_{\beta \in A_\alpha} |b_{\alpha,\beta} - a_{\alpha,\beta}| \le \sum_{\beta \in A_\alpha} |f(b_{\alpha,\beta} - a_{\alpha,\beta})| \le f(b_\alpha) - f(a_\alpha) \le m(b_\alpha - a_\alpha).$$

This then gives

$$M \sum_{\alpha \in A} \sum_{\beta \in A_\alpha} |b_{\alpha,\beta} - a_{\alpha,\beta}| \le m \sum_{\alpha \in A} |b_\alpha - a_\alpha|,$$

or $\Sigma_2 \le \frac{m}{M}\Sigma_1$, where

$$\Sigma_1 = \sum_{\alpha \in A} \sum_{\beta_\alpha \in K_\alpha} |b_{\alpha,\beta} - a_{\alpha,\beta}|, \quad \Sigma_2 = \sum_{\alpha \in A} |b_\alpha - a_\alpha|.$$

Now, this process can be repeated, defining

$$S_{\alpha,\beta,m} = \{x \in [a_{\alpha,\beta}, b_{\alpha,\beta}] \mid \lambda_l(x) < m\},$$

and so on. We then generate a sequence of countable disjoint intervals of total length Σ_α and satisfying

$$\Sigma_{2\alpha} \le \frac{m}{M}\Sigma_{2\alpha-1} \le \left(\frac{m}{M}\right)^\alpha \Sigma_1, \qquad \alpha \in A.$$

It therefore follows that $\lim_{\alpha\to\infty} \Sigma_\alpha = 0$. Thus the set of points

$$S_{M,m} = \{x \in [a, b] \mid m < \lambda_l(x) \text{ and } \Lambda_r(x) > M\}$$

is contained in a set of zero measure provided that $m < M$. Now note that

$$\{x \in [a, b] \mid \lambda_l(x) \ge \Lambda_r(x)\} \subseteq \bigcup \{S_{M,m} \mid m, M \in \mathbb{Q},\ m < M\}.$$

The union on the left is a countable union of sets of zero measure, and so has zero measure itself (by Exercise 2.5.11). This shows that f is differentiable on a set whose complement has zero measure.

To show that $f'(x) \geq 0$ for all points x at which f is differentiable, suppose the converse. Thus suppose that there exists $x \in [a, b]$ such that $f'(x) < 0$. This means that for ϵ sufficiently small and positive,

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} < 0 \quad \Longrightarrow \quad f(x+\epsilon) - f(x) < 0,$$

which contradicts the fact that f is monotonically increasing. This completes the proof of the theorem. ■

Let us give two examples of functions that illustrate the surprisingly strange behaviour that can arise from monotonic functions. These functions are admittedly degenerate, and not something one is likely to encounter in applications. However, they do show that one cannot strengthen the conclusions of Theorem 3.2.26.

Our first example is one of the standard "peculiar" monotonic functions, and its construction relies on the middle-thirds Cantor set constructed in Example 2.5.39.

3.2.27 Example (A continuous increasing function with an almost everywhere zero derivative) Let C_k, $k \in \mathbb{Z}_{>0}$, be the sets, comprised of collections of disjoint closed intervals, used in the construction of the middle-thirds Cantor set of Example 2.5.39. Note that, for $x \in [0, 1]$, the set $[0, x] \cap C_k$ consists of a finite number of intervals. Let $g_k \colon [0, 1] \to [0, 1]$ be defined by asking that $g_{C,k}(x)$ be the sum of the lengths of the intervals comprising $[0, x] \cap C_k$. Then define $f_{C,k} \colon [0, 1] \to [0, 1]$ by $f_{C,k}(x) = \left(\frac{3}{2}\right)^k g_{C,k}(x)$. Thus $f_{C,k}$ is a function that is constant on the complement to the closed intervals comprising C_k, and is linear on those same closed intervals, with a slope determined in such a way that the function is continuous. We then define $f_C \colon [0, 1] \to [0, 1]$ by $f_C(x) = \lim_{k\to\infty} f_{C,k}(x)$. In Figure 3.9 we depict f_C. The reader new to this function should take the requisite moment or two to understand our definition of f_C, perhaps by sketching a couple of the functions $f_{C,k}$, $k \in \mathbb{Z}_{>0}$.

Let us record some properties of the function f_C, which is called the ***Cantor function*** or the ***Devil's staircase***.

1 Lemma f_C *is continuous.*

Proof We prove this by showing that the sequence of functions $(f_{C,k})_{k\in\mathbb{Z}_{>0}}$ converges uniformly, and then using Theorem 3.6.8 to conclude that the limit function is continuous. Note that the functions $f_{C,k}$ and $f_{C,k+1}$ differ only on the closed intervals comprising C_k. Moreover, if $J_{k,j}$, $k \in \mathbb{Z}_{\geq 0}$, $j \in \{1, \dots, 2^k - 1\}$, denotes the set of open intervals forming $[0, 1] \setminus C_k$, numbered from left to right, then the value of $f_{C,k}$ on $J_{k,j}$ is $j2^{-k}$. Therefore,

$$\sup\{|f_{C,k+1}(x) - f_{C,k}(x)| \mid x \in [0, 1]\} < 2^{-k}, \qquad k \in \mathbb{Z}_{\geq 0}.$$

This implies that $(f_{C,k})_{k\in\mathbb{Z}_{>0}}$ is uniformly convergent as in Definition 3.6.4. Thus Theorem 3.6.8 gives continuity of f_C, as desired. ▼

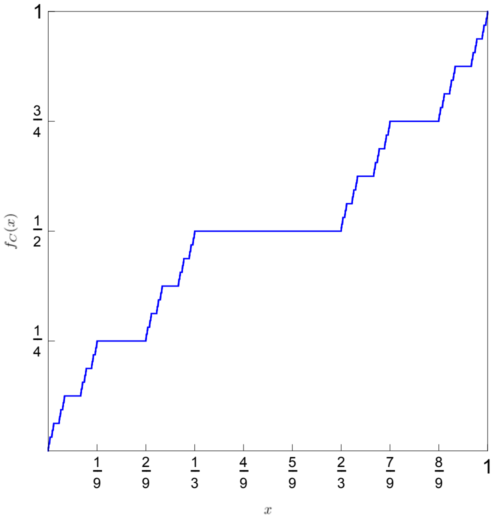


Figure 3.9 A depiction of the Cantor function

2 Lemma f_C *is differentiable at all points in* $[0,1] \setminus C$*, and its derivative, where it exists, is zero.*

Proof Since C is constructed as an intersection of the closed sets C_k, and since such intersections are themselves closed by Exercise 2.5.1, it follows that $[0,1] \setminus C$ is open. Thus if $x \in [0,1] \setminus C$, there exists $\epsilon \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\epsilon, x) \subseteq [0,1] \setminus C$. Since $\mathsf{B}(\epsilon, x)$ contains no endpoints for intervals from the sets C_k, $k \in \mathbb{Z}_{>0}$, it follows that $f_{C,k}|\mathsf{B}(\epsilon, x)$ is constant for sufficiently large k. Therefore $f_C|\mathsf{B}(\epsilon, x)$ is constant, and it then follows that f_C is differentiable at x, and that $f'_C(x) = 0$. ▼

In Example 2.5.39 we showed that C has measure zero. Thus we have a continuous, monotonically increasing function from $[0,1]$ to $[0,1]$ whose derivative is almost everywhere zero. It is perhaps not *a priori* obvious that such a function can exist, since one's first thought might be that zero derivative implies a constant function. The reasons for the failure of this rule of thumb in this example will not become perfectly clear until we examine the notion of absolute continuity in Section III-2.9.6. •

The second example of a "peculiar" monotonic function is not quite as standard in the literature, but is nonetheless interesting since it exhibits somewhat different oddities than the Cantor function.

3.2.28 Example (A strictly increasing function, discontinuous on the rationals, with an almost everywhere zero derivative) We define a strictly monotonically increasing function $f_{\mathbb{Q}}\colon \mathbb{R} \to \mathbb{R}$ as follows. Let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be an enumeration of the rational numbers and for $x \in \mathbb{R}$ define

$$I(x) = \{j \in \mathbb{Z}_{>0} \mid q_j < x\}.$$

Now define

$$f_{\mathbb{Q}}(x) = \sum_{j\in I(x)} \frac{1}{2^j}.$$

Let us record the properties of $f_{\mathbb{Q}}$ in a series of lemmata.

1 **Lemma** $\lim_{x\to-\infty} \mathrm{f}_{\mathbb{Q}}(\mathrm{x}) = 0$ *and* $\lim_{x\to\infty} \mathrm{f}_{\mathbb{Q}}(\mathrm{x}) = 1$.

Proof Recall from Example 2.4.2–1 that $\sum_{j=1}^{\infty} \frac{1}{2^j} = 1$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose $N \in \mathbb{Z}_{>0}$ such that $\sum_{j=N+1}^{\infty} \frac{1}{2^j} < \epsilon$. Now choose $M \in \mathbb{R}_{>0}$ such that $\{q_1, \dots, q_N\} \subseteq [-M, M]$. Then, for $x < M$ we have

$$f_{\mathbb{Q}}(x) = \sum_{j\in I(x)} \frac{1}{2^j} = \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{j\in\mathbb{Z}_{>0}\setminus I(x)} \frac{1}{2^j} \le \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{j=1}^{N} \frac{1}{2^j} < \epsilon.$$

Also, for $x > M$ we have

$$f_{\mathbb{Q}}(x) = \sum_{j\in I(x)} \frac{1}{2^j} \ge \sum_{j=1}^{N} \frac{1}{2^j} > 1 - \epsilon.$$

Thus $\lim_{x\to-\infty} f_{\mathbb{Q}}(x) = 0$ and $\lim_{x\to\infty} f_{\mathbb{Q}}(x) = 1$. ▼

2 **Lemma** $\mathrm{f}_{\mathbb{Q}}$ *is strictly monotonically increasing.*

Proof Let $x, y \in \mathbb{R}$ with $x < y$. Then, by Corollary 2.2.16, there exists $q \in \mathbb{Q}$ such that $x < q < y$. Let $j_0 \in \mathbb{Z}_{>0}$ have the property that $q = q_{j_0}$. Then

$$f_{\mathbb{Q}}(y) = \sum_{j\in I(y)} \frac{1}{2^j} \ge \sum_{j\in I(x)} \frac{1}{2^j} + \frac{1}{2^{j_0}} > f_{\mathbb{Q}}(x),$$

as desired. ▼

3 **Lemma** $\mathrm{f}_{\mathbb{Q}}$ *is discontinuous at each point in* $\mathbb{Q}$.

Proof Let $q \in \mathbb{Q}$ and let $x > q$. Let $j_0 \in \mathbb{Z}_{>0}$ satisfy $q = q_{j_0}$. Then

$$f_{\mathbb{Q}}(x) = \sum_{j\in I(x)} \frac{1}{2^j} \ge \frac{1}{2^{j_0}} + \sum_{j\in I(q)} \frac{1}{2^j} = \frac{1}{2^{j_0}} + \sum_{j\in I(q)} \frac{1}{2^j}.$$

Therefore, $\lim_{x\downarrow q} f_{\mathbb{Q}}(x) \ge \frac{1}{2^{j_0}} + f_{\mathbb{Q}}(q)$, implying that $f_{\mathbb{Q}}$ is discontinuous at q by Theorem 3.1.3. ▼

4 **Lemma** $f_{\mathbb{Q}}$ *is continuous at each point in* $\mathbb{R} \setminus \mathbb{Q}$.

Proof Let $x \in \mathbb{R} \setminus \mathbb{Q}$ and let $\epsilon \in \mathbb{R}_{>0}$. Take $N \in \mathbb{Z}_{>0}$ such that $\sum_{j=N+1}^{\infty} \frac{1}{2^j} < \epsilon$ and define $\delta \in \mathbb{R}_{>0}$ such that $\mathsf{B}(\delta, x) \cap \{q_1, \dots, q_N\} = \varnothing$ (why is this possible?). Now let

$$I(\delta, x) = \{j \in \mathbb{Z}_{>0} \mid q_j \in \mathsf{B}(\delta, x)\}$$

and note that, for $y \in \mathsf{B}(\delta, x)$ with $x < y$, we have

$$\begin{aligned} f_{\mathbb{Q}}(y) - f_{\mathbb{Q}}(x) &= \sum_{j \in I(y)} \frac{1}{2^j} - \sum_{j \in I(x)} \frac{1}{2^j} \leq \sum_{j \in I(\delta,x)} \frac{1}{2^j} = \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{\mathbb{Z}_{>0} \setminus I(\delta,x)} \frac{1}{2^j} \\ &\leq \sum_{j=1}^{\infty} \frac{1}{2^j} - \sum_{j=1}^{N} \frac{1}{2^j} = \sum_{j=N+1}^{\infty} \frac{1}{2^j} < \epsilon. \end{aligned}$$

A similar argument holds for $y < x$ giving $f_{\mathbb{Q}}(x) - f_{\mathbb{Q}}(y) < \epsilon$ in this case. Thus $|f_{\mathbb{Q}}(y) - f_{\mathbb{Q}}(x)| < \epsilon$ for $|y - x| < \delta$, thus showing continuity of f at x. ▼

5 **Lemma** *The set* $\{x \in \mathbb{R} \mid f'_{\mathbb{Q}}(x) \neq 0\}$ *has measure zero.*

Proof The proof relies on some concepts from Section 3.6. For $k \in \mathbb{Z}_{>0}$ define $f_{\mathbb{Q},k} \colon \mathbb{R} \to \mathbb{R}$ by

$$f_{\mathbb{Q},k}(x) = \sum_{j \in I(x) \cap \{1,\dots,k\}} \frac{1}{2^j}.$$

Note that $(f_{\mathbb{Q},k})_{k \in \mathbb{Z}_{>0}}$ is a sequence of monotonically increasing functions with the following properties:

1. $\lim_{k\to\infty} f_{\mathbb{Q},k}(x) = f_{\mathbb{Q}}(x)$ for each $x \in \mathbb{R}$;
2. the set $\{x \in \mathbb{R} \mid f'_{\mathbb{Q},k}(x) \neq 0\}$ is finite for each $k \in \mathbb{Q}$.

The result now follows from Theorem 3.6.25. ▼

Thus we have an example of a strictly monotonically increasing function whose derivative is zero almost everywhere. Note that this function also has the feature that in any neighbourhood of a point where it is differentiable, there lie points where it is not differentiable. This is an altogether peculiar function. •

3.2.6 Convex functions and differentiability

Let us now return to our consideration of convex functions introduced in Section 3.1.6. Here we discuss the differentiability properties of convex functions. The following notation for a function $f \colon I \to \mathbb{R}$ will be convenient:

$$f'(x+) = \lim_{\epsilon \downarrow 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}, \quad f'(x-) = \lim_{\epsilon \downarrow 0} \frac{f(x) - f(x-\epsilon)}{\epsilon},$$

provided that these limits exist.

With this notation, convex functions have the following properties.

3.2.29 Proposition (Properties of convex functions II) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and for a convex function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$, *the following statements hold:*

(i) if I *is open then the limits* $\mathrm{f}'(\mathrm{x}+)$ *and* $\mathrm{f}'(\mathrm{x}-)$ *exist and* $\mathrm{f}'(\mathrm{x}-) \leq \mathrm{f}'(\mathrm{x}+)$ *for each* $\mathrm{x} \in \mathrm{I}$;

(ii) if I *is open then the functions*

$$\mathrm{I} \ni \mathrm{x} \mapsto \mathrm{f}'(\mathrm{x}+), \quad \mathrm{I} \ni \mathrm{x} \mapsto \mathrm{f}'(\mathrm{x}-)$$

are monotonically increasing, and strictly monotonically increasing if f *is strictly convex;*

(iii) if I *is open and if* $\mathrm{x}_1, \mathrm{x}_2 \in \mathrm{I}$ *satisfy* $\mathrm{x}_1 < \mathrm{x}_2$, *then* $\mathrm{f}'(\mathrm{x}_1+) \leq \mathrm{f}'(\mathrm{x}_2-)$;

(iv) f *is differentiable except at a countable number of points in* I.

Proof (i) Since I is open there exists $\epsilon_0 \in \mathbb{R}_{>0}$ such that $[x, x+\epsilon_0) \subseteq I$. Let $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $(0, \epsilon_0)$ converging to 0 and such that $\epsilon_{j+1} < \epsilon_j$ for every $j \in \mathbb{Z}_{>0}$. Then the sequence $(s_f(x, x+\epsilon_j))_{j\in\mathbb{Z}_{>0}}$ is monotonically decreasing. This means that, by Lemma 3.1.33,

$$\frac{f(x+\epsilon_{j+1}) - f(x)}{\epsilon_{j+1}} \leq \frac{f(x+\epsilon_j) - f(x)}{\epsilon_j}$$

for each $j \in \mathbb{Z}_{>0}$. Moreover, if $x' \in I$ satisfies $x' < x$ then we have $s_f(x', x) \leq s_f(x, x+\epsilon_j)$ for each $j \in \mathbb{Z}_{>0}$. Thus the sequence $(\epsilon_j^{-1}(f(x+\epsilon_j) - f(x)))_{j\in\mathbb{Z}_{>0}}$ is decreasing and bounded from below. Thus it must converge, cf. Theorem 2.3.8.

The proof for the existence of the other asserted limit follows that above, *mutatis mutandis*.

To show that $f'(x-) \leq f'(x+)$, note that, for all ϵ sufficiently small,

$$\frac{f(x) - f(x-\epsilon)}{\epsilon} = s_f(x-\epsilon, x) \leq s_f(x, x+\epsilon) = \frac{f(x+\epsilon) - f(x)}{\epsilon}.$$

Taking limits as $\epsilon \downarrow 0$ gives the desired inequality.

(ii) For $x_1, x_2 \in I$ with $x_1 < x_2$ we have

$$f'(x_1+) = \lim_{\epsilon\downarrow 0} s_f(x_1, x_1+\epsilon) \leq \lim_{\epsilon\downarrow 0} s_f(x_2, x_2+\epsilon) = f'(x_2+),$$

using Lemma 3.1.33. A similar computation, *mutatis mutandis*, shows that the other function in this part of the result is also monotonically increasing. Moreover, if f is strictly convex that the inequalities above can be replaced with strict inequalities by (3.2). From this we conclude that $x \mapsto f'(x_+)$ and $x \mapsto f'(x_-)$ are strictly monotonically increasing.

(iii) For $\epsilon \in \mathbb{R}_{>0}$ sufficiently small we have

$$x_1 + \epsilon < x_2 - \epsilon.$$

For all such sufficiently small ϵ we have

$$\frac{f(x_1+\epsilon) - f(x_1)}{\epsilon} = s_f(x_1, x_1+\epsilon) \leq s_f(x_2-\epsilon, x_2) = \frac{f(x_2) - f(x_2-\epsilon)}{\epsilon}$$

by Lemma 3.1.33. Taking limits as $\epsilon \downarrow 0$ gives this part of the result.

(iv) Let A_f be the set of points in I where f is not differentiable. Note that

$$\frac{f(x) - f(x-\epsilon)}{\epsilon} = s_f(x-\epsilon, x) \leq s_f(x, x+\epsilon) = \frac{f(x+\epsilon) - f(x)}{\epsilon}$$

by Lemma 3.1.33. Therefore, if $x \in A_f$, then $f'(x-) < f'(x+)$. We define a map $\phi\colon A_f \to \mathbb{Q}$ as follows. If $x \in A_f$ we use the Axiom of Choice and Corollary 2.2.16 to select $\phi(x) \in \mathbb{Q}$ such that $f'(x-) < \phi(x) < f'(x+)$. We claim that ϕ is injective. Indeed, if $x, y \in A_f$ are distinct (say $x < y$) then, using parts (ii) and (iii),

$$f'(x-) < \phi(x) < f'(x+) < f'(y-) < \phi(y) < f'(y+).$$

Thus $\phi(x) < \phi(y)$ and so ϕ is injective as desired. Thus A_f must be countable. ■

For functions that are sufficiently differentiable, it is possible to conclude convexity from properties of the derivative.

3.2.30 Proposition (Convexity and derivatives) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and for a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *the following statements hold:*

(i) for each $\mathrm{x}_1, \mathrm{x}_2 \in \mathrm{I}$ *with* $\mathrm{x}_1 \neq \mathrm{x}_2$ *we have*

$$\mathrm{f}(\mathrm{x}_2) \geq \mathrm{f}(\mathrm{x}_1) + \mathrm{f}'(\mathrm{x}_1+)(\mathrm{x}_2 - \mathrm{x}_1), \quad \mathrm{f}(\mathrm{x}_2) \geq \mathrm{f}(\mathrm{x}_1) + \mathrm{f}'(\mathrm{x}_1-)(\mathrm{x}_2 - \mathrm{x}_1);$$

(ii) if f *is differentiable, then* f *is convex if and only if* f′ *is monotonically increasing;*

(iii) if f *is differentiable, then* f *is strictly convex if and only if* f′ *is strictly monotonically increasing;*

(iv) if f *is twice continuously differentiable, then it is convex if and only if* $\mathrm{f}''(\mathrm{x}) \geq 0$ *for every* $\mathrm{x} \in \mathrm{I}$*;*

(v) if f *is twice continuously differentiable, then it is strictly convex if and only if* $\mathrm{f}''(\mathrm{x}) > 0$ *for every* $\mathrm{x} \in \mathrm{I}$*.*

Proof (i) Suppose that $x_1 < x_2$. Then, for $\epsilon \in \mathbb{R}_{>0}$ sufficiently small,

$$\frac{f(x_1+\epsilon) - f(x_1)}{\epsilon} \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1}$$

by Lemma 3.1.33. Thus, taking limits as $\epsilon \downarrow 0$,

$$f'(x_1+) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1},$$

and rearranging gives

$$f(x_2) \geq f(x_1) + f'(x_1+)(x_2 - x_1).$$

Since we also have $f'(x_1-) \leq f'(x_1+)$ by Proposition 3.2.29(i), we have both of the desired inequalities in this case.

Now suppose that $x_2 < x_1$. Again, for $\epsilon \in \mathbb{R}_{>0}$ sufficiently small, we have

$$\frac{f(x_1+\epsilon) - f(x_1)}{\epsilon} \geq \frac{f(x_1) - f(x_2)}{x_1 - x_2},$$

and taking the limit as $\epsilon \downarrow 0$ gives

$$f'(x_1+) \geq \frac{f(x_1) - f(x_2)}{x_1 - x_2}.$$

Rearranging gives

$$f(x_2) \geq f(x_1) + f'(x_1+)(x_2 - x_1)$$

and since $f'(x_1-) \leq f'(x_1+)$ the desired inequalities follow in this case.

(ii) From Proposition 3.2.29(ii) we deduce that if f is convex and differentiable then f' is monotonically increasing. Conversely, suppose that f is differentiable and that f' is monotonically increasing. Let $x_1, x_2 \in I$ satisfy $x_1 < x_2$ and let $s \in (0, 1)$. By the Mean Value Theorem there exists $c_1, c_2 \in I$ satisfying

$$x_1 < c_1 < (1 - s)x_1 + sx_2 < d_1 < x_2$$

such that

$$\frac{f((1-s)x_1 + sx_2) - f(x_1)}{(1-s)x_1 + sx_2 - x_1} = f'(c_1) \leq f'(c_2) = \frac{f(x_2) - f((1-s)x_1 + sx_2)}{x_2 - ((1-s)x_1 + sx_2)}. \tag{3.9}$$

Rearranging, we get

$$\frac{f((1-s)x_1 + sx_2) - f(x_1)}{s(x_2 - x_1)} \leq \frac{f(x_2) - f((1-s)x_1 + sx_2)}{(1-s)(x_2 - x_1)},$$

and further rearranging gives

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2),$$

and so f is convex.

(iii) If f is strictly convex, then from Proposition 3.2.29 we conclude that f' is strictly monotonically increasing. Next suppose that f' is strictly monotonically decreasing and let $x_1, x_2 \in I$ satisfy $x_1 < x_2$ and let $s \in (0, 1)$. The proof that f is strictly convex follows as in the preceding part of the proof, noting that, in (3.9), we have $f'(c_1) < f'(c_2)$. Carrying this strict inequality through the remaining computations shows that

$$f((1-s)x_1 + sx_2) \leq (1-s)f(x_1) + sf(x_2),$$

giving strict convexity of f.

(iv) If f'' is nonnegative, then f' is monotonically increasing by Proposition 3.2.23. The result now follows from part (ii).

(iv) If f'' is positive, then f' is strictly monotonically increasing by Proposition 3.2.23. The result now follows from part (iii). ■

Let us consider a few examples illustrating how convexity and differentiability are related.

3.2.31 Examples (Convex functions and differentiability)

1. The convex function $n_{x_0}\colon \mathbb{R} \to \mathbb{R}$ defined by $n_{x_0}(x) = |x - x_0|$ is differentiable everywhere except for $x = x_0$. But at $x = x_0$ the derivatives from the left and right exist. Moreover, $f'(x) = -1$ for $x < x_0$ and $f'(x) = 1$ for $x > x_0$. Thus we see that the derivative is monotonically increasing, although it is not defined everywhere.
2. As we showed in Proposition 3.2.29(iv), a convex function is differentiable except at a countable set of points. Let us show that this conclusion cannot be improved. Let $C \subseteq \mathbb{R}$ be a countable set. We shall construct a convex function $f\colon \mathbb{R} \to \mathbb{R}$ whose derivative exists on $\mathbb{R} \setminus C$ and does not exist on C. In case C is finite, we write $C = \{x_1, \dots, x_k\}$. Then one verifies that the function f defined by

$$f(x) = \sum_{j=1}^{k} |x - x_j|$$

is verified to be convex, being a finite sum of convex functions (see Proposition 3.1.39). It is clear that f is differentiable at points in $\mathbb{R} \setminus C$ and is not differentiable at points in C. Now suppose that C is not finite. Let us write $C = \{x_j\}_{j \in \mathbb{Z}_{>0}}$, i.e., enumerate the points in C. Let us define $c_j = (2^j \max\{1, |x_j|\})^{-1}$, $j \in \mathbb{Z}_{>0}$, and define $f\colon \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \sum_{j=1}^{\infty} c_j |x - x_j|.$$

We shall prove that this function is well-defined, convex, differentiable at points in $\mathbb{R} \setminus C$, and not differentiable at points in C. In proving this, we shall make reference to some results we have not yet proved.

First let us show that f is well-defined.

1 Lemma *For every compact subset* $\mathrm{K} \subseteq \mathbb{R}$, *the series*

$$\sum_{\mathrm{j}=1}^{\infty} \mathrm{c_j}|\mathrm{x} - \mathrm{x_j}|$$

converges uniformly on K *(see Section 3.6.2 for uniform convergence).*

Proof Let $K \subseteq \mathbb{R}$ and let $R \in \mathbb{R}_{>0}$ be large enough that $K \subseteq [-R, R]$. Then, for $x \in K$ we have

$$|c_j|x - x_j|| \leq c_j(|x| + |x_j|) \leq \frac{R+1}{2^j}.$$

By the Weierstrass M-test (Theorem 3.6.15 below) and Example 2.4.2–1 the lemma follows. ▼

It follows immediately from the lemma that the series defining f converges pointwise, and so f is well-defined, and is moreover convex by Theorem 3.6.26. Now we show that f is differentiable at points in $\mathbb{R} \setminus C$.

2 Lemma *The function* f *is differentiable at every point in* $\mathbb{R} \setminus \mathrm{C}$.

Proof Let us denote $g_j(x) = c_j|x - x_j|$. Let $x_0 \in \mathbb{R} \setminus C$ and define, for each $j \in \mathbb{Z}_{>0}$,

$$h_{j,x_0} = \begin{cases} \frac{g_j(x)-g_j(x_0)}{x-x_0}, & x \neq x_0, \\ g_j'(x_0), & x = x_0, \end{cases}$$

noting that the functions g_j, $j \in \mathbb{Z}_{>0}$, are differentiable at points in $\mathbb{R} \setminus C$.

Let $j \in \mathbb{Z}$. We claim that if $x_0 \neq x_j$ then

$$|h_{j,x_0}(x)| \leq \frac{3}{2^j} \tag{3.10}$$

for all $x \in \mathbb{R}$. We consider three cases.

(a) $x = x_0$: Note that g_j is differentiable at $x = x_0$ and that $|g_j'(x_0)| = c_j \leq \frac{1}{2^j} < \frac{3}{2^j}$. Thus the estimate (3.10) holds when $x = x_0$.

(b) $x \neq x_0$ and $(x - x_j)(x_0 - x_j) > 0$: We have

$$|h_{j,x_0}(x)| = c_j \left| \frac{(x - x_j) - (x_0 - x_j)}{x - x_0} \right| = a_j \leq \frac{1}{2^j} < \frac{3}{2^j},$$

giving (3.10) in this case.

(c) $x \neq x_0$ and $(x - x_j)(x_0 - x_j) < 0$: We have

$$|h_{j,x_0}(x)| = c_j \left| \frac{(x - x_j) - (x_j - x_0)}{x - x_0} \right| = c_j \left| 1 + \frac{2(x_0 - x_j)}{x_0 - x} \right| \leq \frac{1}{2^j} \left| 1 + \frac{2(x_0 - x_j)}{x_0 - x} \right|.$$

Since $(x - x_j)$ and $x_0 - x_j$ have opposite sign, this implies that either (1) $x < x_j$ and $x_0 > x_j$ or (2) $x > x_j$ and $x_0 < x_j$. In either case, $|x_0 - x_j| < |x_0 - x|$. This, combined with our estimate above, gives (3.10) in this case.

Now, given (3.10), we can use the Weierstrass M-test (Theorem 3.6.15 below) and Example 2.4.2–1 to conclude that $\sum_{j=1}^{\infty} h_{j,x_0}$ converges uniformly on $\mathbb{R}$ for each $x_0 \in \mathbb{R} \setminus C$.

Now we prove that f is differentiable at $x_0 \in \mathbb{R} \setminus C$. If $x \neq x_0$ then the definition of the functions h_{j,x_0}, $j \in \mathbb{Z}_{>0}$, gives

$$\frac{f(x) - f(x_0)}{x - x_0} = \sum_{j=1}^{\infty} h_{j,x_0}(x),$$

the latter sum making sense since we have shown that it converges uniformly. Moreover, since the functions g_j, $j \in \mathbb{Z}_{>0}$, are differentiable at x_0, it follows that, for each $j \in \mathbb{Z}_{>0}$,

$$\lim_{x \to x_0} h_{j,x_0}(x) = \lim_{x \to x_0} \frac{g_j(x) - g_j(x_0)}{x - x_0} = g_j'(x_0) = h_{j,x_0}(x_0).$$

That is, h_{j,x_0} is continuous at x_0. It is clear that h_{j,x_0} is continuous at all $x \neq x_0$. Thus, since $\sum_{j=1}^{\infty} h_{j,x_0}$ converges uniformly, the limit function is continuous by Theorem 3.6.8. Thus we have

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \to x_0} \sum_{j=1}^{\infty} h_{j,x_0}(x) = \sum_{j=1}^{\infty} h_{j,x_0}(x_0) = \sum_{j=1}^{\infty} g_j'(x_0).$$

This gives the desired differentiability since the last series converges. ▼

Finally, we show that f is not differentiable at points in C.

3 Lemma *The function* f *is not differentiable at every point in* C*.*

Proof For $k \in \mathbb{Z}_{>0}$, let us write

$$f(x) = g_k(x) + \underbrace{\sum_{\substack{j=1\\ j\neq k}} g_j(x)}_{f_j(x)}.$$

The arguments from the proof of the preceding lemma can be applied to show that the function f_j defined by the sum on the right is differentiable at x_k. Since g_k is not differentiable at x_k, we conclude that f cannot be differentiable at x_k by Proposition 3.2.10. ▼

This shows that the conclusions of Proposition 3.2.29(iv) cannot generally be improved. ●

3.2.7 Piecewise differentiable functions

In Section 3.1.7 we considered functions that were piecewise continuous. In this section we consider a class of piecewise continuous functions that have additional properties concerning their differentiability. We let $I \subseteq \mathbb{R}$ be an interval with $f\colon I \to \mathbb{R}$ a function. In Section 3.1.7 we defined the notation $f(x-)$ and $f(x+)$. Here we also define

$$f'(x-) = \lim_{\epsilon \downarrow 0} \frac{f(x-\epsilon) - f(x-)}{-\epsilon}, \quad f'(x+) = \lim_{\epsilon \downarrow 0} \frac{f(x+\epsilon) - f(x+)}{\epsilon}.$$

These limits, of course, may fail to exist, or even to make sense if $x \in \mathrm{bd}(I)$.

Now, recalling the notion of a partition from Definition 2.5.7, we make the following definition.

3.2.32 Definition (Piecewise differentiable function) A function $f\colon [a,b] \to \mathbb{R}$ is ***piecewise differentiable*** if there exists a partition $P = (I_1, \dots, I_k)$, with $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$, of $[a,b]$ with the following properties:

(i) $f|\,\mathrm{int}(I_j)$ is differentiable for each $j \in \{1, \dots, k\}$;

(ii) for $j \in \{1, \dots, k-1\}$, the limits $f(x_j+)$, $f(x_j-)$, $f'(x_j+)$, and $f'(x_j-)$ exist;

(iii) the limits $f(a+)$, $f(b-)$, $f'(a+)$, and $f'(b-)$ exist. •

It is evident that a piecewise differentiable function is piecewise continuous. It is not surprising that the converse is not true, and a simple example of this will be given in the following collection of examples.

3.2.33 Examples (Piecewise differentiable functions)

1. Let $I = [-1, 1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1 + x, & x \in [-1, 0], \\ 1 - x, & (0, 1]. \end{cases}$$

One verifies that f is differentiable on $(-1, 0)$ and $(0, 1)$. Moreover, we compute the limits

$$\begin{gathered} f(-1+) = 0, \quad f'(-1+) = 1, \quad f(1-) = 0, \quad f'(1-) = -1, \\ f(0-) = 1, \quad f(0+) = 1, \quad f'(0-) = 1, \quad f'(0+) = -1. \end{gathered}$$

Thus f is piecewise differentiable. Note that f is also continuous.

2. Let $I = [-1, 1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = \operatorname{sign}(x)$. On $(-1, 0)$ and $(0, 1)$ we note that f is differentiable. Moreover, we compute

$$\begin{gathered} f(-1+) = -1, \quad f'(-1+) = 0, \quad f(1-) = 1, \quad f'(1-) = 0, \\ f(0-) = -1, \quad f(0+) = 1, \quad f'(0-) = 0, \quad f'(0+) = 0. \end{gathered}$$

Note that it is important here to *not* compute the limits $f'(0-)$ and $f'(0+)$ using the formulae

$$\lim_{\epsilon \downarrow 0} \frac{f(0-\epsilon) - f(0)}{-\epsilon}, \quad \lim_{\epsilon \downarrow 0} \frac{f(0+\epsilon) - f(0)}{\epsilon}.$$

Indeed, these limits do not exist, where as the limits $f'(0-)$ and $f'(0+)$ do exist. In any event, f is piecewise differentiable, although it is not continuous.

3. Let $I = [0, 1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = \sqrt{x(1-x)}$. On $(0, 1)$, f is differentiable. Also, the limits $f(0+)$ and $f(1-)$ exist. However, the limits $f'(0+)$ and $f'(1-)$ do not exist, as we saw in Example 3.2.3–3. Thus f is not piecewise differentiable. However, it is continuous, and therefore piecewise continuous, on $[0, 1]$. •

3.2.8 Notes

It was Weierstrass who first proved the existence of a continuous but nowhere differentiable function. The example Weierstrass gave was

$$\tilde{f}(x) = \sum_{j=0}^{\infty} b^n \cos(a^n \pi x),$$

where $b \in (0,1)$ and a satisfies $ab > \frac{3}{2}\pi + 1$. It requires a little work to show that this function is nowhere differentiable. The example we give as Example 3.2.9 is fairly simple by comparison, and is taken from the paper of McCarthy [1953].

Example 3.2.31–2 if from [Siksek and El-Sedy 2004]

Exercises

3.2.1 Let $I \subseteq \mathbb{R}$ be an interval and let $f, g\colon I \to \mathbb{R}$ be differentiable. Is it true that the functions

$$I \ni x \mapsto \min\{f(x), g(x)\} \in \mathbb{R}, \qquad I \ni x \mapsto \max\{f(x), g(x)\} \in \mathbb{R},$$

are differentiable? If it is true provide a proof, if it is not true, give a counterexample.

Section 3.3

ℝ-valued functions of bounded variation

In this section we present a class of functions, functions of so-called bounded variation, that are larger than the set of differentiable functions. However, they are sufficiently friendly that they often play a distinguished rôle in certain parts of signal theory, as evidenced by the theorems of Jordan concerning inversion of Fourier transforms (see Theorems IV-5.2.31 and IV-6.2.24). It is often not obvious after an initial reading on the topic of functions of bounded variation, just why such functions are important. Historically, the class of functions of bounded variation arose out of the desire to understand functions that are sums of functions that are monotonically increasing (see Definition 3.1.27 for the definition). Indeed, as we shall see in Theorem 3.3.3, functions of bounded variation and monotonically increasing functions are inextricably linked. The question about the importance of functions of bounded variation can thus be reduced to the question about the importance of monotonically increasing functions. An intuitive reason why such functions might be interesting is that many of the functions one encounters in practice, while not themselves increasing or decreasing, have intervals on which they *are* increasing or decreasing. Thus one hopes that, by understanding increasing or decreasing functions, one can understand more general functions.

It is also worth mentioning here that the class of functions of bounded variation arise in functional analysis as the topological dual to Banach spaces of continuous functions. In this regard, we refer the reader to Theorem III-2.12.6.

Do I need to read this section? This section should be strongly considered for omission on a first read, and then referred to when the concept of bounded variation comes up in subsequent chapters, namely in Chapters IV-5 and IV-6. Such an omission is suggested, not because the material is unimportant or uninteresting, but rather because it constitutes a significant diversion that might be better left until it is needed. •

3.3.1 Functions of bounded variation on compact intervals

In this section we define functions of bounded variation on intervals that are compact. In the next section we shall extend these ideas to general intervals. For a compact interval I, recall that $\mathrm{Part}(I)$ denotes the set of partitions of I, and that if $P \in \mathrm{Part}(I)$ then $\mathrm{EP}(P)$ denotes the endpoints of the intervals comprising P (see the discussion surrounding Definition 2.5.7).

3.3.1 Definition (Total variation, function of bounded variation) For $I = [a, b]$ a compact interval and $f \colon I \to \mathbb{R}$ a function on I, the ***total variation*** of f is given by

$$\mathrm{TV}(f) = \sup\left\{ \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| \;\middle|\; (x_0, x_1, \dots, x_k) = \mathrm{EP}(P),\ P \in \mathrm{Part}([a, b]) \right\}.$$

If $\mathrm{TV}(f) < \infty$ then f has ***bounded variation***. •

Let us characterise real functions of bounded variation on compact intervals. The principal part of this characterisation is the decomposition of a function of bounded variation into the difference of monotonically increasing functions. However, another interesting characterisation involves the following idea which relies on the notion of the graph of a function, introduced following Definition 1.3.1.

3.3.2 Definition (Arclength of the graph of a function) Let $[a, b]$ be a compact interval and let $f \colon [a, b] \to \mathbb{R}$ be a function. The ***arclength*** of graph(f) is defined to be

$$\ell(\mathrm{graph}(f)) = \sup\left\{ \sum_{j=1}^{k} \left((f(x_j) - f(x_{j-1}))^2 + (x_j - x_{j-1})^2\right)^{1/2} \;\middle|\; (x_0, x_1, \dots, x_k) = \mathrm{EP}(P),\ P \in \mathrm{Part}([a, b]) \right\}. \quad \bullet$$

We now have the following result which characterises functions of bounded variation.

3.3.3 Theorem (Characterisation of functions of bounded variation) *For a compact interval* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *and a function* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$*, the following statements are equivalent:*

(i) f *has bounded variation;*

(ii) *there exists monotonically increasing functions* $\mathrm{f}_+, \mathrm{f}_- \colon \mathrm{I} \to \mathbb{R}$ *such that* $\mathrm{f} = \mathrm{f}_+ - \mathrm{f}_-$ *(**Jordan**[8] **decomposition** of a function of bounded variation);*

(iii) *the graph of* f *has finite arclength in* $\mathbb{R}^2$.

Furthermore, each of the preceding three statements implies the following:

(iv) *the following limits exist:*

(a) f(a+);

(b) f(b−);

(c) f(x+) *and* f(x−) *for all* $\mathrm{x} \in \mathrm{int}(\mathrm{I})$,

(v) f *is continuous except at a countable number of points in* I,

[8] Marie Ennemond Camille Jordan (1838–1922) was a French mathematician who made significant contributions to the areas of algebra, analysis, complex analysis, and topology. He wrote a three volume treatise on analysis entitled *Cours d'analyse de l'École Polytechnique* which was quite influential.

(vi) f *possesses a derivative almost everywhere in* I.

Proof (i) $\Longrightarrow$ (ii) Define $\mathrm{V}(f)(x) = \mathrm{TV}(f|[a,x])$ so that $x \mapsto \mathrm{V}(f)(x)$ is a monotonic function. Let us define

$$f_+(x) = \tfrac{1}{2}(\mathrm{V}(f)(x) + f(x)), \quad f_-(x) = \tfrac{1}{2}(\mathrm{V}(f)(x) - f(x)). \tag{3.11}$$

Since we obviously have $f = f_+ - f_-$, this part of the theorem will follow if f_+ and f_- can be shown to be monotonic. Let $\xi_2 > \xi_1$ and let $(x_0, x_1, \dots, x_k)$ be the endpoints of a partition of $[a, \xi_1]$. Then $(x_0, x_1, \dots, x_k, x_{k+1} = \xi_2)$ are the endpoints of a partition of $[a, \xi_2]$. We have the inequalities

$$\mathrm{V}(f)(\xi_2) \geq \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| + |f(\xi_2) - f(\xi_1)|.$$

Since this is true for any partition of $[a, \xi_1]$ we have

$$\mathrm{V}(f)(\xi_2) \geq \mathrm{V}(f)(\xi_1) + |f(\xi_2) - f(\xi_1)|.$$

We then have

$$\begin{aligned} 2f_+(\xi_2) &= \mathrm{V}(f)(\xi_2) + f(\xi_2) \\ &\geq \mathrm{V}(f)(\xi_1) + f(\xi_1) + |f(\xi_2) - f(\xi_1)| + f(\xi_2) - f(\xi_1) \\ &\geq \mathrm{V}(f)(\xi_1) + f(\xi_1) = 2f_+(\xi_1) \end{aligned}$$

and

$$\begin{aligned} 2f_-(\xi_2) &= \mathrm{V}(f)(\xi_2) - f(\xi_2) \\ &\geq \mathrm{V}(f)(\xi_1) - f(\xi_1) + |f(\xi_2) - f(\xi_1)| - f(\xi_2) + f(\xi_1) \\ &\geq \mathrm{V}(f)(\xi_1) - f(\xi_1) = 2f_+(\xi_1), \end{aligned}$$

giving this part of the theorem.

(ii) $\Longrightarrow$ (i) If f is monotonically increasing and if $(x_0, x_1, \dots, x_k)$ are the endpoints for a partition of $[a, b]$, then

$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| = \sum_{j=1}^{k} (f(x_j) - f(x_{j-1})) = f(b) - f(a).$$

Thus monotonically increasing functions, and similarly monotonically decreasing functions, have bounded variation. Now consider two functions f and g, both of bounded variation. By part (i) of Proposition 3.3.12, $f + g$ is also of bounded variation. In particular, the sum of a monotonically increasing and a monotonically decreasing function will be a function of bounded variation.

(i) $\Longleftrightarrow$ (iii) First we note that, for any $a, b \in \mathbb{R}$,

$$(|a| + |b|)^2 = a^2 + b^2 + 2|a||b|,$$

from which we conclude that $(a^2 + b^2)^{1/2} \leq |a| + |b|$. Therefore, if $(x_0, x_1, \dots, x_k)$ are the endpoints of a partition of $[a, b]$, then

$$\begin{aligned}\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| &\leq \sum_{j=1}^{k} \Big((f(x_j) - f(x_{j-1}))^2 + (x_j - x_{j-1})^2 \Big)^{1/2} \\ &\leq \sum_{j=1}^{k} \Big(|f(x_j) - f(x_{j-1})| + |x_j - x_{j-1}| \Big) = \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| + b - a. \quad (3.12)\end{aligned}$$

This implies that

$$\mathrm{TV}(f) \leq \ell(\mathrm{graph}(f)) \leq \mathrm{TV}(f) + b - a,$$

from which this part of the result follows.

(iv) Let f_+ and f_- be monotonically increasing functions as per part (ii). By Theorem 3.1.28 we know that the limits asserted in this part of the theorem hold for both f_+ and f_-. This part of the theorem now follows from Propositions 2.3.23 and 2.3.29.

(v) This follows from Theorem 3.1.28 and Proposition 3.1.15, using the decomposition $f = f_+ - f_-$ from part (ii).

(vi) Again using the decomposition $f = f_+ - f_-$ from part (ii), this part of the theorem follows from Theorem 3.2.26 and Proposition 3.2.10. ■

3.3.4 Remark We comment the converses of parts (iv), (v), and (vi) of Theorem 3.3.3 do not generally hold. This is because, as we shall see in Example 3.3.5–4, continuous functions are not necessarily of bounded variation. •

Let us give some examples of functions that have and do not have bounded variation.

3.3.5 Examples (Functions of bounded variation on compact intervals)

1. On $[0, 1]$ define $f\colon [0, 1] \to \mathbb{R}$ by $f(x) = c$, for $c \in \mathbb{R}$. We easily see that $\mathrm{TV}(f) = 0$, so f has bounded variation.
2. On $[0, 1]$ consider the function $f\colon [0, 1] \to \mathbb{R}$ defined by $f(x) = x$. We claim that f has bounded variation. Indeed, if $(x_0, x_1, \dots, x_k)$ are the endpoints of a partition of $[0, 1]$, then we have

$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| = \sum_{j=1}^{k} |x_j - x_{j-1}| = 1 - 0 = 1,$$

thus giving f as having bounded variation.

Note that f is itself a monotonically increasing function, so that for part (ii) of Theorem 3.3.3 we may take $f_+ = f$ and f_- to be the zero function. However, we can also write $f = g_+ - g_-$ where $g_+(x) = 2x$ and $g_-(x) = x$. Thus the decomposition of part (ii) of Theorem 3.3.3 is not unique.

3. On $I = [0, 1]$ consider the function

$$f(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}] \\ -1, & x \in (\frac{1}{2}, 1]. \end{cases}$$

We claim that $\mathrm{TV}(f) = 1$. Let $(x_0, x_1, \dots, x_k)$ be the endpoints of a partition of $[0, 1]$. Let $\bar{k}$ be the least element in $\{1, \dots, k\}$ for which $x_{\bar{k}} > \frac{1}{2}$. Then we have

$$\begin{aligned}\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| = \sum_{j=1}^{\bar{k}-1} |f(x_j) - f(x_{j-1})| + \sum_{j=\bar{k}+1}^{k} |f(x_j) - f(x_{j-1})| \\ + |f(x_{\bar{k}}) - f(x_{\bar{k}-1})| = 1.\end{aligned}$$

This shows that $\mathrm{TV}(f) = 1$ and so f has bounded variation. Note that this also shows that functions of bounded variation need not be continuous. This, along with the next example, shows that the relationship between continuity and bounded variation is not a straightforward one.

4. Consider the function on $I = [0, 1]$ defined by

$$f(x) = \begin{cases} x \sin \frac{1}{x}, & x \in (0, 1], \\ 0, & x = 0. \end{cases}$$

We first claim that f is continuous. Clearly it is continuous at x provided that $x \neq 0$. To show continuity at $x = 0$, let $\epsilon \in \mathbb{R}_{>0}$ and note that, if $x < \epsilon$, we have $|f(x)| < \epsilon$, thus showing continuity.

However, f does not have bounded variation. Indeed, for $j \in \mathbb{Z}_{>0}$ denote $\xi_j = \frac{1}{(j+\frac{1}{2})\pi}$. Then, for $k \in \mathbb{Z}_{>0}$, consider the partition with endpoints

$$(x_0 = 0, x_1 = \xi_k, \dots, x_k + = \xi_1, x_{k+1} = 1).$$

Direct computation then gives

$$\begin{aligned}\sum_{j=1}^{k+1} |f(x_j) - f(x_{j-1})| &\geq \frac{2}{\pi} \sum_{j=1}^{k} \left| \frac{(-1)^j}{2j+1} - \frac{(-1)^{j-1}}{2j-1} \right| \\ &= \frac{2}{\pi} \sum_{j=1}^{k} \left| \frac{1}{2j+1} + \frac{1}{2j-1} \right| \geq \frac{2}{\pi} \sum_{j=1}^{k} \left| \frac{2}{2j+1} \right|.\end{aligned}$$

Thus

$$\mathrm{TV}(f) \geq \frac{2}{\pi} \sum_{j=1}^{\infty} \left| \frac{2}{2j+1} \right| = \infty,$$

showing that f has unbounded variation. •

3.3.2 Functions of bounded variation on general intervals

Now, with the definitions and properties of bounded variation for functions defined on compact intervals, we can sensibly define notions of variation for general intervals.

3.3.6 Definition (Bounded variation, locally bounded variation) Let I be an interval with $f\colon I \to \mathbb{R}$ a function.

(i) If $f|[a,b]$ is a function of bounded variation for every compact interval $[a,b] \subseteq I$, then f is a function of ***locally bounded variation***.

(ii) If $\sup\{\mathrm{TV}(f|[a,b]) \mid [a,b] \subseteq I\} < \infty$, then f is a function of ***bounded variation***. •

3.3.7 Remark (Properties of functions of locally bounded variation) We comment that the characterisations of functions of bounded variation given in Theorem 3.3.3 carry over to functions of locally bounded variation in the sense that the following statements are equivalent for a function $f\colon I \to \mathbb{R}$ defined on a general interval I:

1. f has locally bounded variation;
2. there exists monotonically increasing functions $f_+, f_-\colon I \to \mathbb{R}$ such that $f = f_+ - f_-$.

Furthermore, each of the preceding two statements implies the following:

3. the following limits exist:
   (a) $f(a+)$;
   (b) $f(b-)$;
   (c) $f(x+)$ and $f(x-)$ for all $x \in \mathrm{int}(I)$,
4. f is continuous except at a countable number of points in I,
5. f possesses a derivative almost everywhere in I.

These facts follow easily from the definition of locally bounded variation, along with facts about countable sets, and sets of measure zero. We leave the details to the reader as Exercise 3.3.4. •

3.3.8 Notation ("Locally bounded variation" versus "bounded variation") These extended definitions agree with the previous ones in that, when I is compact, (1) the new definition of a function of bounded variation agrees with that of Definition 3.3.1 and (2) the definition of a function of bounded variation agrees with the definition of a function of locally bounded variation. The second point is particularly important to remember, because most of the results in the remainder of this section will be stated for functions of locally bounded variation. Our observation here is that these results automatically apply to functions of bounded variation, as per Definition 3.3.1. For this reason, we will generally default from now on to using "locally bounded variation" in place of "bounded variation," reserving the latter for when it is intended in its distinct place when the interval of definition of a function is compact. •

Let us give some examples of functions that do and no not have locally bounded variation.

3.3.9 Examples (Functions of locally bounded variation on general intervals)

1. Let $I \subseteq \mathbb{R}$ be an arbitrary interval, let $c \in \mathbb{R}$, and consider the function $f \colon I \to \mathbb{R}$ defined by $f(x) = c$. Applying the definition shows that $\mathrm{TV}(f|[a,b])(x) = 0$ for all compact intervals $[a,b] \subseteq I$, no matter the character of I. Thus constant functions, unsurprisingly, have locally bounded variation.
2. Let us consider the function $f \colon I \to \mathbb{R}$ on $I = [0,\infty)$ defined by $f(x) = x$. We claim that f has locally bounded variation. Indeed, let $[a,b] \subseteq I$ and consider a partition of $[a,b]$ with endpoints $(x_0, x_1, \dots, x_k)$. We have
$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| = \sum_{j=1}^{k} (x_j - x_{j-1}) = b - a.$$
This shows that f has locally bounded variation. However, since $b - a$ can be arbitrarily large, f does not have bounded variation.
3. On the interval $I = (0,1]$ consider the function $f \colon I \to \mathbb{R}$ defined by $f(x) = \frac{1}{x}$. Note that, for $[a,b] \subseteq (0,1]$, the function $f|[a,b]$ is monotonically decreasing, and so has bounded variation. We can thus conclude that f is a function of locally bounded variation. We claim that f does not have bounded variation. To see this, note that if $(x_0, x_1, \dots, x_k)$ are the endpoints of a partition of $[a,b] \subseteq (0,1]$, then it is easy to see that, since f is strictly monotonically decreasing and continuous that $(f(x_k), \dots, f(x_1), f(x_0))$ are the endpoints of a partition of $[f(x_k), f(x_0)]$. We thus have
$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| = f(x_0) - f(x_k).$$
Since $f(x_0)$ can be made arbitrarily large by choosing a small, it follows that f cannot have bounded variation. •

We close this section by introducing the notion of the variation of a function, and giving a useful property of this concept.

3.3.10 Definition (Variation of a function of bounded variation) Let $I \subseteq \mathbb{R}$ be an interval, let $a \in I$, let $f \colon I \to \mathbb{R}$ be a function of locally bounded variation, and define $\mathrm{V}_a(f) \colon I \to \mathbb{R}_{>0}$ by
$$\mathrm{V}_a(f)(x) = \begin{cases} \mathrm{TV}(f|[x,a]), & x < a, \\ 0, & x = a, \\ \mathrm{TV}(f|[a,x]), & x > a. \end{cases}$$
The function $\mathrm{V}_a(f)$ is the ***variation*** of f with reference point a. •

One can easily check that the choice of a in the definition of $\mathrm{V}_a(f)$ serves only to shift the values of the function. Thus the essential features of the variation are independent of the reference point.

When a function of bounded variation is continuous, so too is its variation.

3.3.11 Proposition (The variation of a continuous function is continuous and vice versa) *Let $I \subseteq \mathbb{R}$ be an interval, let $a \in I$, and let $f\colon I \to \mathbb{R}$ be a function of locally bounded variation. Then f is continuous at $x \in I$ if and only if $\mathrm{V}_a(f)$ is continuous at x. Moreover, if f is a continuous function of bounded variation, then $f = f_+ - f_-$ where f_+ and f_- are continuous monotonically increasing functions.*

Proof The general result follows easily from the case where $I = [a, b]$ is compact. Furthermore, in this case it suffices to consider the variation of f with reference points a or b. We shall consider only the reference point a, since the other case follows in much the same manner.

Suppose that f is continuous at $x_0 \in I$ and let $\epsilon \in \mathbb{R}_{>0}$. First suppose that $x_0 \in [a, b)$, and let $\delta \in \mathbb{R}_{>0}$ be chosen such that $x \in \mathsf{B}(\delta, x_0) \cap I$ implies that $|f(x) - f(x_0)| < \frac{\epsilon}{2}$. Choose a partition of $[x_0, b]$ with endpoints $(x_0, x_1, \dots, x_k)$ such that

$$\mathrm{TV}(f|[x_0, b]) - \tfrac{\epsilon}{2} \le \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})|. \tag{3.13}$$

We may without loss of generality suppose that $x_1 - x_0 < \delta$. Indeed, if this is not the case, we may add a new endpoint to our partition, noting that the estimate (3.13) will hold for the new partition. We then have

$$\begin{aligned}\mathrm{TV}(f|[x_0, b]) - \tfrac{\epsilon}{2} &\le |f(x_1) - f(x_0)| + \sum_{j=2}^{k} |f(x_j) - f(x_{j-1})| \\ &\le \tfrac{\epsilon}{2} + \sum_{j=2}^{k} |f(x_j) - f(x_{j-1})| \le \tfrac{\epsilon}{2} + \mathrm{TV}(f|[x_1, b]).\end{aligned}$$

This then gives

$$\mathrm{TV}(f|[x_0, b]) - \mathrm{TV}(f|[x_1, b]) = \mathrm{V}_a(f)(x_1) - \mathrm{V}_a(f)(x_0) < \epsilon.$$

Since this holds for any partition for which $x_1 - x_0 < \delta$, it follows that $\lim_{x \downarrow x_0} \mathrm{V}_a(f)(x) = \mathrm{V}_a(f)(x_0)$ for every $x_0 \in [a, b)$ at which f is continuous. One can similarly show that $\lim_{x \uparrow x_0} \mathrm{V}_a(f)(x) = \mathrm{V}_a(f)(x_0)$ for every $x_0 \in (a, b]$ at which f is continuous. This gives the result by Theorem 3.1.3.

Suppose that $\mathrm{V}_a(f)$ is continuous at $x_0 \in I$ and let $\epsilon \in \mathbb{R}_{>0}$. Choose $\delta \in \mathbb{R}_{>0}$ such that $|\mathrm{V}_a(f)(x) - \mathrm{V}_a(f)(x_0)| < \epsilon$ for $x \in \overline{\mathsf{B}}(2\delta, x_0)$. Then, for $x \in \overline{\mathsf{B}}(2\delta, x_0)$ with $x > x_0$,

$$|f(x) - f(x_0)| \le \mathrm{TV}(f|[x_0, x]) = \mathrm{V}_a(f)(x) - \mathrm{V}_a(f)(x_0) < \epsilon,$$

using the fact that (x_0, x) are the endpoints of a partition of $[x_0, x]$. In like manner, if $x \in \overline{\mathsf{B}}(2\delta, x_0)$ with $x > x_0$, then

$$|f(x) - f(x_0)| \le \mathrm{TV}(f|[x, x_0]) = \mathrm{V}_a(f)(x_0) - \mathrm{V}_a(f)(x) < \epsilon.$$

Thus $|f(x) - f(x_0)| < \epsilon$ for every $x \in \overline{\mathsf{B}}(2\delta, x_0)$, and so for every $x \in \mathsf{B}(\delta, x_0)$, giving continuity of f at x_0.

The final assertion follows from the definition of the Jordan decomposition given in (3.11). ■

3.3.3 Bounded variation and operations on functions

In this section we illustrate how functions of locally bounded variation interact with the usual operations one performs on functions.

3.3.12 Proposition (Addition and multiplication, and locally bounded variation) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be functions of locally bounded variation. Then the following statements hold:*

(i) $\mathrm{f} + \mathrm{g}$ *is a function of locally bounded variation;*

(ii) fg *is a function of locally bounded variation;*

(iii) if additionally there exists $\alpha \in \mathbb{R}_{>0}$ *such that* $|\mathrm{g}(\mathrm{x})| \geq \alpha$ *for all* $\mathrm{x} \in \mathrm{I}$*, then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is a function of locally bounded variation.*

Proof Without loss of generality we may suppose that $I = [a, b]$ is a compact interval.

(i) Let $(x_0, x_1, \dots, x_k)$ be the endpoints for a partition of $[a, b]$ and compute

$$\sum_{j=1}^{k} |f(x_j) + g(x_j) - f(x_{j-1}) - g(x_{j-1})| \leq \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| + \sum_{j=1}^{k} |g(x_j) - g(x_{j-1})|$$

using the triangle inequality. It then follows from Proposition 2.2.27 that $\mathrm{TV}(f + g) \leq \mathrm{TV}(f) + \mathrm{TV}(g)$, and so $f + g$ has locally bounded variation.

(ii) Let

$$M_f = \sup\{|f(x)| \mid x \in [a, b]\}, \quad M_g = \sup\{|g(x)| \mid x \in [a, b]\}.$$

Then, for a partition of $[a, b]$ with endpoints $(x_0, x_1, \dots, x_k)$, compute

$$\begin{aligned}\sum_{j=1}^{k} |f(x_j)g(x_j) - f(x_{j-1})g(x_{j-1})| &\leq \sum_{j=1}^{k} |f(x_j)g(x_j) - f(x_{j-1})g(x_j)| \\ &\quad + \sum_{j=1}^{k} |f(x_{j-1})g(x_j) - f(x_{j-1})g(x_{j-1})| \\ &\leq \sum_{j=1}^{k} M_g |f(x_j) - f(x_{j-1})| + \sum_{j=1}^{k} M_f |g(x_j) - g(x_{j-1})| \\ &\leq M_g \, \mathrm{TV}(f) + M_f \, \mathrm{TV}(g),\end{aligned}$$

giving the result.

(iii) Let $(x_0, x_1, \dots, x_k)$ be a partition of $[a, b]$ and compute

$$\sum_{j=1}^{k} \left| \frac{1}{g(x_j)} - \frac{1}{g(x_{j-1})} \right| = \sum_{j=1}^{k} \left| \frac{g(x_{j-1}) - g(x_j)}{g(x_j)g(x_{j-1})} \right| \leq \sum_{j=1}^{k} \left| \frac{g(x_j) - g(x_{j-1})}{\alpha^2} \right| \leq \frac{\mathrm{TV}(g)}{\alpha^2}.$$

Thus $\frac{1}{g}$ has locally bounded variation, and this part of the result follows from part (ii). ■

Next we show that to determine whether a function has locally bounded variation, one can break up the interval of definition into subintervals.

3.3.13 Proposition (Locally bounded variation on disjoint subintervals) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{I} = \mathrm{I}_1 \cup \mathrm{I}_2$*, where* $\mathrm{I}_1 \cap \mathrm{I}_2 = \{\mathrm{c}\}$*, where* c *is the right endpoint of* I_1 *and the left endpoint of* I_2*. Then* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *has locally bounded variation if and only if* $\mathrm{f}|\mathrm{I}_1$ *and* $\mathrm{f}|\mathrm{I}_2$ *have locally bounded variation.*

Proof It suffices to consider the case where $I = [a, b]$, $I_1 = [a, c]$, and $I_2 = [c, b]$. First let $(x_0, x_1, \dots, x_k)$ be the endpoints of a partition of $[a, c]$ and let $(y_0, y_1, \dots, y_l)$ be the endpoints of a partition of $[c, b]$. Then

$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| + \sum_{j=1}^{l} |f(y_j) - f(y_{j-1})| \leq \mathrm{TV}(f),$$

which shows that $\mathrm{TV}(f|[a, c]) + \mathrm{TV}(f|[c, b]) \leq \mathrm{TV}(f)$. Now let $(x_0, x_1, \dots, x_k)$ be the endpoints of a partition of $[a, b]$. If c is not one of the endpoints, then let $m \in \{1, \dots, k-1\}$ satisfy $x_{m-1} < c < x_m$, and define a new partition with endpoints

$$(y_0 = x_0, y_1 = x_1, \dots, ym - 1 = x_{m-1}, y_m = c, y_{m+1} = x_m, \dots, y_{k+1} = x_k).$$

Then

$$\begin{aligned} \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| &\leq \sum_{j=1}^{k+1} |f(y_j) - f(y_{j-1})| \\ &\leq \sum_{j=1}^{m} |f(y_j) - f(y_{j-1})| + \sum_{j=m}^{m+1} |f(y_j) - f(y_{j-1})| \\ &\leq \mathrm{TV}([a, c]) + \mathrm{TV}(f|[c, b]). \end{aligned}$$

This shows that $\mathrm{TV}(f) \leq \mathrm{TV}(f|[a, c]) + \mathrm{TV}(f|[c, b])$, which gives the result when combined with our previous estimate $\mathrm{TV}(f|[a, c]) + \mathrm{TV}(f|[c, b]) \leq \mathrm{TV}(f)$. ■

While Examples Example 3.3.5–3 and 4 illustrate that functions of locally bounded variation need not be continuous, and that continuous functions need not have locally bounded variation, the story for differentiability is more pleasant.

3.3.14 Proposition (Differentiable functions have locally bounded variation) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if the function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is differentiable with the derivative* f' *being locally bounded, then* f *has locally bounded variation. In particular, if* f *is of class* C^1*, then* f *is of locally bounded variation.*

Proof The general result follows from the case where $I = [a, b]$, so we suppose in the proof that I is compact. Let $(x_0, x_1, \dots, x_k)$ be a partition of $[a, b]$. By the Mean Value Theorem, for each $j \in \{1, \dots, k\}$ there exists $y_j \in (x_{j-1}, x_j)$ such that

$$f(x_j) - f(x_{j-1}) = f'(y_j)(x_j - x_{j-1}).$$

Moreover, since f' is bounded, let $M \in \mathbb{R}_{>0}$ satisfy $|f'(x)| < M$ for each $x \in [a,b]$. Then

$$\sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| = \sum_{j=1}^{k} |f'(y_j)||x_j - x_{j-1}| \leq \sum_{j=1}^{k} M|x_j - x_{j-1}| = M(b-a).$$

The final assertion follows since, if f is of class C^1, then f' is continuous and so bounded by Theorem 3.1.22. ■

In the preceding result we asked that the derivative be locally bounded. This condition is essential, as the following example shows.

3.3.15 Example (A differentiable function that does not have bounded variation) We take $f\colon [-1,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x^2 \sin(\frac{1}{x^2}), & x \neq 0, \\ 0, & x = 0. \end{cases}$$

We will show that this function is differentiable but does not have bounded variation. The differentiability of f at $x \neq 0$ follows from the product rule and the Chain Rule since the functions $x \mapsto x^2$, $x \mapsto \frac{1}{x^2}$, and sin are all differentiable away from zero. Indeed, by the product and Chain Rule we have

$$f'(x) = 2x \sin(\tfrac{1}{x^2}) - \tfrac{2}{x} \cos(\tfrac{1}{x^2}).$$

For differentiability at $x = 0$ we compute

$$\lim_{h \to 0} \frac{f(0+h) - f(0)}{h} = \lim_{h \to 0} \frac{h^2 \sin(\frac{1}{h^2}) - 0}{h} = \lim_{h \to 0} h \sin(\tfrac{1}{h^2}) = 0,$$

giving the derivative at $x = 0$ to be zero.

To show that f does not have bounded variation, for $j \in \mathbb{Z}_{>0}$ define

$$\xi_j = \frac{1}{\sqrt{(j + \frac{1}{2})\pi}}.$$

For $k \in \mathbb{Z}_{>0}$ define a partition of $[0,1]$ by asking that it have endpoints $(x_0, x_1 = \xi_k, \dots, x_k = \xi_1, x_{k+1})$. Then

$$\begin{aligned} \sum_{j=1}^{k+1} |f(x_j) - f(x_{j-1})| \geq \sum_{j=1}^{k} |f(x_j) - f(x_{j-1})| &= \frac{2}{\pi} \sum_{j=1}^{k} \left| \frac{(-1)^j}{2j+1} - \frac{(-1)^{j-1}}{2j-1} \right| \\ &\geq \frac{2}{\pi} \sum_{j=1}^{k} \left| \frac{1}{2j+1} + \frac{1}{2j-1} \right| \geq \frac{2}{\pi} \sum_{j=1}^{k} \left| \frac{2}{2j+1} \right|. \end{aligned}$$

Thus

$$\mathrm{TV}(f) \geq \frac{2}{\pi}\sum_{j=1}^{\infty}\left|\frac{2}{2j+1}\right| = \infty,$$

giving our assertion that f does not have bounded variation.

Note that it follows from Proposition 3.3.14 that f' is not bounded. This can be verified explicitly as well. •

While the composition of continuous functions is again a continuous function, and the composition of differentiable functions is again a differentiable function, the same assertion does not hold for functions of locally bounded variation.

3.3.16 Example (Compositions of functions of locally bounded variation need not be functions of locally bounded variation) Let $I = [-1, 1]$ and define $f, g\colon I \to \mathbb{R}$ by $f(x) = x^{1/3}$ and

$$g(x) = \begin{cases} x^3(\sin\frac{1}{x})^3, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

We claim that f and g are functions of bounded variation. To show that f has bounded variation, we note that f is monotonically increasing, and so necessarily of bounded variation by Theorem 3.3.3(ii). To show that g is of bounded variation, we shall show that it is of class C^1, and then use Proposition 3.3.14. Clearly g is differentiable with continuous derivative on the intervals $[-1, 0)$ and $(0, 1]$. Thus we need to show that g is differentiable at 0 with continuous derivative there. To see that g is differentiable at 0, we compute

$$\lim_{x\to 0}\frac{g(x) - g(0)}{x - 0} = \lim_{x\to 0} x^2(\sin\tfrac{1}{x})^{1/3} = 0,$$

since $\left|(\sin\frac{1}{x})^{1/3}\right| \leq 1$. Thus $g'(0) = 0$. We also can readily compute that $\lim_{x\downarrow 0} g'(x) = \lim_{x\uparrow 0} g'(x) = 0$. Thus g' is also continuous at 0, so showing that g has bounded variation.

However, note that

$$f \circ g(x) = \begin{cases} x\sin\frac{1}{x}, & x \neq 0, \\ 0, & x = 0, \end{cases}$$

and in Example 3.3.5–4 we showed that this function does not have bounded variation on the interval $[0, 1]$. Therefore, it cannot have bounded variation on the interval $[-1, 1]$. This gives our desired conclusion that $f \circ g$ is not a function of bounded variation, even though both f and g are. •

3.3.4 Saltus functions

As we saw in part (v) of Theorem 3.3.3, a function of locally bounded variation is discontinuous at a countable set of points. Moreover, part (iv) of the same theorem indicates that all discontinuities are jump discontinuities. In the next section we shall see that it is possible to separate out these discontinuities into a single function which, when subtracted from a function of locally bounded variation, leaves a *continuous* function of locally bounded variation.

First we give a general definition, unrelated specifically to functions of locally bounded variation. For this definition we recall from Section 2.4.7 our discussion of sums over arbitrary index sets.

3.3.17 Definition (Saltus function) Let $I \subseteq \mathbb{R}$ be an interval and let I' be the interval obtained by removing the right endpoint from I, if I indeed contains its right endpoint; otherwise take $I' = I$. A ***saltus function***[9] on I is a function $j\colon I \to \mathbb{R}$ of the form

$$j(x) = \sum_{\xi\in(-\infty,x)\cap I} r_\xi + \sum_{\xi\in(-\infty,x]\cap I} l_\xi,$$

where $(r_\xi)_{\xi\in I'}$ and $(l_\xi)_{\xi\in I}$ are summable families of real numbers. •

This definition seems mildly ridiculous at a first read, in that there seems to be no reason why such a function should be of any interest. However, as we shall see, every function of locally bounded variation naturally gives rise to a saltus function. Before we get to this, let us look at some properties of saltus function. It might be helpful to note that the function of Example 3.2.28 is a saltus function, as is easily seen from its definition. Many of the general properties of saltus functions follow in the same manner as they did for that example.

3.3.18 Proposition (Continuity of saltus functions) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{j}\colon \mathrm{I} \to \mathbb{R}$ *is a saltus function given by*

$$\mathrm{j}(\mathrm{x}) = \sum_{\xi\in(-\infty,\mathrm{x})\cap\mathrm{I}} \mathrm{r}_\xi + \sum_{\xi\in(-\infty,\mathrm{x}]\cap\mathrm{I}} \mathrm{l}_\xi,$$

then for $\mathrm{x} \in \mathrm{I}$ *the following statements are equivalent:*

(i) j *is continuous at* x*;*

(ii) $\mathrm{r}_\mathrm{x} = \mathrm{l}_\mathrm{x} = 0$*.*

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and note that, as can be deduced from our proof of Proposition 2.4.33, there exists a finite set $A_\epsilon \subseteq I$ such that

$$\sum_{x\in I'\setminus A_\epsilon} |r_x| + \sum_{x\in I\setminus A_\epsilon} |l_x| \leq \epsilon,$$

where $I' = I \setminus \{b\}$ is I is an interval containing its right endpoint b, and $I' = I$ otherwise. Now, for $x \in I$, let $\delta \in \mathbb{R}_{>0}$ have the property that $\mathsf{B}(\delta, x) \cap A_\epsilon$ is either empty, or contains

[9] "Saltus" is a Latin word meaning "to leap." Indeed, a saltus function is also frequently referred to as a ***jump function***.

only x. For $y \in \mathsf{B}(\delta, x) \cap I$ with $y < x$ we have

$$|j(y) - j(x) - l_x| = \left| \sum_{\xi \in [y,x)} r_\xi + \sum_{\xi \in [y,x)} l_\xi \right| \leq \sum_{\xi \in I' \setminus A_\epsilon} |r_\xi| + \sum_{\xi \in I \setminus A_\epsilon} |l_\xi| < \epsilon.$$

Also, for $x < y$ we have

$$|j(y) - (j(x) + r_x)| = \left| \sum_{\xi \in (x,y)} r_\xi + \sum_{\xi \in (x,y]} l_\xi \right| \leq \sum_{\xi \in I' \setminus A_\epsilon} |r_\xi| + \sum_{\xi \in I \setminus A_\epsilon} |l_\xi| < \epsilon.$$

This gives $j(x-) = j(x) - l_x$ provided that x is not the left endpoint of I and $j(x+) = j(x) + r_x$ provided that x is not the right endpoint of I. Thus j is continuous at x if and only if $r_x = l_x = 0$. ■

3.3.19 Proposition (Saltus functions are of locally bounded variation) *If* I *is an interval and if* j: I → ℝ *is a saltus function, then* j *is a function of locally bounded variation.*

Proof We may without loss of generality suppose that $I = [a, b]$. Let us write

$$j(x) = \sum_{\xi \in (-\infty, x) \cap I} r_\xi + \sum_{\xi \in (-\infty, x] \cap I} l_\xi.$$

Let $x, y \in [a, b]$ with $x < y$. Then

$$j(y) - j(x) = r_x + l_y + \sum_{\xi \in (x,y)} (r_\xi + l_\xi).$$

Thus

$$|j(y) - j(x)| \leq \sum_{\xi \in [x,y)} |r_\xi| + \sum_{\xi \in (x,y]} |l_\xi|.$$

Now let $(x_0, x_1, \dots, x_m)$ be the endpoints of a partition of $[a, b]$. Then we compute

$$\sum_{k=1}^{m} |j(x_k) - j(x_{k-1})| \leq \sum_{k=1}^{m} \left(\sum_{\xi \in [x_{k-1}, x_k)} |u_\xi| + \sum_{\xi \in (x_{k-1}, x_k]} |l_\xi| \right) \leq \sum_{\xi \in [a,b)} |r_\xi| + \sum_{\xi \in (a,b]} |l_\xi|,$$

which gives the result. ■

Note then that we may now attribute to saltus functions all of the properties associated to functions of locally bounded variation, as presented in Theorem 3.3.3. In particular, a saltus function is differentiable almost everywhere. However, about the derivative of a saltus function, more can be said.

3.3.20 Proposition (Saltus functions have a.e. zero derivative) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{j}\colon \mathrm{I} \to \mathbb{R}$ *is a saltus function, then the set* $\{\mathrm{x} \in \mathrm{I} \mid \mathrm{j}'(\mathrm{x}) \neq 0\}$ *has measure zero.*

Proof Since j is of locally bounded variation, by Theorem 3.3.3(ii) we may write $j = j_+ - j_-$ for monotonically increasing functions j_+ and j_-. It then suffices to prove the result for the case when j is monotonically increasing, since the derivative is linear (Proposition 3.2.10) and since the union of two sets of measure zero is a set of measure zero (Exercise 2.5.11). As we saw in the proof of Proposition 3.3.18, $j(x-) = j(x) - l_x$ and $j(x+) = j(x) + r_x$. Therefore, if j is monotonically increasing, then $r_x \geq 0$ for all $x \in I'$ and $l_x \geq 0$ for all $x \in I$.

By Proposition 2.4.33 we may write

$$\{x \in I' \mid r_x \neq 0\} = \cup_{a \in A}\{\xi_a\}, \quad \{x \in I \mid l_x \neq 0\} = \cup_{b \in B}\{\eta_b\},$$

where the sets A and B are countable. For $x \in I$ define

$$A(x) = \{a \in A \mid \xi_a < x\}, \quad B(x) = \{b \in B \mid \eta_b \leq x\}.$$

Then we have

$$\sum_{\xi \in (-\infty, x) \cap I} r_\xi = \sum_{a \in A(x)} r_{\xi_a}, \quad \sum_{\xi \in (-\infty, x] \cap I} l_\xi = \sum_{b \in B(x)} r_{\eta_b}.$$

Now let us suppose that the sets A and B are well ordered and for $k \in \mathbb{Z}_{>0}$ define

$$A_k = \{a \in A \mid a \leq k\}, \quad B_k = \{b \in B \mid b \leq k\}$$

and

$$A_k(x) = \{a \in A_k \mid \xi_a < x\}, \quad B_k(x) = \{b \in B_k \mid \eta_b \leq x\}.$$

We then define $j_k\colon I \to \mathbb{R}$ by

$$j_k(x) = \sum_{a \in A_k(x)} r_{\xi_a} + \sum_{b \in B_k(x)} r_{\eta_b}.$$

Now we use some facts from Section 3.6. Note the following facts:

1. for each $k \in \mathbb{Z}_{>0}$, the functions j_k are monotonically increasing since $r_x \geq 0$ for all $x \in I'$ and $l_x \geq 0$ for each $x \in I$;
2. for each $k \in \mathbb{Z}_{>0}$, the set $\{x \in I \mid j_k'(x) \neq 0\}$ is finite;
3. $\lim_{k \to \infty} j_k(x) = j(x)$ for each $x \in I$.

Therefore, we may apply Theorem 3.6.25 below to conclude that $j'(x) = 0$ almost everywhere. ■

3.3.21 Remark (Functions with a.e. zero derivative need not be saltus functions) Note that the Cantor function of Example 3.2.27 is a function with a derivative that is zero almost everywhere. However, since this function is continuous, it is not a saltus function. More precisely, according to Proposition 3.3.18, the Cantor function is a saltus function where the two families of summable numbers used to define it are both identically zero. That is to say, it is not an interesting saltus function. This observation will be important when we discuss the Lebesgue decomposition of a function of bounded variation in Theorem III-2.9.27. •

3.3.5 The saltus function for a function of locally bounded variation

Now that we have outlined the general definition and properties of saltus functions, let us indicate how they arise from an attempt to generally characterise functions of locally bounded variation. Since functions of locally bounded variation are so tightly connected with monotonically increasing functions, we begin by constructing a saltus function associated to a monotonically increasing function.

3.3.22 Proposition (Saltus function of a monotonically increasing function) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be monotonically increasing. Define two families* $(\mathrm{r}_{\mathrm{f},\mathrm{x}})_{\mathrm{x}\in\mathrm{I}'}$ *and* $(\mathrm{l}_{\mathrm{f},\mathrm{x}})_{\mathrm{x}\in\mathrm{I}}$ *of real numbers by*

$$\begin{aligned}&\mathrm{r}_{\mathrm{f},\mathrm{x}} = \mathrm{f}(\mathrm{x}+) - \mathrm{f}(\mathrm{x}), \qquad \mathrm{x} \in [\mathrm{a}, \mathrm{b}),\\ &\mathrm{l}_{\mathrm{f},\mathrm{a}} = 0,\ \mathrm{l}_{\mathrm{f},\mathrm{x}} = \mathrm{f}(\mathrm{x}) - \mathrm{f}(\mathrm{x}-), \qquad \mathrm{x} \in (\mathrm{a}, \mathrm{b}],\end{aligned}$$

and let $\mathrm{j}_\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be defined by*

$$\mathrm{j}_\mathrm{f}(\mathrm{x}) = \sum_{\xi\in(-\infty,\mathrm{x})\cap\mathrm{I}} \mathrm{r}_{\mathrm{f},\xi} + \sum_{\xi\in(-\infty,\mathrm{x}]\cap\mathrm{I}} \mathrm{l}_{\mathrm{f},\xi}.$$

Then j_f *is a monotonically increasing saltus function, and the function* $\mathrm{f} - \mathrm{j}_\mathrm{f}$ *is a continuous monotonically increasing function.*

Proof Note that since f is monotonically increasing, $r_{f,x} \geq 0$ for all $x \in [a, b)$ and $l_{f,x} \geq 0$ for all $x \in [a, b]$. To show that j_f is a saltus function, it suffices to show that $(r_{f,x})_{x\in I'}$ and $(l_{f,x})_{x\in I}$ are summable. Let $(x_1, \dots, x_k)$ be a finite family of elements of $[a, b]$ (not necessarily the endpoints of a partition) and compute

$$\sum_{j=1}^{k} (r_{f,x_j} + l_{f,x_j}) = \sum_{j=1}^{k} (f(x_j+) - f(x_j-)) \leq f(b) - f(a).$$

Since this holds for every finite family $(x_1, \dots, x_k)$, we can assert that both families $(r_{f,x})_{x\in I'}$ and $(l_{f,x})_{x\in I}$ are summable.

Now let $x, y \in [a, b]$ with $x < y$. Take a partition of $[x, y]$ with endpoints $(x_0, x_1, \dots, x_k)$ and compute

$$\begin{aligned}f(x+) - f(x) + \sum_{j=1}^{k} (f(x_j+) - f(x_j-)) + f(y) - f(y-),\\ = f(y) - f(x) + \sum_{j=1}^{k+1} (f(x_j-) - f(x_{j-1}+)) \leq f(y) - f(x).\end{aligned}$$

Taking the supremum over all partitions of $[x, y]$ we have

$$f(x+) - f(x) + \sum_{\xi\in(x,y)}^{k} (f(x+) - f(x-)) + f(y) - f(y-) \leq f(y) - f(x),$$

from which we deduce that

$$j_f(y) - j_f(x) = f(x+) - f(x) + \sum_{\xi \in (x,y)}^{k} (f(x+) - f(x-)) + f(y) - f(y-) \leq f(y) - f(x).$$

This shows that $j_f(y) \geq j_f(x)$ and that $f(y) - j_f(y) \geq f(x) - j_f(x)$, showing that j_f and $f - j_f$ are monotonically increasing.

Now note that, as we saw in the proof of Proposition 3.3.18,

$$\begin{aligned} j_f(x+) - j_f(x) &= r_{f,x}, \qquad x \in [a,b), \\ j_f(x) - j_f(x-) &= l_{f,x}, \qquad x \in (a,b]. \end{aligned}$$

We also have $j_f(a) = 0$. Thus, for $x \in [a,b)$, we have

$$(f(x) - j_f(x)) - (f(x-) - j_f(x-)) = f(x) - f(x-) - l_{f,x} = 0$$

and, for $x \in (a,b]$, we have

$$(f(x+) - j_f(x+)) - (f(x) - j_f(x)) = f(x+) - f(x) - r_{f,x} = 0.$$

Thus $f - j_f$ is continuous, as claimed. ■

This gives the following corollary which follows more or less directly from Theorem 3.3.3(ii).

3.3.23 Corollary (Saltus function of a function of bounded variation) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be of bounded variation. Define two families* $(\mathrm{r}_{\mathrm{f},\mathrm{x}})_{\mathrm{x}\in\mathrm{I}}$, *and* $(\mathrm{l}_{\mathrm{f},\mathrm{x}})_{\mathrm{x}\in\mathrm{I}}$ *of real numbers by*

$$\begin{aligned} \mathrm{r}_{\mathrm{f},\mathrm{x}} &= \mathrm{f}(\mathrm{x}+) - \mathrm{f}(\mathrm{x}), \qquad \mathrm{x} \in [\mathrm{a}, \mathrm{b}), \\ \mathrm{l}_{\mathrm{f},\mathrm{a}} &= 0,\ \mathrm{l}_{\mathrm{f},\mathrm{x}} = \mathrm{f}(\mathrm{x}) - \mathrm{f}(\mathrm{x}-), \qquad \mathrm{x} \in (\mathrm{a}, \mathrm{b}], \end{aligned}$$

and let $\mathrm{j}_\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be defined by*

$$\mathrm{j}_\mathrm{f}(\mathrm{x}) = \sum_{\xi \in (-\infty, \mathrm{x}) \cap \mathrm{I}} \mathrm{r}_{\mathrm{f},\xi} + \sum_{\xi \in (-\infty, \mathrm{x}] \cap \mathrm{I}} \mathrm{l}_{\mathrm{f},\xi}.$$

Then j_f *is a function of bounded variation, and the function* $\mathrm{f} - \mathrm{j}_\mathrm{f}$ *is a continuous function of bounded variation.*

Of course, the preceding two results carry over, with some notational complications at endpoints, to functions of locally bounded variation defined on general intervals.

Note that Examples 3.2.27 and 3.2.28 illustrate some of the features of saltus functions and functions of locally bounded variation. Indeed, the Cantor function of Example 3.2.27 is a function of locally bounded variation for which the associated saltus function is zero, while the function of Example 3.2.28 is “all” saltus function. Perhaps it is also useful to give a more mundane example to illustrate the decomposition of a function of locally bounded variation into its saltus and continuous part.

3.3.24 Example (Saltus function of a function of locally bounded variation) Let $I = [0,1]$ and consider three functions $f_1, f_2, f_3 \colon I \to \mathbb{R}$ defined by

$$f_1(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}], \\ -1, & x \in (\frac{1}{2}, 1], \end{cases}$$

$$f_2(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}], \\ 0, & x = \frac{1}{2}, \\ -1, & x \in (\frac{1}{2}, 1], \end{cases}$$

$$f_3(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}), \\ -1, & x \in [\frac{1}{2}, 1]. \end{cases}$$

In Example 3.3.5–3 we explicitly showed that f_1 is a function of locally bounded variation, and a similar argument shows that f_2 and f_3 are also functions of locally bounded variation. A direct application of the definition of Corollary 3.3.23 gives

$$j_{f_1}(x) = \begin{cases} 0, & x \in [0, \frac{1}{2}], \\ -2, & x \in (\frac{1}{2}, 1], \end{cases}$$

$$j_{f_2}(x) = \begin{cases} 0, & x \in [0, \frac{1}{2}), \\ -1, & x = \frac{1}{2}, \\ -2, & x \in (\frac{1}{2}, 1], \end{cases}$$

$$j_{f_3}(x) = \begin{cases} 0, & x \in [0, \frac{1}{2}), \\ -2, & x \in [\frac{1}{2}, 1]. \end{cases}$$

For $k \in \{1, 2, 3\}$ we have $f_k(x) = j_{f_k}(x) = 1$, $x \in [0, 1]$. •

One might think that this is all that can be done as far as goes the decomposition of a function with locally bounded variation. However, this is not so. However, to further refine our present decomposition requires the notion of the integral as we consider it in Chapter III-2. Thus we postpone a more detailed discussion of functions of locally bounded variation until Theorem III-2.9.27.

Exercises

3.3.1 Show that if $I \subseteq \mathbb{R}$ is an interval and if $f \colon I \to \mathbb{R}$ is continuous then the following statements are equivalent:

1. f is injective;
2. f is either strictly monotonically increasing or strictly monotonically decreasing.

3.3.2 On the interval $I = [-1, 1]$ consider the function $f \colon I \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \frac{1}{2}x + x^2 \sin \frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

(a) Show that f is differentiable at $x = 0$ and has a positive derivative there.

(b) Show that for every $\epsilon \in \mathbb{R}_{>0}$ the restriction of f to $[-\epsilon, \epsilon]$ is neither monotonically decreasing (not surprisingly) nor monotonically increasing (surprisingly).

(c) Why is this not in contradiction with Proposition 3.2.23?

3.3.3 Give an example of an interval I and a function $f\colon I \to \mathbb{R}$ that is continuous, strictly monotonically increasing, but not differentiable.

3.3.4 Prove the assertions of Remark 3.3.7.

3.3.5 Let I be an interval and suppose that $I = I_1 \cup I_2$ where $I_1 \cap I_2 = \{x_0\}$ for some $x_0 \in \mathbb{R}$. If $f\colon I \to \mathbb{F}$ then

$$\mathrm{V}(f)(x) = \begin{cases} \mathrm{V}(f|I_1)(x), & x \in I_1, \\ \mathrm{V}(f|I_2)(x) + \mathrm{V}(f|I_1)(x_0), & x \in I_2 \end{cases}$$

if I_1 is finite,

$$\mathrm{V}(f)(x) = \begin{cases} \mathrm{V}(f|I_1)(x) - \mathrm{V}(f|I_2)(x_0), & x \in I_1, \\ \mathrm{V}(f|I_2)(x), & x \in I_2 \end{cases}$$

if I_1 is infinite and $x_0 < 0$, and

$$\mathrm{V}(f)(x) = \begin{cases} \mathrm{V}(f|I_1)(x), & x \in I_1, \\ \mathrm{V}(f|I_2)(x) + \mathrm{V}(f|I_1)(x_0), & x \in I_2 \end{cases}$$

if I_1 is infinite and $x_0 \geq 0$.

Section 3.4

The Riemann integral

Opposite to the derivative, in a sense made precise by Theorem 3.4.30, is the notion of integration. In this section we describe a “simple” theory of integration, called Riemann integration,[10] that typically works insofar as computations go. In Chapter III-2 we shall see that the Riemann integration suffers from a defect somewhat like the defect possessed by rational numbers. That is to say, just like there are sequences of rational numbers that seem like they should converge (i.e., are Cauchy) but do not, there are sequences of functions possessing a Riemann integral which do not converge to a function possessing a Riemann integral (see Example III-2.1.11). This has some deleterious consequences for developing a general theory based on the Riemann integral, and the most widely used fix for this is the Lebesgue integral of Chapter III-2. However, for now let us stick to the more pedestrian, and more easily understood, Riemann integral.

As we did with differentiation, we suppose that the reader has had the sort of calculus course where they learn to compute integrals of common functions. Indeed, while we do not emphasise the art of computing integrals, we do not intend this to mean that this art should be ignored. The reader should know the basic integrals and the basic tricks and techniques for computing them.

Do I need to read this section? The best way to think of this section is as a setup for the general developments of Chapter III-2. Indeed, we begin Chapter III-2 with essentially a deconstruction of what we do in this section. For this reason, this chapter should be seen as preparatory to Chapter III-2, and so can be skipped until one wants to learn Lebesgue integration in a serious way. At that time, a reader may wish to be prepared by understanding the slightly simpler Riemann integral. •

3.4.1 Step functions

Our discussion begins by our considering intervals that are compact. In Section 3.4.4 we consider the case of noncompact intervals.

In a theme that will be repeated when we consider the Lebesgue integral in Chapter III-2, we first introduce a simple class of functions whose integral is “obvious.” These functions are then used to approximate a more general class of functions which are those that are considered “integrable.” For the Riemann integral, the simple class of functions are defined as being constant on the intervals forming a partition. We recall from Definition 2.5.7 the notion of a partition and from the

[10] After Georg Friedrich Bernhard Riemann, 1826–1866. Riemann made important and long lasting contributions to real analysis, geometry, complex function theory, and number theory, to name a few areas. The presently unsolved Riemann Hypothesis is one of the outstanding problems in modern mathematics.

discussion surrounding the definition the notion of the endpoints associated with a partition.

3.4.1 Definition (Step function) Let $I = [a, b]$ be a compact interval. A function $f\colon I \to \mathbb{R}$ is a ***step function*** if there exists a partition $P = (I_1, \dots, I_k)$ of I such that

(i) $f|\operatorname{int}(I_j)$ is a constant function for each $j \in \{1, \dots, k\}$,

(ii) $f(a+) = f(a)$ and $f(b-) = f(b)$, and

(iii) for each $x \in \mathrm{EP}(P) \setminus \{a, b\}$, either $f(x-) = f(x)$ or $f(x+) = f(x)$. •

In Figure 3.10 we depict a typical step function. Note that at discontinuities we allow the function to be continuous from either the right or the left. In the development we undertake, it does not really matter which it is.

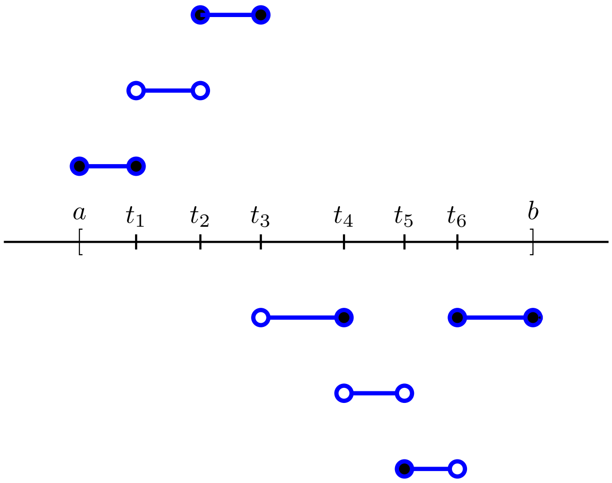


Figure 3.10 A step function

The idea of the integral of a function is that it measures the "area" below the graph of a function. If the value of the function is negative, then the area is taken to be negative. For step functions, this idea of the area under the graph is clear, so we simply define this to be the integral of the function.

3.4.2 Definition (Riemann integral of a step function) Let $I = [a, b]$ and let $f\colon I \to \mathbb{R}$ be a step function defined using the partition $P = (I_1, \dots, I_k)$ with endpoints $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$. Suppose that the value of f on $\operatorname{int}(I_j)$ is c_j for $j \in \{1, \dots, k\}$. The ***Riemann integral*** of f is

$$A(f) = \sum_{j=1}^{k} c_j(x_j - x_{j-1}). \qquad \bullet$$

The notation $A(f)$ is intended to suggest "area."

3.4.2 The Riemann integral on compact intervals

Next we define the Riemann integral of a function that is not necessarily a step function. We do this by approximating a function by step functions.

3.4.3 Definition (Lower and upper step functions) Let $I = [a, b]$ be a compact interval, let $f \colon I \to \mathbb{R}$ be a bounded function, and let $P = (I_1, \dots, I_k)$ be a partition of I.

(i) The ***lower step function*** associated to f and P is the function $s_-(f, P) \colon I \to \mathbb{R}$ defined according to the following:

(a) if $x \in I$ lies in the interior of an interval I_j, $j \in \{1, \dots, k\}$, then $s_-(f, P)(x) = \inf\{f(x) \mid x \in \mathrm{cl}(I_j)\}$;

(b) $s_-(f, P)(a) = s_-(f, P)(a+)$ and $s_-(f, P)(b) = s_-(f, P)(b-)$;

(c) for $x \in \mathrm{EP}(P) \setminus \{a, b\}$, $s_-(f, P)(x) = s_-(f, P)(x+)$.

(ii) The ***upper step function*** associated to f and P is the function $s_+(f, P) \colon I \to \mathbb{R}$ defined according to the following:

(a) if $x \in I$ lies in the interior of an interval I_j, $j \in \{1, \dots, k\}$, then $s_+(f, P)(x) = \sup\{f(x) \mid x \in \mathrm{cl}(I_j)\}$;

(b) $s_+(f, P)(a) = s_+(f, P)(a+)$ and $s_+(f, P)(b) = s_+(f, P)(b-)$;

(c) for $x \in \mathrm{EP}(P) \setminus \{a, b\}$, $s_+(f, P)(x) = s_+(f, P)(x+)$. •

Note that both the lower and upper step functions are well-defined since f is bounded. Note also that at the middle endpoints for the partition, we ask that the lower and upper step functions be continuous from the right. This is an arbitrary choice. Finally, note that for each $x \in [a, b]$ we have

$$s_-(f, P)(x) \le f(x) \le s_+(f, P)(x).$$

That is to say, for any bounded function f, we have defined two step functions, one bounding f from below and one bounding f from above.

Next we associate to the lower and upper step functions their integrals, which we hope to use to define the integral of the function f.

3.4.4 Definition (Lower and upper Riemann sums) Let $I = [a, b]$ be a compact interval, let $f \colon I \to \mathbb{R}$ be a bounded function, and let $P = (I_1, \dots, I_k)$ be a partition of I.

(i) The ***lower Riemann sum*** associated to f and P is $A_-(f, P) = A(s_-(f, P))$.

(ii) The ***upper Riemann sum*** associated to f and P is $A_+(f, P) = A(s_+(f, P))$. •

Now we define the best approximations of the integral of f using the lower and upper Riemann sums.

3.4.5 Definition (Lower and upper Riemann integral) Let $I = [a, b]$ be a compact interval and let $f\colon I \to \mathbb{R}$ be a bounded function.

(i) The ***lower Riemann integral*** of f is

$$I_-(f) = \sup\{A_-(f, P) \mid P \in \mathrm{Part}(I)\}.$$

(ii) The ***upper Riemann integral*** of f is

$$I_+(f) = \inf\{A_+(f, P) \mid P \in \mathrm{Part}(I)\}. \qquad \bullet$$

Note that since f is bounded, it follows that the sets

$$\{A_-(f, P) \mid P \in \mathrm{Part}(I)\}, \quad \{A_+(f, P) \mid P \in \mathrm{Part}(I)\}$$

are bounded (why?). Therefore, the lower and upper Riemann integral always exist. So far, then, we have made a some constructions that apply to *any* bounded function. That is to say, for any bounded function, it is possible to define the lower and upper Riemann integral. What is not clear is that these two things should be equal. In fact, they are *not* generally equal, which leads to the following definition.

3.4.6 Definition (Riemann integrable function on a compact interval) A bounded function $f\colon [a, b] \to \mathbb{R}$ on a compact interval is ***Riemann integrable*** if $I_-(f) = I_+(f)$. We denote

$$\int_a^b f(x)\,\mathrm{d}x = I_-(f) = I_+(f),$$

which is the ***Riemann integral*** of f. The function f is called the ***integrand***. •

3.4.7 Notation (Swapping limits of integration) In the expression $\int_a^b f(x)\,\mathrm{d}x$, "$a$" is the ***lower limit of integration*** and "b" is the ***upper limit of integration***. We have tacitly assumed that $a < b$ in our constructions to this point. However, we can consider the case where $b < a$ by adopting the convention that

$$\int_b^a f(x)\,\mathrm{d}x = -\int_a^b f(x)\,\mathrm{d}x. \qquad \bullet$$

Let us provide an example which illustrates that, in principle, it is possible to use the definition of the Riemann integral to perform computations, even though this is normally tedious. A more common method for computing integrals is to use the Fundamental Theorem of Calculus to "reverse engineer" the process.

3.4.8 Example (Computing a Riemann integral) Let $I = [0,1]$ and define $f\colon I \to \mathbb{R}$ by $f(x) = x$. Let $P = (I_1,\dots,I_k)$ be a partition with $s_-(f,P)$ and $s_+(f,P)$ the associated lower and upper step functions, respectively. Let $\mathrm{EP}(P) = (x_0,x_1,\dots,x_k)$ be the endpoints of the intervals of the partition. One can then see that, for $j \in \{1,\dots,k\}$, $s_-(f,P)|\operatorname{int}(I_j) = x_{j-1}$ and $s_+(f,P)|\operatorname{int}(I_j) = x_j$. Therefore,

$$A_-(f,P) = \sum_{j=1}^{k} x_{j-1}(x_j - x_{j-1}), \quad A_+(f,P) = \sum_{j=1}^{k} x_j(x_j - x_{j-1}).$$

We claim that $I_-(f) \geq \frac{1}{2}$ and that $I_+(f) \leq \frac{1}{2}$, and note that, once we prove this, it follows that f is Riemann integrable and that $I_-(f) = I_+(f) = \frac{1}{2}$ (why?).

For $k \in \mathbb{Z}_{>0}$ consider the partition P_k with endpoints $\mathrm{EP}(P_k) = \{\frac{j}{k} \mid j \in \{0,1,\dots,k\}\}$. Then, using the formula $\sum_{j=1}^{l} j = \frac{1}{2}l(l+1)$, we compute

$$A_-(f,P_k) = \sum_{j=1}^{k} \frac{j-1}{k^2} = \frac{k(k-1)}{2k^2}, \qquad A_+(f,P_k) = \sum_{j=1}^{k} \frac{j}{k^2} = \frac{k(k+1)}{2k^2}.$$

Therefore,

$$\lim_{k\to\infty} A_-(f,P_k) = \tfrac{1}{2}, \qquad \lim_{k\to\infty} A_+(f,P_k) = \tfrac{1}{2}.$$

This shows that $I_-(f) \geq \frac{1}{2}$ and that $I_+(f) \leq \frac{1}{2}$, as desired. •

3.4.3 Characterisations of Riemann integrable functions on compact intervals

In this section we provide some insightful characterisations of the notion of Riemann integrability. First we provide four equivalent characterisations of the Riemann integral. Each of these captures, in a slightly different manner, the notion of the Riemann integral as a limit. It will be convenient to introduce the language that a ***selection*** from a partition $P = (I_1,\dots,I_k)$ is a family $\boldsymbol{\xi} = (\xi_1,\dots,\xi_k)$ of points such that $\xi_j \in \mathrm{cl}(I_j)$, $j \in \{1,\dots,k\}$.

3.4.9 Theorem (Riemann, Darboux,[11] and Cauchy characterisations of Riemann integrable functions) *For a compact interval* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$ *and a bounded function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, the following statements are equivalent:*

(i) f *is Riemann integrable;*

(ii) for every $\epsilon \in \mathbb{R}_{>0}$*, there exists a partition* P *such that* $\mathrm{A}_+(\mathrm{f},\mathrm{P}) - \mathrm{A}_-(\mathrm{f},\mathrm{P}) < \epsilon$ *(**Riemann's condition**);*

[11] Jean Gaston Darboux (1842–1917) was a French mathematician. His made important contributions to analysis and differential geometry.

(iii) there exists $\mathrm{I(f)} \in \mathbb{R}$ *such that, for every* $\epsilon \in \mathbb{R}_{>0}$ *there exists* $\delta \in \mathbb{R}_{>0}$ *such that, if* $\mathrm{P} = (\mathrm{I}_1, \dots, \mathrm{I}_k)$ *is a partition for which* $|\mathrm{P}| < \delta$ *and if* $(\xi_1, \dots, \xi_k)$ *is a selection from* P*, then*

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \epsilon,$$

where $\mathrm{EP(P)} = (x_0, x_1, \dots, x_k)$ *(**Darboux' condition**);*

(iv) for each $\epsilon \in \mathbb{R}_{>0}$ *there exists* $\delta \in \mathbb{R}_{>0}$ *such that, for any partitions* $\mathrm{P} = (\mathrm{I}_1, \dots, \mathrm{I}_k)$ *and* $\mathrm{P}' = (\mathrm{I}'_1, \dots, \mathrm{I}'_{k'})$ *with* $|\mathrm{P}|, |\mathrm{P}'| < \delta$ *and for any selections* $(\xi_1, \dots, \xi_k)$ *and* $(\xi'_1, \dots, \xi'_{k'})$ *from* P *and* P′*, respectively, we have*

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - \sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1})\right| < \epsilon,$$

where $\mathrm{EP(P)} = (x_0, x_1, \dots, x_k)$ *and* $\mathrm{EP(P')} = (x'_0, x'_1, \dots, x'_{k'})$ *(**Cauchy's condition**).*

Proof First let us prove a simple lemma about lower and upper Riemann sums and refinements of partitions.

1 Lemma *Let* $\mathrm{I} = [a, b]$*, let* $f\colon \mathrm{I} \to \mathbb{R}$ *be bounded, and let* P_1 *and* P_2 *be partitions of* I *with* P_2 *a refinement of* P_1*. Then*

$$A_-(f, P_2) \geq A_-(f, P_1), \quad A_+(f, P_2) \leq A_+(f, P_1).$$

Proof Let $x_1, x_2 \in \mathrm{EP}(P_1)$ and denote by $y_1, \dots, y_l$ the elements of $\mathrm{EP}(P_2)$ that satisfy

$$x_1 \leq y_1 < \dots < y_l \leq x_2.$$

Then

$$\begin{aligned}\sum_{j=1}^{l}(y_j - y_{j-1}) \inf\{f(y) \mid y \in [y_j, y_{j-1}]\} &\geq \sum_{j=1}^{l}(y_j - y_{j-1}) \inf\{f(x) \mid x \in [x_1, x_2]\} \\ &= (x_2 - x_1) \inf\{f(x) \mid x \in [x_1, x_2]\}.\end{aligned}$$

Now summing over all consecutive pairs of endpoints for P_1 gives $A_-(f, P_2) \geq A_-(f, P_1)$. A similar argument gives $A_+(f, P_2) \leq A_+(f, P_1)$. ▼

The following trivial lemma will also be useful.

2 Lemma $\mathrm{I}_-(f) \leq \mathrm{I}_+(f)$.

Proof Since, for any two partitions P_1 and P_2, we have

$$s_-(f, P_1) \leq f(x) \leq s_+(f, P_2),$$

it follows that

$$\sup\{A_-(f, P) \mid P \in \mathrm{Part}(I)\} \leq \inf\{A_+(f, P) \mid P \in \mathrm{Part}(I)\},$$

which is the result. ▼

(i) $\implies$ (ii) Suppose that f is Riemann integrable and let $\epsilon \in \mathbb{R}_{>0}$. Then there exists partitions P_- and P_+ such that

$$A_-(f, P_-) > I_-(f) - \tfrac{\epsilon}{2}, \quad A_+(f, P_+) < I_+(f) + \tfrac{\epsilon}{2}.$$

Now let P be a partition that is a refinement of both P_1 and P_2 (obtained, for example, by asking that $\mathrm{EP}(P) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$). By Lemma 1 it follows that

$$A_+(f, P) - A_-(f, P) \le A_+(f, P_+) - A_-(f, P_-) < I_+(f) + \tfrac{\epsilon}{2} - I_-(f) + \tfrac{\epsilon}{2} = \epsilon.$$

(ii) $\implies$ (i) Now suppose that $\epsilon \in \mathbb{R}_{>0}$ and let P be a partition such that $A_+(f, P) - A_-(f, P) < \epsilon$. Since we additionally have $I_-(f) \le I_+(f)$ by Lemma 2, it follows that

$$A_-(f, P) \le I_-(f) \le I_+(f) \le A_+(f, P),$$

from which we deduce that

$$0 \le I_+(f) - I_-(f) < \epsilon.$$

Since ϵ is arbitrary, we conclude that $I_-(f) = I_+(f)$, as desired.

(i) $\implies$ (iii) We first prove a lemma about partitions of compact intervals.

3 Lemma *If* $\mathrm{P} = (\mathrm{I}_1, \dots, \mathrm{I}_k)$ *is a partition of* $[\mathrm{a}, \mathrm{b}]$ *and if* $\epsilon \in \mathbb{R}_{>0}$*, then there exists* $\delta \in \mathbb{R}_{>0}$ *such that, if* $\mathrm{P}' = (\mathrm{I}'_1, \dots, \mathrm{I}'_{k'})$ *is a partition with* $|\mathrm{P}'| < \delta$ *and if*

$$\{\mathrm{j}'_1, \dots, \mathrm{j}'_r\} = \{\mathrm{j}' \in \{1, \dots, \mathrm{k}'\} \mid \mathrm{cl}(\mathrm{I}'_{\mathrm{j}'}) \not\subset \mathrm{cl}(\mathrm{I}_\mathrm{j}) \textit{ for any } \mathrm{j} \in \{1, \dots, \mathrm{k}\}\},$$

then

$$\sum_{l=1}^{r} |\mathrm{x}_{\mathrm{j}'_l} - \mathrm{x}_{\mathrm{j}'_l - 1}| < \epsilon,$$

where $\mathrm{EP}(\mathrm{P}') = (\mathrm{x}_0, \mathrm{x}_1, \dots, \mathrm{x}_{\mathrm{k}'})$.

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and take $\delta = \frac{\epsilon}{k+1}$. Let $P' = (I'_1, \dots, I'_{k'})$ be a partition with endpoints $(x_0, x_1, \dots, x_{k'})$ and satisfying $|P'| < \delta$. Define

$$K_1 = \{j' \in \{1, \dots, k'\} \mid \mathrm{cl}(I'_{j'}) \not\subset \mathrm{cl}(I_j) \text{ for any } j \in \{1, \dots, k\}\}.$$

If $j' \in K_1$ then $I'_{j'}$ is not contained in any interval of P and so $I'_{j'}$ must contain at least one endpoint from P. Since P has $k+1$ endpoints we obtain $\mathrm{card}(K_1) \le k+1$. Since the intervals $I'_{j'}$, $j' \in K_1$, have length at most δ we have

$$\sum_{j' \in K_1} (x_{j'} - x_{j'-1}) \le (k+1)\delta \le \epsilon,$$

as desired. ▼

Now let $\epsilon \in \mathbb{R}_{>0}$ and define $M = \sup\{|f(x)| \mid x \in I\}$. Denote by $I(f)$ the Riemann integral of f. Choose partitions P_- and P_+ such that

$$I(f) - A_-(f, P_-) < \tfrac{\epsilon}{2}, \quad A_+(f, P_+) - I(f) < \tfrac{\epsilon}{2}.$$

If $P = (I_1, \dots, I_k)$ is chosen such that $\mathrm{EP}(P) = \mathrm{EP}(P_-) \cup \mathrm{EP}(P_+)$, then

$$I(f) - A_-(f, P) < \tfrac{\epsilon}{2}, \quad A_+(f, P) - I(f) < \tfrac{\epsilon}{2}.$$

By Lemma 3 choose $\delta \in \mathbb{R}_{>0}$ such that if P' is any partition for which $|P'| < \delta$ then the sum of the lengths of the intervals of P' not contained in some interval of P does not exceed $\frac{\epsilon}{2M}$. Let $P' = (I'_1, \dots, I'_{k'})$ be a partition with endpoints $(x_0, x_1, \dots, x_{k'})$ and satisfying $|P'| < \delta$. Denote

$$K_1 = \{j' \in \{1, \dots, k'\} \mid I'_{j'} \not\subset I_j \text{ for some } j \in \{1, \dots, k\}\}$$

and $K_2 = \{1, \dots, k'\} \setminus K_1$. Let $(\xi_1, \dots, \xi_{k'})$ be a selection of P'. Then we compute

$$\begin{aligned}\sum_{j=1}^{k'} f(\xi_j)(x_j - x_{j-1}) &= \sum_{j \in K_1} f(\xi_j)(x_j - x_{j-1}) + \sum_{j \in K_2} f(\xi_j)(x_j - x_{j-1}) \\ &\leq A_+(f, P) + M\frac{\epsilon}{2M} < I(f) + \epsilon.\end{aligned}$$

In like manner we show that

$$\sum_{j=1}^{k'} f(\xi_j)(x_j - x_{j-1}) > I(f) - \epsilon.$$

This gives

$$\left|\sum_{j=1}^{k'} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \epsilon,$$

as desired.

(iii) $\Longrightarrow$ (ii) Let $\epsilon \in \mathbb{R}_{>0}$ and let $P = (I_1, \dots, I_k)$ be a partition for which

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \frac{\epsilon}{4}$$

for every selection $(\xi_1, \dots, \xi_k)$ from P. Now particularly choose a selection such that

$$|f(\xi_j) - \sup\{f(x) \mid x \in \mathrm{cl}(I_j)\}| < \frac{\epsilon}{4k(x_j - x_{j-1})}.$$

Then

$$\begin{aligned}|A_+(f, P) - I(f)| &\leq \left|A_+(f, P) - \sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1})\right| + \left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| \\ &< \sum_{j=1}^{k} \frac{\epsilon}{4k(x_j - x_{j-1})}(x_j - x_{j-1}) + \frac{\epsilon}{4} < \frac{\epsilon}{2}.\end{aligned}$$

In like manner one shows that $|A_-(f,P) - I(f)| < \frac{\epsilon}{2}$. Therefore,

$$|A_+(f,P) - A_-(f,P)| \leq |A_+(f,P) - I(f)| + |I(f) - A_-(f,P)| < \epsilon,$$

as desired.

(iii) $\Longrightarrow$ (iv) Let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that, whenever $P = (I_1, \dots{}_k)$ is a partition satisfying $|P| < \delta$ and $(\xi_1, \dots, \xi_k)$ is a selection from P, it holds that

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| < \frac{\epsilon}{2}.$$

Now let $P = (I_1, \dots, I_k)$ and $P' = (I'_1, \dots, I'_{k'})$ be two partitions with $|P|, |P'| < \delta$, and let $(\xi_1, \dots, \xi_k)$ and $(\xi'_1, \dots, \xi'_{k'})$ selections from P and P', respectively. Then we have

$$\begin{aligned}\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - \sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1})\right| \\ \leq \left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - I(f)\right| + \left|\sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1}) - I(f)\right| < \epsilon,\end{aligned}$$

which gives this part of the result.

(iv) $\Longrightarrow$ (iii) Let $(P_j = (I_{j,1}, \dots, I_{j,k_j}))_{j \in \mathbb{Z}_{>0}}$ be a sequence of partitions for which $\lim_{j\to\infty}|P_j| = 0$. Then, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that

$$\left|\sum_{j=1}^{k_l} f(\xi_{l,j})(x_{l,j} - x_{l,j-1}) - \sum_{j=1}^{k_m} f(\xi_{m,j})(x_{m,j} - x_{m,j-1})\right| < \epsilon,$$

for $l, m \geq N$, where $\boldsymbol{\xi}_j = (\xi_{j,1}, \dots, \xi_{j,k_j})$, is a selection from P_j, $j \in \mathbb{Z}_{>0}$, and where $\mathrm{EP}(P_j) = (x_{j,0}, x_{j,1}, \dots, x_{j,k_j})$, $j \in \mathbb{Z}_{>0}$. If we define

$$A(f, P_j, \boldsymbol{\xi}_j) = \sum_{r=1}^{k_j} f(\xi_r)(x_{j,r} - x_{j,r-1}),$$

then the sequence $(A(f, P_j, \boldsymbol{\xi}_j))_{j \in \mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ for any choices of points $\boldsymbol{\xi}_j$, $j \in \mathbb{Z}_{>0}$. Denote the resulting limit of this sequence by $I(f)$. We claim that $I(f)$ is the Riemann integral of f. To see this, let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ be such that

$$\left|\sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}) - \sum_{j=1}^{k'} f(\xi'_j)(x'_j - x'_{j-1})\right| < \frac{\epsilon}{2}$$

for any two partitions P and P' satisfying $|P|, |P'| < \delta$ and for any selections $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ from P and P', respectively. Now let $N \in \mathbb{Z}_{>0}$ satisfy $|P_j| < \delta$ for every $j \geq N$. Then, if P is any partition with $|P| < \delta$ and if $\boldsymbol{\xi}$ is any selection from P, we have

$$|A(f, P, \boldsymbol{\xi}) - I(f)| \leq |A(f, P, \boldsymbol{\xi}) - A(f, P_N, \boldsymbol{\xi}_N)| + |A(f, P_N, \boldsymbol{\xi}_N) - I(f)| < \epsilon,$$

for any selection $\boldsymbol{\xi}_N$ of P_N. This shows that $I(f)$ is indeed the Riemann integral of f, and so gives this part of the theorem. ■

A consequence of the proof is that, of course, the quantity $I(f)$ in part (iii) of the theorem is nothing other than the Riemann integral of f.

Many of the functions one encounters in practice are, in fact, Riemann integrable. However, not all functions are Riemann integrable, as the following simple examples shows.

3.4.10 Example (A function that is not Riemann integrable) Let $I = [0,1]$ and let $f\colon I \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} 1, & x \in \mathbb{Q} \cap I \\ 0, & x \notin \mathbb{Q} \cap I. \end{cases}$$

Thus f takes the value 1 at all rational points, and is zero elsewhere. Now let $s_+, s_-\colon I \to \mathbb{R}$ be any step functions satisfying $s_-(x) \leq f(x) \leq s_+(x)$ for all $x \in I$. Since any nonempty subinterval of I contains infinitely many irrational numbers, it follows that $s_-(x) \leq 0$ for every $x \in I$. Since every nonempty subinterval of I contains infinitely many rational numbers, it follows that $s_+(x) \geq 1$ for every $x \in I$. Therefore, $A(s_+) - A(s_-) \geq 1$. It follows from Theorem 3.4.9 that f is not Riemann integrable. While this example may seem pointless and contrived, it will be used in Examples II-1.7.7(1) and Example III-2.1.11 to exhibit undesirable features of the Riemann integral. •

The following result provides an interesting characterisation of Riemann integrable functions, illustrating precisely the sorts of functions whose Riemann integrals may be computed.

3.4.11 Theorem (Riemann integrable functions are continuous almost everywhere, and vice versa) *For a compact interval* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$, *a bounded function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is Riemann integrable if and only if the set*

$$\mathrm{D_f} = \{\mathrm{x} \in \mathrm{I} \mid \mathrm{f} \textit{ is discontinuous at } \mathrm{x}\}$$

has measure zero.

Proof Recall from Definition 3.1.10 the notion of the oscillation ω_f for a function f, and that $\omega_f(x) = 0$ if and only if f is continuous at x. For $k \in \mathbb{Z}_{>0}$ define

$$D_{f,k} = \left\{x \in I \;\middle|\; \omega_f(x) \geq \tfrac{1}{k}\right\}.$$

Then Proposition 3.1.11 implies that $D_f = \cup_{k \in \mathbb{Z}_{>0}} D_{f,k}$. By Exercise 2.5.11 we can assert that D_f has measure zero if and only if each of the sets $D_{f,k}$ has measure zero, $k \in \mathbb{Z}_{>0}$.

Now suppose that $D_{f,k}$ does not have measure zero for some $k \in \mathbb{Z}_{>0}$. Then there exists $\epsilon \in \mathbb{R}_{>0}$ such that, if a family $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$ of open intervals has the property that

$$D_{f,k} \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j),$$

then

$$\sum_{j=1}^{\infty} |b_j - a_j| \geq \epsilon.$$

Now let P be a partition of I and denote $\mathrm{EP}(P) = (x_0, x_1, \dots, x_m)$. Now let $\{j_1, \dots, j_l\} \subseteq \{1, \dots, m\}$ be those indices for which $j_r \in \{j_1, \dots, j_l\}$ implies that $D_{f,k} \cap (x_{j_r-1}, x_{j_r}) \neq \emptyset$. Note that it follows that the set $\bigcup_{r=1}^{l}(x_{j_r-1}, x_{j_r})$ covers $D_{f,k}$ with the possible exception of a finite number of points. It then follows that one can enlarge the length of each of the intervals (x_{j_r-1}, x_{j_r}), $r \in \{1, \dots, l\}$, by $\frac{\epsilon}{2l}$, and the resulting intervals will cover $D_{f,k}$. The enlarged intervals will have total length at least ϵ, which means that

$$\sum_{r=1}^{l} |x_{j_r} - x_{j_r-1}| \geq \frac{\epsilon}{2}.$$

Moreover, for each $r \in \{1, \dots, l\}$,

$$\sup\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\} - \inf\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\} \geq \tfrac{1}{k}$$

since $D_{f,k} \cap (x_{j_r-1}, x_{j_r}) \neq \emptyset$ and by definition of $D_{f,k}$ and ω_f. It now follows that

$$\begin{aligned} A_+(f,P) - A_-(f,P) &= \sum_{j=1}^{m} (x_j - x_{j-1}) \Big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\ &\qquad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big) \\ &\geq \sum_{r=1}^{l} (x_{j_r} - x_{j_r-1}) \Big(\sup\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\} \\ &\qquad - \inf\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\}\Big) \\ &\geq \tfrac{\epsilon}{2k}. \end{aligned}$$

Since this must hold for every partition, it follows that f is not Riemann integrable.

Now suppose that D_f has measure zero. Since f is bounded, let $M = \sup\{|f(x)| \mid x \in I\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and for brevity define $\epsilon' = \frac{\epsilon}{b-a+2}$. Choose a sequence $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$ of open intervals such that

$$D_f \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} I_j, \quad \sum_{j=1}^{\infty} |b_j - a_j| < \tfrac{\epsilon'}{M}.$$

Define $\delta \colon I \to \mathbb{R}_{>0}$ such that the following properties hold:

1. if $x \notin D_f$ then $\delta(x)$ is taken such that, if $y \in I \cap \mathsf{B}(\delta(x), x)$, then $|f(y) - f(x)| < \frac{\epsilon'}{2}$;
2. if $x \in D_f$ then $\delta(x)$ is taken such that $\mathsf{B}(\delta(x), x) \subseteq I_j$ for some $j \in \mathbb{Z}_{>0}$.

Now, by Proposition 2.5.10, let $((c_1, I_1), \dots, (c_k, I_k))$ be a δ-fine tagged partition with $P = (I_1, \dots, I_k)$ the associated partition. Now partition the set $\{1, \dots, k\}$ into two sets K_1

and K_2 such that $j \in K_1$ if and only if $c_j \notin D_f$. Then we compute

$$\begin{aligned}
A_+(f,P) - A_-(f,P) &= \sum_{j=1}^{k}(x_j - x_{j-1})\big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
&\quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\big) \\
&= \sum_{j\in K_1}(x_j - x_{j-1})\big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
&\quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\big) \\
&\quad + \sum_{j\in K_2}(x_j - x_{j-1})\big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\} \\
&\quad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\big) \\
&\le \sum_{j\in K_1}\epsilon'(x_j - x_{j-1}) + \sum_{j\in K_2} 2M(x_j - x_{j-1}) \\
&\le \epsilon'(b-a) + 2M\sum_{j=1}^{\infty}|b_j - a_j| \\
&< \epsilon'(b-a+2) = \epsilon.
\end{aligned}$$

This part of the result now follows by Theorem 3.4.9. ■

The theorem indicates why the function of Example 3.4.10 is not Riemann integrable. Indeed, the function in that example is discontinuous at *all* points in $[0,1]$ (why?). The theorem also has the following obvious corollary which illustrates why so many functions in practice are Riemann integrable.

3.4.12 Corollary (Continuous functions are Riemann integrable) *If* $f\colon [a,b] \to \mathbb{R}$ *is continuous, then it is Riemann integrable.*

By virtue of Theorem 3.3.3, we also have the following result, giving another large class of Riemann integrable functions, distinct from those that are continuous.

3.4.13 Corollary (Functions of bounded variation are Riemann integrable) *If* $f\colon [a,b] \to \mathbb{R}$ *has bounded variation, then* f *is Riemann integrable.*

3.4.4 The Riemann integral on noncompact intervals

Up to this point in this section we have only considered the Riemann integral for bounded functions defined on compact intervals. In this section we extend the notion of the Riemann integral to allow its definition for unbounded functions and for general intervals. There are complications that arise in this situation that do not arise in the case of a compact interval in that one has two possible notions of what one might call a Riemann integrable function. In all cases, we use the existing definition of the Riemann integral for compact intervals as our basis, and allow the other cases as limits.

3.4.14 Definition (Positive Riemann integrable function on a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}_{\geq 0}$ be a function whose restriction to every compact subinterval of I is Riemann integrable.

(i) If $I = [a, b]$ then the Riemann integral of f is as defined in the preceding section.

(ii) If $I = (a, b]$ then define

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{r_a \downarrow a} \int_{r_a}^b f(x)\,\mathrm{d}x.$$

(iii) If $I = [a, b)$ then define

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{r_b \uparrow b} \int_a^{r_b} f(x)\,\mathrm{d}x.$$

(iv) If $I = (a, b)$ then define

$$\int_a^b f(x)\,\mathrm{d}x = \lim_{r_a \downarrow a} \int_{r_a}^c f(x)\,\mathrm{d}x + \lim_{r_b \uparrow b} \int_c^{r_b} f(x)\,\mathrm{d}x$$

for some $c \in (a, b)$.

(v) If $I = (-\infty, b]$ then define

$$\int_{-\infty}^b f(x)\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^b f(x)\,\mathrm{d}x.$$

(vi) If $I = (-\infty, b)$ then define

$$\int_{-\infty}^b f(x)\,\mathrm{d}x = \lim_{R \to \infty} \int_{-R}^c f(x)\,\mathrm{d}x + \lim_{r_b \uparrow b} \int_c^{r_b} f(x)\,\mathrm{d}x$$

for some $c \in (-\infty, b)$.

(vii) If $I = [a, \infty)$ then define

$$\int_a^\infty f(x)\,\mathrm{d}x = \lim_{R \to \infty} \int_a^R f(x)\,\mathrm{d}x.$$

(viii) If $I = (a, \infty)$ then define

$$\int_a^\infty f(x)\,\mathrm{d}x = \lim_{r_a \downarrow a} \int_{r_a}^c f(x)\,\mathrm{d}x + \lim_{R \to \infty} \int_c^R f(x)\,\mathrm{d}x$$

for some $c \in (a, \infty)$.

(ix) If $I = \mathbb{R}$ then define

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \lim_{R\to\infty} \int_{-R}^{c} f(x)\,\mathrm{d}x + \lim_{R\to\infty} \int_{c}^{R} f(x)\,\mathrm{d}x$$

for some $c \in \mathbb{R}$.

If, for a given I and f, the appropriate of the above limits exists, then f is ***Riemann integrable*** on I, and the ***Riemann integral*** is the value of the limit. Let us denote by

$$\int_I f(x)\,\mathrm{d}x$$

the Riemann integral. •

One can easily show that where, in the above definitions, one must make a choice of c, the definition is independent of this choice (cf. Proposition 3.4.26).

The above definition is intended for functions taking nonnegative values. For more general functions we have the following definition.

3.4.15 Definition (Riemann integrable function on a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function whose restriction to any compact subinterval of I is Riemann integrable. Define $f_+, f_-\colon I \to \mathbb{R}_{\geq 0}$ by

$$f_+(x) = \max\{0, f(x)\}, \quad f_-(x) = -\min\{0, f(x)\}$$

so that $f = f_+ - f_-$. The function f is ***Riemann integrable*** if both f_+ and f_- are Riemann integrable, and the ***Riemann integral*** of f is

$$\int_I f(x)\,\mathrm{d}x = \int_I f_+(x)\,\mathrm{d}x - \int_I f_-(x)\,\mathrm{d}x. \qquad \bullet$$

At this point, if I is compact, we have potentially competing definitions for the Riemann integral of a bounded function $I\colon f \to \mathbb{R}$. One definition is the direct one of Definition 3.4.6. The other definition involves computing the Riemann integral, as per Definition 3.4.6, of the positive and negative parts of f, and then take the difference of these. Let us resolve the equivalence of these two notions.

3.4.16 Proposition (Consistency of definition of Riemann integral on compact intervals) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$, *let* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$, *and let* $\mathrm{f}_+, \mathrm{f}_-\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}_{\geq 0}$ *be the positive and negative parts of* f. *Then the following two statements are equivalent:*

(i) f *is integrable as per Definition 3.4.6 with Riemann integral* $\mathrm{I}(\mathrm{f})$;

(ii) f_+ *and* f_- *are Riemann integrable as per Definition 3.4.6 with Riemann integrals* $\mathrm{I}(\mathrm{f}_+)$ *and* $\mathrm{I}(\mathrm{f}_-)$.

Moreover, if one, and therefore both, of parts (i) and (ii) hold, then $\mathrm{I}(\mathrm{f}) = \mathrm{I}(\mathrm{f}_+) - \mathrm{I}(\mathrm{f}_-)$.

Proof We shall refer ahead to the results of Section 3.4.5.

(i) $\implies$ (ii) Define continuous functions $g_+, g_- \colon \mathbb{R} \to \mathbb{R}$ by

$$g_+(x) = \max\{0, x\}, \quad g_-(x) = -\min\{0, x\}$$

so that $f_+ = g_+ \circ f$ and $f_- = g_- \circ f$. By Proposition 3.4.23 (noting that the proof of that result is valid for the Riemann integral as per Definition 3.4.6) it follows that f_+ and f_- are Riemann integrable as per Definition 3.4.6.

(ii) $\implies$ (i) Note that $f = f_+ - f_-$. Also note that the proof of Proposition 3.4.22 is valid for the Riemann integral as per Definition 3.4.6. Therefore, f is Riemann integrable as per Definition 3.4.6.

Now we show that $I(f) = I(f_+) - I(f_-)$. This, however, follows immediately from Proposition 3.4.22. ■

It is not uncommon to see the general integral as we have defined it called the ***improper Riemann integral***.

The preceding definitions may appear at first to be excessively complicated. The following examples illustrate the rationale behind the care taken in the definitions.

3.4.17 Examples (Riemann integral on a general interval)

1. Let $I = (0, 1]$ and let $f(x) = x^{-1}$. Then, if $r_a \in (0, 1)$, we compute the proper Riemann integral
$$\int_{r_a}^{1} f(x)\,\mathrm{d}x = -\log r_a,$$
where log is the natural logarithm. Since $\lim_{r_a \downarrow} \log r_a = -\infty$ this function is not Riemann integrable on $(0, 1]$.
2. Let $I = (0, 1]$ and let $f(x) = x^{-1/2}$. Then, if $r_a \in (0, 1)$, we compute the proper Riemann integral
$$\int_{r_a}^{1} f(x)\,\mathrm{d}x = 2 - 2\sqrt{r_a}.$$
In this case the function is Riemann integrable on $(0, 1]$ and the value of the Riemann integral is 2.
3. Let $I = \mathbb{R}$ and define $f(x) = (1 + x^2)^{-1}$. In this case we have
$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{1 + x^2}\,\mathrm{d}x &= \lim_{R\to\infty} \int_{-R}^{0} \frac{1}{1 + x^2}\,\mathrm{d}x + \lim_{R\to\infty} \int_{0}^{R} \frac{1}{1 + x^2}\,\mathrm{d}x \\ &= \lim_{R\to\infty} \arctan R + \lim_{R\to\infty} \arctan R = \pi.\end{aligned}$$
Thus this function is Riemann integrable on $\mathbb{R}$ and has a Riemann integral of π.

4. The next example we consider is $I = \mathbb{R}$ and $f(x) = x(1+x^2)^{-1}$. In this case we compute

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x &= \lim_{R\to\infty} \int_{-R}^{0} \frac{x}{1+x^2}\,\mathrm{d}x + \lim_{R\to\infty} \int_{0}^{R} \frac{x}{1+x^2}\,\mathrm{d}x \\ &= \lim_{R\to\infty} \frac{1}{2}\log(1+R^2) - \lim_{R\to\infty} \frac{1}{2}\log(1+R^2).\end{aligned}$$

Now, it is not permissible to say here that $\infty - \infty = 0$. Therefore, we are forced to conclude that f is not Riemann integrable on $\mathbb{R}$.

5. To make the preceding example a little more dramatic, and to more convincingly illustrate why we should not cancel the infinities, we take $I = \mathbb{R}$ and $f(x) = x^3$. Here we compute

$$\int_{-\infty}^{\infty} x^3\,\mathrm{d}x = \lim_{R\to\infty} \frac{1}{4}R^4 - \lim_{R\to\infty} \frac{1}{4}R^4.$$

In this case again we must conclude that f is not Riemann integrable on $\mathbb{R}$. Indeed, it seems unlikely that one *would* wish to conclude that such a function was Riemann integrable since it is so badly behaved as $|t| \to \infty$. However, if we reject this function as being Riemann integrable, we must also reject the function of Example 4, even though it is not as ill behaved as the function here. •

Note that the above constructions involved first separating a function into its positive and negative parts, and then integrating these separately. However, there is not *a priori* reason why we could not have defined the limits in Definition 3.4.14 directly, and not just for positive functions. One can do this in fact. However, as we shall see, the two ensuing constructions of the integral are not equivalent.

3.4.18 Definition (Conditionally Riemann integrable functions on a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function whose restriction to any compact subinterval of I is Riemann integrable. Then f is ***conditionally Riemann integrable*** if the limit in the appropriate of the nine cases of Definition 3.4.14 exists. This limit is called the ***conditional Riemann integral*** of f. If f is conditionally integrable we write

$$\oint_I f(x)\,\mathrm{d}x$$

as the conditional Riemann integral. •

Before we explain the differences between conditionally integrable and integrable functions via examples, let us provide the relationship between the two notions.

3.4.19 Proposition (Relationship between integrability and conditional integrability) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, then the following statements hold:*

(i) if f *is Riemann integrable then it is conditionally Riemann integrable;*

(ii) if I *is additionally compact then, if* f *is conditionally Riemann integrable it is Riemann integrable.*

Proof In the proof it is convenient to make use of the results from Section 3.4.5.

(i) Let f_+ and f_- be the positive and negative parts of f. Since f is Riemann integrable, then so are f_+ and f_- by Definition 3.4.15. Moreover, since Riemann integrability and conditional Riemann integrability are clearly equivalent for nonnegative functions, it follows that f_+ and f_- are conditionally Riemann integrable. Therefore, by Proposition 3.4.22, it follows that $f = f_+ - f_-$ is conditionally Riemann integrable.

(ii) This follows from Definition 3.4.15 and Proposition 3.4.16. ■

Let us show that conditional Riemann integrability and Riemann integrability are not equivalent.

3.4.20 Example (A conditionally Riemann integrable function that is not Riemann integrable) Let $I = [1, \infty)$ and define $f(x) = \frac{\sin x}{x}$. Let us first show that f is conditionally Riemann integrable. We have, using integration by parts (Proposition 3.4.28),

$$\begin{aligned}\int_1^\infty \frac{\sin x}{x}\,\mathrm{d}x &= \lim_{R\to\infty}\int_1^R \frac{\sin x}{x}\,\mathrm{d}x = \lim_{R\to\infty}\left(-\frac{\cos x}{x}\Big|_1^R - \int_1^R \frac{\cos x}{x^2}\,\mathrm{d}x\right)\\ &= \cos 1 - \lim_{R\to\infty}\int_1^R \frac{\cos x}{x^2}\,\mathrm{d}x.\end{aligned}$$

We claim that the last limit exists. Indeed,

$$\left|\int_1^R \frac{\cos x}{x^2}\,\mathrm{d}x\right| \le \int_1^R \frac{|\cos x|}{x^2}\,\mathrm{d}x \le \int_1^R \frac{1}{x^2}\,\mathrm{d}x = 1 - \frac{1}{R},$$

and the limit as $R \to \infty$ is then 1. This shows that the limit defining the conditional integral is indeed finite, and so f is conditionally Riemann integrable on $[1, \infty)$.

Now let us show that this function is not Riemann integrable. By Proposition 3.4.25, f is Riemann integrable if and only if $|f|$ is Riemann integrable. For $R > 0$ let $N_R \in \mathbb{Z}_{>0}$ satisfy $R \in [N_R\pi, (N_R + 1)\pi]$. We then have

$$\begin{aligned}\int_1^R \left|\frac{\sin x}{x}\right|\mathrm{d}x &\ge \int_\pi^{N_R\pi} \left|\frac{\sin x}{x}\right|\mathrm{d}x\\ &\ge \sum_{j=1}^{N_R-1} \frac{1}{j\pi}\int_{j\pi}^{(j+1)\pi} |\sin x|\,\mathrm{d}x = \frac{2}{\pi}\sum_{j=1}^{N_R-1}\frac{1}{j}.\end{aligned}$$

By Example 2.4.2–2, the last sum diverges to ∞ as $N_R \to \infty$, and consequently the integral on the left diverges to ∞ as $R \to \infty$, giving the assertion. •

3.4.21 Remark ("Conditional Riemann integral" versus "Riemann integral") The previous example illustrates that one needs to exercise some care when talking about the Riemann integral. Adding to the possible confusion here is the fact that there is no established convention concerning what is intended when one says "Riemann integral." Many authors use "Riemann integrability" where we use "conditional Riemann integrability" and then use "absolute Riemann integrability" where we use "Riemann integrability." There is a good reason to do this.

1. One can think of integrals as being analogous to sums. When we talked about convergence of sums in Section 2.4 we used "convergence" to talk about that concept which, for the Riemann integral, is analogous to "conditional Riemann integrability" in our terminology. We used the expression "absolute convergence" for that concept which, for the Riemann integral, is analogous to "Riemann integrability" in our terminology. Thus the alternative terminology of "Riemann integrability" for "conditional Riemann integrability" and "absolute Riemann integrability" for "Riemann integrability" is more in alignment with the (more or less) standard terminology for sums.

However, there is also a good reason to use the terminology we use. However, the reasons here have to do with terminology attached to the Lebesgue integral that we discuss in Chapter III-2. However, here is as good a place as any to discuss this.

2. For the Lebesgue integral, the most natural notion of integrability is analogous to the notion of "Riemann integrability" in our terminology. That is, the terminology "Lebesgue integrability" is a generalisation of "Riemann integrability." The notion of "conditional Riemann integrability" is not much discussed for the Lebesgue integral, so there is not so much an established terminology for this. However, if there were an established terminology it would be "conditional Lebesgue integrability."

In Table 3.1 we give a summary of the preceding discussion, noting that apart from overwriting some standard conventions, there is no optimal way to choose what language to use. Our motivation for the convention we use is that it is best

Table 3.1 "Conditional" versus "absolute" terminology. In the top row we give our terminology, in the second row we give the alternative terminology for the Riemann integral, in the third row we give the analogous terminology for sums, and in the fourth row we give the terminology for the Lebesgue integral.

	Riemann integrable	conditionally Riemann integrable
Alternative	absolutely Riemann integrable	Riemann integrable
Sums	absolutely convergent	convergent
Lebesgue integral	Lebesgue integrable	conditionally Lebesgue integrable

that “Lebesgue integrability” should generalise “Riemann integrability.” But it is necessary to understand what one is reading and what is intended in any case. •

3.4.5 The Riemann integral and operations on functions

In this section we consider the interaction of integration with the usual algebraic and other operations on functions. We will consider both Riemann integrability and conditional Riemann integrability. If we wish to make a statement that we intend to hold for both notions, we shall write “(conditionally) Riemann integrable” to connote this. We will also write

$$(\mathrm{C})\int_I f(x)\,\mathrm{d}x$$

to denote either the Riemann integral or the conditional Riemann integral in cases where we wish for both to apply. The reader should also keep in mind that Riemann integrability and conditional Riemann integrability agree for compact intervals.

3.4.22 Proposition (Algebraic operations and the Riemann integral) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be (conditionally) Riemann integrable functions, and let* $\mathrm{c} \in \mathbb{R}$*. Then the following statements hold:*

(i) $\mathrm{f} + \mathrm{g}$ *is (conditionally) Riemann integrable and*

$$(\mathrm{C})\int_{\mathrm{I}}(\mathrm{f}+\mathrm{g})(\mathrm{x})\,\mathrm{dx} = (\mathrm{C})\int_{\mathrm{I}}\mathrm{f}(\mathrm{x})\,\mathrm{dx} + (\mathrm{C})\int_{\mathrm{I}}\mathrm{g}(\mathrm{x})\,\mathrm{dx};$$

(ii) cf *is (conditionally) Riemann integrable and*

$$(\mathrm{C})\int_{\mathrm{I}}(\mathrm{cf})(\mathrm{x})\,\mathrm{dx} = \mathrm{c}(\mathrm{C})\int_{\mathrm{I}}\mathrm{f}(\mathrm{x})\,\mathrm{dx};$$

(iii) if I *is additionally compact, then* fg *is Riemann integrable;*

(iv) if I *is additionally compact and if there exists* $\alpha \in \mathbb{R}_{>0}$ *such that* $\mathrm{g}(\mathrm{x}) \geq \alpha$ *for each* $\mathrm{x} \in \mathrm{I}$*, then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is Riemann integrable.*

Proof (i) We first suppose that $I = [a, b]$ is a compact interval. Let $\epsilon \in \mathbb{R}_{>0}$ and by Theorem 3.4.9 we let P_f and P_g be partitions of $[a, b]$ such that

$$A_+(f, P_f) - A_-(f, P_f) < \tfrac{\epsilon}{2}, \quad A_+(g, P_g) - A_-(g, P_g) < \tfrac{\epsilon}{2},$$

and let P be a partition for which $(x_0, x_1, \dots, x_k) = \mathrm{EP}(P) = \mathrm{EP}(P_f) \cup \mathrm{EP}(P_g)$. Then, using Proposition 2.2.27,

$$\sup\{f(x) + g(x) \mid x \in [x_{j-1}, x_j]\} = \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} + \sup\{g(x) \mid x \in [x_{j-1}, x_j]\}$$

and

$$\inf\{f(x) + g(x) \mid x \in [x_{j-1}, x_j]\} = \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} + \inf\{g(x) \mid x \in [x_{j-1}, x_j]\}$$

for each $j \in \{1, \dots, k\}$. Thus

$$A_+(f+g, P) - A_-(f+g, P) \leq A_+(f, P) + A_+(g, P) - A_-(f, P) - A_-(g, P) < \epsilon,$$

using Lemma 1 from the proof of Theorem 3.4.9. This shows that $f + g$ is Riemann integrable by Theorem 3.4.9.

Now let P_f and P_g be any two partitions and let P satisfy $(x_0, x_1, \dots, x_k) = \mathrm{EP}(P) = \mathrm{EP}(P_f) \cup \mathrm{EP}(P_g)$. Then

$$A_+(f, P_f) + A_+(g, P_g) \geq A_+(f, P) + A_+(g, P) \geq A_+(f+g, P) \geq I_+(f+g).$$

We then have

$$I_+(f+g) \leq A_+(f, P_f) + A_+(g, P_g) \quad \Longrightarrow \quad I_+(f+g) \leq I_+(f) + I_+(g).$$

In like fashion we obtain the estimate

$$I_-(f+g) \geq I_-(f) + I_-(g).$$

Combining this gives

$$I_-(f) + I_-(g) \leq I_-(f+g) = I_+(f+g) \leq I_+(f) + I_+(g),$$

which implies equality of these four terms since $I_-(f) = I_+(f)$ and $I_-(g) = I_+(g)$. This gives this part of the result when I is compact. The result follows for general intervals from the definition of the Riemann integral for such intervals, and by applying Proposition 2.3.23.

(ii) As in part (i), the result will follow if we can prove it when I is compact. When $c = 0$ the result is trivial, so suppose that $c \neq 0$. First consider the case $c > 0$. For $\epsilon \in \mathbb{R}_{>0}$ let P be a partition for which $A_+(f, P) - A_-(f, P) < \frac{\epsilon}{c}$. Since $A_-(cf, P) = cA_-(f, P)$ and $A_+(cf, P) = cA_+(f, P)$ (as is easily checked), we have $A_+(cf, P) - A_-(cf, P) < \epsilon$, showing that cf is Riemann integrable. The equalities $A_-(cf, P) = cA_-(f, P)$ and $A_+(cf, P) = cA_+(f, P)$ then directly imply that $I_-(cf) = cI_-(f)$ and $I_+(cf) = cI_+(f)$, giving the result for $c > 0$. For $c < 0$ a similar argument holds, but asking that P be a partition for which $A_+(f, P) - A_-(f, P) < -\frac{\epsilon}{c}$.

(iii) First let us show that if I is compact then f^2 is Riemann integrable if f is Riemann integrable. This, however, follows from Proposition 3.4.23 by taking $g\colon I \to \mathbb{R}$ to be $g(x) = x^2$. To show that a general product fg of Riemann integrable functions on a compact interval is Riemann integrable, we note that

$$fg = \tfrac{1}{2}((f+g)^2 - f^2 - g^2).$$

By part (i) and using the fact that the square of a Riemann integrable function is Riemann integrable, the function on the right is Riemann integrable, so giving the result.

(iv) That $\frac{1}{g}$ is Riemann integrable follows from Proposition 3.4.23 by taking $g\colon I \to \mathbb{R}$ to be $g(x) = \frac{1}{x}$. ■

In parts (iii) and (iv) we asked that the interval be compact. It is simple to find counterexamples which indicate that compactness of the interval is generally necessary (see Exercise 3.4.3).

We now consider the relationship between composition and Riemann integration.

3.4.23 Proposition (Function composition and the Riemann integral) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval, if* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is a Riemann integrable function satisfying* $\mathrm{image}(\mathrm{f}) \subseteq [\mathrm{c}, \mathrm{d}]$*, and if* $\mathrm{g}\colon [\mathrm{c}, \mathrm{d}] \to \mathbb{R}$ *is continuous, then* $\mathrm{g} \circ \mathrm{f}$ *is Riemann integrable.*

Proof Denote $M = \sup\{|g(y)| \mid y \in [c, d]\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and write $\epsilon' = \frac{\epsilon}{2M+d-c}$. Since g is uniformly continuous by the Heine–Cantor Theorem, let $\delta \in \mathbb{R}$ be chosen such that $0 < \delta < \epsilon'$ and such that, $|y_1 - y_2| < \delta$ implies that $|g(y_1) - g(y_2)| < \epsilon'$. Then choose a partition P of $[a, b]$ such that $A_+(f, P) - A_-(f, P) < \delta^2$. Let $(x_0, x_1, \dots, x_k)$ be the endpoints of P and define

$$\begin{aligned}
A &= \{j \in \{1, \dots, k\} \mid \ \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} < \delta\},\\
B &= \{j \in \{1, \dots, k\} \mid \ \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} \geq \delta\}.
\end{aligned}$$

For $j \in A$ we have $|f(\xi_1) - f(\xi_2)| < \delta$ for every $\xi_1, \xi_2 \in [x_{j-1}, x_j]$ which implies that $|g \circ f(\xi_1) - g \circ f(\xi_2)| < \epsilon'$ for every $\xi_1, \xi_2 \in [x_{j-1}, x_j]$. For $j \in B$ we have

$$\begin{aligned}
\delta \sum_{j \in B} (x_j - x_{j-1}) \leq \sum_{j \in B} \Big(&\sup\{f(x) \mid x \in [x_{j-1}, x_j]\}\\
&- \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big)(x_j - x_{j-1})\\
\leq{}& A_+(f, P) - A_-(f, P) < \delta^2.
\end{aligned}$$

Therefore we conclude that

$$\sum_{j \in B} (x_j - x_{j-1}) \leq \epsilon'.$$

Thus

$$\begin{aligned}
A_+(g \circ f, P) - A_-(g \circ f, P) &= \sum_{j=1}^{k} \Big(\sup\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\}\\
&\quad - \inf\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\}\Big)(x_j - x_{j-1})\\
&= \sum_{j \in A} \Big(\sup\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\}\\
&\quad - \inf\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\}\Big)(x_j - x_{j-1})\\
&\quad + \sum_{j \in B} \Big(\sup\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\}\\
&\quad - \inf\{g \circ f(x) \mid x \in [x_{j-1}, x_j]\}\Big)(x_j - x_{j-1})\\
&< \epsilon'(d - c) + 2\epsilon' M < \epsilon,
\end{aligned}$$

giving the result by Theorem 3.4.9. ■

The Riemann integral also has the expected properties relative to the partial order and the absolute value function on $\mathbb{R}$.

3.4.24 Proposition (Riemann integral and total order on $\mathbb{R}$) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{f}, \mathrm{g} \colon \mathrm{I} \to \mathbb{R}$ *be (conditionally) Riemann integrable functions for which* $\mathrm{f}(\mathrm{x}) \leq \mathrm{g}(\mathrm{x})$ *for each* $\mathrm{x} \in \mathrm{I}$*. Then*

$$(\mathrm{C}) \int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\,\mathrm{d}\mathrm{x} \leq (\mathrm{C}) \int_{\mathrm{I}} \mathrm{g}(\mathrm{x})\,\mathrm{d}\mathrm{x}.$$

Proof Note that by part (i) of Proposition 3.4.22 it suffices to take $f = 0$ and then show that $\int_I g(x)\,\mathrm{d}x \geq 0$. In the case where $I = [a, b]$ we have

$$\int_a^b g(x)\,\mathrm{d}x \geq (b-a)\inf\{g(x) \mid x \in [a,b]\} \geq 0,$$

which gives the result in this case. The result for general intervals follows from the definition, and the fact the a limit of nonnegative numbers is nonnegative. ■

3.4.25 Proposition (Riemann integral and absolute value on $\mathbb{R}$) *Let* I *be an interval, let* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$*, and define* $|\mathrm{f}| \colon \mathrm{I} \to \mathbb{R}$ *by* $|\mathrm{f}|(\mathrm{x}) = |\mathrm{f}(\mathrm{x})|$*. Then the following statements hold:*

(i) if f *is Riemann integrable then* $|\mathrm{f}|$ *is Riemann integrable;*

(ii) if I *is compact and if* f *is conditionally Riemann integrable then* $|\mathrm{f}|$ *is conditionally Riemann integrable.*

Moreover, if the hypotheses of either part hold then

$$\left|\int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\,\mathrm{d}\mathrm{x}\right| \leq \int_{\mathrm{I}} |\mathrm{f}|(\mathrm{x})\,\mathrm{d}\mathrm{x}.$$

Proof (i) If f is Riemann integrable then f_+ and f_- are Riemann integrable. Since $|f| = f_+ + f_-$ it follows from Proposition 3.4.22 that $|f|$ is Riemann integrable.

(ii) When I is compact, the statement follows since conditional Riemann integrability is equivalent to Riemann integrability.

The inequality in the statement of the proposition follows from Proposition 3.4.24 since $f(x) \leq |f(x)|$ for all $x \in I$. ■

We comment that the preceding result is, in fact, not true if one removes the condition that I be compact. We also comment that the converse of the result is false, in that the Riemann integrability of $|f|$ does not imply the Riemann integrability of f. The reader is asked to sort this out in Exercise 3.4.4.

The Riemann integral also behaves well upon breaking an interval into two intervals that are disjoint except for a common endpoint.

3.4.26 Proposition (Breaking the Riemann integral in two) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{I} = \mathrm{I}_1 \cup \mathrm{I}_2$*, where* $\mathrm{I}_1 \cap \mathrm{I}_2 = \{\mathrm{c}\}$*, where* c *is the right endpoint of* I_1 *and the left endpoint of* I_2*. Then* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$ *is (conditionally) Riemann integrable if and only if* $\mathrm{f}|\mathrm{I}_1$ *and* $\mathrm{f}|\mathrm{I}_2$ *are (conditionally) Riemann integrable. Furthermore, we have*

$$(\mathrm{C}) \int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\,\mathrm{d}\mathrm{x} = (\mathrm{C}) \int_{\mathrm{I}_1} \mathrm{f}(\mathrm{x})\,\mathrm{d}\mathrm{x} + (\mathrm{C}) \int_{\mathrm{I}_2} \mathrm{f}(\mathrm{x})\,\mathrm{d}\mathrm{x}.$$

Proof We first consider the case where $I_1 = [a,c]$ and $I_2 = [c,b]$.

Let us suppose that f is Riemann integrable and let $(x_0, x_1, \dots, x_k)$ be endpoints of a partition of $[a,b]$ for which $A_+(f,P) - A_-(f,P) < \epsilon$. If $c \in (x_0, x_1, \dots, x_k)$, say $c = x_j$, then we have

$$A_-(f,P) = A_-(f|I_1, P_1) + A_-(f|I_2, P_2), \quad A_+(f,P) = A_+(f|I_1, P_1) + A_+(f|I_2, P_2),$$

where $\mathrm{EP}(P_1) = (x_0, x_1, \dots, x_j)$ are the endpoints of a partition of $[a,c]$ and $\mathrm{EP}(P_2) = (x_j, \dots, x_k)$ is a partition of $[c,b]$. From this we directly deduce that

$$A_+(f|I_1, P_1) - A_-(f|I_1, P_1) < \epsilon, \quad A_+(f|I_2, P_2) - A_-(f|I_2, P_2) < \epsilon. \tag{3.14}$$

If c is not an endpoint of P, then one can construct a new partition P' of $[a,b]$ with c as an extra endpoint. By Lemma 1 of Theorem 3.4.9 we have $A_+(f,P') - A_-(f,P') < \epsilon$. The argument then proceeds as above to show that (3.14) holds. Thus $f|I_1$ and $f|I_2$ are Riemann integrable by Theorem 3.4.9.

To prove the equality of the integrals in the statement of the proposition, we proceed as follows. Let P_1 and P_2 be partitions of I_1 and I_2, respectively. From these construct a partition $P(P_1, P_2)$ of I by asking that $\mathrm{EP}(P(P_1,P_2)) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$. Then

$$A_+(f|I_1, P_1) + A_+(f|I_2, P_2) = A_+(f, P(P_1, P_2)).$$

Thus

$$\begin{aligned} \inf\{A_+(f|I_1, P_1) \mid P_1 \in \mathrm{Part}(I_1)\} &+ \inf\{A_+(f|I_2, P_2) \mid P_2 \in \mathrm{Part}(I_2)\} \\ &\geq \inf\{A_+(f,P) \mid P \in \mathrm{Part}(I)\}. \end{aligned} \tag{3.15}$$

Now let P be a partition of I and construct partitions $P_1(P)$ and $P_2(P)$ of I_1 and I_2 respectively by adding defining, if necessary, a new partition P' of I with c as the (say) jth endpoint, and then defining $P_1(P)$ such that $\mathrm{EP}(P_1(P))$ are the first $j+1$ endpoints of P' and then defining $P_2(P)$ such that $\mathrm{EP}(P_2(P))$ are the last $k-j$ endpoints of P'. By Lemma 1 of Theorem 3.4.9 we then have

$$A_+(f,P) \geq A_+(f,P') = A_+(f|I_1, P_1(P)) + A_+(f|I_2, P_2(P)).$$

This gives

$$\begin{aligned} &\inf\{A_+(f,P) \mid P \in \mathrm{Part}(I)\} \\ &\qquad \geq \inf\{A_+(f|I_1, P_1) \mid P_1 \in \mathrm{Part}(I_1)\} + \inf\{A_+(f|I_2, P_2) \mid P_2 \in \mathrm{Part}(I_2)\}. \end{aligned}$$

Combining this with (3.15) gives

$$\begin{aligned} &\inf\{A_+(f,P) \mid P \in \mathrm{Part}(I)\} \\ &\qquad = \inf\{A_+(f|I_1, P_1) \mid P_1 \in \mathrm{Part}(I_1)\} + \inf\{A_+(f|I_2, P_2) \mid P_2 \in \mathrm{Part}(I_2)\}, \end{aligned}$$

which is exactly the desired result.

The result for a general interval follows from the general definition of the Riemann integral, and from Proposition 2.3.23. ■

The next result gives a useful tool for evaluating integrals, as well as a being a result of some fundamental importance.

3.4.27 Proposition (Change of variables for the Riemann integral) *Let* $[a,b]$ *be a compact interval and let* $u\colon [a,b] \to \mathbb{R}$ *be differentiable with* u' *Riemann integrable. Suppose that* $\mathrm{image}(u) \subseteq [c,d]$ *and that* $f\colon [c,d] \to \mathbb{R}$ *is Riemann integrable and that* $f = F'$ *for some differentiable function* $F\colon [c,d] \to \mathbb{R}$*. Then*

$$\int_a^b f \circ u(x)u'(x)\,dx = \int_{u(a)}^{u(b)} f(y)\,dy.$$

Proof Let $G\colon [a,b] \to \mathbb{R}$ be defined by $G = F \circ u$. Then $G' = (f \circ u)u'$ by the Chain Rule. Moreover, G' is Riemann integrable by Propositions 3.4.22 and 3.4.23. Thus, twice using Theorem 3.4.30 below,

$$\int_a^b f \circ u(x)u'(x)\,dx = G(b) - G(a) = F \circ u(b) - F \circ u(a) = \int_{u(a)}^{u(b)} f(y)\,dy,$$

as desired. ■

As a final result in this section, we prove the extremely valuable integration by parts formula.

3.4.28 Proposition (Integration by parts for the Riemann integral) *If* $[a,b]$ *is a compact interval and if* $f, g\colon [a,b] \to \mathbb{R}$ *are differentiable functions with* f' *and* g' *Riemann integrable, then*

$$\int_a^b f(x)g'(x)\,dx + \int_a^b f'(x)g(x)\,dx = f(b)g(b) - f(a)g(a).$$

Proof By Proposition 3.2.10 it holds that fg is differentiable and that $(fg)' = f'g + fg'$. Thus, by Proposition 3.4.22, fg is differentiable with Riemann integrable derivative. Therefore, by Theorem 3.4.30 below,

$$\int_a^b (fg)(x)\,dx = f(b)g(b) - f(a)g(a),$$

and the result follows directly from the formula for the product rule. ■

3.4.6 The Fundamental Theorem of Calculus and the Mean Value Theorems

In this section we begin to explore the sense in which differentiation and integration are inverses of one another. This is, in actuality, and somewhat in contrast to the manner in which one considers this question in introductory calculus courses, a quite complicated matter. Indeed, we will not fully answer this question until Section III-2.9.7, after we have some knowledge of the Lebesgue integral. Nevertheless, in this section we give some simple results, and some examples which illustrate the value and the limitations of these results. We also present the Mean Value Theorems for integrals.

The following language is often used in conjunction with the Fundamental Theorem of Calculus.

3.4.29 Definition (Primitive) If $I \subseteq \mathbb{R}$ is an interval and if $f\colon I \to \mathbb{R}$ is a function, a ***primitive*** for f is a function $F\colon I \to \mathbb{R}$ such that $F' = f$. •

Note that primitives are not unique since if one adds a constant to a primitive, the resulting function is again a primitive.

The basic result of this section is the following.

3.4.30 Theorem (Fundamental Theorem of Calculus for Riemann integrals) *For a compact interval* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$*, the following statements hold:*

(i) if $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is Riemann integrable with primitive* $\mathrm{F}\colon \mathrm{I} \to \mathbb{R}$*, then*

$$\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\,\mathrm{d}\mathrm{x} = \mathrm{F}(\mathrm{b}) - \mathrm{F}(\mathrm{a});$$

(ii) if $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is Riemann integrable, and if* $\mathrm{F}\colon \mathrm{I} \to \mathbb{R}$ *is defined by*

$$\mathrm{F}(\mathrm{x}) = \int_{\mathrm{a}}^{\mathrm{x}} \mathrm{f}(\xi)\,\mathrm{d}\xi,$$

then

(a) F *is continuous and*

(b) at each point $\mathrm{x} \in \mathrm{I}$ *for which* f *is continuous,* F *is differentiable and* $\mathrm{F}'(\mathrm{x}) = \mathrm{f}(\mathrm{x})$*.*

Proof (i) Let $(P_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of partitions for which $\lim_{j\to\infty}|P_j| = 0$. Denote by $(x_{j,0}, x_{j,1}, \dots, x_{j,k_j})$ the endpoints of P_j, $j \in \mathbb{Z}_{>0}$. By the Mean Value Theorem, for each $j \in \mathbb{Z}_{>0}$ and for each $r \in \{1, \dots, k_r\}$, there exists $\xi_{j,r} \in [x_{j,r-1}, x_{j,r}]$ such that $F(x_{j,r}) - F(x_{j,r-1}) = f(\xi_{j,r})(x_{j,r} - x_{j,r-1})$. Since f is Riemann integrable we have

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x &= \lim_{j\to\infty} \sum_{r=1}^{k_j} f(\xi_{j,r})(x_{j,r} - x_{j,r-1})\\ &= \lim_{j\to\infty} \sum_{r=1}^{k_j} (F(x_{j,r}) - F(x_{j,r-1}))\\ &= \lim_{j\to\infty} (F(b) - F(a)) = F(b) - F(a),\end{aligned}$$

as desired.

(ii) Let $x \in (a, b)$ and note that, for h sufficiently small,

$$F(x + h) - F(x) = \int_x^{x+h} f(\xi)\,\mathrm{d}\xi,$$

using Proposition 3.4.26. By Proposition 3.4.24 it follows that

$$h \inf\{f(y) \mid y \in [a, b]\} \le \int_x^{x+h} f(\xi)\,\mathrm{d}\xi \le h \sup\{f(y) \mid y \in [a, b]\},$$

provided that $h > 0$. This shows that

$$\lim_{h\downarrow 0}\int_x^{x+h} f(\xi)\,\mathrm{d}\xi = 0.$$

A similar argument can be fashioned for the case when $h < 0$ to show also that

$$\lim_{h\uparrow 0}\int_x^{x+h} f(\xi)\,\mathrm{d}\xi = 0,$$

so showing that F is continuous at point in (a, b). A slight modification to this argument shows that F is also continuous at a and b.

Now suppose that f is continuous at x. Let $h > 0$. Again using Proposition 3.4.24 we have

$$\begin{aligned} & h\inf\{f(y) \mid y \in [x, x+h]\} \le \int_x^{x+h} f(\xi)\,\mathrm{d}\xi \le h\sup\{f(y) \mid y \in [x, x+h]\} \\ \implies \quad & \inf\{f(y) \mid y \in [x, x+h]\} \le \frac{F(x+h)-F(x)}{h} \le \sup\{f(y) \mid y \in [x, x+h]\}. \end{aligned}$$

Continuity of f at x gives

$$\lim_{h\downarrow 0}\inf\{f(y) \mid y \in [x, x+h]\} = f(x), \quad \lim_{h\downarrow 0}\sup\{f(y) \mid y \in [x, x+h]\} = f(x).$$

Therefore,

$$\lim_{h\downarrow 0}\frac{F(x+h)-F(x)}{h} = f(x).$$

A similar argument can be made for $h < 0$ to give

$$\lim_{h\uparrow 0}\frac{F(x+h)-F(x)}{h} = f(x),$$

so proving this part of the theorem. ■

Let us give some examples that illustrate what the Fundamental Theorem of Calculus says and does not say.

3.4.31 Examples (Fundamental Theorem of Calculus)

1. Let $I = [0, 1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} x, & x \in [0, \frac{1}{2}], \\ 1 - x, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then

$$F(x) \triangleq \int_0^x f(\xi)\,\mathrm{d}\xi = \begin{cases} \frac{1}{2}x^2, & x \in [0, \frac{1}{2}], \\ -\frac{1}{2}x^2 + x - \frac{1}{8}, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then, for any $x \in [a, b]$, we see that

$$\int_0^x f(\xi)\,\mathrm{d}\xi = F(x) - F(0).$$

This is consistent with part (i) of Theorem 3.4.30, whose hypotheses apply since f is continuous, and so Riemann integrable.

2. Let $I = [0, 1]$ and define $f\colon I \to \mathbb{R}$ by

$$f(x) = \begin{cases} 1, & x \in [0, \frac{1}{2}], \\ -1, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then

$$F(x) \triangleq \int_0^x f(\xi)\,\mathrm{d}\xi = \begin{cases} x, & x \in [0, \frac{1}{2}], \\ 1 - x, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Then, for any $x \in [a, b]$, we see that

$$\int_0^x f(\xi)\,\mathrm{d}\xi = F(x) - F(0).$$

In this case, we have the conclusions of part (i) of Theorem 3.4.30, and indeed the hypotheses hold, since f is Riemann integrable.

3. Let I and f be as in Example 1 above. Then f is Riemann integrable, and we see that F is continuous, as per part (ii) of Theorem 3.4.30, and that F is differentiable, also as per part (ii) of Theorem 3.4.30.
4. Let I and f be as in Example 2 above. Then f is Riemann integrable, and we see that F is continuous, as per part (ii) of Theorem 3.4.30. However, f is not continuous at $x = \frac{1}{2}$, and we see that, correspondingly, F is not differentiable at $x = \frac{1}{2}$.
5. The next example we consider is one with which, at this point, we can only be sketchy about the details. Consider the Cantor function $f_C\colon [0, 1] \to \mathbb{R}$ of Example 3.2.27. Note that f_C' is defined and equal to zero, except at points in the Cantor set C; thus except at points forming a set of measure zero. It will be clear when we discuss the Lebesgue integral in Section III-2.9 that this ensures that $\int_0^x f_C'(\xi)\,\mathrm{d}\xi = 0$ for every $x \in [0, 1]$, where the integral in this case is the Lebesgue integral. (By defining f_C' arbitrarily on C, we can also use the Riemann integral by virtue of Theorem 3.4.11.) This shows that the conclusions of part (i) of Theorem 3.4.30 can fail to hold, even when the derivative of F is defined almost everywhere.
6. The last example we give is the most significant, in some sense, and is also the most complicated. The example we give is of a function $F\colon [0, 1] \to \mathbb{R}$ that is differentiable with bounded derivative, but whose derivative $f = F'$

is not Riemann integrable. Thus f possesses a primitive, but is not Riemann integrable.

To define F, let $G\colon \mathbb{R}_{>0} \to \mathbb{R}$ be the function

$$G(x) = \begin{cases} x^2 \sin\frac{1}{x}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

For $c > 0$ let $x_c > 0$ be defined by

$$x_c = \sup\{x \in \mathbb{R}_{>0} \mid G'(x) = 0,\ x \leq c\},$$

and define $G_c\colon (0, c] \to \mathbb{R}$ by

$$G_c(x) = \begin{cases} G(x), & x \in (0, x_c], \\ G(x_c), & x \in (x_c, x]. \end{cases}$$

Now, for $\epsilon \in (0, \frac{1}{2})$, let $C_\epsilon \subseteq [0, 1]$ be a fat Cantor set as constructed in Example 2.5.42. Define F as follows. If $x \in C_\epsilon$ we take $F(x) = 0$. If $x \notin C_\epsilon$, then, since C_ϵ is closed, by Proposition 2.5.6 x lies in some open interval, say (a, b). Then take $c = \frac{1}{2}(b - a)$ and define

$$F(x) = \begin{cases} G_c(x - a), & x \in (a, \frac{1}{2}(a + b)), \\ G_c(b - x), & x \in [\frac{1}{2}(a + b), b). \end{cases}$$

Note that $F|(a, b)$ is designed so that its derivative will oscillate wildly in the limit as the endpoints of (a, b) are approached, but be nicely behaved at all points in (a, b). This is, as we shall see, the key feature of F.

Let us record some properties of F in a sequence of lemmata.

1 **Lemma** *If* x ∈ C_ϵ, *then* F *is differentiable at* x *and* F′(x) = 0.

Proof Let $y \in [0, 1] \setminus \{x\}$. If $y \in C_\epsilon$ then

$$\frac{f(y) - f(x)}{y - x} = 0.$$

If $y \notin C_\epsilon$, then y must lie in an open interval, say (a, b). Let d be the endpoint of (a, b) nearest y and let $c = \frac{1}{2}(b - a)$. Then

$$\begin{aligned} \left|\frac{f(y) - f(x)}{y - x}\right| = \frac{f(y)}{y - x} \leq \frac{f(y)}{y - d} &= \frac{G_c(|y - d|)}{y - d} \\ &\leq \frac{|y - d|^2}{y - d} = |y - d| \leq |y - x|. \end{aligned}$$

Thus

$$\lim_{y \to x} \frac{f(y) - f(x)}{y - x} = 0,$$

giving the lemma. ▼

2 Lemma *If* $\mathrm{x} \notin \mathrm{C}_\epsilon$*, then* F *is differentiable at* x *and* $|\mathrm{F}'(\mathrm{x})| \leq 3$.

Proof By definition of F for points not in C_ϵ we have

$$|F'(x)| \leq \left|2y \sin \tfrac{1}{y} - \cos \tfrac{1}{y}\right| \leq 3,$$

for some $y \in [0,1]$. ▼

3 Lemma $\mathrm{C}_\epsilon \subseteq \mathrm{D}_{\mathrm{F}'}$.

Proof By construction of C_ϵ, if $x \in C_\epsilon$ then there exists a sequence $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$ of open intervals in $[0,1] \setminus C_\epsilon$ having the property that $\lim_{j\to\infty} a_j = \lim_{j\to\infty} b_j = x$. Note that $\limsup_{y\downarrow 0} g'(y) = 1$. Therefore, by the definition of F on the open intervals (a_j, b_j), $j \in \mathbb{Z}_{>0}$, it holds that $\limsup_{y\downarrow a_j} F'(y) = \limsup_{y\uparrow b_j} F'(y) = 1$. Therefore, $\limsup_{y\to x} F'(y) = 1$. Since $F'(x) = 0$, it follows that F' is discontinuous at x. ▼

Since F' is discontinuous at all points in C_ϵ, and since C_ϵ does not have measure zero, it follows from Theorem 3.4.11 that F' is not Riemann integrable. Therefore, the function $f = F'$ possesses a primitive, namely F, but is not Riemann integrable. •

Finally we state two results that, like the Mean Value Theorem for differentiable functions, relate the integral to the values of a function.

3.4.32 Proposition (First Mean Value Theorem for Riemann integrals) *Let* $[\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}, \mathrm{g}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be functions with* f *continuous and with* g *nonnegative and Riemann integrable. Then there exists* $\mathrm{c} \in [\mathrm{a}, \mathrm{b}]$ *such that*

$$\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\mathrm{g}(\mathrm{x})\,\mathrm{dx} = \mathrm{f}(\mathrm{c}) \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{g}(\mathrm{x})\,\mathrm{dx}$$

Proof Let

$$m = \inf\{f(x) \mid x \in [a,b]\}, \quad M = \sup\{f(x) \mid x \in [a,b]\}.$$

Since g is nonnegative we have

$$mg(x) \leq f(x)g(x) \leq Mg(x), \qquad x \in [a,b],$$

from which we deduce that

$$m\int_a^b g(x)\,\mathrm{d}x \leq \int_a^b f(x)g(x)\,\mathrm{d}x \leq M\int_a^b g(x)\,\mathrm{d}x.$$

Continuity of f and the Intermediate Value Theorem gives $c \in [a,b]$ such that the result holds. ■

3.4.33 Proposition (Second Mean Value Theorem for Riemann integrals) *Let* $[a, b]$ *be a compact interval and let* $f, g\colon [a, b] \to \mathbb{R}$ *be functions with*

(i) g *Riemann integrable and having the property that there exists* G *such that* $g = G'$, *and*

(ii) f *differentiable with Riemann integrable, nonnegative derivative.*

Then there exists $c \in [a, b]$ *so that*

$$\int_a^b f(x)g(x)\,dx = f(a)\int_a^c g(x)\,dx + f(b)\int_c^b g(x)\,dx.$$

Proof Without loss of generality we may suppose that

$$G(x) = \int_a^x g(\xi)\,d\xi,$$

since all we require is that $G' = g$. We then compute

$$\begin{aligned}\int_a^b f(x)g(x)\,dx = \int_a^b f(x)G'(x)\,dx &= f(b)G(b) - \int_a^b f'(x)G(x)\,dx \\ &= f(b)G(b) - G(c)\int_a^b f'(x)\,dx,\end{aligned}$$

for some $c \in [a, b]$, using integration by parts and Proposition 3.4.32. Now using Theorem 3.4.30,

$$\int_a^b f(x)g(x)\,dx = f(b)G(b) - G(c)(f(b) - f(a)),$$

which gives the desired result after using the definition of G and after some rearrangement. ■

3.4.7 The Cauchy principal value

In Example 3.4.17 we explored some of the nuances of the improper Riemann integral. There we saw that for integrals that are defined using limits, one often needs to make the definitions in a particular way. The principal value integral is intended to relax this, and enable one to have a meaningful notion of the integral in cases where otherwise one might not. To motivate our discussion we consider an example.

3.4.34 Example Let $I = [-1, 2]$ and consider the function $f\colon I \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} \frac{1}{x}, & x \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

This function has a singularity at $x = 0$, and the integral $\int_{-1}^{2} f(x)\,\mathrm{d}x$ is actually divergent. However, for $\epsilon \in \mathbb{R}_{>0}$ note that

$$\int_{-1}^{-\epsilon} \frac{1}{x}\,\mathrm{d}x + \int_{\epsilon}^{2} \frac{1}{x}\,\mathrm{d}x = -\log x|_{\epsilon}^{1} + \log x|_{\epsilon}^{2} = \log 2.$$

Thus we can devise a way around the singularity in this case, the reason being that the singular behaviour of the function on either side of the function "cancels" that on the other side. •

With this as motivation, we give a definition.

3.4.35 Definition (Cauchy principal value) Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function. Denote $a = \inf I$ and $b = \sup I$, allowing that $a = -\infty$ and $b = \infty$.

(i) If, for $x_0 \in \operatorname{int}(I)$, there exists $\epsilon_0 \in \mathbb{R}_{>0}$ such that the functions $f|(a, x_0 - \epsilon]$ and $f|[x_0 + \epsilon, b)$ are Riemann integrable for all $\epsilon \in (0, \epsilon_0]$, then the ***Cauchy principal value*** for f is defined by

$$\mathrm{pv}\int_I f(x)\,\mathrm{d}x = \lim_{\epsilon \to 0}\left(\int_a^{x_0-\epsilon} f(x)\,\mathrm{d}x + \int_{x_0+\epsilon}^{b} f(x)\,\mathrm{d}x\right).$$

(ii) If $a = -\infty$ and $b = \infty$ and if for each $R \in \mathbb{R}_{>0}$ the function $f|[-R, R]$ is Riemann integrable, then the ***Cauchy principal value*** for f is defined by

$$\mathrm{pv}\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \lim_{R \to \infty}\int_{-R}^{R} f(x)\,\mathrm{d}x.$$

•

3.4.36 Remarks

1. If f is Riemann integrable on I then the Cauchy principal value is equal to the Riemann integral.
2. The Cauchy principal value is allowed to be infinite by the preceding definition, as the following examples will show.
3. It is not standard to define the Cauchy principal value in part (ii) of the definition. In many texts where the Cauchy principal value is spoken of, it is part (i) that is being used. However, we will find the definition from part (ii) useful. •

3.4.37 Examples (Cauchy principal value)

1. For the example of Example 3.4.34 we have

$$\mathrm{pv}\int_{-1}^{2} \frac{1}{x}\,\mathrm{d}x = \log 2.$$

2. For $I = \mathbb{R}$ and $f(x) = x(1+x^2)^{-1}$ we have

$$\mathrm{pv}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x = \lim_{R\to\infty}\int_{-R}^{R}\frac{x}{1+x^2}\,\mathrm{d}x = \lim_{R\to\infty}\Big(\frac{1}{2}\log(1+R^2) - \frac{1}{2}\log(1+R^2)\Big) = 0.$$

   Note that in Example 3.4.17–4 we showed that this function was not Riemann integrable.

3. Next we consider $I = \mathbb{R}$ and $f(x) = |x|(1+x^2)$. In this case we compute

$$\mathrm{pv}\int_{-\infty}^{\infty} \frac{|x|}{1+x^2}\,\mathrm{d}x = \lim_{R\to\infty}\int_{-R}^{R}\frac{|x|}{1+x^2}\,\mathrm{d}x = \lim_{R\to\infty}\Big(\frac{1}{2}\log(1+R^2) + \frac{1}{2}\log(1+R^2)\Big) = \infty.$$

   We see then that there is no reason why the Cauchy principal value may not be infinite. •

3.4.8 Notes

The definition we give for the Riemann integral is actually that used by Darboux, and the condition given in part (iii) of Theorem 3.4.9 is the original definition of Riemann. What Darboux showed was that the two definitions are equivalent. It is not uncommon to instead use the Darboux definition as the standard definition because, unlike the definition of Riemann, it does not rely on an arbitrary selection of a point from each of the intervals forming a partition.

Exercises

3.4.1 Let $I \subseteq \mathbb{R}$ be an interval and let $f\colon I \to \mathbb{R}$ be a function that is Riemann integrable and satisfies $f(x) \geq 0$ for all $x \in I$. Show that $\int_I f(x)\,\mathrm{d}x \geq 0$.

3.4.2 Let $I \subseteq \mathbb{R}$ be an interval, let $f, g\colon I \to \mathbb{R}$ be functions, and define $D_{f,g} = \{x \in I \mid f(x) \neq g(x)\}$.

   (a) Show that, if $D_{f,g}$ is finite and f is Riemann integrable, then g is Riemann integrable and $\int_I f(x)\,\mathrm{d}x = \int_I g(x)\,\mathrm{d}x$.

   (b) Is it true that, if $D_{f,g}$ is countable and f is Riemann integrable, then g is Riemann integrable and $\int_I f(x)\,\mathrm{d}x = \int_I g(x)\,\mathrm{d}x$? If it is true, give a proof; if it is not true, give a counterexample.

3.4.3 Do the following:

   (a) find an interval I and functions $f, g\colon I \to \mathbb{R}$ such that f and g are both Riemann integrable, but fg is not Riemann integrable;

   (b) find an interval I and functions $f, g\colon I \to \mathbb{R}$ such that f and g are both Riemann integrable, but $g \circ f$ is not Riemann integrable.

3.4.4 Do the following:

   (a) find an interval I and a conditionally Riemann integrable function $f\colon I \to \mathbb{R}$ such that $|f|$ is not Riemann integrable;

(b) find a function $f\colon [0,1] \to \mathbb{R}$ such that $|f|$ is Riemann integrable, but f is not Riemann integrable.

3.4.5 Show that, if $f\colon [a,b] \to \mathbb{R}$ is continuous, then there exists $c \in [a,b]$ such that

$$\int_a^b f(x)\,\mathrm{d}x = f(c)(b-a).$$

Section 3.5

The Riemann–Stieltjes integral

In this section we consider a generalisation of the Riemann integral that has some important applications, some of which will come up during the course of our presentation. The character of the development is slightly different than for the Riemann integral, and we will point out the differences and the reasons for these differences as they arise.

Do I need to read this section? This section can be skipped until it is needed in reading other material in the text. The principal references to the Riemann–Stieltjes integral will arise in our characterisation of a certain dual space in Theorem III-2.12.6 and in characterising certain distributions in Section IV-3.4.5. •

3.5.1 The Riemann–Stieltjes integral on compact intervals

The definition of the Riemann–Stieltjes[12] integral is done in a manner that is more or less direct, as compared to the Riemann integral, where one introduces step functions, then lower and upper sums, etc. The reason for this is that, as we shall see, the characterisation using lower and upper sums is a little more subtle for the Riemann–Stieltjes integral than for the Riemann integral. Thus we essentially define the Riemann–Stieltjes integral in a manner similar to the Darboux characterisation for the Riemann integral as given in Theorem 3.4.9.

We first define the approximation used in the definition of the Riemann–Stieltjes integral.

3.5.1 Definition (Riemann–Stieltjes sum) Let $I = [a, b]$ be a compact interval and let P be a partition with endpoints $(x_0, x_1, \dots, x_k)$. Let $f, \varphi\colon [a, b] \to \mathbb{R}$ be bounded functions. The ***Riemann–Stieltjes sum*** of f with respect to φ associated to a partition P and a selection ξ from P is

$$A(f, \varphi, P, \xi) = \sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})). \qquad \bullet$$

Now we can directly define the Riemann–Stieltjes integral.

3.5.2 Definition (Riemann–Stieltjes integral on compact intervals) Let $I = [a, b]$ be a compact interval and let $f, \varphi\colon [a, b] \to \mathbb{R}$ be bounded functions. Let $I(f, \varphi) \in \mathbb{R}$. If,

[12] Thomas Jan Stieltjes 1856–1894 was a Dutch mathematician who worked in the areas of analysis, number theory, and complex function theory.

for each $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for every partition P of $[a, b]$ with $|P| < \delta$ and for every selection $\boldsymbol{\xi}$ of P, we have

$$|A(f, \varphi, P, \boldsymbol{\xi}) - I(f, \varphi)| < \epsilon,$$

then f is ***Riemann–Stieltjes integrable*** with respect to φ. We denote

$$I(f, \varphi) = \int_a^b f(x)\,\mathrm{d}\varphi(x),$$

which is the ***Riemann–Stieltjes integral*** of f with respect to φ. The function f is called the ***integrand*** and the function φ is called the ***integrator***. •

The Darboux characterisation of Riemann integrability given in Theorem 3.4.9 immediately gives the following result.

3.5.3 Theorem (The Riemann–Stieltjes integral generalises the Riemann integral) *Let* $[\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}, \varrho\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be functions with* f *bounded and with* $\varrho(\mathrm{x}) = \mathrm{x}$*. Then* f *is Riemann integrable if and only if it is Riemann–Stieltjes integrable with respect to* ϱ*, and, moreover,*

$$\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\,\mathrm{dx} = \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\,\mathrm{d}\varrho(\mathrm{x}).$$

Unlike (for the most part) the Riemann integral, there are essential subtleties in the definition of the Riemann–Stieltjes integral that can lead to confusion. Let us address one of these subtleties head-on by giving an alternative definition of the Riemann–Stieltjes. The definition we give above, which is the one we shall use, is essentially the classical version. The one we give now is commonly encountered in more modern treatments of the Riemann–Stieltjes integral.

3.5.4 Definition (Generalised Riemann–Stieltjes integral on compact intervals) Let $I = [a, b]$ be a compact interval and let $f, \varphi\colon [a, b] \to \mathbb{R}$ be bounded functions. Let $I_g(f, \varphi) \in \mathbb{R}$. If, for each $\epsilon \in \mathbb{R}_{>0}$ there exists a partition P such that, for every refinement P' of P and for every selection $\boldsymbol{\xi}'$ of P', we have

$$|A(f, \varphi, P, \boldsymbol{\xi}) - I_g(f, \varphi)| < \epsilon,$$

then f is ***generalised Riemann–Stieltjes integrable*** with respect to φ. We denote

$$I_g(f, \varphi) = G\int_a^b f(x)\,\mathrm{d}\varphi(x),$$

which is the ***generalised Riemann–Stieltjes integral*** of f with respect to φ. •

It is fairly evident that, if f is Riemann–Stieltjes integrable with respect to φ, then it is also generalised Riemann–Stieltjes integrable with respect to φ. The converse is not true, however. Before we get to a counterexample, let us also consider the matter of using lower and upper sums in the definition of the Riemann–Stieltjes integral.

3.5.5 Definition (Lower and upper Riemann–Stieltjes sums) Let $I = [a, b]$ be a compact interval, let $f, \varphi \colon I \to \mathbb{R}$ be bounded functions on I, and let $P = (I_1, \dots, I_k)$ be a partition with endpoints $(x_0, x_1, \dots, x_k)$.

(i) The ***lower Riemann–Stieltjes sum*** of f with respect to φ and the partition P is

$$A_-(f, \varphi, P) = \sum_{j=1}^{k} \Big(\inf\{f(x) \mid x \in \mathrm{cl}(I_j)\}\Big)(\varphi(x_j) - \varphi(x_{j-1})).$$

(ii) The ***upper Riemann–Stieltjes sum*** of f with respect to φ and the partition P is

$$A_+(f, \varphi, P) = \sum_{j=1}^{k} \Big(\sup\{f(x) \mid x \in \mathrm{cl}(I_j)\}\Big)(\varphi(x_j) - \varphi(x_{j-1})).$$ •

As we did for the Riemann integral, we can now define the lower and upper Riemann–Stieltjes integrals.

3.5.6 Definition (Lower and upper Riemann–Stieltjes integral) Let $I = [a, b]$ be a compact interval and let $f, \varphi \colon I \to \mathbb{R}$ be bounded functions.

(i) The ***lower Riemann–Stieltjes integral*** of f with respect to φ is

$$I_-(f, \varphi) = \sup\{A_-(f, \varphi, P) \mid P \in \mathrm{Part}(I)\}.$$

(ii) The ***upper Riemann–Stieltjes integral*** of f with respect to φ is

$$I_+(f, \varphi) = \inf\{A_+(f, \varphi, P) \mid P \in \mathrm{Part}(I)\}.$$ •

Unlike what we saw for Riemann sums, the sets

$$\{A_-(f, \varphi, P) \mid P \in \mathrm{Part}(I)\}, \quad \{A_+(f, \varphi, P) \mid P \in \mathrm{Part}(I)\}$$

may not be bounded even though f and φ. However, one does have the following result which apart from giving conditions for the boundedness of these sets, gives us our first glimpse of why the notion of bounded variation should come up for the integrator in the Riemann–Stieltjes integral.

3.5.7 Proposition (Existence of lower and upper Riemann–Stieltjes integrals) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}, \varphi \colon \mathrm{I} \to \mathbb{R}$ *be functions with* f *bounded and* φ *of bounded variation. Then both* $\mathrm{I}_-(\mathrm{f}, \varphi)$ *and* $\mathrm{I}_+(\mathrm{f}, \varphi)$ *are finite.*

Proof Let $M = \sup\{|f(x)| \mid x \in [a,b]\}$. Then

$$\begin{aligned}|I_-(f,\varphi)| &= |\sup\{A_-(f,\varphi,P) \mid P \in \mathrm{Part}(I)\}| \\ &= \left|\sup\left\{\sum_{j=1}^{k} \inf\{f(x) \mid x \in [x_{j-1},x_j]\}(\varphi(x_j)-\varphi(x_{j-1}))\right|\right. \\ &\qquad (x_0,x_1,\dots,x_k) = \mathrm{EP}(P),\ P \in \mathrm{Part}(I)\}| \\ &\le \sup\left\{\sum_{j=1}^{k} M|\varphi(x_j)-\varphi(x_{j-1})| \;\middle|\; (x_0,x_1,\dots,x_k) = \mathrm{EP}(P),\ P \in \mathrm{Part}(I)\right\} \\ &= M\,\mathrm{TV}(\varphi).\end{aligned}$$

A similar computation shows that $|I_+(f,\varphi)| \le M\,\mathrm{TV}(\varphi)$, giving the result. ■

Were life with the Riemann–Stieltjes integrals as they are with Riemann integrals, we would now show that Riemann–Stieltjes integrability is equivalent to the equality of the upper and lower Riemann–Stieltjes integrals. However, this is not true, so we instead give an example that shows this, as well as showing the distinction between the Riemann–Stieltjes integral and the generalised Riemann–Stieltjes integral.

3.5.8 Example (A simple but subtle example) Consider the interval $[0,1]$ and the functions $f,\varphi\colon [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 0, & x \in [0,\frac{1}{2}], \\ 1, & x \in (\frac{1}{2},1], \end{cases} \qquad \varphi(x) = \begin{cases} 0, & x \in [0,\frac{1}{2}), \\ 1, & x \in [\frac{1}{2},1]. \end{cases}$$

If P is a partition having $\frac{1}{2}$ as an endpoint, then it follows immediately that $A_+(f,\varphi,P) = A_-(f,\varphi,P) = 0$. For such a partition we also have $A(f,\varphi,P,\boldsymbol{\xi}) = 0$ for every selection $\boldsymbol{\xi}$ from P. If P is a partition not having $\frac{1}{2}$ as an endpoint then $A_+(f,\varphi,P) = 1$ and $A_-(f,\varphi,P) = 0$. For such a partition let I_j be the interval containing $\frac{1}{2}$ in its interior. Then we have

$$A(f,\varphi,P,\boldsymbol{\xi}) = \begin{cases} 0, & \xi_j \le \frac{1}{2}, \\ 1, & \xi_j > \frac{1}{2}. \end{cases}$$

These calculations give the following conclusions.

1. The function f is not Riemann–Stieltjes integrable with respect to φ. To see this, note that, for any $\delta \in \mathbb{R}_{>0}$, there exists partitions P_1 and P_2 with selections $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$, respectively, such that $|P_1|, |P_2| < \delta$ and such that $A(f,\varphi,P_1,\boldsymbol{\xi}_1) = 0$ and $A(f,\varphi,P_2,\boldsymbol{\xi}_2) = 1$.
2. The function f is generalised Riemann–Stieltjes integrable with respect to φ and $G\int_a^b f(x)\,\mathrm{d}\varphi(x) = 0$. Indeed, let P have endpoints $\{0,\frac{1}{2},1\}$. Then, for any refinement P' of P and for any selection $\boldsymbol{\xi}'$ from P', $A(f,\varphi,P',\boldsymbol{\xi}') = 0$.

3. The lower and upper Riemann–Stieltjes integrals exist and are equal, and are both equal to the generalised Riemann–Stieltjes integral. •

This example shows that things are clearly different with the Riemann–Stieltjes integral(s) than they are with the Riemann integral. In the next section we shall consider some results which serve to further illustrate this dichotomy by showing that some of the conditions of Theorem 3.4.9 for the Riemann integral apply to the Riemann–Stieltjes integral, while others apply to the generalised Riemann–Stieltjes integral.

The preceding presentation of the definitions and basic properties of the Riemann–Stieltjes integral is one that probably deserves multiple readings to fully appreciate. Moreover, the impact of some of the ideas may not become clear until our discussions in the next section. One of the principle causes for the complication is our multiple, and distinct, definitions of Riemann–Stieltjes integrals. The reason we do this is that *both* definitions are in common use, so it is best to explicitly address this. Each of the definitions has its advantages in terms of the clarity of presentation. For example, Proposition 3.5.9 gives the Riemann–Stieltjes a nicer correspondence with the some of the characterisations of Riemann integral from Theorem 3.4.9, while Theorem 3.5.12 gives nicer correspondence of the generalised Riemann–Stieltjes integral with the other of the characterisations of the Riemann integral from Theorem 3.4.9. Moreover, in Theorem 3.5.18 we shall see that one has an in some sense nicer characterisation of conditions for the existence of Riemann–Stieltjes integrals than one has for generalised Riemann–Stieltjes integrals. Furthermore, our primary interest in the Riemann–Stieltjes integral (e.g., in Theorem III-2.12.6) will be restricted to situations where the Riemann–Stieltjes and the generalised Riemann–Stieltjes integrals exist, and so agree.

3.5.2 Characterisations of Riemann–Stieltjes integrable functions on compact intervals

Before we get to the matter of providing conditions for (generalised) Riemann–Stieltjes integrability, let us provide useful characterisations, along the lines of the Cauchy condition of Theorem 3.4.9.

3.5.9 Proposition (Cauchy condition for Riemann–Stieltjes integral) *Let* $I = [a, b]$ *be a compact interval and let* $f, \varphi\colon [a, b] \to \mathbb{R}$ *be functions. Then the following two statements are equivalent:*

(i) f *is Riemann–Stieltjes integrable with respect to* φ*;*

(ii) for each $\epsilon \in \mathbb{R}_{>0}$ *there exists* $\delta \in \mathbb{R}_{>0}$ *such that, for any partitions* $P = (I_1, \dots, I_k)$ *and* $P' = (I'_1, \dots, I'_{k'})$ *with* $|P|, |P'| < \delta$ *and for any selections* $(\xi_1, \dots, \xi_k)$ *and* $(\xi'_1, \dots, \xi'_{k'})$ *from* P *and* P'*, respectively, we have*

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - \sum_{j=1}^{k'} f(\xi'_j)(\varphi(x'_j) - \varphi(x'_{j-1}))\right| < \epsilon,$$

where $\mathrm{EP(P)} = (\mathrm{x}_0, \mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k})$ *and* $\mathrm{EP(P')} = (\mathrm{x}'_0, \mathrm{x}'_1, \dots, \mathrm{x}'_{\mathrm{k}'})$.

Proof The proof here mirrors, up to necessary modifications of notation, the proof of the equivalence of parts (iii) and (iv) in Theorem 3.4.9. We leave to the reader the routine matter of checking that this is indeed the case. ■

For the generalised Riemann–Stieltjes integral, we have the following formulation.

3.5.10 Proposition (Cauchy condition for generalised Riemann–Stieltjes integral) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}, \varphi \colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be functions. Then the following two statements are equivalent:*

(i) f *is Riemann–Stieltjes integrable with respect to* φ*;*

(ii) for each $\epsilon \in \mathbb{R}_{>0}$ *there exists a partition* P_0 *such that, for any refinements* $\mathrm{P} = (\mathrm{I}_1, \dots, \mathrm{I}_\mathrm{k})$ *and* $\mathrm{P}' = (\mathrm{I}'_1, \dots, \mathrm{I}'_{\mathrm{k}'})$ *of* P_0 *and for any selections* $(\xi_1, \dots, \xi_\mathrm{k})$ *and* $(\xi'_1, \dots, \xi'_{\mathrm{k}'})$ *from* P *and* P′*, respectively, we have*

$$\left|\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{f}(\xi_\mathrm{j})(\varphi(\mathrm{x}_\mathrm{j}) - \varphi(\mathrm{x}_{\mathrm{j}-1})) - \sum_{\mathrm{j}=1}^{\mathrm{k}'} \mathrm{f}(\xi'_\mathrm{j})(\varphi(\mathrm{x}'_\mathrm{j}) - \varphi(\mathrm{x}'_{\mathrm{j}-1}))\right| < \epsilon,$$

where $\mathrm{EP(P)} = (\mathrm{x}_0, \mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k})$ *and* $\mathrm{EP(P')} = (\mathrm{x}'_0, \mathrm{x}'_1, \dots, \mathrm{x}'_{\mathrm{k}'})$.

Proof Suppose that f is Riemann–Stieltjes integrable with respect to φ, and let $I(f, \varphi)$ denote the value of the integral. Let $\epsilon \in \mathbb{R}_{>0}$ and let P_0 be a partition such that, whenever $P = (I_1, \dots_k)$ is a refinement of P_0 and $(\xi_1, \dots, \xi_k)$ is a selection from P, it holds that

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - I(f, \varphi)\right| < \frac{\epsilon}{2}.$$

Now let $P = (I_1, \dots, I_k)$ and $P' = (I'_1, \dots, I'_{k'})$ be two refinements of P_0 and let $(\xi_1, \dots, \xi_k)$ and $(\xi'_1, \dots, \xi'_{k'})$ selections from P and P', respectively. Then we have

$$\begin{aligned}&\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - \sum_{j=1}^{k'} f(\xi'_j)(\varphi(x'_j) - \varphi(x'_{j-1}))\right| \\ &\qquad \leq \left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - I(f)\right| + \left|\sum_{j=1}^{k'} f(\xi'_j)(\varphi(x'_j) - \varphi(x'_{j-1})) - I(f)\right| < \epsilon,\end{aligned}$$

which gives this part of the result.

For the converse, let $(P_{j,0})_{j \in \mathbb{Z}_{>0}}$ be a sequence of partitions for which, if P_j and P'_j are refinements of $P_{j,0}$ and if $\boldsymbol{\xi}_j$ and $\boldsymbol{\xi}'_j$ are selections from P_j and P'_j, respectively, then

$$|A(f, \varphi, P_j, \boldsymbol{\xi}_j) - A(f, \varphi, P'_j, \boldsymbol{\xi}'_j)| < \tfrac{1}{j}.$$

We claim that $(A(f, \varphi, P_{j,0}, \boldsymbol{\xi}_j))_{j \in \mathbb{Z}_{>0}}$ is a Cauchy sequence for any sequence $(\boldsymbol{\xi}_{j,0})_{j \in \mathbb{Z}_{>0}}$ for which $\boldsymbol{\xi}_{j,0}$ is a selection from $P_{j,0}$, $j \in \mathbb{Z}_{>0}$. Indeed, for $\epsilon \in \mathbb{R}_{>0}$ choose $N \in \mathbb{Z}_{>0}$ such that, if $j,k \geq N$, then $\max\{\frac{1}{j},\frac{1}{k}\} < \frac{\epsilon}{2}$. Then, if P is any refinement of both $P_{j,0}$ and $P_{k,0}$ and if ξ is a selection from P, we have

$$\begin{aligned}&|A(f,\varphi,P_{j,0},\xi_{j,0}) - A(f,\varphi,P_{k,0},\xi_{k,0})|\\ &\qquad \leq |A(f,\varphi,P_{j,0},\xi_{j,0}) - A(f,\varphi,P,\xi)| + |A(f,\varphi,P_{k,0},\xi_{k,0}) - A(f,\varphi,P,\xi)| \leq \tfrac{1}{j} + \tfrac{1}{k} < \epsilon.\end{aligned}$$

Thus $(A(f,\varphi,P_{j,0},\xi_j))_{j\in\mathbb{Z}_{>0}}$ is indeed a Cauchy sequence, and so is convergent. Denote its limit by $I(f,\varphi)$. We claim that $I(f,\varphi)$ is the generalised Riemann–Stieltjes integral of f with respect to φ. To see this, let $\epsilon \in \mathbb{R}_{>0}$ and let $j \in \mathbb{Z}_{>0}$ be such that $P_{j,0}$,

1. if P and P' are refinements of $P_{j,0}$ and if ξ and ξ' are selections from P and P', respectively, then
$$|A(f,\varphi,P,\xi) - A(f,\varphi,P',\xi')| < \tfrac{\epsilon}{2},$$
and
2. $|A(f,\varphi,P_{j,0},\xi_{j,0}) - I(f,\varphi)| < \frac{\epsilon}{2}$.

Now let P be a refinement of P_0 and let ξ be a selection from P. Then we have

$$\begin{aligned}&|A(f,\varphi,P,\xi) - I(f,\varphi)|\\ &\qquad \leq |A(f,\varphi,P,\xi) - A(f,\varphi,P_{j,0},\xi_{j,0})| + |A(f,\varphi,P_{j,0},\xi_{j,0}) - I(f,\varphi)| < \epsilon.\end{aligned}$$

This shows that $I(f,\varphi)$ is indeed the generalised Riemann–Stieltjes integral of f with respect to φ, and so gives this part of the result. ■

The matter of ascertaining the most general properties of f and φ such that f is Riemann–Stieltjes integrable with respect to φ is difficult. Let us focus on a specific and interesting version of this problem in order to get at something useful. The question we ask is, "What conditions must be satisfied by φ in order to ensure that, for every continuous function f, f is Riemann–Stieltjes integrable with respect to φ?"

3.5.11 Theorem (Riemann–Stieltjes integrability of continuous functions implies an integrator of bounded variation) *Let* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$ *be a compact interval and let* $\varphi\colon [\mathrm{a},\mathrm{b}] \to \mathbb{R}$ *be a bounded function. If every continuous function* $\mathrm{f}\colon [\mathrm{a},\mathrm{b}] \to \mathbb{R}$ *is Riemann–Stieltjes integrable with respect to* φ*, then* φ *has bounded variation.*

Proof For a bounded function φ that does not have bounded variation, we will construct a continuous function f that is not Riemann–Stieltjes integrable with respect to φ. We will use a series of technical lemmata.

1 Lemma *If* $\varphi\colon [\mathrm{a},\mathrm{b}] \to \mathbb{R}$ *is a bounded function not of bounded variation, then there exists* $\mathrm{c} \in [\mathrm{a},\mathrm{b}]$ *such that at least one of the following two statements holds:*

(i) if $\mathrm{J} \subseteq [\mathrm{a},\mathrm{b}]$ *is a closed interval with* c *as its left endpoint, then* $\varphi|\mathrm{J}$ *is not of bounded variation;*

(ii) if $\mathrm{J} \subseteq [\mathrm{a},\mathrm{b}]$ *is a closed interval with* c *as its right endpoint, then* $\varphi|\mathrm{J}$ *is not of bounded variation.*

Proof First we show that there exists $c \in [a, b]$ such that, if $J \subseteq [a, b]$ is an interval containing c, then $\varphi|J$ is not of bounded variation. Suppose that no such point exists. Then, for each $x \in [a, b]$ there exists an interval J_x containing x for which $\varphi|J_x$ has bounded variation. Then, by the Heine–Borel Theorem there exists a finite set, $J_{x_1}, \dots, J_{x_k}$, of these intervals such that $[a, b] = \cup_{j=1}^{k} J_{x_j}$. Now, by Proposition 3.3.13 it follows that φ has bounded variation.

Now we show that c can be assumed to be a right or left endpoint of the intervals on which φ has unbounded variation. Suppose that there is pair of closed intervals J_1 and J_2 with c a right endpoint of J_1 and a left endpoint of J_2 such that $\varphi|J_1$ and $\varphi|J_2$ both have bounded variation. Then $\varphi|J_1 \cup J_2$ has bounded variation by Proposition 3.3.13, which is a contradiction of the property for c we proved above. ▼

2 Lemma *If* $\varphi\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is a bounded function not of bounded variation then there exists a sequence* $(\mathrm{c_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *of points in* $[\mathrm{a}, \mathrm{b}]$ *with the following properties:*

(i) the sequence $(\mathrm{c_j})_{\mathrm{j} \in \mathbb{Z}_{>0}}$ *is either strictly monotonically increasing or strictly monotonically decreasing;*

(ii) the series $\sum_{\mathrm{j}=1}^{\infty} |\varphi(\mathrm{c_{j+1}}) - \varphi(\mathrm{c_j})|$ *diverges.*

Proof Let $c \in [a, b]$ be a point as in Lemma 1, and for concreteness suppose that every interval J having c as a right endpoint has the property that $\varphi|J$ is of unbounded variation. A similar argument can be fashioned for the case when every interval J having c as a left endpoint has the property that $\varphi|J$ has unbounded variation. Now let $(x_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence in $[a, b]$ that is strictly monotonically increasing and which converges to c. There are two possibilities.

1. The set
$$A = \{ j \in \mathbb{Z}_{>0} \mid \varphi|[x_j, x_{j+1}] \text{ has unbounded variation} \}$$
is infinite: Let $M \in \mathbb{R}_{>0}$. For $j \in A$ let $x_{j,0}, x_{j,1}, \dots, x_{j,k_j}$ be such that
$$x_j = x_{j,0} < x_{j,1} < \dots < x_{j,k_j} = x_{j+1}$$
and for which $\sum_{l=1}^{k_j} |\varphi(x_j, l) - \varphi(x_{j,l-1})| > M$. One now defines the sequence $(c_j)_{j \in \mathbb{Z}_{>0}}$ by asking that
$$(c_j)_{j \in \mathbb{Z}_{>0}} = \cup_{j \in A} \cup_{l=1}^{k_j} (x_{j,l}).$$

2. The set
$$\{ j \in \mathbb{Z}_{>0} \mid \varphi|[x_j, x_{j+1}] \text{ has unbounded variation} \}$$
is finite: Let $N \in \mathbb{Z}_{>0}$ have the property that $\varphi|[x_j, x_{j+1}]$ has bounded variation for all $j \geq N$. Define $y_j = x_{N+j-1}$, $j \in \mathbb{Z}_{>0}$, so that $\varphi|[y_j, y_{j+1}]$ has bounded variation for $j \in \mathbb{Z}_{>0}$. If $v_j = \mathrm{TV}(\varphi|[x_j, x_{j+1}])$, we claim that the series $\sum_{j=1}^{\infty} v_j$ is divergent. Suppose it converges to s. Let $M \in \mathbb{R}_{>0}$ satisfy $|\varphi(x)| < M$ for $x \in [a, b]$ and take points $p_0, p_1, \dots, p_k \in [a, b]$ satisfying $a = p_0 < p_1 < \dots < p_k = c$ and such that $\sum_{j=1}^{k} |\varphi(p_j) - \varphi(p_{j-1})| > s + 2M$. This is possible since $\varphi|[a, c]$ has unbounded variation. Now denote by $q_1, \dots, q_m$ the set
$$\{q_0, q_1, \dots, q_m\} = \{p_0, p_1, \dots, p_k\} \cup \left\{ y \in \{y_j \mid j \in \mathbb{Z}_{>0}\} \;\middle|\; y < p_{k-1} \right\},$$

and ordered such that $a = q_0 < q_1 < \cdots < q_m = c$. We then have

$$\sum_{j=1}^{m} |\varphi(q_j) - \varphi(q_{j-1})| > s + 2M.$$

But we also have

$$\sum_{j=1}^{m-1} |\varphi(q_j) - \varphi(q_{j-1})| < \sum_{j=1}^{\infty} v_j = s,$$

and

$$|\varphi(q_m) - \varphi(q_{m-1})| \leq |\varphi(q_m)| + |\varphi(q_{m-1})| < 2M.$$

Combining these last two expressions gives

$$\sum_{j=1} |\varphi(p_j) - \varphi(p_{j-1})| \leq \sum_{j=1}^{m-1} |\varphi(q_j) - \varphi(q_{j-1})| \leq s + 2M,$$

which is a contradiction. Thus $\sum_{j=1}^{\infty} v_j$ indeed diverges. Now let $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of positive numbers such that $\sum_{j=1}^{\infty} \epsilon_j$ converges and such that $v_j - \epsilon_j \geq 0$. Thus the series $\sum_{j=1}^{\infty}(v_j - \epsilon_j)$ diverges. For each $j \in \mathbb{Z}_{>0}$ take points $y_{j,0}, y_{j,1}, \ldots, y_{j,k_j}$ such that $y_j = y_{j,0} < y_{j,1} < \cdots < y_{j,k_j} = y_{j+1}$ and such that

$$\sum_{l=1}^{k_j} |\varphi(y_{j,l}) - \varphi(y_{j,l-1})| \geq v_j - \epsilon_j.$$

Now define $(c_j)_{j\in\mathbb{Z}_{>0}}$ such that

$$\{c_j \mid j \in \mathbb{Z}_{>0}\} = \cup_{j=1}^{\infty} \cup_{l=1}^{k_j} \{y_{j,l}\}$$

to give the proof of the lemma. ▼

3 Lemma *If* $(\mathrm{a}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is a sequence in* $\mathbb{R}_{>0}$ *such that the series* $\sum_{\mathrm{j}=1}^{\infty} \mathrm{a}_\mathrm{j}$ *diverges, then there exists a sequence* (ϵ_j) *in* $\mathbb{R}_{>0}$*, converging to zero, such that the series* $\sum_{\mathrm{j}=1}^{\infty} \mathrm{a}_\mathrm{j}\epsilon_\mathrm{j}$ *diverges.*

Proof Let $(A_k)_{k\in\mathbb{Z}_{>0}}$ denote the sequence of partial sums for $\sum_{j=1}^{\infty} a_j$, and note that the sequence $(A_k)_{k\in\mathbb{Z}_{>0}}$ is strictly monotonically increasing and divergent. For $k \in \mathbb{Z}_{>0}$ define $\epsilon_k = \frac{1}{A_k}$. Clearly $\epsilon_k \geq 0$ for each $k \in \mathbb{Z}_{>0}$ and $\lim_{k\to\infty} \epsilon_k = 0$. Now compute

$$\epsilon_1 a_1 = 1, \quad \epsilon_j a_j = \frac{A_j - A_{j-1}}{A_j}, \qquad j \geq 2.$$

We claim that the series

$$\sum_{j=2}^{\infty} \frac{A_j - A_{j-1}}{A_j}$$

diverges. To see this, let $k, l \in \mathbb{Z}_{>0}$ with $2 \geq k < l$ and compute

$$\sum_{j=k}^{l} \frac{A_j - A_{j-1}}{A_j} > \sum_{j=k}^{l} \frac{A_j - A_{j-1}}{A_l}$$

since $(A_j)_{j \in \mathbb{Z}_{>0}}$ is strictly monotonically increasing. But

$$\sum_{j=k}^{l} \frac{A_j - A_{j-1}}{A_l} = \frac{A_l - A_{k-1}}{A_l}.$$

Since the sequence $(A_j)_{j \in \mathbb{Z}_{>0}}$ diverges to ∞, for $k \geq 2$ we can choose l_k sufficiently large that $A_l \geq 2A_{k-1}$. Therefore

$$\sum_{j=k}^{l} \frac{A_j - A_{j-1}}{A_j} \geq 1 - \frac{1}{2} = \frac{1}{2}.$$

Now let $M \in \mathbb{R}_{>0}$ and let $r \in \mathbb{Z}_{>0}$ satisfy $\frac{1}{2}r > M$. Then define $k_0 = 2$, and define k_1 to be sufficiently large that

$$\sum_{j=k_0}^{k_1} \frac{A_j - A_{j-1}}{A_j} \geq \frac{1}{2}.$$

Then define k_2 sufficiently large that

$$\sum_{j=k_1+1}^{k_2} \frac{A_j - A_{j-1}}{A_j} \geq \frac{1}{2}.$$

Repeat this to define $r + 1$ positive integers $k_0, k_1, \dots, k_r$. Then we have

$$\sum_{j=2}^{k_r} \frac{A_j - A_{j-1}}{A_j} \geq \frac{1}{2}r > M.$$

Since $M \in \mathbb{R}_{>0}$ is arbitrary, we have shown that $\sum_{j=2}^{\infty} \frac{A_j - A_{j-1}}{A_j}$ diverges, as desired. ▼

Now let $c \in [a, b]$ be as in Lemma 1, and suppose that for any interval $J \subseteq [a, b]$ with c as a right endpoint, $\varphi|J$ is of unbounded variation. The case where every interval J possessing c as a left endpoint has the property that $\varphi|J$ is of unbounded variation is treated similarly. Now let $(c_j)_{j \in \mathbb{Z}_{>0}}$ be an increasing sequence of points converging to c as in Lemma 2. Let $(\epsilon_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence of positive numbers, converging to zero, for which the series

$$\sum_{j=1}^{\infty} |\varphi(c_{j+1}) - \varphi(c_j)| \epsilon_j$$

diverges, as in Lemma 3. Define $f \colon [a, b] \to \mathbb{R}$ according to the following:

1. $f(c_j) = \operatorname{sign}(\varphi(c_{j+1}) - \varphi(c_j))\epsilon_j$, $j \in \mathbb{Z}_{>0}$;

2. for $x \in (c_j, c_{j+1})$, $j \in \mathbb{Z}_{>0}$,

$$f(x) = \frac{c_{j+1}f(c_j) - c_j f(c_{j+1}) + (f(c_{j+1}) - f(c_j))x}{c_{j+1} - c_j};$$

3. $f(x) = f(c_1)$ for $x \leq c_1$;
4. $f(x) = 0$ for $x \geq c$.

Since f is linear in the intervals between the points c_{j+1} and c_j for each $j \in \mathbb{Z}_{>0}$, and since $\lim_{j\to\infty} f(c_j) = 0$, it follows that f is continuous.

Now let $\delta \in \mathbb{R}_{>0}$ have the property that, if P is a partition for which $|P| < \delta$, then $\mathrm{EP}(P)$ must contain a point to the left of c. Then, if P is a partition for which $|P| < \delta$, take x_k to be the greatest element of $\mathrm{EP}(P)$ which lies to the left of c and let c_m be the least element of $(c_j)_{j\in\mathbb{Z}_{>0}}$ lying to the right of x_k. Now for $r \in \mathbb{Z}_{>0}$ define a partition P_r by adding to the endpoints of P the points $c_m, c_{m+1}, \dots, c_{m+r}$. Note that $|P_r| < \delta$. For $j \in \{m, m+1, \dots, m+r\}$ define $\eta_j = c_{m+j}$ and compute

$$\sum_{j=1}^{r} f(\eta_j)(\varphi(c_{m+j}) - \varphi(c_{m+j-1})) = \sum_{j=1}^{r} |\varphi(c_{m+j}) - \varphi(c_{m+j-1})|\epsilon_{m+j-1}.$$

By our choice of the sequence $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ it follows that

$$\lim_{r\to\infty} \sum_{j=1}^{r} f(\eta_j)(\varphi(c_{m+j}) - \varphi(c_{m+j-1})) = \infty.$$

Now denote the endpoints of P_r by $(x_{r,0}, x_{r,1}, \dots, x_{r,k_r})$ and let $(\xi_{r,1}, \dots, \xi_{r,k_r})$ be a selection of P_r such that $\xi_{r,l} \in (\eta_1, \dots, \eta_r)$ whenever this is possible. It then follows that

$$\lim_{r\to\infty} \sum_{j=1}^{k_r} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) = \infty,$$

showing that f is not Riemann–Stieltjes integrable with respect to φ. ■

Thus we see that there is further motivation, beyond that coming from Proposition 3.5.7, for assuming that the integrator φ has bounded variation. Let us next consider some characterisations of the generalised Riemann–Stieltjes integral that mirror those for the Riemann integral given in Theorem 3.4.9.

3.5.12 Theorem (The generalised Riemann–Stieltjes integral for integrators of bounded variation) *Let $I = [a,b]$ be a compact interval and let $f, \varphi\colon [a,b] \to \mathbb{R}$ be functions with f bounded and φ of bounded variation. Then the following three statements are equivalent:*

(i) f is generalised Riemann–Stieltjes integrable with respect to φ;

(ii) $I_-(f,\varphi) = I_+(f,\varphi)$;

(iii) for any $\epsilon \in \mathbb{R}_{>0}$ there exists a partition P of I such that $A_+(f,\varphi,P) - A_-(f,\varphi,P) < \epsilon$.

Moreover, if either of the above conditions are satisfied, then

$$\mathrm{G}\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\, \mathrm{d}\varphi(\mathrm{x}) = \mathrm{I}_-(\mathrm{f},\varphi) = \mathrm{I}_+(\mathrm{f},\varphi).$$

Proof We begin the proof by assuming that φ is monotonically increasing. When φ is monotonically increasing, one can prove the following lemmata, analogous to those in Theorem 3.4.9.

1 Lemma *Let* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$*, let* $\mathrm{f},\varphi\colon \mathrm{I} \to \mathbb{R}$ *be functions with* f *bounded and* φ *monotonically increasing, and let* P_1 *and* P_2 *be partitions of* I *with* P_2 *a refinement of* P_1*. Then*

$$\mathrm{A}_-(\mathrm{f},\varphi,\mathrm{P}_2) \geq \mathrm{A}_-(\mathrm{f},\varphi,\mathrm{P}_1), \quad \mathrm{A}_+(\mathrm{f},\varphi,\mathrm{P}_2) \leq \mathrm{A}_+(\mathrm{f},\varphi,\mathrm{P}_1).$$

Proof Let $x_1, x_2 \in \mathrm{EP}(P_1)$ and denote by $y_1,\dots,y_l$ the elements of $\mathrm{EP}(P_2)$ that satisfy

$$x_1 \leq y_1 < \cdots < y_l \leq x_2.$$

Then

$$\begin{aligned}\sum_{j=1}^{l}(\varphi(y_j) - \varphi(y_{j-1}))\inf\{f(y) \mid y \in [y_j, y_{j-1}]\}\\ &\geq \sum_{j=1}^{l}(\varphi(y_j) - \varphi(y_{j-1}))\inf\{f(x) \mid x \in [x_1,x_2]\}\\ &= (\varphi(x_2) - \varphi(x_1))\inf\{f(x) \mid x \in [x_1,x_2]\}.\end{aligned}$$

Now summing over all consecutive pairs of endpoints for P_1 gives $A_-(f,\varphi,P_2) \geq A_-(f,\varphi,P_1)$. A similar argument gives $A_+(f,\varphi,P_2) \leq A_+(f,\varphi,P_1)$. ▼

2 Lemma *Let* $\mathrm{I} = [\mathrm{a},\mathrm{b}]$*, let* $\mathrm{f},\varphi\colon \mathrm{I} \to \mathbb{R}$ *be functions with* f *bounded and* φ *monotonically increasing. Then* $\mathrm{I}_-(\mathrm{f},\varphi) \leq \mathrm{I}_+(\mathrm{f},\varphi)$.

Proof Consider two partitions P_1 and P_2 and let P be a partition such that $\mathrm{EP}(P) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$. Then, by Lemma 1,

$$A_-(f,\varphi,P) \geq A_-(f,\varphi,P_1), \quad A_+(f,\varphi,P) \leq A_+(f,\varphi(P_2).$$

Since obviously $A_-(f,\varphi,P) \leq A_+(f,\varphi,P)$ this gives $A_-(f,\varphi,P_1) \leq A_+(f,\varphi,P_2)$. From this it follows that

$$\sup\{A_-(f,P) \mid P \in \mathrm{Part}(I)\} \leq \inf\{A_+(f,P) \mid P \in \mathrm{Part}(I)\},$$

which is the result. ▼

(i) $\Longrightarrow$ (iii) Let $\epsilon \in \mathbb{R}_{>0}$ and let $P = (I_1,\dots,I_k)$ be a partition for which

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - I(f,\varphi)\right| < \frac{\epsilon}{4}$$

for every choice of $\xi_j \in \mathrm{cl}(I_j)$, $j \in \{1,\dots,k\}$. Partition the set $\{1,\dots,k\}$ into sets K_1 and K_2 such that $j \in K_1$ if and only if $\varphi(x_{j-1}) \neq \varphi(x_j)$. Now choose $\xi_j \in \mathrm{cl}(I_j)$, $j \in \{1,\dots,k\}$, such that for $j \in K_1$ we have

$$|f(\xi_j) - \sup\{f(x) \mid x \in \mathrm{cl}(I_j)\}| < \frac{\epsilon}{4\,\mathrm{card}(K_1)(\varphi(x_j) - \varphi(x_{j-1}))},$$

and choose ξ_j arbitrarily for $j \in K_2$. Then

$$\begin{aligned}|A_+(f,\varphi,P) - I(f,\varphi)| &\leq \left|A_+(f,\varphi,P) - \sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1}))\right| \\ &\quad + \left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - I(f,\varphi)\right| \\ &< \sum_{j\in K_1} \frac{\epsilon}{4\,\mathrm{card}(K_1)(\varphi(x_j) - \varphi(x_{j-1}))}(\varphi(x_j) - \varphi(x_{j-1})) + \frac{\epsilon}{4} < \frac{\epsilon}{2}.\end{aligned}$$

In like manner one shows that $|A_-(f,\varphi,P) - I(f,\varphi)| < \frac{\epsilon}{2}$. Therefore,

$$|A_+(f,\varphi,P) - A_-(f,\varphi,P)| \leq |A_+(f,\varphi,P) - I(f,\varphi)| + |I(f,\varphi) - A_-(f,\varphi,P)| < \epsilon,$$

as desired.

(ii) $\implies$ (i) Suppose that $I_-(f,\varphi) = I_+(f,\varphi)$, and let $I_g(f,\varphi)$ denote this quantity. Let $\epsilon \in \mathbb{R}_{>0}$. Let P_- and P_+ be partitions of I such that

$$I_g(f,\varphi) - A_-(f,\varphi,P_-) < \epsilon, \quad A_+(f,\varphi,P_+) - I_g(f) < \epsilon.$$

If P' is a partition for which $\mathrm{EP}(P') = \mathrm{EP}(P_-) \cup \mathrm{EP}(P_+)$, then

$$I_g(f,\varphi) - A_-(f,\varphi,P) < \epsilon, \quad A_+(f,\varphi,P) - I_g(f,\varphi) < \epsilon$$

for any refinement P of P' by Lemma 1. Denote such a refinement by $P = (I_1,\dots,I_k)$ and let $\xi_j \in \mathrm{cl}(I_j)$ for $j \in \{1,\dots,k\}$. Then

$$\begin{aligned}\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) &\leq \sum_{j=1}^{k} \sup\{f(x) \mid x \in \mathrm{cl}(I_j)\}(\varphi(x_j) - \varphi(x_{j-1})) \\ &\leq A_+(f,\varphi,P) < \epsilon + I_g(f,\varphi).\end{aligned}$$

In like manner

$$\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) > I_g(f,\varphi) - \epsilon.$$

This gives

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - I_g(f,\varphi)\right| < \epsilon,$$

as desired. Since this holds for any refinement P of P' and for any selection ξ of P, it follows that f is generalised Riemann–Stieltjes integrable with respect to φ.

(ii) $\Longrightarrow$ (iii) Suppose that $I_-(f,\varphi) = I_+(f,\varphi)$ and let $\epsilon \in \mathbb{R}_{>0}$. Then there exists partitions P_- and P_+ such that

$$A_-(f,\varphi,P_-) > I_-(f,\varphi) - \tfrac{\epsilon}{2}, \quad A_+(f,\varphi,P_+) < I_+(f,\varphi) + \tfrac{\epsilon}{2}.$$

Now let P be a partition that is a refinement of both P_1 and P_2 (obtained, for example, by asking that $\mathrm{EP}(P) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$). By Lemma 1 it follows that

$$A_+(f,\varphi,P) - A_-(f,\varphi,P) \le A_+(f,\varphi,P_+) - A_-(f,\varphi,P_-) < I_+(f,\varphi) + \tfrac{\epsilon}{2} - I_-(f,\varphi) + \tfrac{\epsilon}{2} = \epsilon.$$

(iii) $\Longrightarrow$ (ii) Now suppose that $\epsilon \in \mathbb{R}_{>0}$ and let P be a partition such that $A_+(f,\varphi,P) - A_-(f,\varphi,P) < \epsilon$. Since we additionally have $I_-(f,\varphi) \le I_+(f,\varphi)$ by Lemma 2, it follows that

$$A_-(f,\varphi,P) \le I_-(f,\varphi) \le I_+(f,\varphi) \le A_+(f,\varphi,P),$$

from which we deduce that

$$0 \le I_+(f,\varphi) - I_-(f,\varphi) < \epsilon.$$

Since ϵ is arbitrary, we conclude that $I_-(f,\varphi) = I_+(f,\varphi)$, as desired.

The above arguments prove the theorem when φ is monotonically increasing. If φ is not monotonically increasing, then, by part (ii) of Theorem 3.3.3, we can write $\varphi = \varphi_+ - \varphi_-$ where both φ_+ and φ_- are monotonically increasing. By Proposition 3.5.24 we have

$$\int_a^b f(x)\,\mathrm{d}\varphi(x) = \int_a^b f(x)\,\mathrm{d}\varphi_+(x) - \int_a^b f(x)\,\mathrm{d}\varphi_-(x),$$

and the two integrals on the right exist if and only if the integral on the left exists. Moreover, it is clear that

$$\begin{aligned} A(f,\varphi,P,\xi) &= A(f,\varphi_+,P,\xi) - A(f,\varphi_-,P,\xi), \\ A_-(f,\varphi,P) &= A_-(f,\varphi_+,P) - A_-(f,\varphi_-,P), \\ A_+(f,\varphi,P) &= A_+(f,\varphi_+,P) - A_+(f,\varphi_-,P) \end{aligned}$$

for every partition P of I and every selection ξ from P, and that

$$\begin{aligned} I_-(f,\varphi) &= I_-(f,\varphi_+) - I_-(f,\varphi_-), \\ I_+(f,\varphi) &= I_+(f,\varphi_+) - I_+(f,\varphi_-). \end{aligned}$$

With these equalities at hand, it is easy to complete the proof of the theorem for general functions of bounded variation, and we leave the elementary details to the reader. ■

Let us now ask, "If φ has bounded variation, which functions f are Riemann–Stieltjes integrable with respect to φ?" Let us begin our consideration of this question by first looking at a specific class of functions of bounded variation, namely saltus functions. The reader may wish to refer to Section 3.3.4 for the definition and properties of saltus functions.

3.5.13 Proposition (Riemann–Stieltjes integral with respect to a saltus function) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}, \varphi\colon \mathrm{I} \to \mathbb{R}$ *be functions with* f *bounded and* φ *a saltus function defined by the summable families* $(\mathrm{r}_\xi)_{\xi\in[\mathrm{a},\mathrm{b})}$ *and* $(\mathrm{l}_\xi)_{\xi\in[\mathrm{a},\mathrm{b}]}$*. Let*

$$\begin{aligned}\mathrm{D}_\mathrm{f} &= \{\mathrm{x} \in \mathrm{I} \mid \ \mathrm{f} \textit{ is discontinuous at } \mathrm{x}\}\\ \mathrm{D}_\varphi &= \{\mathrm{x} \in \mathrm{I} \mid \ \varphi \textit{ is discontinuous at } \mathrm{x}\}.\end{aligned}$$

Then the following statements hold:

(i) if $\mathrm{D}_\mathrm{f} \cap \mathrm{D}_\varphi = \emptyset$*, then* f *is Riemann–Stieltjes integrable with respect to* φ *and*

$$\int_\mathrm{a}^\mathrm{b} \mathrm{f}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}) = \sum_{\xi\in[\mathrm{a},\mathrm{b})} \mathrm{f}(\xi)\mathrm{r}_\xi + \sum_{\xi\in[\mathrm{a},\mathrm{b}]} \mathrm{f}(\xi)\mathrm{l}_\xi;$$

(ii) if $\mathrm{D}_\mathrm{f} \cap \mathrm{D}_\varphi \neq \emptyset$*, then* f *is not Riemann–Stieltjes integrable with respect to* φ*.*

Proof (i) Let

$$\begin{aligned}M_1 &= \sup\{|f(x)| \mid \ x \in [a,b]\},\\ M_2 &= \sup\{|r_x| \mid \ x \in [a,b]\},\\ M_3 &= \sup\{|l_x| \mid \ x \in [a,b]\},\end{aligned}$$

and take $M = \max\{M_1, M_2, M_3\}$, let $\epsilon \in \mathbb{R}_{>0}$, and choose points $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_m\} \subseteq D_\varphi$ such that

$$\sum_{\xi\in[a,b)-\boldsymbol{\eta}} |r_\xi| < \epsilon_1, \qquad \sum_{\xi\in[a,b]-\boldsymbol{\eta}} |l_\xi| < \epsilon_1,$$

where $\epsilon_1 = \frac{\epsilon}{12M}$. We will need to allow the possibility that $b \in \{\eta_1, \dots, \eta_m\}$ or not, so let us define $\boldsymbol{\eta}' = \boldsymbol{\eta} - \{b\}$. Now let $\delta \in \mathbb{R}_{>0}$ have the properties that

1. if P is a partition with $|P| < \delta$, then there is at most one point from $\boldsymbol{\eta}$ in each of the intervals comprising P, and
2. if, for any $j \in \{1, \dots, m\}$, $x \in \mathsf{B}(\delta, \eta_j) \cap [a,b]$, then $|f(x) - f(\eta_j)| < \epsilon_2$, where $\epsilon_2 = \frac{\epsilon}{8mM}$ (this uses the fact that f is continuous at points in D_φ).

Let P be a partition with $|P| < \delta$, let $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$, and let $\boldsymbol{\xi}$ be a selection of P. We compute

$$\sum_{j=1}^k f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) = \sum_{j=1}^k f(\xi_j)\left(l_{x_j} + r_{x_{j-1}} + \sum_{\xi\in(x_{j-1},x_j)} (r_\xi + l_\xi)\right).$$

Denote $\mathrm{EP}'(P) = (x_1, \dots, x_k)$ and introduce the notation

$$A_x = f(\xi_j)l_{x_j}, \quad B_x = f(\xi_j)r_{x_{j-1}}$$

where $x = x_j \in \mathrm{EP}'(P)$. Using the fact that f is continuous at points in $\boldsymbol{\eta}$ and the properties of δ, it then holds that

$$\begin{aligned}\sum_{\xi\in\mathrm{EP}'(P)\cap\boldsymbol{\eta}} |A_\xi - f(\xi)l_\xi| &\leq \sum_{j=1}^{m} \epsilon_2 M = mM\epsilon_2\\ \sum_{\xi\in\mathrm{EP}'(P)\cap\boldsymbol{\eta}'} |B_\xi - f(\xi)r_\xi| &\leq \sum_{j=1}^{m} \epsilon_2 M = mM\epsilon_2\\ \sum_{j=1}^{k}\sum_{\xi\in(x_{j-1},x_j)\cap\boldsymbol{\eta}} |(f(\xi_j) - f(\xi))(r_\xi + l_\xi)| &\leq \sum_{j=1}^{m} 2\epsilon_2 M = 2mM\epsilon_2.\end{aligned} \tag{3.16}$$

Using the properties of $\boldsymbol{\eta}$ we compute

$$\begin{aligned}\sum_{\xi\in\mathrm{EP}'(P)-\boldsymbol{\eta}} |A_\xi| &\leq M\epsilon_1\\ \sum_{\xi\in\mathrm{EP}'(P)-\boldsymbol{\eta}} |B_\xi| &\leq M\epsilon_1\\ \sum_{j=1}^{k}\sum_{\xi\in(x_{j-1},x_j)-\boldsymbol{\eta}} |f(\xi_j)(r_\xi + l_\xi)| &\leq 2M\epsilon_1.\end{aligned} \tag{3.17}$$

Then we directly compute

$$\begin{aligned}&\left|\sum_{\xi\in[a,b)} f(\xi)r_\xi + \sum_{\xi\in[a,b]} f(\xi)l_\xi - \sum_{\xi\in\boldsymbol{\eta}'} f(\xi)r_\xi - \sum_{\xi\in\boldsymbol{\eta}} f(\xi)l_\xi\right|\\ &\qquad\leq \sum_{\xi\in[a,b)-\boldsymbol{\eta}'} |f(\xi)r_\xi| + \sum_{\xi\in[a,b]-\boldsymbol{\eta}} |f(\xi)l_\xi| < 2M\epsilon_1\end{aligned}$$

and, with the aid of (3.16) and (3.17), we compute

$$\begin{aligned}&\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) - \sum_{\xi\in\boldsymbol{\eta}'} f(\xi)r_\xi - \sum_{\xi\in\boldsymbol{\eta}} f(\xi)l_\xi\right|\\ &= \left|\sum_{\xi\in\mathrm{EP}'(P)} (A_\xi + B_\xi) + \sum_{j=1}^{k}\sum_{\xi\in(x_{j-1},x_j)} f(\xi_j)(r_\xi + l_\xi) - \sum_{\xi\in\boldsymbol{\eta}'} f(\xi)r_\xi - \sum_{\xi\in\boldsymbol{\eta}} f(\xi)l_\xi\right|\\ &= \left|\sum_{\xi\in\mathrm{EP}'(P)\cap\boldsymbol{\eta}} (A_\xi - f(\xi)l_\xi) + \sum_{\xi\in\mathrm{EP}'(P)\cap\boldsymbol{\eta}'} (B_\xi - f(\xi)r_\xi) + \sum_{\xi\in\mathrm{EP}'(P)-\boldsymbol{\eta}} (A_\xi + B_\xi)\right.\\ &\qquad\left. + \sum_{j=1}^{k}\sum_{\xi\in(x_{j-1},x_j)\cap\boldsymbol{\eta}} (f(\xi_j) - f(\xi))(r_\xi + l_\xi) + \sum_{j=1}^{k}\sum_{\xi\in(x_{j-1},x_j)-\boldsymbol{\eta}} f(\xi_j)(r_\xi + l_\xi)\right|\\ &\leq 4mM\epsilon_2 + 4M\epsilon_1.\end{aligned}$$

Combining these preceding computations gives

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j)-\varphi(x_{j-1})) - \sum_{\xi\in[a,b)} f(\xi)r_\xi - \sum_{\xi\in(a,b]} f(\xi)l_\xi\right| < 4mM\epsilon_2 + 6M\epsilon_1 = \epsilon.$$

Since P is any partition with $|P| < \delta$ and since the selection $\boldsymbol{\xi}$ is arbitrary, this part of the result follows.

(ii) Suppose that there exists $\epsilon \in \mathbb{R}_{>0}$ and $x_0 \in [a,b)$ such that $|r_{x_0} - l_{x_0}| > \epsilon$ and such that $\omega_f(x_0) > \epsilon$. That is to say, suppose that f and φ are both discontinuous at x_0. Then there exists $\delta_0 \in \mathbb{R}_{>0}$ such that if $x_1, x_2 \in \mathsf{B}(\delta_0, x_0)$ with $x_1 < x_0 < x_2$, then

$$|f(x_1) - f(x_2)| > \tfrac{\epsilon}{2}, \quad |\varphi(x_1) - \varphi(x_2)| > \tfrac{\epsilon}{2},$$

this being possible by the definition of ω_f and since the limits $\varphi(x_0+)$ and $\varphi(x_0-)$ exist (why does this follow?). This means that for any $\delta \in (0, \delta_0)$ there exists a partition $P = (I_1, \dots, I_k)$ with the following properties:

1. $|P| = \delta$;
2. there exists $j_0 \in \{1, \dots, k\}$ such that $x_0 \in \operatorname{int}(I_{j_0})$;
3. there are points $\xi_{j_0}, \xi'_{j_0} \in \operatorname{cl}(I_{j_0})$ such that $|\xi_{j_0} - \xi'_{j_0}| > \frac{\epsilon}{2}$.

Now take selections $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{j_0}, \dots, \xi_k)$ and $\boldsymbol{\xi}' = (\xi'_1, \dots, \xi'_{j_0}, \dots, \xi'_k)$ from P such that $\xi'_j = \xi_j$ for $j \neq j_0$. Then compute

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j)-\varphi(x_{j-1})) - \sum_{j=1}^{k} f(\xi'_j)(\varphi(x_j)-\varphi(x_{j-1}))\right|$$
$$= |f(\xi_{j_0}) - f(\xi'_{j_0})||\varphi(x_j) - \varphi(x_{j-1})| > \tfrac{\epsilon^2}{4}.$$

Since this equality holds for P for which $|P|$ is arbitrarily small, it follows that f is not Riemann–Stieltjes integrable with respect to φ. ■

3.5.14 Remark (Generalised Riemann–Stieltjes integral with respect to a saltus function) Note that if, in the setup and notation of the preceding result, $D_f \cap D_\varphi = \emptyset$, then f is generalised Riemann–Stieltjes integrable with respect to φ. The converse, while holding for the Riemann–Stieltjes integral, does not hold for the generalised Riemann–Stieltjes integral, the functions of Example 3.5.8 being a counterexample. The characterisation of generalised Riemann–Stieltjes integrability with respect to a saltus function has to do not only with the shared discontinuities of f and φ, but also the exact value of the functions at the discontinuities. We do not go into this since it is rather beside the point of our objective. We merely point the reader to Exercise 3.5.4. •

This general result about saltus functions has the following corollary, whose importance is perhaps best understood in the context of its relationship to the idea of a δ-function introduced in Example IV-3.2.11–3. For readers who are unfamiliar with δ-functions and are not yet ready to take on Chapter IV-3, the following result will still make sense, but won't have much impact.

3.5.15 Corollary (A Riemann–Stieltjes interpretation of the δ-function) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{c} \in [\mathrm{a}, \mathrm{b}]$*. If* $\varphi_{\mathrm{c}} \colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is given by*

$$\varphi_{\mathrm{c}}(\mathrm{x}) = \begin{cases} 0, & \mathrm{x} \in [\mathrm{a}, \mathrm{c}], \\ 1, & \mathrm{x} \in (\mathrm{c}, \mathrm{b}] \end{cases}$$

and if $\mathrm{f} \colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is continuous at* c*, then*

$$\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\, \mathrm{d}\varphi_{\mathrm{c}}(\mathrm{x}) = \mathrm{f}(\mathrm{c}).$$

Proof This follows from Proposition 3.5.13 after noting that φ_c is the saltus function defined by the families $(r_\xi)_{\xi \in [a,b)}$ and $(l_\xi)_{\xi \in [a,b]}$ given by

$$r_\xi = \begin{cases} 1, & \xi = c, \\ 0, & \xi \neq c, \end{cases}, \quad l_\xi = 0, \qquad \xi \in [a, b]. \qquad \blacksquare$$

As the reader may verify in Exercise 3.5.3, various other definitions of φ_c can be used to get the same conclusion. Essentially, the exact value of φ_c at c is inconsequential in evaluating the Riemann–Stieltjes integral.

To fully understand this question of which functions are Riemann–Stieltjes integrable with respect to an integrator φ of bounded variation, one needs an additional concept.

3.5.16 Definition (φ-null set) Let $I = [a, b]$ be a compact interval and let $\varphi \colon [a, b] \to \mathbb{R}$ be a function of bounded variation. Extend φ to a function $\bar{\varphi} \colon \mathbb{R} \to \mathbb{R}$ by

$$\bar{\varphi}(x) = \begin{cases} \varphi(x), & x \in [a, b], \\ \varphi(a), & x < a, \\ \varphi(b), & x > b. \end{cases}$$

A set $A \subseteq \mathbb{R}$ is ***φ-null*** if

$$\inf \left\{ \sum_{j=1}^{\infty} \mathrm{TV}(\bar{\varphi}|[a_j, b_j]) \;\middle|\; A \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j) \right\} = 0.$$

(We allow the possibility of a finite sum in the equation above by allowing the possibility that $a_j = b_j$.)

If $A \subseteq \mathbb{R}$ and if $P \colon A \to \{\text{true}, \text{false}\}$ is a property defined on A, then the property P holds ***φ-almost everywhere***, ***φ-a.e.***, or ***for φ-almost every $\mathbf{x} \in \mathbf{A}$*** if the set $\{x \in A \mid P(x) = \text{false}\}$ is φ-null. •

3.5.17 Remarks (φ-null sets)

1. If $\varphi(x) = x$, then $\mathrm{TV}(\varphi|[a_j, b_j]) = b_j - a_j$ (cf. Example 3.3.5–2), and so set is φ-null in this case if and only if it has measure zero.
2. A countable union of φ-null sets is again φ-null. This can be proved in exactly the same way as one proves the analogous statement for sets of measure zero (see Exercise 2.5.11). •

With this notion of a φ-null set, we may state the analogue for Riemann–Stieltjes integrals of Theorem 3.4.11.

3.5.18 Theorem (Riemann–Stieltjes integrable functions are continuous φ-almost everywhere, and vice versa) *For a compact interval* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *and a function* $\varphi\colon \mathrm{I} \to \mathbb{R}$ *of bounded variation, a bounded function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, is Riemann–Stieltjes integrable with respect to* φ *if and only if the set*

$$\mathrm{D_f} = \{\mathrm{x} \in \mathrm{I} \mid \mathrm{f} \textit{ is discontinuous at } \mathrm{x}\}$$

is φ*-null.*

Proof Recall from the proof of Theorem 3.4.11 the definition of c_f and $D_{f,k}$, $k \in \mathbb{Z}_{>0}$. Suppose that $D_{f,k}$ is not φ-null for some $k \in \mathbb{Z}_{>0}$. Let us for the moment suppose that φ is continuous and of bounded variation. By part (ii) of Theorem 3.3.3 and by Proposition 3.5.24 we can also suppose that φ is monotonically increasing. There exists $\epsilon \in \mathbb{R}_{>0}$ such that, if

$$D_{f,k} \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j),$$

then

$$\sum_{j=1}^{\infty} \mathrm{TV}(\bar{\varphi}|[a_j, b_j]) \geq \epsilon.$$

Let P be a partition of I and let $(x_0, x_1, \dots, x_m) = \mathrm{EP}(P)$. Let $\{j_1, \dots, j_l\} \subseteq \{1, \dots, m\}$ be those indices for which $j_r \in \{j_1, \dots, j_l\}$ implies that $D_{f,k} \cap (x_{j_r-1}, x_{j_r}) \neq \emptyset$. Note that $\cup_{r=1}^{l} (x_{j_r-1}, x_{j_r})$ then covers all but a finite number of points of D_{f_k}. It then follows that one can enlarge the lengths of the intervals (x_{j_r-1}, x_{j_r}), $r \in \{1, \dots, l\}$, such that the resulting intervals cover $D_{f,k}$. The sum of the variations of $\bar{\varphi}$ restricted to these enlarged intervals then necessarily is at least ϵ. Since $\mathrm{V}(f)$ is continuous by Proposition 3.3.11, the intervals can be enlarged slightly enough that

$$\sum_{r=1}^{l} \mathrm{TV}(\bar{\varphi}|[x_{j_r}, x_{j_r-1}]) \geq \tfrac{\epsilon}{2}.$$

For each $r \in \{1, \dots, l\}$,

$$\sup\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\} - \inf\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\} \geq \tfrac{1}{k}$$

since $D_{f,k} \cap (x_{j_r-1}, x_{j_r}) \neq \emptyset$ and by definition of $D_{f,k}$ and c_f. Then we have

$$\begin{aligned}A_+(f,\varphi,P) - A_-(f,\varphi,P) &= \sum_{j=1}^{m} (\varphi(x_j) - \varphi(x_{j-1}))\Big(\sup\{f(x) \mid x \in [x_{j-1}, x_j]\}\\ &\qquad - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big)\\ &\geq \sum_{r=1}^{l} (\varphi(x_{j_r}) - \varphi(x_{j_r-1}))\Big(\sup\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\}\\ &\qquad - \inf\{f(x) \mid x \in [x_{j_r-1}, x_{j_r}]\}\Big)\\ &\geq \tfrac{\epsilon}{2k},\end{aligned}$$

using the fact that φ is monotonically increasing, so that

$$\mathrm{TV}(\bar{\varphi}|[x_{j_r}, x_{j_r-1}]) = \varphi(x_{j_r}) - \varphi(x_{j_r-1}).$$

By Theorem 3.5.12 this shows that f is not generalised Riemann–Stieltjes integrable, and so not Riemann–Stieltjes integrable, with respect to φ.

Now we allow that φ be discontinuous. By Proposition 3.3.22 we write $\varphi = g_\varphi + j_\varphi$ where g_φ is continuous and j_φ is a saltus function. Suppose that f is Riemann–Stieltjes integrable with respect to φ. We claim that this implies that f and φ have no discontinuities in common, i.e., that $D_f \cap D_\varphi = 0$. This can be proved in exactly the same manner as the second part of Proposition 3.5.13; indeed, the proof there really does not rely on the fact that the integrator is a saltus function. Now, since the discontinuities of φ are exactly the discontinuities of j_φ, this implies that if f is Riemann–Stieltjes integrable with respect to j_φ by Proposition 3.5.13. By Proposition 3.5.24 we know then that f is Riemann–Stieltjes integrable with respect to g_φ. By our above arguments D_f is g_φ-null. We also claim that D_f is j_φ-null. To see this, let $\epsilon \in \mathbb{R}_{>0}$ and choose $\boldsymbol{\eta} = (\eta_1, \dots, \eta_m) \subseteq D_{j_\varphi}$ such that

$$\sum_{\xi \in [a,b] \setminus \boldsymbol{\eta}} |\varphi(\xi+) - \varphi(\xi-)| < \epsilon \tag{3.18}$$

(recalling from Proposition 3.3.22 the definition of the summable families associated to the saltus function a function of bounded variation). Now note that $[a,b] \setminus \boldsymbol{\eta}$ is a finite collection of open intervals (or, more precisely, a finite collection of open intervals intersected with $[a,b]$). Each of these open intervals can be written as the countable union of a sequence of intervals of strictly increasing length. In this way we write $[a,b] \setminus \boldsymbol{\eta}$ as a countable collection of open intervals, and let us denote this family of intervals by $((a_j, b_j))_{j \in \mathbb{Z}_{>0}}$. By construction,

$$D_f \subseteq \bigcup_{j \in \mathbb{Z}_{>0}} (a_j, b_j).$$

By (3.18), and since the total variation of a saltus function on any interval is the sum of the magnitudes of the jumps, we have

$$\sum_{j=1}^{\infty} \mathrm{TV}(j_\varphi|[a_j, b_j]) < \epsilon.$$

This shows that D_f is j_φ-null. Finally, we claim that, if D_f is both g_φ-null and j_φ-null, then it is φ-null. To see this, let $\epsilon \in \mathbb{R}_{>0}$ and choose countable collections $((a_j, b_j))_{j\in A}$ and $((c_k, d_k))_{k\in B}$ of open intervals such that

$$D_f \subseteq \bigcup_{j\in A}(a_j, b_j), \quad D_f \subseteq \bigcup_{k\in B}(c_k, d_k),$$

and such that

$$\sum_{j\in A} \mathrm{TV}(g_\varphi|[a_j, b_j]) < \frac{\epsilon}{2}, \quad \sum_{k\in B} \mathrm{TV}(j_\varphi|[c_k, d_k]) < \frac{\epsilon}{2}.$$

Then the set $(\cup_{j\in A}(a_j, b_j)) \cap (\cup_{k\in B}(c_k, d_k))$ is open by Exercise 2.5.1 and so is a countable union of open intervals by Proposition 2.5.6. Thus we can write

$$\left(\bigcup_{j\in A}(a_j, b_j)\right) \cap \left(\bigcup_{k\in B}(c_k, d_k)\right) = \bigcup_{l\in C}(\alpha_l, \beta_l)$$

for a countable collection $((\alpha_l, \beta_l))_{l\in C}$ of open intervals. Moreover,

$$\sum_{l\in C} \mathrm{TV}(\varphi|[\alpha_l, \beta_l]) \leq \sum_{l\in C} \mathrm{TV}(g_\varphi|[c_k, d_k]) + \sum_{l\in C} \mathrm{TV}(j_\varphi|[c_k, d_k]) < \epsilon,$$

using the fact that the total variation of a sum of functions is bounded above by the sum of the total variations of the two functions (as was shown during the course of the proof of Proposition 3.3.12(i)) and that

$$\bigcup_{l\in C}(\alpha_l, \beta_l) \subseteq \bigcup_{j\in A}(a_j, b_j), \quad \bigcup_{l\in C}(\alpha_l, \beta_l) \subseteq \bigcup_{k\in B}(c_k, d_k).$$

This completes the proof of this part of the theorem.

Now we show the converse, and so assume that D_f is φ-null. In this part of the proof we drop the assumption that φ is monotonically increasing, since our proof does not require this. We first prove a technical lemma which gives a sufficient condition for Riemann–Stieltjes integrability. If $J \subseteq [a, b]$ is an interval, then

$$\omega_f(J) = \sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in J\},$$

where the notation is suggestive of the notion of the oscillation of a function as introduced in Definition 3.1.10.

1 **Lemma** *Let* $[\mathrm{a}, \mathrm{b}]$ *be a compact interval, let* $\mathrm{f}, \varphi\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be functions with* f *bounded and* φ *of bounded variation, and suppose that, for every* $\epsilon \in \mathbb{R}_{>0}$*, there exists* $\delta \in \mathbb{R}_{>0}$ *such that, if* $\mathrm{P} = (\mathrm{I}_1, \dots, \mathrm{I}_\mathrm{k})$ *is a partition with* $|\mathrm{P}| < \delta$*, then*

$$\sum_{\mathrm{j}=1}^{\mathrm{k}} \omega_{\mathrm{f}}(\mathrm{cl}(\mathrm{I}_\mathrm{j}))\, \mathrm{TV}(\varphi|\, \mathrm{cl}(\mathrm{I}_\mathrm{j})) < \epsilon.$$

Then f *is Riemann–Stieltjes integrable with respect to* φ*.*

Proof Let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ be such that

$$\sum_{j=1}^{k} \omega_f(\mathrm{cl}(I_j))\,\mathrm{TV}(\varphi|\,\mathrm{cl}(I_j)) < \frac{\epsilon}{2}$$

for any partition $P = (I_1, \dots, I_k)$ satisfying $|P| < \delta$. Now let P_1 and P_2 be partitions with $|P_1|, |P_2| < \delta$, and let P be a partition for which $\mathrm{EP}(P) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$. Denote $\mathrm{EP}(P_r) = (x_{r,0}, x_{r,1}, \dots, x_{r,k_r})$ for $r \in \{1,2\}$ and denote $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$. Let $\boldsymbol{\xi}_r = (\xi_{a,1}, \dots, \xi_{r,k_r})$ be selections from P_r, $r \in \{1,2\}$, and let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_k)$ be a selection from P. We now perform a computation with $r \in \{1,2\}$ fixed but arbitrary. For each $m \in \{0, 1, \dots, k_r\}$ there exists a unique $l(m) \in \{0, 1, \dots, k\}$ such that $x_{l(m)} = x_{r,m}$. We then have

$$f(\xi_{r,m})(\varphi(x_{r,m}) - \varphi(x_{r,m-1})) - \sum_{j=l(m-1)+1}^{l(m)} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) \\ = \sum_{j=l(m-1)+1}^{l(m)} (f(\xi_{r,m}) - f(\xi_j))(\varphi(x_j) - \varphi(x_{j-1}))$$

for each $m \in \{1, \dots, k_r\}$. Therefore,

$$\left| f(\xi_{r,m})(\varphi(x_{r,m}) - \varphi(x_{r,m-1})) - \sum_{j=l(m-1)+1}^{l(m)} f(\xi_j)(\varphi(x_j) - \varphi(x_{j-1})) \right| \\ \leq \sum_{j=l(m-1)+1}^{l(m)} |f(\xi_{r,m}) - f(\xi_j)||\varphi(x_j) - \varphi(x_{j-1})| \\ \leq \omega_f([x_{r,m-1}, x_{r,m}]) \sum_{j=l(m-1)+1}^{l(m)} |\varphi(x_j) - \varphi(x_{j-1})| \\ \leq \omega_f([x_{r,m-1}, x_{r,m}])\,\mathrm{TV}(\varphi|[x_{r,m-1}, x_{r,m}]).$$

Summing this last estimate over $m \in \{1, \dots, k_r\}$ we obtain

$$|A(f, \varphi, P_r, \boldsymbol{\xi}_r) - A(f, \varphi, P, \boldsymbol{\xi})| \leq \sum_{m=1}^{k_r} \omega_f([x_{r,m-1}, x_{r,m}])\,\mathrm{TV}(\varphi|[x_{r,m-1}, x_{r,m}]) < \frac{\epsilon}{2}.$$

Now we have

$$|A(f, \varphi, P_1, \boldsymbol{\xi}_1) - A(f, \varphi, P_2, \boldsymbol{\xi}_2)| \\ \leq |A(f, \varphi, P_1, \boldsymbol{\xi}_r) - A(f, \varphi, P, \boldsymbol{\xi})| + |A(f, \varphi, P_2, \boldsymbol{\xi}_r) - A(f, \varphi, P, \boldsymbol{\xi})| < \epsilon.$$

Now let $(P_j)_{j \in \mathbb{Z}_{>0}}$ be a sequence of partitions for which $\lim_{j\to\infty} |P_j| = 0$. By our above computations, for each $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that,

$$|A(f, \varphi, P_j, \boldsymbol{\xi}_j) - A(f, \varphi, P_k, \boldsymbol{\xi}_k)| < \epsilon$$

for $j, k \geq N$ and for any selections ξ_l of P_l, $l \in \mathbb{Z}_{>0}$. It this follows that the sequence $(A(f, \varphi, P_j, \xi_j))_{j \in \mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ for any selections ξ_j of P_j, $j \in \mathbb{Z}_{>0}$. Denote the resulting limit of this sequence by $I(f, \varphi)$. We claim that $I(f, \varphi)$ is the Riemann–Stieltjes integral of f with respect to φ. To see this, let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ be such that

$$|A(f, \varphi, P, \xi) - A(f, \varphi, P', \xi')| < \tfrac{\epsilon}{2}$$

for any two partitions P and P' satisfying $|P|, |P'| < \delta$ and for any selections ξ and ξ' from P and P', respectively. Now let $N \in \mathbb{Z}_{>0}$ satisfy $|P_j| < \delta$ for every $j \geq N$. Then, if P is any partition with $|P| < \delta$ and if ξ is any selection from P, we have

$$|A(f, \varphi, P, \xi) - I(f, \varphi)| \leq |A(f, \varphi, P, \xi) - A(f, \varphi, P_N, \xi_N)| + |A(f, \varphi, P_N, \xi_N) - I(f, \varphi)| < \epsilon,$$

for any selection ξ_N of P_N. This shows that $I(f, \varphi)$ is indeed the Riemann–Stieltjes integral of f with respect to φ, and so gives the lemma. ▼

Now, equipped with this lemma, the theorem will be proved if we can show that the assumption that D_f is φ-null implies that, for any $\epsilon \in \mathbb{R}_{>0}$, there exists $\delta \in \mathbb{R}_{>0}$ such that, for any partition $P = (I_1, \dots, I_k)$ with $|P| < \delta$, we have

$$\sum_{j=1}^{k} \omega_f(\mathrm{cl}(I_j))\,\mathrm{TV}(\varphi|\,\mathrm{cl}(I_j)) < \epsilon.$$

For $\epsilon \in \mathbb{R}_{>0}$ let $\epsilon_1 = \frac{\epsilon}{2\,\mathrm{TV}(\varphi)}$ and define

$$D_{f,\epsilon_1} = \{x \in [a, b] \mid \omega_f(x) \geq \epsilon_1\},$$

and recall from Proposition 3.1.13 that D_{f,ϵ_1} is closed. Since D_f is φ-null, D_{f,ϵ_1} is also φ-null. Thus there exists a countable family $((a_\alpha, b_\alpha))_{\alpha \in A}$ of open intervals such that

$$D_{f,\epsilon_1} \subseteq \bigcup_{\alpha \in A} (a_\alpha, b_\alpha)$$

and such that

$$\sum_{\alpha \in A} \mathrm{TV}(\varphi|(a_\alpha, b_\alpha)) < \epsilon_2,$$

where $\epsilon_2 = \frac{\epsilon}{4M}$. By the Heine–Borel Theorem, we may furthermore assume that the index set A is finite, by virtue of D_{f,ϵ_1} being closed and bounded. The set $[a, b] \setminus D_{f,\epsilon_1}$ is open in $[a, b]$. Then for each $x \in [a, b] \setminus D_{f,\epsilon_1}$ there exists a neighbourhood U_x of x such that

$$\sup\{|f(x_1) - f(x_2)| \mid x_1, x_2 \in U_x \cap [a, b]\} < \epsilon_1.$$

Now we note that $((a_\alpha, b_\alpha))_{\alpha \in A} \cup (U_x)_{x \in [a,b] \setminus D_{f,\epsilon_1}}$ is an open cover of $[a, b]$. Thus, by Theorem 2.5.30 there exists $\delta \in \mathbb{R}_{>0}$ such that each set $\mathsf{B}(\delta, x) \cap [a, b]$ is contained in at least one of the elements of the open cover. Now let P be a partition for which $|P| < \delta$, and denote $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$. Partition the set $\{1, \dots, k\}$ as $K_1 \cup K_2$ such that if

$j \in K_1$ then $[x_{j-1}, x_j] \subseteq (a_\alpha, b_\alpha)$ for some $\alpha \in A$ and such that if $j \in K_2$ then $[x_{j-1}, x_j] \subseteq U_x$ for some $x \in [a, b]$. Let $M = \sup\{|f(x)| \mid x \in [a, b]\}$. Then we have

$$\begin{aligned}\sum_{j=1}^{m} \omega_f([x_{j-1}, x_j])\,\mathrm{TV}(\varphi|[x_{j-1}, x_j]) &= \sum_{j\in K_1} \omega_f([x_{j-1}, x_j])\,\mathrm{TV}(\varphi|[x_{j-1}, x_j]) \\ &\quad + \sum_{j\in K_2} \omega_f([x_{j-1}, x_j])\,\mathrm{TV}(\varphi|[x_{j-1}, x_j]) \\ &\le 2M\epsilon_2 + \mathrm{TV}(\varphi)\epsilon_1 < \epsilon.\end{aligned}$$

The result now follows from Lemma 1, as desired. ■

The result has the following immediate corollaries.

3.5.19 Corollary (Characterisation of generalised Riemann–Stieltjes integrable functions) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval, let* $\mathrm{f}, \varphi\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be functions with* f *bounded and* φ *of bounded variation, and define*

$$\mathrm{D}_\mathrm{f} = \{\mathrm{x} \in \mathrm{I} \mid \mathrm{f} \textit{ is discontinuous at } \mathrm{x}\}.$$

Then the following statements hold:

(i) if D_f *is* φ*-null, then* f *is generalised Riemann–Stieltjes integrable;*

(ii) if φ *is continuous and if* f *is generalised Riemann–Stieltjes integrable, then* D_f *is* φ*-null.*

Proof The first assertion follows directly from Theorem 3.4.11, while the second was proved during the course of the proof of Theorem 3.4.11. ■

3.5.20 Corollary (Riemann–Stieltjes integrability for continuous integrators of bounded variation) *Let* $[\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\varphi\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be a continuous function of bounded variation. Then a bounded function* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is Riemann–Stieltjes integrable with respect to* φ *if and only if it is generalised Riemann–Stieltjes integrable with respect to* φ*.*

The next result is the most commonly encountered integrability condition of any generality.

3.5.21 Corollary (Continuous integrands are Riemann–Stieltjes integrable with respect to integrators of bounded variation) *Let* $[\mathrm{a}, \mathrm{b}]$ *be a compact interval. If* $\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is continuous and* $\varphi\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is of bounded variation, then* f *is Riemann–Stieltjes integrable with respect to* φ*.*

3.5.3 The Riemann–Stieltjes integral on noncompact intervals

Next we do as we did with the Riemann integral, and give the extension of the Riemann–Stieltjes integral to general intervals. The reader will recall from Section 3.4.4 that for the general Riemann integral we had two notions, those of

Riemann integrability and conditional Riemann integrability. We do not duplicate this for the Riemann–Stieltjes integral, instead giving what is analogous to the conditional Riemann integral.

3.5.22 Definition (Riemann–Stieltjes integrable functions of a general interval) Let $I \subseteq \mathbb{R}$ be an interval and let $f, \varphi \colon I \to \mathbb{R}$ be functions such that $f|J$ is Riemann–Stieltjes integrable with respect to $\varphi|J$ for every compact subinterval J of I.

(i) If $I = [a, b]$ then the Riemann–Stieltjes integral of f with respect to φ is as defined in the preceding section.

(ii) If $I = (a, b]$ then define

$$\int_a^b f(x)\,\mathrm{d}\varphi(x) = \lim_{r_a \downarrow a} \int_{r_a}^b f(x)\,\mathrm{d}\varphi(x).$$

(iii) If $I = [a, b)$ then define

$$\int_a^b f(x)\,\mathrm{d}\varphi(x) = \lim_{r_b \uparrow a} \int_a^{r_b} f(x)\,\mathrm{d}\varphi(x).$$

(iv) If $I = (a, b)$ then define

$$\int_a^b f(x)\,\mathrm{d}\varphi(x) = \lim_{r_a \downarrow a} \int_{r_a}^c f(x)\,\mathrm{d}\varphi(x) + \lim_{r_b \uparrow b} \int_c^{r_b} f(x)\,\mathrm{d}\varphi(x)$$

for any $c \in (a, b)$.

(v) If $I = (-\infty, b]$ then define

$$\int_{-\infty}^b f(x)\,\mathrm{d}\varphi(x) = \lim_{R \to \infty} \int_{-R}^b f(x)\,\mathrm{d}\varphi(x).$$

(vi) If $I = (-\infty, b)$ then define

$$\int_{-\infty}^b f(x)\,\mathrm{d}\varphi(x) = \lim_{R \to \infty} \int_{-R}^c f(x)\,\mathrm{d}\varphi(x) + \lim_{r_b \uparrow b} \int_c^{r_b} f(x)\,\mathrm{d}\varphi(x).$$

(vii) If $I = [a, \infty)$ then define

$$\int_a^\infty f(x)\,\mathrm{d}\varphi(x) = \lim_{R \to \infty} \int_a^R f(x)\,\mathrm{d}\varphi(x).$$

(viii) If $I = (a, \infty)$ then define

$$\int_a^\infty f(x)\,\mathrm{d}\varphi(x) = \lim_{r_a \downarrow a} \int_{r_a}^c f(x)\,\mathrm{d}\varphi(x) + \lim_{R \to \infty} \int_c^R f(x)\,\mathrm{d}\varphi(x)$$

for some $c \in (a, \infty)$.

(ix) If $I = (-\infty, \infty)$ then define

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}\varphi(x) = \lim_{R\to\infty} \int_{-R}^{c} f(x)\,\mathrm{d}\varphi(x) + \lim_{R\to\infty} \int_{c}^{R} f(x)\,\mathrm{d}\varphi(x)$$

for some $c \in \mathbb{R}$.

If, for a given I, f, and φ, the appropriate of the above limits exists, then f is ***Riemann–Stieltjes integrable*** with respect to φ on I, and the Riemann–Stieltjes integral is the value of the limit.

By replacing $\int_a^b$ with $G\int_a^b$, we also define the notion of a f as being ***generalised Riemann–Stieltjes integrable*** with respect to φ on I, and the value of the appropriate limit is the generalised Riemann–Stieltjes integral. •

As with the Riemann integral, the above definitions are independent of the choice of the point c, when such a choice must be made (cf. Propositions 3.5.29 and 3.5.31).

3.5.23 Notation (General notation for the Riemann–Stieltjes integral) If $I \subseteq \mathbb{R}$ is an interval and if $f, \varphi \colon I \to \mathbb{R}$ are such that f is Riemann–Stieltjes integrable with respect to φ, then it is convenient to denote by

$$\int_I f(x)\,\mathrm{d}\varphi(x), \quad G\int_I f(x)\,\mathrm{d}\varphi(x)$$

the values of the Riemann–Stieltjes integral and the generalised Riemann–Stieltjes integral, respectively, of f on I as a shorthand for any one of the pieces of notation of Definition 3.5.22. •

We forgo explicit examples for the Riemann–Stieltjes integral on general intervals, since the corresponding examples for the Riemann integral also apply here, and generally serve to illustrate the desired phenomenon.

3.5.4 The Riemann–Stieltjes integral and operations on functions

In this section we present the usual formulae concerning the relationship between the Riemann–Stieltjes integral and the usual operations one performs on functions and subsets of $\mathbb{R}$. The only (possibly) big surprise here is the result giving the relationship with partitioning of intervals (Proposition 3.5.29). Unlike the case with the Riemann integral, the implication only goes one way for the Riemann–Stieltjes integral. However, for the generalised Riemann–Stieltjes integral, things are as they are for the Riemann integral (Proposition 3.5.31). We advise the reader that the hypotheses we place on the integrator vary; some of the results are general, some require the integrator to have bounded variation, and some require the integrator to be monotonically increasing. Thus some attention is required so as to not apply the results improperly. We also comment that, in all cases, any conditions placed on the integrator are necessary for the result to be true, although we do not provide counterexamples for this assertion in all cases.

Let us begin by considering the algebraic operations on functions. There are two results to consider here, one for the integrand and one for the integrator.

3.5.24 Proposition (Algebraic operations on the integrand and the Riemann–Stieltjes integral) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{f}, \mathrm{g}\colon \mathrm{I} \to \mathbb{R}$ *be functions that are (generalised) Riemann–Stieltjes integrable with respect to* $\varphi\colon \mathrm{I} \to \mathbb{R}$*, and let* $\mathrm{c} \in \mathbb{R}$*. Then the following statements hold:*

(i) $\mathrm{f}+\mathrm{g}$ *is (generalised) Riemann–Stieltjes integrable with respect to* φ *and*

$$(\mathrm{G})\int_{\mathrm{I}}(\mathrm{f}+\mathrm{g})(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}) = (\mathrm{G})\int_{\mathrm{I}}\mathrm{f}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}) + (\mathrm{G})\int_{\mathrm{I}}\mathrm{g}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x});$$

(ii) cf *is (generalised) Riemann–Stieltjes integrable with respect to* φ *and*

$$(\mathrm{G})\int_{\mathrm{I}}(\mathrm{cf})(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}) = \mathrm{c}(\mathrm{G})\int_{\mathrm{I}}\mathrm{f}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x});$$

(iii) if I *is additionally compact and if* φ *has bounded variation, then* fg *is (generalised) Riemann–Stieltjes integrable with respect to* φ*;*

(iv) if I *is additionally compact, if* φ *has bounded variation, and if there exists* $\alpha \in \mathbb{R}_{>0}$ *such that* $\mathrm{g}(\mathrm{x}) \geq \alpha$ *for each* $\mathrm{x} \in \mathrm{I}$*, then* $\frac{\mathrm{f}}{\mathrm{g}}$ *is (generalised) Riemann–Stieltjes integrable with respect to* φ*.*

Proof Just as was the case for the corresponding results for the Riemann integral in Proposition 3.4.22, we can assume, without loss of generality, that I is compact in the first two parts of the result.

(i) Abbreviate the Riemann–Stieltjes integrals for f and g by $I(f,\varphi)$ and $I(g,\varphi)$, respectively. For any partition P and any selection ξ from P, we have

$$A(f+g,\varphi,P,\xi) = A(f,\varphi,P,\xi) + A(g,\varphi,P,\xi).$$

For $\epsilon \in \mathbb{R}_{>0}$ choose $\delta \in \mathbb{R}_{>0}$ such that

$$|A(f,\varphi,P,\xi) - I(f,\varphi)| < \tfrac{\epsilon}{2}, \quad |A(g,\varphi,P,\xi) - I(g,\varphi)| < \tfrac{\epsilon}{2}$$

for any partition P satisfying $|P| < \delta$ and for any selection ξ from P. Then

$$|A(f+g,\varphi,P,\xi) - I(f,\varphi) - I(g,\varphi)| \\ \leq |A(f,\varphi,P,\xi) - I(f,\varphi)| + |A(g,\varphi,P,\xi) - I(g,\varphi)|$$

whenever P is a partition satisfying $|P| < \delta$ and for any selection ξ from P. This gives the result.

For the corresponding result for the generalised Riemann–Stieltjes integral, we note that, by Theorem 3.5.12, the generalised Riemann–Stieltjes integral is characterised by the equality of the upper and lower integrals (part (ii) of the theorem), and also by the close approximation of the integral by the lower and upper sums (part (iii) of the theorem). Our proof of part (i) of Proposition 3.4.22 relied on just these characterisations of the Riemann integral. Therefore, the proof for the Riemann integral carries over verbatim to the generalised Riemann–Stieltjes integral.

(ii) Let $I(f, \varphi)$ denote the Riemann–Stieltjes integral of f with respect to φ. Since the result is clear when $c = 0$, we suppose first that $c > 0$. For $\epsilon \in \mathbb{R}_{>0}$ let $\delta \in \mathbb{R}_{>0}$ have the property that

$$|A(f, \varphi, P, \xi) - I(f, \varphi)| < \tfrac{\epsilon}{c}$$

for any partition P satisfying $|P| < \delta$ and for any selection ξ from P. Then, noting that $A(cf, \varphi, P, \xi) = cA(f, \varphi, P, \xi)$, we have

$$|A(cf, \varphi, P, \xi) - cI(f, \varphi)| < \epsilon$$

for any partition P satisfying $|P| < \delta$ and for any selection ξ from P. This gives the result for $c > 0$, and a similar argument holds for $c < 0$.

For the generalised Riemann–Stieltjes integral, the proof goes just like that in part (ii) of Proposition 3.4.22 in the same way that part (i) above follows from part (i) of Proposition 3.4.22.

(iii) By taking $g(y) = y^2$ in Proposition 3.5.26 below, we conclude that f^2 is (generalised) Riemann–Stieltjes integrable with respect to φ whenever f is (generalised) Riemann–Stieltjes integrable with respect to φ. Since

$$fg = \tfrac{1}{2}((f+g)^2 - f^2 - g^2),$$

this part of the result then follows from part (i) above.

(iv) This follows from Proposition 3.5.26 below by taking $g(y) = \frac{1}{y}$. ■

3.5.25 Proposition (Algebraic operations on the integrator and the Riemann–Stieltjes integral) *Let $I \subseteq \mathbb{R}$ be an interval, let $f\colon I \to \mathbb{R}$ be a function that is (generalised) Riemann–Stieltjes integrable with respect to both $\varphi, \psi\colon I \to \mathbb{R}$, and let $c \in \mathbb{R}$. Then the following statements hold:*

(i) f is (generalised) Riemann–Stieltjes integrable with respect to $\varphi + \psi$ and

$$(G)\int_I f(x)\,d(\varphi + \psi)(x) = (G)\int_I f(x)\,d\varphi(x) + (G)\int_I f(x)\,d\psi(x);$$

(ii) f is (generalised) Riemann–Stieltjes integrable with respect to $c\varphi$ and

$$(G)\int_I f(x)\,d(c\varphi)(x) = c(G)\int_I f(x)\,d\varphi(x).$$

Proof The proof here follows, *mutatis mutandis*, as do the corresponding proofs in Proposition 3.5.24. Alternatively, one can use integration by parts, Theorem 3.5.32 below. ■

We now consider now the case of composition of functions.

3.5.26 Proposition (Function composition and the Riemann–Stieltjes integral) *If* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *is a compact interval, if* $\mathrm{f}, \varphi\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *are functions such that*

(i) φ *of bounded variation,*

(ii) f *(generalised) Riemann–Stieltjes integrable with respect to* φ*, and*

(iii) $\mathrm{image}(\mathrm{f}) \subseteq [\mathrm{c}, \mathrm{d}]$,

and if $\mathrm{g}\colon [\mathrm{c}, \mathrm{d}] \to \mathbb{R}$ *is continuous, then* $\mathrm{g} \circ \mathrm{f}$ *is (generalised) Riemann–Stieltjes integrable.*

Proof First we consider the case of the Riemann–Stieltjes integral. Since f is Riemann–Stieltjes integrable with respect to φ, the set D_f of discontinuities of f is φ-null by Theorem 3.5.18. By Proposition 3.1.16, the set $D_{g\circ f}$ of discontinuities of $g \circ f$ is contained in D_f, and so is also φ-null. Again by Theorem 3.5.18, $g \circ f$ is then Riemann–Stieltjes integrable.

For the generalised Riemann–Stieltjes integral, we proceed as follows. We suppose, without loss of generality by Theorem 3.3.3 and by Proposition 3.5.25, that φ is monotonically increasing. Denote $M = \sup\{|g(y)| \mid y \in [c, d]\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and write $\epsilon' = \frac{\epsilon}{2M+\varphi(b)-\varphi(a)}$. Since g is uniformly continuous by the Heine–Cantor Theorem, let $\delta \in \mathbb{R}$ be chosen such that $0 < \delta < \epsilon'$ and such that, $|y_1 - y_2| < \delta$ implies that $|g(y_1) - g(y_2)| < \epsilon'$. Then choose a partition P of $[a, b]$ such that $A_+(f, \varphi, P) - A_-(f, \varphi, P) < \delta^2$. Let $(x_0, x_1, \dots, x_k)$ be the endpoints of P and define

$$\begin{aligned}
A &= \{j \in \{1, \dots, k\} \mid \ \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} < \delta\},\\
B &= \{j \in \{1, \dots, k\} \mid \ \sup\{f(x) \mid x \in [x_{j-1}, x_j]\} - \inf\{f(x) \mid x \in [x_{j-1}, x_j]\} \geq \delta\}.
\end{aligned}$$

For $j \in A$ we have $|f(\xi_1) - f(\xi_2)| < \delta$ for every $\xi_1, \xi_2 \in [x_{j-1}, x_j]$ which implies that $|g \circ f(\xi_1) - g \circ f(\xi_2)| < \epsilon'$ for every $\xi_1, \xi_2 \in [x_{j-1}, x_j]$. For $j \in B$ we have

$$\begin{aligned}
\delta \sum_{j\in B} (\varphi(x_j) - \varphi(x_{j-1})) \leq \sum_{j\in B} \Big(&\sup\{f(x) \mid x \in [x_{j-1}, x_j]\}\\
&- \inf\{f(x) \mid x \in [x_{j-1}, x_j]\}\Big)(\varphi(x_j) - \varphi(x_{j-1}))\\
\leq{}& A_+(f, \varphi, P) - A_-(f, \varphi, P) < \delta^2.
\end{aligned}$$

Therefore we conclude that

$$\sum_{j\in B} (\varphi(x_j) - \varphi(x_{j-1})) \leq \epsilon',$$

using the fact that φ is monotonically increasing. Thus

$$\begin{aligned}
A_+(g\circ f,\varphi,P) - A_-(g\circ f,\varphi,P) &= \sum_{j=1}^{k}\Big(\sup\{g\circ f(x) \mid x\in[x_{j-1},x_j]\} \\
&\qquad - \inf\{g\circ f(x) \mid x\in[x_{j-1},x_j]\}\Big)(\varphi(x_j)-\varphi(x_{j-1})) \\
&= \sum_{j\in A}\Big(\sup\{g\circ f(x) \mid x\in[x_{j-1},x_j]\} \\
&\qquad - \inf\{g\circ f(x) \mid x\in[x_{j-1},x_j]\}\Big)(\varphi(x_j)-\varphi(x_{j-1})) \\
&\qquad + \sum_{j\in B}\Big(\sup\{g\circ f(x) \mid x\in[x_{j-1},x_j]\} \\
&\qquad - \inf\{g\circ f(x) \mid x\in[x_{j-1},x_j]\}\Big)(\varphi(x_j)-\varphi(x_{j-1})) \\
&< \epsilon'(\varphi(b)-\varphi(a)) + 2\epsilon' M < \epsilon,
\end{aligned}$$

giving the result by Theorem 3.5.12. ■

With respect to the natural total order on $\mathbb{R}$ and the absolute value function on $\mathbb{R}$, we have the following two results.

3.5.27 Proposition (Riemann–Stieltjes integral and total order on $\mathbb{R}$) *Let* $\mathrm{I}\subseteq\mathbb{R}$ *be an interval and let* $\mathrm{f},\mathrm{g},\varphi\colon\mathrm{I}\to\mathbb{R}$ *be functions for which*

(i) *φ is monotonically increasing,*

(ii) *f and g are both (generalised) Riemann–Stieltjes integrable with respect to φ, and*

(iii) $\mathrm{f}(\mathrm{x})\le\mathrm{g}(\mathrm{x})$ *for each* $\mathrm{x}\in\mathrm{I}$.

Then

$$(\mathrm{G})\int_{\mathrm{I}}\mathrm{f}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}) \le (\mathrm{G})\int_{\mathrm{I}}\mathrm{g}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}).$$

Proof By part (i) of Proposition 3.5.24, it suffices to consider the case when $f = 0$. First take the case where $I = [a,b]$. If P is a partition of $[a,b]$ with $\mathrm{EP}(P) = (x_0,x_1,\dots,x_k)$ and if ξ is a selection from P then

$$\sum_{j=1}^{k} g(\xi_j)(\varphi(x_j)-\varphi(x_{j-1})) \ge 0.$$

This allows us to conclude that both

$$\int_a^b g(x)\,\mathrm{d}\varphi(x), \quad G\int_a^b g(x)\,\mathrm{d}\varphi(x)$$

are positive, so giving the result in the case when I is compact. For general intervals, the result follows from the definition of the (generalised) Riemann–Stieltjes integral in these cases. ■

3.5.28 Proposition (Riemann–Stieltjes integral and absolute value on $\mathbb{R}$) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $\mathrm{f}, \varphi \colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *be functions with* φ *monotonically increasing and with* f *(generalised) Riemann–Stieltjes integrable with respect to* φ*. Then the function* $|\mathrm{f}| \colon \mathrm{x} \mapsto |\mathrm{f}(\mathrm{x})|$ *is (generalised) Riemann–Stieltjes integrable with respect to* φ *and*

$$\left|(\mathrm{G}) \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\, \mathrm{d}\varphi(\mathrm{x})\right| \leq (\mathrm{G}) \int_{\mathrm{a}}^{\mathrm{b}} |\mathrm{f}|(\mathrm{x})\, \mathrm{d}\varphi(\mathrm{x}).$$

Proof Since φ is monotonically increasing, it also has bounded variation, and so the result follows from Propositions 3.5.26 and 3.5.27 by taking $g(y) = |y|$ in Proposition 3.5.26. ■

Now we turn to describing the manner in which the Riemann–Stieltjes integral can be broken up into two integrals over disjoint intervals. Here we see one of the more important distinctions between the Riemann–Stieltjes integral and the generalised Riemann–Stieltjes integral, since the result that holds in the latter case is the better one.

3.5.29 Proposition (Breaking the Riemann–Stieltjes integral in two) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval and let* $\mathrm{I} = \mathrm{I}_1 \cup \mathrm{I}_2$*, where* $\mathrm{I}_1 \cap \mathrm{I}_2 = \{\mathrm{c}\}$*, where* c *is the right endpoint of* I_1 *and the left endpoint of* I_2*. If* $\mathrm{f}, \varphi \colon \mathrm{I} \to \mathbb{R}$ *are functions with* f *Riemann–Stieltjes integrable with respect to* φ*, then* $\mathrm{f}|\mathrm{I}_1$ *and* $\mathrm{f}|\mathrm{I}_2$ *are Riemann–Stieltjes integrable with respect to* φ*. Furthermore, we have*

$$\int_{\mathrm{I}} \mathrm{f}(\mathrm{x})\, \mathrm{d}\varphi(\mathrm{x}) = \int_{\mathrm{I}_1} \mathrm{f}(\mathrm{x})\, \mathrm{d}\varphi(\mathrm{x}) + \int_{\mathrm{I}_2} \mathrm{f}(\mathrm{x})\, \mathrm{d}\varphi(\mathrm{x}).$$

Proof It suffices to consider the case where $I = [a, b]$ and where $c \in (a, b)$; the general case follows from this case by using the definition of the Riemann–Stieltjes integral for general intervals.

Let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that, for each pair of partitions P and P' of $[a, b]$ satisfying $|P|, |P'| < \delta$ and for each pair of selections ξ and ξ' of P and P', respectively, we have

$$|A(f, \varphi, P, \xi) - A(f, \varphi, P', \xi')| < \epsilon,$$

this being possible by Proposition 3.5.9. Now let $\tilde{P}$ and $\tilde{P}'$ be partitions of $[a, c]$ satisfying $|\tilde{P}|, |\tilde{P}'| < \delta$ and let $\tilde{\xi}$ and $\tilde{\xi}'$ be selections from $\tilde{P}$ and $\tilde{P}'$, respectively. Now let $\hat{P}$ be a partition of $[c, b]$ for which $|\hat{P}| < \delta$ and let $\hat{\xi}$ be a selection from $\hat{P}$. Define partitions P and P' of $[a, b]$ satisfying $\mathrm{EP}(P) = \mathrm{EP}(\tilde{P}) \cup \mathrm{EP}(\hat{P})$ and $\mathrm{EP}(P') = \mathrm{EP}(\tilde{P}') \cup \mathrm{EP}(\hat{P})$. Define selections ξ and ξ' of P and P', respectively, by $\xi = \tilde{\xi} \cup \hat{\xi}$ and $\xi' = \tilde{\xi}' \cup \hat{\xi}$. Then we have

$$|A(f, \varphi, P, \xi) - A(f, \varphi, P', \xi')| = |A(f|[a, c], \varphi|[a, c], \tilde{P}, \tilde{\xi}) - A(f|[a, c], \varphi|[a, c], \tilde{P}', \tilde{\xi}')| < \epsilon.$$

By Proposition 3.5.9 it follows that $f|[a, c]$ is Riemann–Stieltjes integrable with respect to $\varphi|[a, c]$. An entirely similar argument shows that $f|[c, b]$ is Riemann–Stieltjes integrable with respect to $\varphi|[c, b]$.

Now let $\epsilon \in \mathbb{R}_{>0}$ and choose $\delta \in \mathbb{R}_{>0}$ such that if P, P_1, and P_2 are partitions of $[a, b]$, $[a, c]$, and $[c, b]$, respectively, and if ξ, ξ_1, and ξ_2 are selections from P, P_1, and P_2, respectively, then

$$\begin{aligned}|I(f,\varphi) - A(f,\varphi,P,\xi)| &< \tfrac{\epsilon}{3},\\ |I(f|[a,c],\varphi|[a,c]) - A(f|[a,c],\varphi|[a,c],P,\xi)| &< \tfrac{\epsilon}{3},\\ |I(f|[c,b],\varphi|[c,b]) - A(f|[c,b],\varphi|[c,b],P,\xi)| &< \tfrac{\epsilon}{3}\end{aligned}$$

provided that $|P|, |P_1|, |P_2| < \delta$. Now let P, P_1, and P_2 be partitions satisfying $|P|, |P_1|, |P_2| < \delta$. Without changing the requirement that $|P| < \delta$ we can assume that $c \in \mathrm{EP}(P)$, and thus we assume this. Then we let ξ, ξ_1, and ξ_2 be selections from P, P_1, and P_2, respectively. Noting that

$$A(f,\varphi,P,\xi) = A(f|[a,c],\varphi|[a,c],P_1,\xi_1) + A(f|[c,b],\varphi|[c,b],P_2,\xi_2),$$

we then compute

$$\begin{aligned}|I(f,\varphi)-&I(f|[a,c],\varphi|[a,c]) - I(f|[c,b],\varphi|[c,b])|\\ &= |I(f,\varphi) - I(f|[a,c],\varphi|[a,c]) - I(f|[c,b], f|[c,b])\\ &\quad - A(f,\varphi,P,\xi) + A(f|[a,c],\varphi|[a,c],P_1,\xi_1) + A(f|[c,b],\varphi|[c,b],P_2,\xi_2)|\\ &\le |I(f,\varphi) - A(f,\varphi,P,\xi)| + |I(f|[a,c],\varphi|[a,c]) - A(f|[a,c],\varphi|[a,c],P_1,\xi_1)|\\ &\quad + |I(f|[c,b],\varphi|[c,b]) - A(f|[c,b],\varphi|[c,b],P_2,\xi_2)| < \epsilon,\end{aligned}$$

which gives the desired equality in the statement of the result. ■

3.5.30 Example (A counterexample for the converse of Proposition 3.5.29) We take $I = [0, 1]$ and f and φ as in Example 3.5.8. In that example we saw that f was not Riemann–Stieltjes integrable with respect to φ. Nonetheless, we claim that $f|[0, \frac{1}{2}]$ is Riemann–Stieltjes integrable with respect to $\varphi|[0, \frac{1}{2}]$ and that $f|[\frac{1}{2}, 1]$ is Riemann–Stieltjes integrable with respect to $\varphi|[\frac{1}{2}, 1]$. Indeed, if P is any partition of $[0, \frac{1}{2}]$ and if ξ is any selection from P, we see directly that $A(f|[0, \frac{1}{2}], \varphi|[0, \frac{1}{2}], P, \xi) = 0$. Thus $f|[0, \frac{1}{2}]$ is Riemann–Stieltjes integrable with respect to $\varphi|[0, \frac{1}{2}]$. It is similarly directly computed that, if P is any partition of $[\frac{1}{2}, 1]$ and if ξ is any selection from P, then $A(f|[\frac{1}{2}, 1], \varphi|[\frac{1}{2}, 1], P, \xi) = 0$. Thus $f|[\frac{1}{2}, 1]$ is Riemann–Stieltjes integrable with respect to $\varphi|[\frac{1}{2}, 1]$. Thus the converse of Proposition 3.5.29 fails to hold for this example. •

The interesting fact is now that the sharp result concerning the splitting of the domain *does* hold for the generalised Riemann–Stieltjes integral.

3.5.31 Proposition (Breaking the generalised Riemann–Stieltjes integral in two) *Let* $I \subseteq \mathbb{R}$ *be an interval and let* $I = I_1 \cup I_2$*, where* $I_1 \cap I_2 = \{c\}$*, where* c *is the right endpoint of* I_1 *and the left endpoint of* I_2*. If* $f, \varphi \colon I \to \mathbb{R}$ *are functions, then* f *is generalised Riemann–Stieltjes integrable with respect to* φ *if and only if* $f|I_1$ *and* $f|I_2$ *are generalised Riemann–Stieltjes integrable with respect to* φ*. Furthermore, we have*

$$G\int_I f(x)\,d\varphi(x) = G\int_{I_1} f(x)\,d\varphi(x) + G\int_{I_2} f(x)\,d\varphi(x).$$

Proof Suppose that f is generalised Riemann–Stieltjes integrable with respect to φ. The implication that both $f|I_1$ and $f|I_2$ are generalised Riemann–Stieltjes integrable, and that the value of the integral over I is the sum of the integrals over I_1 and I_2 follows from Proposition 3.5.10 in the same manner, with minor modifications, as Proposition 3.5.29 follows from Proposition 3.5.9. We leave to the reader the straightforward notational substitutions.

Now suppose that $f|I_1$ and $f|I_2$ are generalised Riemann–Stieltjes integrable with respect to $\varphi|I_1$ and $\varphi|I_2$, respectively. We can and do assume, without loss of generality, that $I = [a, b]$, $I_1 = [a, c]$, and $I_2 = [c, b]$. Let the Riemann–Stieltjes integrals be denoted by $I(f|[a, c], \varphi|[a, c])$ and $I(f|[c, b], \varphi|[c, b])$. Let $\epsilon \in \mathbb{R}_{>0}$ and choose partitions P_1 of $[a, c]$ and P_2 of $[c, b]$ such that, if P'_1 is a refinement of P_1, if P'_2 is a refinement of P_2, if ξ'_1 is a selection from P_1, and if ξ'_2 is a selection from P_2, then

$$|A(f|[a, c], \varphi|[a, c], P'_1, \xi'_1) - I(f|[a, c], \varphi|[a, c])| < \tfrac{\epsilon}{2},$$
$$|A(f|[c, b], \varphi|[c, b], P'_2, \xi'_2) - I(f|[c, b], \varphi|[c, b])| < \tfrac{\epsilon}{2}.$$

Let P_0 be a partition of $[a, b]$ for which $\mathrm{EP}(P_0) = \mathrm{EP}(P_1) \cup \mathrm{EP}(P_2)$. Then, if P' is a refinement of P_0, there exists refinements P'_1 and P'_2 of P_1 and P_2, respectively, such that $\mathrm{EP}(P') = \mathrm{EP}(P'_1) \cup \mathrm{EP}(P'_2)$. For a refinement P' of P_0 and a selection ξ' of P we compute

$$\begin{aligned}
&|A(f, \varphi, P', \xi') - I(f|[a, c], \varphi|[a, c]) - I(f|[c, b], \varphi|[c, b])| \\
&\qquad \le |A(f|[a, c], \varphi|[a, c], P'_1, \xi'_1) - I(f|[a, c], \varphi|[a, c])| \\
&\qquad\qquad + |A(f|[c, b], \varphi|[c, b], P'_2, \xi'_2) - I(f|[c, b], \varphi|[c, b])| < \epsilon.
\end{aligned}$$

Thus the Riemann–Stieltjes integral of f with respect to φ exists and is equal to $I(f|[a, c], \varphi|[a, c]) + I(f|[c, b], \varphi|[c, b])$. ■

There is an integration by parts formula for the Riemann–Stieltjes integral which is striking in its generality.

3.5.32 Theorem (Integration by parts for the Riemann–Stieltjes integral) *Let* $I = [a, b]$ *be a compact interval and let* $f, \varphi \colon [a, b] \to \mathbb{R}$ *be functions. Then* f *is (generalised) Riemann–Stieltjes integrable with respect to* φ *if and only if* φ *is (generalised) Riemann–Stieltjes integrable with respect to* f*, and if either of these statements hold, then*

$$(G)\int_a^b f(x)\, d\varphi(x) + (G)\int_a^b \varphi(x)\, df(x) = f(b)\varphi(b) - f(a)\varphi(a).$$

Proof Consider first the case of the Riemann–Stieltjes integral. We prove a lemma concerning partitions.

1 Lemma *Let* $I = [a, b]$ *be a compact interval and let* $f, \varphi \colon [a, b] \to \mathbb{R}$ *be functions. If* P *is a partition of* $[a, b]$ *and if* ξ *is a selection of* P*, then there exists a partition* P' *and a selection* ξ' *from* P' *such that*

(i) $|P'| \le |P|$ *and*

(ii) $A(\varphi, f, P, \xi) + A(f, \varphi, P', \xi') = f(b)\varphi(b) - f(a)\varphi(a)$.

Proof Let P be a partition of $[a, b]$ and denote $\mathrm{EP}(P) = (x_0, x_1, \dots, x_k)$. Let $\boldsymbol{\xi} = (\xi_1, \dots, \xi_k)$ be a selection from P. We then have the elementary equalities

$$\begin{aligned} f(b)\varphi(b) - f(a)\varphi(a) &= \sum_{j=1}^{k} f(x_j)(\varphi(x_j) - \varphi(x_{j-1})), \\ A(\varphi, f, P, \boldsymbol{\xi}) &= \sum_{j=1}^{k} \varphi(\xi_j)(f(x_j) - f(x_{j-1})). \end{aligned}$$

Subtracting these gives

$$\begin{aligned} f(b)\varphi(b) - f(a)\varphi(a) - &A(\varphi, f, P, \boldsymbol{\xi}) \\ &= \sum_{j=1}^{k} f(x_j)(\varphi(x_j) - \varphi(x_{j-1})) - \sum_{j=1}^{k} \varphi(\xi_j)(f(x_j) - f(x_{j-1})) \\ &= \sum_{j=1}^{k} f(x_j)(\varphi(x_j) - \varphi(\xi_j)) + \sum_{j=1}^{k} f(x_{j-1})(\varphi(\xi_j) - \varphi(x_{j-1})). \end{aligned} \tag{3.19}$$

Define a partition P' having the property that $\mathrm{EP}(P') = \mathrm{EP}(P) \cup \boldsymbol{\xi}$, and note that $|P'| \le |P|$. Note that we allow the case that $\mathrm{card}(\mathrm{EP}(P')) < \mathrm{card}(\mathrm{EP}(P)) + \mathrm{card}(\boldsymbol{\xi})$, since some of points from the selection $\boldsymbol{\xi}$ might agree with endpoints from P. We then define a selection $\boldsymbol{\xi}'$ from P' by taking ξ'_j to be the left endpoint of the jth interval of P'. One can easily directly check that

$$A(f, \varphi, P', \boldsymbol{\xi}') = \sum_{j=1}^{k} f(x_j)(\varphi(x_j) - \varphi(\xi_j)) + \sum_{j=1}^{k} f(x_{j-1})(\varphi(\xi_j) - \varphi(x_{j-1})), \tag{3.20}$$

again allowing that some of the terms in the sum will be zero, corresponding to the cases where points from $\boldsymbol{\xi}$ agree with endpoints of P. The lemma follows by combining (3.19) and (3.20). ▼

Let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that, if P' is a partition of $[a, b]$ with $|P'| < \frac{\delta}{2}$ and if $\boldsymbol{\xi}'$ is a selection from P,

$$\left| A(f, \varphi, P', \boldsymbol{\xi}') - \int_a^b f(x)\,\mathrm{d}\varphi(x) \right| < \epsilon.$$

Now let P be any partition with $|P| < \delta$ and let $\boldsymbol{\xi}$ be a selection from P. Let P' and $\boldsymbol{\xi}'$ be as guaranteed by Lemma 1. We then have

$$\begin{aligned} &\left| A(\varphi, f, P, \boldsymbol{\xi}) - f(b)\varphi(b) + f(a)\varphi(a) - \int_a^b f(x)\,\mathrm{d}\varphi(x) \right| \\ &\qquad\qquad = \left| A(f, \varphi, P', \boldsymbol{\xi}') - \int_a^b f(x)\,\mathrm{d}\varphi(x) \right| < \epsilon. \end{aligned}$$

This shows that φ is Riemann–Stieltjes integrable with respect to f, and that the value of the integral is as claimed. The argument can clearly be reversed, giving the desired result for the Riemann–Stieltjes integral.

For the generalised Riemann–Stieltjes integral, we prove the following lemma, playing the rôle of Lemma 1 above.

2 Lemma *Let* $I = [a, b]$ *be a compact interval and let* $f, \varphi\colon [a, b] \to \mathbb{R}$ *be functions. If* P_0 *is a partition of* $[a, b]$ *then, given a refinement* P *of* P_0 *and a selection* ξ *from* P*, there exists a refinement* P' *of* P_0 *and a selection* ξ' *from* P' *such that* $A(\varphi, f, P, \xi) + A(f, \varphi, P', \xi') = f(b)\varphi(b) - f(a)\varphi(a)$.

Proof Let P_0 be a partition and let P be a refinement of P_0 and let ξ be a selection from P. Note that a partition P' having the property that $\mathrm{EP}(P') = \mathrm{EP}(P) \cup \xi$ is necessarily a refinement of P_0 since it is a refinement of P, which is itself a refinement of P_0. Thus the result follows by taking P' and ξ' to be as in the proof of Lemma 1. ▼

Assume that f is generalised Riemann–Stieltjes integrable with respect to φ. Let $\epsilon \in \mathbb{R}_{>0}$ and let P_0 be a partition with property that, if P' is a refinement of P_0 and if ξ' is a selection from P', then

$$\left|A(f, \varphi, P', \xi') - G\int_a^b f(x)\,\mathrm{d}\varphi(x)\right| < \epsilon.$$

Now let P be a refinement of P_0 and let ξ be a selection from P, and let P' and ξ' the partition and selection satisfying

$$A(\varphi, f, P, \xi) + A(f, \varphi, P', \xi') = f(b)\varphi(b) - f(a)\varphi(a),$$

as per Lemma 2. Then we have

$$\begin{aligned}\left|A(\varphi, f, P, \xi) - f(b)\varphi(b) + f(a)\varphi(a) - G\int_a^b f(x)\,\mathrm{d}\varphi(x)\right|& \\ = \left|A(f, \varphi, P', \xi') - G\int_a^b f(x)\,\mathrm{d}\varphi(x)\right| &< \epsilon.\end{aligned}$$

This φ is generalised Riemann–Stieltjes integrable with respect to f, with the value of the integral as claimed. The argument can be reversed to show that the generalised Riemann–Stieltjes integrability of φ with respect to f implies the generalised Riemann–Stieltjes integrability of f with respect to φ. ■

To close this section, we indicate a relationship between the Riemann–Stieltjes integral and the Riemann integral when the integrator has certain properties.

3.5.33 Proposition (Riemann–Stieltjes integral for differentiable integrators) *Let* $I = [a, b]$ *be a compact interval, let* $f, \varphi\colon [a, b] \to \mathbb{R}$ *be functions such that*

(i) f *is Riemann integrable and*

(ii) φ *is differentiable and* φ' *is Riemann integrable.*

Then f *is (generalised) Riemann–Stieltjes integrable with respect to* φ *if and only if* $\mathrm{f}\varphi'$ *is Riemann integrable. Furthermore, if either of these statements holds, then*

$$(\mathrm{G})\int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\,\mathrm{d}\varphi(\mathrm{x}) = \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{f}(\mathrm{x})\varphi'(\mathrm{x})\,\mathrm{dx}.$$

Proof Since φ is differentiable and has Riemann integrable derivative, it is necessarily continuous by Proposition 3.2.7 and of bounded variation by Proposition 3.3.14. Therefore, by Corollaries 3.5.19 and 3.5.20, the following statements are equivalent:

1. f is Riemann–Stieltjes integrable with respect to φ;
2. f is generalised Riemann–Stieltjes integrable with respect to φ;
3. D_f is φ-null.

Now suppose that f and φ are as stated and that $f\varphi'$ is Riemann integrable. Let $M = \sup\{|\varphi'(x)| \mid x \in [a,b]\}$. Since D_f has measure zero, for $\epsilon \in \mathbb{R}_{>0}$ choose a countable family $((a_\alpha, b_\alpha))_{\alpha\in A}$ of open intervals such that

$$\sum_{\alpha\in A} |b_\alpha - a_\alpha| < \frac{\epsilon}{M}$$

and such that

$$D_{f\varphi'} \subseteq \bigcup_{\alpha\in A} (a_\alpha, b_\alpha).$$

If P_α is a partition of $[a_\alpha, b_\alpha]$ with $\mathrm{EP}(P_\alpha) = (x_{\alpha,0}, x_{\alpha,1}, \dots, x_{\alpha,k_\alpha})$ then, by the Mean Value Theorem, for $j \in \{1, \dots, k_\alpha\}$ there exists $\xi_{\alpha,j} \in (x_{\alpha,j-1}, x_{\alpha,j})$ such that

$$\varphi(x_{\alpha,j}) - \varphi(x_{\alpha,j-1}) = \varphi'(\xi_{\alpha,j})(x_{\alpha,j} - x_{\alpha,j-1}).$$

Then we compute

$$\begin{aligned}\sum_{j=1}^{k_\alpha} |\varphi(x_{\alpha,j}) - \varphi(x_{\alpha,j-1})| &= \sum_{j=1}^{k_\alpha} |\varphi'(\xi_{\alpha,j})||x_{\alpha,j} - x_{\alpha,j-1}| \\ &\le M \sum_{j=1}^{k_\alpha} |x_{\alpha,j} - x_{\alpha,j-1}| = M(b_\alpha - a_\alpha).\end{aligned}$$

Therefore,

$$\sum_{\alpha\in A} \mathrm{TV}(\varphi|[a_\alpha, b_\alpha]) \le M \sum_{\alpha\in A} |b_\alpha - a_\alpha| < \epsilon.$$

This shows that D_f is φ-null, and so f is (generalised) Riemann–Stieltjes integrable with respect to φ. Also note that the hypotheses of the result immediately imply that $f\varphi'$ is Riemann integrable.

Finally, we show the equality of the integrals. Denote by $I(f\varphi')$ and $I(f,\varphi)$ the Riemann integral of $f\varphi'$ and the Riemann–Stieltjes integral of f with respect to φ, respectively. Let $\epsilon \in \mathbb{R}_{>0}$ and let $\delta \in \mathbb{R}_{>0}$ have the property that

$$\left|\sum_{j=1}^{k} f(\xi_j)\varphi'(\xi_j)(x_j - x_{j-1}) - I(f\varphi')\right| < \frac{\epsilon}{2}$$

and

$$\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j)-\varphi(x_{j-1}))-I(f,\varphi)\right|<\frac{\epsilon}{2}$$

for every partition P with $|P|<\delta$ (we are denoting $\mathrm{EP}(P)=(x_0,x_1,\dots,x_k)$) and for every selection $(\xi_1,\dots,\xi_k)$ of P. Then let P be a partition of $[a,b]$ with $|P|<\delta$ and let $\mathrm{EP}(P)=(x_0,x_1,\dots,x_k)$. By the Mean Value Theorem, for $j\in\{1,\dots,k\}$ there exists $\xi_j\in(x_{j-1},x_j)$ such that

$$\varphi(x_j)-\varphi(x_{j-1})=\varphi'(\xi_j)(x_j-x_{j-1}).$$

Then

$$|I(f\varphi')-I(f,\varphi)|\le\left|\sum_{j=1}^{k} f(\xi_j)\varphi'(\xi_j)(x_j-x_{j-1})-I(f\varphi')\right| \\ +\left|\sum_{j=1}^{k} f(\xi_j)(\varphi(x_j)-\varphi(x_{j-1}))-I(f,\varphi)\right|<\epsilon.$$

Thus $I(f\varphi')=I(f,\varphi)$, as desired. ■

3.5.5 The Fundamental Theorem of Calculus and the Mean Value Theorems

In this section we give an analogue of the Fundamental Theorem of Calculus, stated as Theorem 3.4.30 for the Riemann integral, for the Riemann–Stieltjes integral. We also state the integral Mean Value Theorems for the Riemann–Stieltjes integral; these theorems are, in fact, somewhat more natural in the Riemann–Stieltjes setting than in the Riemann setting for integration.

3.5.34 Proposition (Fundamental Theorem of Calculus for Riemann–Stieltjes integrals) *Let* $\mathrm{I}=[\mathrm{a},\mathrm{b}]$ *be a compact interval and let* $\mathrm{f},\varphi\colon[\mathrm{a},\mathrm{b}]\to\mathbb{R}$ *be functions with* f *continuous and* φ *monotonically increasing. If* $\mathrm{F}\colon[\mathrm{a},\mathrm{b}]\to\mathbb{R}$ *is defined by*

$$\mathrm{F}(\mathrm{x})=\int_{\mathrm{a}}^{\mathrm{x}}\mathrm{f}(\xi)\,\mathrm{d}\varphi(\xi),$$

then F *is differentiable at points where* φ *is differentiable, and, if* x *is such a point, then* $\mathrm{F}'(\mathrm{x})=\mathrm{f}(\mathrm{x})\varphi'(\mathrm{x})$.

Proof For $x\in[a,b)$, suppose that φ is differentiable at x. For $\epsilon\in\mathbb{R}_{>0}$ sufficiently small we have

$$F(x+\epsilon)-F(x)=\int_{x}^{x+\epsilon}f(\xi)\,\mathrm{d}\varphi(\xi)=f(\xi_\epsilon)(\varphi(x+\epsilon)-\varphi(x))$$

for some $\xi_\epsilon\in[x,x+\epsilon]$, by Proposition 3.5.35 and using Proposition 3.5.29. Then

$$\lim_{\epsilon\downarrow 0}\frac{F(x+\epsilon)-F(x)}{\epsilon}=\lim_{\epsilon\downarrow 0}f(\xi_\epsilon)\frac{(\varphi(x+\epsilon)-\varphi(x))}{\epsilon}.$$

A similar argument shows that, if $x \in (a, b]$, then

$$\lim_{\epsilon \downarrow 0} \frac{F(x) - F(x-\epsilon)}{\epsilon} = \lim_{\epsilon \downarrow 0} f(\xi_\epsilon) \frac{\varphi(x) - \varphi(x-\epsilon)}{\epsilon}$$

for some $\xi_\epsilon \in [x-\epsilon, x]$. From Propositions 2.3.23 and 2.3.29, along with the fact that f is continuous so that $\lim_{\epsilon \to I} f(\xi_\epsilon) = f(x)$, we conclude that $F'(x) = f(x)\varphi'(x)$. ■

3.5.35 Proposition (First Mean Value Theorem for Riemann–Stieltjes integrals) *Let* $[a, b]$ *be a compact interval and let* $f, \varphi \colon [a, b] \to \mathbb{R}$ *be functions such that* f *is continuous and* φ *is monotonically increasing. Then there exists* $c \in [a, b]$ *such that*

$$\int_a^b f(x)\,d\varphi(x) = f(c)(\varphi(b) - \varphi(a)).$$

Proof First note that the integral exists by Corollary 3.5.21. Define

$$m = \inf\{f(x) \mid x \in [a, b]\}, \quad M = \sup\{f(x) \mid x \in [a, b]\}.$$

Then

$$m(\varphi(b) - \varphi(a)) \leq \int_a^b f(x)\,d\varphi(x) \leq M(\varphi(b) - \varphi(a))$$

by Exercise 3.5.2. Therefore, there exists $\mu \in [m, M]$ such that

$$\int_a^b f(x) = \mu(\varphi(b) - \varphi(a)).$$

That there exists $c \in [a, b]$ such that $f(c) = \mu$ follows from the Intermediate Value Theorem. ■

3.5.36 Proposition (Second Mean Value Theorem for Riemann–Stieltjes integrals) *Let* $[a, b]$ *be a compact interval and let* $f, \varphi \colon [a, b] \to \mathbb{R}$ *be functions with* f *monotonically increasing and with* φ *continuous. Then there exists* $c \in [a, b]$ *so that*

$$\int_a^b f(x)\,d\varphi(x) = f(a)(\varphi(c) - \varphi(a)) + f(b)(\varphi(b) - \varphi(c)).$$

Proof Using integration by parts,

$$\int_a^b f(x)\,d\varphi(x) = f(b)\varphi(b) - f(a)\varphi(a) - \int_a^b \varphi(x)\,df(x).$$

(This shows, incidentally, that the integral in the statement of the result exists.) By Proposition 3.5.35 there exists $c \in [a, b]$ such that

$$\int_a^b \varphi(x)\,df(x) = \varphi(c)(f(b) - f(a)),$$

and the result immediately follows. ■

3.5.6 Notes

Pollard [1920] shows that, if every continuous f is Riemann–Stieltjes integrable with respect to φ, then φ has bounded variation.

Young [1913] states Theorem 3.5.18, but does not give a convincing proof. The first complete proof seems to be that given by Bliss [1917].

It is tempting to write the Riemann integral as something like

$$\lim_{|P|\to 0} \sum_{j=1}^{k} f(\xi_j)(x_j - x_{j-1}),$$

where it is understood that $\{x_0, x_1, \dots, x_k\}$ are the endpoints of a partition, and that $\{\xi_1, \dots, \xi_k\}$ is a selection from a partition. However, after thinking a little about this limit, it becomes clear that it is not a limit in any way we have thus far encountered. However, it *is* possible, using the language of nets to precisely formulate the Riemann integral as a limit. One of the advantages of the Darboux characterisation using upper and lower integrals is that it obviates the need to precisely define such a limit, as it instead uses the fact that the lower and upper sums increase and decrease, respectively, monotonically as one decreases the mesh of a partition. However, the need to reconsider the idea of a limit gets revisited when one turns to the Riemann–Stieltjes integral. Since only the generalised Riemann–Stieltjes integral has a valid definition in terms of lower and upper integrals, one cannot employ this device for the Riemann–Stieltjes integral. Moreover, the subtle differences in the definitions of the Riemann–Stieltjes integral and the generalised Riemann–Stieltjes integral points out the need to take care if one wishes to define these integrals as limits. We do not address this here, but refer to [Hobson 1957] for a formulation of the Riemann–Stieltjes integral and the generalised Riemann–Stieltjes integrals as limits using nets.

Exercises

3.5.1 For a compact interval $[a, b]$ and a bounded function $\varphi\colon [a, b] \to \mathbb{R}$, show that $\int_a^b d\varphi(x) = \varphi(b) - \varphi(a)$.

3.5.2 Let $[a, b]$ be a compact interval and let $f, \varphi\colon [a, b] \to \mathbb{R}$ be functions with f bounded and φ monotonically increasing. Show that

$$m(\varphi(b) - \varphi(a)) \le \int_a^b f(x)\, d\varphi(x) \le M(\varphi(b) - \varphi(a)),$$

where

$$m = \inf\{f(x) \mid x \in [a, b]\}, \quad M = \sup\{f(x) \mid x \in [a, b]\}.$$

3.5.3 For a compact interval $[a, b]$, a point $c \in [a, b]$, the functions $\varphi_{1,c}, \varphi_{2,c}\colon [a, b] \to$

$\mathbb{R}$ given by

$$\varphi_{1,c}(x) = \begin{cases} 0, & x \in [a, c), \\ 1, & x \in [c, b], \end{cases} \quad \varphi_{2,c}(x) = \begin{cases} 0, & x \in [a, c), \\ \frac{1}{2}, & x = c, \\ 1, & x \in (c, b], \end{cases}$$

and for a function $f\colon [a, b] \to \mathbb{R}$ that is continuous at c, show that

$$\int_a^b f(x)\,\mathrm{d}\varphi_{1,c}(x) = \int_a^b f(x)\,\mathrm{d}\varphi_{2,c}(x) = f(c).$$

3.5.4 Take $I = [0, 1]$ and the functions $f, \varphi\colon [0, 1] \to \mathbb{R}$ defined by

$$f(x) = \varphi(x) = \begin{cases} 0, & x \in [0, \frac{1}{2}], \\ 1, & x \in (\frac{1}{2}, 1]. \end{cases}$$

Show that f is not generalised Riemann–Stieltjes integrable with respect to φ.

Section 3.6

Sequences and series of ℝ-valued functions

In this section we present for the first time the important topic of sequences and series of functions and their convergence. One of the reasons why convergence of sequences of functions is important is that is allows us to classify sets of functions. The idea of classifying sets of functions according to their possessing certain properties leads to the general idea of a "function space." Function spaces are important to understand when developing any systematic theory dealing with functions, since sets of general functions are simply too unstructured to allow much useful to be said. On the other hand, if one restricts the set of functions in the wrong way (e.g., by asking that they all be continuous), then one can end of with a framework with unpleasant properties. But this is getting a little ahead of the issue directly at hand, which is to consider convergence of sequences of functions.

Do I need to read this section? The material in this section is basic, particularly the concepts of pointwise convergence and uniform convergence and the distinction between them. However, it is possible to avoid reading this section until the material becomes necessary, as it will in Chapters IV-3, IV-4, IV-5, and IV-6, for example. •

3.6.1 Pointwise convergent sequences

The first type of convergence we deal with is probably what a typical first-year student, at least the rare one who understood convergence for summations of numbers, would proffer as a good candidate for convergence. As we shall see, it often leaves something to be desired.

In the discussion of pointwise convergence, one needs no assumptions on the character of the functions, as one is essentially talking about convergence of numbers.

3.6.1 Definition (Pointwise convergence of sequences) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of ℝ-valued functions on I.

(i) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges pointwise*** to a function $f\colon I \to \mathbb{R}$ if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f(x) - f_j(x)| < \epsilon$ provided that $j \geq N$.

(ii) The function f in the preceding part of the definition is the ***limit function*** for the sequence.

(iii) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ is ***pointwise Cauchy*** if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ provided that $j, k \geq N$.

•

Let us immediately establish the equivalence of pointwise convergent and pointwise Cauchy sequences. As is clear in the proof of the following result, the key fact is completeness of $\mathbb{R}$.

3.6.2 Theorem (Pointwise convergent equals pointwise Cauchy) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an interval and if* $(\mathrm{f}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is a sequence of* $\mathbb{R}$*-valued functions on* I *then the following statements are equivalent:*

(i) there exists a function $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *such that* $(\mathrm{f}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *converges pointwise to* f*;*

(ii) $(\mathrm{f}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is pointwise Cauchy.*

Proof This merely follows from the following facts.

1. If the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ converges to $f(x)$ then the sequence is Cauchy by Proposition 2.3.3.
2. If the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is Cauchy then there exists a number $f(x) \in \mathbb{R}$ such that $\lim_{j\to\infty} f_j(x) = f(x)$ by Theorem 2.3.5. ■

Based on the preceding theorem we shall switch freely between the notions of pointwise convergent and pointwise Cauchy sequences of functions.

Pointwise convergence is essentially the most natural form of convergence for a sequence of functions in that it depends in a trivial way on the basic notion of convergence of sequences in $\mathbb{R}$. However, as we shall see later in this section, and in Chapters III-3 and III-6, other forms of convergence of often more useful.

3.6.3 Example (Pointwise convergence) Consider the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ of $\mathbb{R}$-valued functions defined on $[0,1]$ by

$$f_j(x) = \begin{cases} 1, & x \in [0, \frac{1}{j}], \\ 0, & x \in (\frac{1}{j}, 1]. \end{cases}$$

Note that $f_j(0) = 1$ for every $j \in \mathbb{Z}_{>0}$, so that the sequence $(f_j(0))_{j\in\mathbb{Z}_{>0}}$ converges, trivially, to 1. For any $x_0 \in (0,1]$, provided that $j > x_0^{-1}$, then $f_j(x_0) = 0$. Thus $(f_j(x_0))_{j\in\mathbb{Z}_{>0}}$ converges, as a sequence of real numbers, to 0 for each $x_0 \in (0,1]$. Thus this sequence converges pointwise, and the limit function is

$$f(x) = \begin{cases} 1, & x = 0, \\ 0, & x \in (0,1]. \end{cases}$$

If N is the smallest natural number with the property that $N > x_0^{-1}$, then we observe, trivially, that this number does indeed depend on x_0. As x_0 gets closer and closer to 0 we have to wait longer and longer in the sequence $(f_j(x_0))_{j\in\mathbb{Z}_{>0}}$ for the arrival of zero. •

3.6.2 Uniformly convergent sequences

Let us first say what we mean by uniform convergence.

3.6.4 Definition (Uniform convergence of sequences) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of ℝ-valued functions on I.

(i) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges uniformly*** to a function $f\colon I \to \mathbb{R}$ if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f(x) - f_j(x)| < \epsilon$ for all $x \in I$, provided that $j \geq N$.

(ii) The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ is ***uniformly Cauchy*** if, for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ for all $x \in I$, provided that $j, k \geq N$. •

Let us immediately give the equivalence of the preceding notions of convergence.

3.6.5 Theorem (Uniformly convergent equals uniformly Cauchy) *For an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and a sequence of* ℝ*-valued functions* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *on* I *the following statements are equivalent:*

(i) there exists a function $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *such that* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *converges uniformly to* f*;*

(ii) $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is uniformly Cauchy.*

Proof First suppose that $(f_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly Cauchy. Then, for each $x \in I$ the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is Cauchy and so by Theorem 2.3.5 converges to a number that we denote by $f(x)$. This defines the function $f\colon I \to \mathbb{R}$ to which the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges pointwise. Let $\epsilon \in \mathbb{R}_{>0}$ and let $N_1 \in \mathbb{Z}_{>0}$ have the property that $|f_j(x) - f_k(x)| < \frac{\epsilon}{2}$ for $j, k \geq N_1$ and for each $x \in I$. Now let $x \in I$ and let $N_2 \in \mathbb{Z}_{>0}$ have the property that $|f_k(x) - f(x)| < \frac{\epsilon}{2}$ for $k \geq N_2$. Then, for $j \geq N_1$, we compute

$$|f_j(x) - f(x)| \leq |f_j(x) - f_k(x)| + |f_k(x) - f(x)| < \epsilon,$$

where $k \geq \max\{N_1, N_2\}$, giving the first implication.

Now suppose that, for $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f(x)| < \epsilon$ for all $j \geq N$ and for all $x \in I$. Then, for $\epsilon \in \mathbb{R}_{>0}$ let $N \in \mathbb{Z}_{>0}$ satisfy $|f_j(x) - f(x)| < \frac{\epsilon}{2}$ for $j \geq N$ and $x \in I$. Then, for $j, k \geq N$ and for $x \in I$, we have

$$|f_j(x) - f_k(x)| \leq |f_j(x) - f(x)| + |f_k(x) - f(x)| < \epsilon,$$

giving the sequence as uniformly Cauchy. ■

Compare this definition to that for pointwise convergence. They sound similar, but there is a fundamental difference. For pointwise convergence, the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is examined separately for convergence at each value of x. As a consequence of this, the value of N might depend on both ϵ and x. For uniform convergence, however, we ask that for a given ϵ, the convergence is tested over all of I. In Figure 3.11 we depict the idea behind uniform convergence. The distinction between uniform and pointwise convergence is subtle on a first encounter, and it is sometimes difficult to believe that pointwise convergence is possible without uniform convergence. However, this is indeed the case, and an example illustrates this readily.

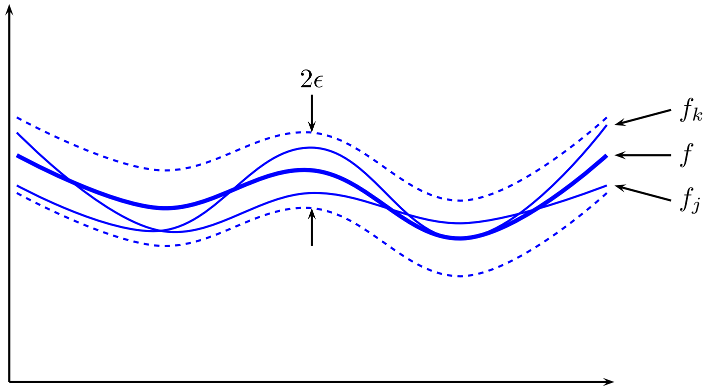


Figure 3.11 The idea behind uniform convergence

3.6.6 Example (Uniform convergence) On $[0,1]$ we consider the sequence of $\mathbb{R}$-valued functions defined by

$$f_j(x) = \begin{cases} 2jx, & x \in [0, \frac{1}{2j}], \\ -2jx + 2, & x \in (\frac{1}{2j}, \frac{1}{j}], \\ 0, & x \in (\frac{1}{j}, 1]. \end{cases}$$

In Figure 3.12 we graph f_j for $j \in \{1, 3, 10, 50\}$. The astute reader will see the point, but let's go through it just to make sure we see how this works.

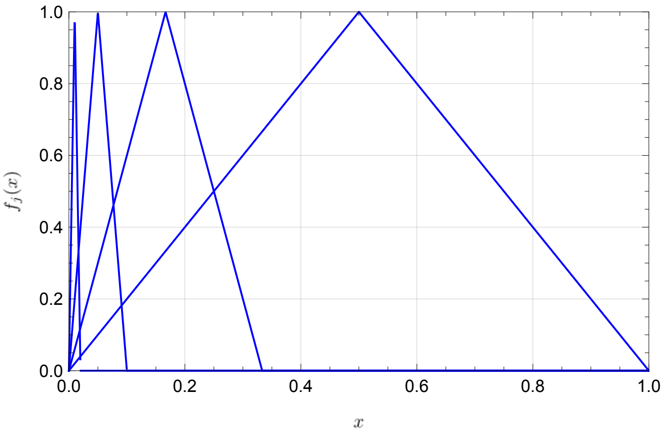


Figure 3.12 A sequence of functions converging pointwise, but not uniformly

First of all, we claim that the sequence converges pointwise to the limit function $f(x) = 0$, $x \in [0,1]$. Since $f_j(0) = 0$ for all $j \in \mathbb{Z}_{>0}$, obviously the sequence converges to 0 at $x = 0$. For $x \in (0,1]$, if $N \in \mathbb{Z}_{>0}$ satisfies $\frac{1}{N} < x$ then we have $f_j(x) = 0$ for $j \geq N$. Thus we do indeed have pointwise convergence.

We also claim that the sequence does not converge uniformly. Indeed, for any positive $\epsilon < 1$, we see that $f_j(\frac{1}{2j}) = 1 > \epsilon$ for every $j \in \mathbb{Z}_{>0}$. This prohibits our asserting the existence of $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ for every $x \in [0, 1]$, provided that $j, k \geq N$. Thus convergence is indeed not uniform. •

As we say, this is perhaps subtle, at least until one comes to grips with, after which point it makes perfect sense. You should not stop thinking about this until it makes perfect sense. If you overlook this distinction between pointwise and uniform convergence, you will be missing one of the most important topics in the theory of frequency representations of signals.

3.6.7 Remark (On "uniformly" again) In Remark 3.1.6 we made some comments on the notion of what is meant by "uniformly." Let us reinforce this here. In Definition 3.1.5 we introduced the notion of uniform continuity, which meant that the "δ" could be chosen so as to be valid on the entire domain. Here, with uniform convergence, the idea is that "N" can be chosen to be valid on the entire domain. Similar uses will occasionally be made of the word "uniformly" throughout the text, and it is hoped that the meaning should be clear from the context. •

Now we prove an important result concerning uniform convergence. The significance of this result is perhaps best recognised in a more general setting, such as that of Theorem III-1.9.1, where the idea of completeness is clear. However, even in the simple setting of our present discussion, the result is important enough.

3.6.8 Theorem (The uniform limit of bounded, continuous functions is bounded and continuous) *Let* I $\subseteq$ $\mathbb{R}$ *be an interval with* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *a sequence of continuous bounded functions on* I *that converge uniformly. Then the limit function is continuous and bounded. In particular, a uniformly convergent sequence of continuous functions defined on a compact interval converges to a continuous limit function.*

Proof Let $x \in I$ define $f(x) = \lim_{j\to\infty} f_j(x)$. This pointwise limit exists since $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is a Cauchy sequence in $\mathbb{R}$ (why?). We first claim that f is bounded. To see this, for $\epsilon \in \mathbb{R}_{>0}$, let $N \in \mathbb{Z}_{>0}$ have the property that $|f(x) - f_N(x)| < \epsilon$ for every $x \in I$. Then

$$|f(x)| \leq |f(x) - f_N(x)| + |f_N(x)| \leq \epsilon + \sup\{f_N(x) \mid x \in I\}.$$

Since the expression on the right is independent of x, this gives the desired boundedness of f.

Now we prove that the limit function f is continuous. Since $(f_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly convergent, for any $\epsilon \in \mathbb{R}_{>0}$ there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f(x)| < \frac{\epsilon}{3}$ for all $x \in I$ and $j \geq N$. Now fix $x_0 \in I$, and consider the $N \in \mathbb{Z}_{>0}$ just defined. By continuity of f_N, there exists $\delta \in \mathbb{R}_{>0}$ such that, if $x \in I$ satisfies $|x - x_0| < \delta$, then $|f_N(x) - f_N(x_0)| < \frac{\epsilon}{3}$. Then, for $x \in I$ satisfying $|x - x_0| < \delta$, we have

$$\begin{aligned}|f(x) - f(x_0)| &= |(f(x) - f_N(x)) + (f_N(x) - f_N(x_0)) + (f_N(x_0) - f(x_0))| \\ &\leq |f(x) - f_N(x)| + |f_N(x) - f_N(x_0)| + |f_N(x_0) - f(x_0)| \\ &< \tfrac{\epsilon}{3} + \tfrac{\epsilon}{3} + \tfrac{\epsilon}{3} = \epsilon,\end{aligned}$$

where we have again used the triangle inequality. Since this argument is valid for any $x_0 \in I$, it follows that f is continuous. ■

Note that the hypothesis that the functions be bounded is essential for the conclusions to hold. As we shall see, the contrapositive of this result is often helpful. That is, it is useful to remember that if a sequence of continuous functions defined on a closed bounded interval converges to a *dis*continuous limit function, then the convergence is *not* uniform.

3.6.3 Dominated and bounded convergent sequences

Bounded convergence is a notion that is particularly useful when discussing convergence of function sequences on noncompact intervals.

3.6.9 Definition (Dominated and bounded convergence of sequences) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I. For a function $g\colon I \to \mathbb{R}_{>0}$, the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges dominated by* g** if

(i) $f_j(x) \leq g(x)$ for every $j \in \mathbb{Z}_{>0}$ and for every $x \in I$ and

(ii) if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f_k(x)| < \epsilon$ for $j, k \geq N$.

If, moreover, g is a constant function, then a sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ that converges dominated by g ***converges boundedly***. •

It is clear that dominated convergence implies pointwise convergence. Indeed, bounded convergence is merely pointwise convergence with the extra hypothesis that all functions be bounded by the same positive function.

Let us give some examples that distinguish between the notions of convergence we have.

3.6.10 Examples (Pointwise, bounded, and uniform convergence)

1. The sequence of functions in Example 3.6.3 converges pointwise, boundedly, but not uniformly.
2. The sequence of functions in Example 3.6.6 converges pointwise, boundedly, but not uniformly.
3. Consider now a new sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ defined on $I = [0,1]$ by

$$f_j(x) = \begin{cases} 2j^2x, & x \in [0, \frac{1}{2j}], \\ -2j^2x + 2j, & x \in (\frac{1}{2j}, \frac{1}{j}], \\ 0, & \text{otherwise.} \end{cases}$$

A few members of the sequence are shown in Figure 3.13. This sequence converges pointwise to the zero function. Moreover, one can easily check that

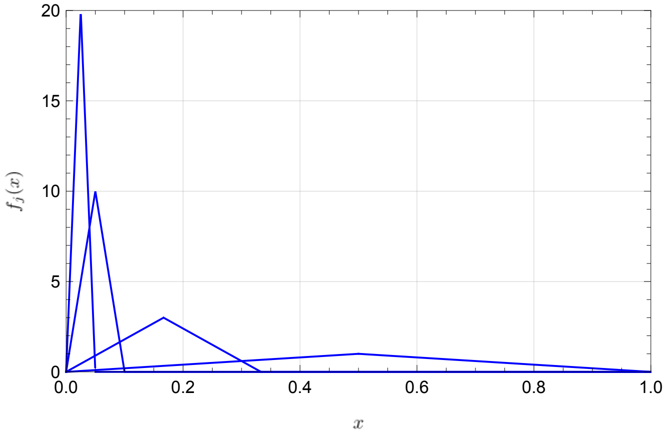


Figure 3.13 A sequence converging pointwise but not boundedly (shown are f_j, $j \in \{1, 5, 10, 20\}$)

the convergence is dominated by the function $g\colon [0,1] \to \mathbb{R}$ defined by

$$g(x) = \begin{cases} \frac{1}{x}, & x \in (0,1], \\ 1, & x = 0. \end{cases}$$

The sequence converges neither boundedly nor uniformly.

4. On $I = \mathbb{R}$ consider the sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ defined by $f_j(x) = x^2 + \frac{1}{j}$. This sequence clearly converges uniformly to $f\colon x \mapsto x^2$. However, it does not converge boundedly. Of course, the reason is simply that f is itself not bounded. We shall see that uniform convergence to a bounded function implies bounded convergence, in a certain sense. •

We have the following relationship between uniform and bounded convergence.

3.6.11 Proposition (Relationship between uniform and bounded convergence) *If a sequence* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *defined on an interval* I *converges uniformly to a bounded function* f*, then there exists* $\mathrm{N} \in \mathbb{Z}_{>0}$ *such that the sequence* $(\mathrm{f_{N+j}})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *converges boundedly to* f*.*

Proof Let $M \in \mathbb{R}_{>0}$ have the property that $|f(x)| < \frac{M}{2}$ for each $x \in I$. Since $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges uniformly to f there exists $N \in \mathbb{Z}_{>0}$ such that $|f(x) - f_j(x)| < \frac{M}{2}$ for all $x \in I$ and for $j > N$. It then follows that

$$|f_j(x)| \le |f(x) - f_j(x)| + |f(x)| < M$$

provided that $j > N$. From this the result follows since pointwise convergence of $(f_j)_{j\in\mathbb{Z}_{>0}}$ to f implies pointwise convergence of $(f_{N+j})_{j\in\mathbb{Z}_{>0}}$ to f. ■

3.6.4 Series of $\mathbb{R}$-valued functions

In the previous sections we considered the general matter of sequences of functions. Of course, this discussion carries over to ***series*** of functions, by which we mean expressions of the form $S(x) = \sum_{j=1}^{\infty} f_j(x)$. This is done in the usual manner by considering the partial sums. Let us do this formally.

3.6.12 Definition (Convergence of series) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I. Let $F(x) = \sum_{j=1}^{\infty} f_j(x)$ be a series. The corresponding sequence of ***partial sums*** is the sequence $(F_k)_{k\in\mathbb{Z}_{>0}}$ of $\mathbb{R}$-valued functions on I defined by

$$S_k(x) = \sum_{j=1}^{k} f_j(x).$$

Let $g\colon I \to \mathbb{R}_{>0}$. The series:

(i) ***converges pointwise*** if the sequence of partial sums converges pointwise;

(ii) ***converges uniformly*** if the sequence of partial sums converges uniformly;

(iii) ***converges dominated by* g** if the sequence of partial sums converges dominated by g;

(iv) ***converges boundedly*** if the sequence of partial sums converges boundedly. •

A fairly simple extension of pointwise convergence of series is the following notion which is unique to series (as opposed to sequences).

3.6.13 Definition (Absolute convergence of series) Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of $\mathbb{R}$-valued functions on I. The sequence $(f_j)_{j\in\mathbb{Z}_{>0}}$ ***converges absolutely*** if, for each $x \in I$ and for each $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that $||f_j(x)| - |f_k(x)|| < \epsilon$ provided that $j, k \geq N$. •

Thus an absolutely convergent sequence is one where, for each $x \in I$, the sequence $(|f_j(x)|)_{j\in\mathbb{Z}_{>0}}$ is Cauchy, and hence convergent. In other words, for each $x \in I$, the sequence $(f_j(x))_{j\in\mathbb{Z}_{>0}}$ is absolutely convergent. It is clear, then, that an absolutely convergent sequence of functions is pointwise convergent. Let us give some examples that illustrate the difference between pointwise and absolute convergence.

3.6.14 Examples (Absolute convergence)

1. The sequence of functions of Example 3.6.3 converges absolutely since the functions all take positive values.

2. For $j \in \mathbb{Z}_{>0}$, define $f_j\colon [0,1] \to \mathbb{R}$ by $f_j(x) = \frac{(-1)^{j+1}x}{j}$. Then, by Example 2.4.2–3, the series $S(x) = \sum_{j=1}^{\infty} f_j(x)$ is absolutely convergent if and only $x = 0$. But in Example 2.4.2–3 we showed that the series is pointwise convergent. •

3.6.5 Some results on uniform convergence of series

At various times in our development, we will find it advantageous to be able to refer to various standard results on uniform convergence, and we state these here.

Let us first recall the Weierstrass M-test.

3.6.15 Theorem (Weierstrass M-test) *If* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is a sequence of* $\mathbb{R}$*-valued functions defined on an interval* $\mathrm{I} \subseteq \mathbb{R}$ *and if there exists a sequence of positive constants* $(\mathrm{M_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *such that*

(i) $|\mathrm{f_j}(\mathrm{x})| \leq \mathrm{M_j}$ *for all* $\mathrm{x} \in \mathrm{I}$ *and for all* $\mathrm{j} \in \mathbb{Z}_{>0}$ *and*

(ii) $\sum_{\mathrm{j}=1}^{\infty} \mathrm{M_j} < \infty$,

then the series $\sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j}$ *converges uniformly and absolutely.*

Proof For $\epsilon \in \mathbb{R}_{>0}$, there exists $N \in \mathbb{Z}_{>0}$ such that, if $l \geq N$, we have

$$|M_l + \cdots + M_{l+k}| < \epsilon$$

for every $k \in \mathbb{Z}_{>0}$. Therefore, by the triangle inequality,

$$\left|\sum_{j=l}^{l+k} f_j(x)\right| \leq \sum_{j=l}^{l+k} |f_j(x)| \leq \sum_{j=l}^{l+k} M_j.$$

This shows that, for every $\epsilon \in \mathbb{R}_{>0}$, the tail of the series $\sum_{j=1}^{\infty} f_j$ can be made smaller than ϵ, and uniformly in x. This implies uniform and absolute convergence. ■

Next we present Abel's test.

3.6.16 Theorem (Abel's test) *Let* $(\mathrm{g_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of* $\mathbb{R}$*-valued functions on an interval* $\mathrm{I} \subseteq \mathbb{R}$ *for which* $\mathrm{g_{j+1}}(\mathrm{x}) \leq \mathrm{g_j}(\mathrm{x})$ *for all* $\mathrm{j} \in \mathbb{Z}_{>0}$ *and* $\mathrm{x} \in \mathrm{I}$*. Also suppose that there exists* $\mathrm{M} \in \mathbb{R}_{>0}$ *such that* $\mathrm{g_j}(\mathrm{x}) \leq \mathrm{M}$ *for all* $\mathrm{x} \in \mathrm{I}$ *and* $\mathrm{j} \in \mathbb{Z}_{>0}$*. Then, if the series* $\sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j}$ *converges uniformly on* I*, then so too does the series* $\sum_{\mathrm{j}=1}^{\infty} \mathrm{g_j f_j}$.

Proof Denote

$$F_k(x) = \sum_{j=1}^{k} f_j(x), \quad G_k(x) = \sum_{j=1}^{k} g_j(x) f_j(x)$$

as the partial sums. Using Abel's partial summation formula (Proposition 2.4.16), for $0 < k < l$ we write

$$G_l(x) - G_k(x) = (F_l(x) - F_k(x))G_1(x) + \sum_{j=k+1}^{l} (F_l(x) - F_j(x))(g_{j+1}(x) - g_j(x)).$$

An application of the triangle inequality gives

$$|G_l(x) - G_k(x)| = |(F_l(x) - F_k(x))|\,|G_1(x)| + \sum_{j=k+1}^{l} \left|(F_l(x) - F_j(x))\right| (g_{j+1}(x) - g_j(x)),$$

since $|g_{j+1}(x) - g_j(x)| = g_{j+1}(x) - g_j(x)$. Now, given $\epsilon \in \mathbb{R}_{>0}$, let $N \in \mathbb{Z}_{>0}$ have the property that

$$|F_l(x) - F_k(x)| \leq \frac{\epsilon}{3M}$$

for all $k, l \geq N$. Then we have

$$\begin{aligned} |G_l(x) - G_k(x)| &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3M} \sum_{j=k+1}^{l} (g_{j+1}(x) - g_j(x)) \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3M}(g_{k+1}(x) - g_{l+1}(x)) \\ &\leq \frac{\epsilon}{3} + \frac{\epsilon}{3M}(|g_{k+1}(x)| + |g_{l+1}(x)|) \leq \epsilon. \end{aligned}$$

Thus the sequence $(G_j)_{j\in\mathbb{Z}_{>0}}$ is uniformly Cauchy, and hence uniformly convergent. ■

The final result on general uniform convergence we present is the Dirichlet test.[13]

3.6.17 Theorem (Dirichlet's test) *Let* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *and* $(\mathrm{g_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be sequences of* $\mathbb{R}$*-valued functions on an interval* I *and satisfying the following conditions:*

(i) *there exists* $\mathrm{M} \in \mathbb{R}_{>0}$ *such that the partial sums*

$$\mathrm{F_k(x)} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{f_j(x)}$$

satisfy $|\mathrm{F_k(x)}| \leq \mathrm{M}$ *for all* $\mathrm{k} \in \mathbb{Z}_{>0}$ *and* $\mathrm{x} \in \mathrm{I}$*;*

(ii) $\mathrm{g_j(x)} \geq 0$ *for all* $\mathrm{j} \in \mathbb{Z}_{>0}$ *and* $\mathrm{x} \in \mathrm{I}$*;*

(iii) $\mathrm{g_{j+1}(x)} \leq \mathrm{g_j(x)}$ *for all* $\mathrm{j} \in \mathbb{Z}_{>0}$ *and* $\mathrm{x} \in \mathrm{I}$*;*

(iv) *the sequence* $(\mathrm{g_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *converges uniformly to the zero function.*

Then the series $\sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j g_j}$ *converges uniformly on* I*.*

Proof We denote

$$F_k(x) = \sum_{j=1}^{k} f_j(x), \quad G_k(x) = \sum_{j=1}^{k} f_j(x)g_j(x).$$

We use again the Abel partial summation formula, Proposition 2.4.16, to write

$$G_l(x) - G_k(x) = F_l(x)g_{l+1}(x) - F_k(x)g_{k+1}(x) - \sum_{j=k+1}^{l} F_j(x)(g_{l+1}(x) - g_l(x)).$$

Now we compute

$$\begin{aligned} |G_l(x) - G_k(x)| &\leq M(g_{l+1}(x) + g_{k+1}(x)) + M \sum_{j=k+1}^{l} (g_j(x) - g_{j+1}(x)) \\ &= 2Mg_{k+1}(x). \end{aligned}$$

[13] Johann Peter Gustav Lejeune Dirichlet 1805–1859 was born in what is now Germany. His mathematical work was primarily in the areas of analysis, number theory and mechanics. For the purposes of these volumes, Dirichlet was gave the first rigorous convergence proof for the trigonometric series of Fourier. These and related results are presented in Section IV-5.2.

Now, for $\epsilon \in \mathbb{R}_{>0}$, if one chooses $N \in \mathbb{Z}_{>0}$ such that $g_k(x) \leq \frac{\epsilon}{2M}$ for all $x \in I$ and $k \geq N$, then it follows that $|G_l(x) - G_k(x)| \leq \epsilon$ for $k, l \geq N$ and for all $x \in I$. From this we deduce that the sequence of partial sums $(G_j)_{j \in \mathbb{Z}_{>0}}$ is uniformly Cauchy, and hence uniformly convergent. ■

3.6.6 The Weierstrass Approximation Theorem

In this section we prove an important result in analysis. The theorem is one on approximating continuous functions with a certain class of easily understood functions. The idea, then, is that if one say something about the class of easily understood functions, it may be readily also ascribed to continuous functions. Let us first describe the class of functions we wish to use to approximate continuous functions.

3.6.18 Definition (Polynomial functions) A function $P\colon \mathbb{R} \to \mathbb{R}$ is a ***polynomial function*** if

$$P(x) = a_k x^k + \cdots + a_1 x + a_0$$

for some $a_0, a_1, \dots, a_k \in \mathbb{R}$. The ***degree*** of the polynomial function P is the largest $j \in \{0, 1, \dots, k\}$ for which $a_j \neq 0$. •

We shall have a great deal to say about polynomials in an algebraic setting in Section 4.4. Here we will only think about the most elementary features of polynomials.

Our constructions are based on a special sort of polynomial. We recall the notation

$$\binom{m}{k} \triangleq \frac{m!}{k!(m-k)!}$$

which are the ***binomial coefficients***.

3.6.19 Definition (Bernstein polynomial, Bernstein approximation) For $m \in \mathbb{Z}_{\geq 0}$ and $k \in \{0, 1, \dots, m\}$ the polynomial function

$$P_k^m(x) = \binom{m}{k} x^k (1-x)^{m-k}$$

is a ***Bernstein polynomial***. For a continuous function $f\colon [a, b] \to \mathbb{R}$ the **m*th Bernstein approximation*** of f is the function $B_m^{[a,b]} f\colon [a, b] \to \mathbb{R}$ defined by

$$B_m^{[a,b]} f(x) = \sum_{k=0}^{m} f(a + \tfrac{k}{m}(b-a)) P_k^m(\tfrac{x-a}{b-a}).$$ •

In Figure 3.14 we depict some of the Bernstein polynomials. The way to imagine the point of these functions is as follows. The polynomial P_k^m on the interval $[0, 1]$ has a single maximum at $\frac{k}{m}$. By letting m vary over $\mathbb{Z}_{\geq 0}$ and letting $k \in \{0, 1, \dots, m\}$,

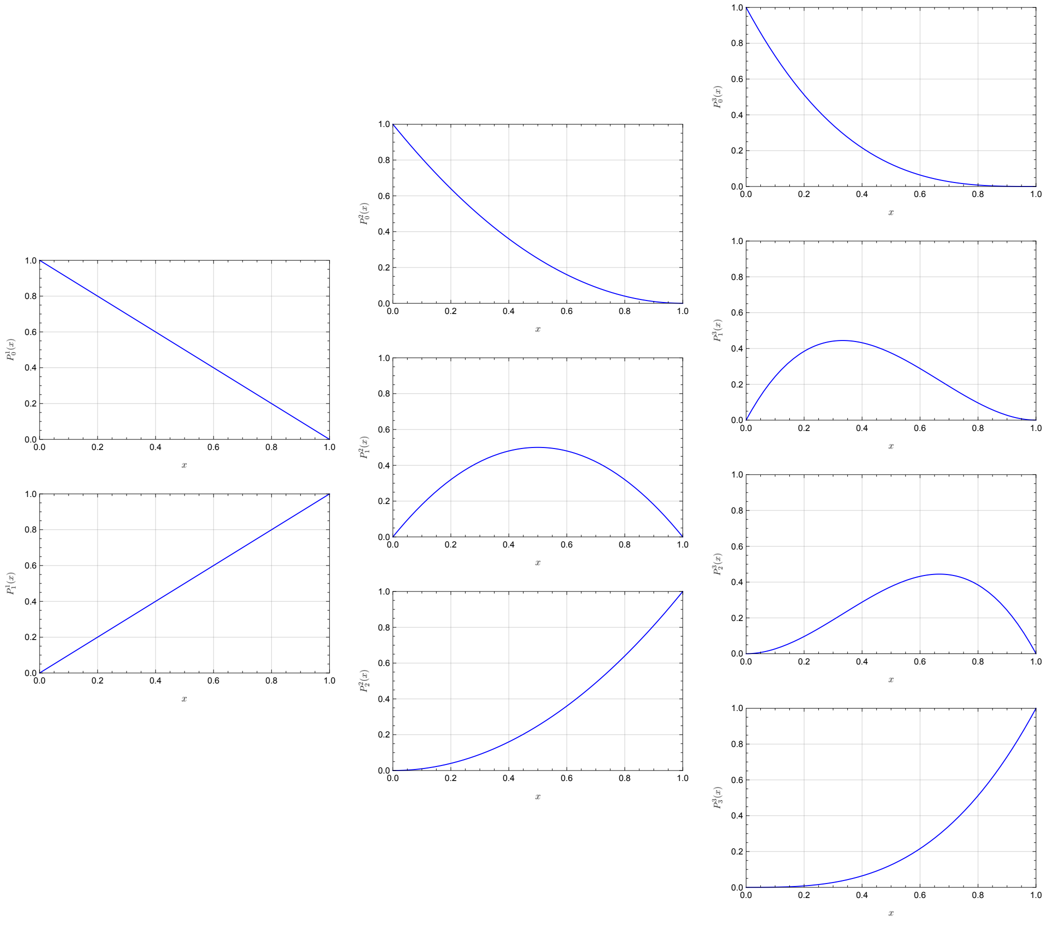


Figure 3.14 The Bernstein polynomials P_0^1 and P_1^1 (left), P_0^2, P_1^2, and P_2^2 (middle), and P_0^3, P_1^3, P_2^3, and P_3^3 (right)

the points of the form $\frac{k}{m}$ will get arbitrarily close to any point in $[0,1]$. The function $f(\frac{k}{m})P_k^m$ thus has a maximum at $\frac{k}{m}$ and the behaviour of f away from $\frac{k}{m}$ is thus (sort of) attenuated. In fact, for large m the behaviour of the function P_k^m becomes increasingly "focussed" at $\frac{k}{m}$. Thus, as m gets large, the function $f(\frac{k}{m})P_k^m$ starts looking like the function taking the value $f(\frac{k}{m})$ at $\frac{k}{m}$ and zero elsewhere. Now, using the identity

$$\sum_{k=0}^{m} \binom{m}{k} x^k(1-x)^m = 1 \tag{3.21}$$

which can be derived using the Binomial Theorem (see Exercise 2.2.1), this means that for large m, $B_m^{[0,1]}f(\frac{k}{m})$ approaches the value $f(\frac{k}{m})$. This is the idea of the Bernstein approximation.

That being said, let us prove some basic facts about Bernstein approximations.

3.6.20 Lemma (Properties of Bernstein approximations) *For continuous functions* $\mathrm{f}, \mathrm{g}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$*, for* $\alpha \in \mathbb{R}$*, and for* $\mathrm{m} \in \mathbb{Z}_{\geq 0}$*, the following statements hold:*

(i) $\mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}(\mathrm{f}+\mathrm{g}) = \mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{f} + \mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{g}$*;*

(ii) $\mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}(\alpha\mathrm{f}) = \alpha\mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{f}$*;*

(iii) $\mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{f}(\mathrm{x}) \geq 0$ *for all* $\mathrm{x} \in [\mathrm{a}, \mathrm{b}]$ *if* $\mathrm{f}(\mathrm{x}) \geq 0$ *for all* $\mathrm{x} \in [\mathrm{a}, \mathrm{b}]$*;*

(iv) $\mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{f}(\mathrm{x}) \leq \mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{g}(\mathrm{x})$ *for all* $\mathrm{x} \in [\mathrm{a}, \mathrm{b}]$ *if* $\mathrm{f}(\mathrm{x}) \leq \mathrm{g}(\mathrm{x})$ *for all* $\mathrm{x} \in [\mathrm{a}, \mathrm{b}]$*;*

(v) $|\mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{f}(\mathrm{x})| \leq \mathrm{B}_{\mathrm{m}}^{[\mathrm{a},\mathrm{b}]}\mathrm{g}(\mathrm{x})$ *for all* $\mathrm{x} \in [\mathrm{a}, \mathrm{b}]$ *if* $|\mathrm{f}(\mathrm{x})| \leq \mathrm{g}(\mathrm{x})$ *for all* $\mathrm{x} \in [\mathrm{a}, \mathrm{b}]$*;*

(vi) for $\mathrm{k}, \mathrm{m} \in \mathbb{Z}_{\geq 0}$ *we have*

$$(\mathrm{B}_{\mathrm{m}+\mathrm{k}}^{[\mathrm{a},\mathrm{b}]})^{(\mathrm{k})}(\mathrm{x}) = \frac{(\mathrm{m}+\mathrm{k})!}{\mathrm{m}!}\frac{1}{(\mathrm{b}-\mathrm{a})^{\mathrm{k}}}\sum_{\mathrm{j}=0}^{\mathrm{m}} \Delta_{\mathrm{h}}^{\mathrm{k}}\mathrm{f}(\mathrm{a} + \tfrac{\mathrm{j}}{\mathrm{k}+\mathrm{m}}(\mathrm{b}-\mathrm{a}))\mathrm{P}_{\mathrm{j}}^{\mathrm{m}}(\tfrac{\mathrm{x}-\mathrm{a}}{\mathrm{b}-\mathrm{a}}),$$

where $\mathrm{h} = \frac{1}{\mathrm{k}+\mathrm{m}}$ *and where* $\Delta_{\mathrm{h}}^{\mathrm{k}}\mathrm{f}\colon [\mathrm{a}, \mathrm{b}] \to \mathbb{R}$ *is defined by*

$$\Delta_{\mathrm{h}}^{\mathrm{k}}\mathrm{f}(\mathrm{x}) = \sum_{\mathrm{j}=0}^{\mathrm{k}} (-1)^{\mathrm{k}-\mathrm{j}}\binom{\mathrm{k}}{\mathrm{j}}\mathrm{f}(\mathrm{x}+\mathrm{jh})$$

(vii) if we define $\mathrm{f}_0, \mathrm{f}_1, \mathrm{f}_2\colon [0, 1] \to \mathbb{R}$ *by*

$$\mathrm{f}_0(\mathrm{x}) = 1, \quad \mathrm{f}_1(\mathrm{x}) = \mathrm{x}, \quad \mathrm{f}_2(\mathrm{x}) = \mathrm{x}^2, \qquad \mathrm{x} \in [0, 1],$$

then

$$\mathrm{B}_{\mathrm{m}}^{[0,1]}\mathrm{f}_0(\mathrm{x}) = 1, \quad \mathrm{B}_{\mathrm{m}}^{[0,1]}\mathrm{f}_1(\mathrm{x}) = \mathrm{x}, \quad \mathrm{B}_{\mathrm{m}}^{[0,1]}\mathrm{f}_2(\mathrm{x}) = \mathrm{x}^2 + \tfrac{1}{\mathrm{m}}(\mathrm{x} - \mathrm{x}^2)$$

for $\mathrm{x} \in [0, 1]$ *and* $\mathrm{m} \in \mathbb{Z}_{\geq 0}$.

Proof Let $\hat{f}\colon [0, 1] \to \mathbb{R}$ be defined by $\hat{f}(y) = f(a + \frac{y}{(}b - a))$. One can verify that if the lemma holds for $\hat{f}$ then it immediately follows for f, and so without loss of generality we suppose that $[a, b] = [0, 1]$. We also abbreviate $B_m^{[0,1]} = B_m$.

(i)–(iv) These assertions follow directly from the definition of the Bernstein approximations.

(v) If $|f(x)| \leq g(x)$ for all $x \in [0, 1]$ then

$$\begin{aligned} & -f(x) \leq g(x) \leq f(x), && x \in [0, 1] \\ \implies & -B_m f(x) \leq B_m g(x) \leq B_m f(x), && x \in [0, 1], \end{aligned}$$

using the fourth assertion.

(vi) Note that

$$B_{m+k}(x) = \sum_{j=0}^{m+k} f(\tfrac{j}{m+k})\binom{m+k}{j}x^j(1-x)^{m+k-j}.$$

Let $g_j(x) = x^j$ and $h_j(x) = (1-x)^{m+k-j}$ and compute

$$g_j^{(r)}(x) = \begin{cases} \frac{j!}{(j-r)!}x^{j-r}, & j - r \geq 0, \\ 0, & j - r < 0 \end{cases}$$

and

$$h_j^{(k-r)}(x) = \begin{cases} (-1)^{k-r}\frac{(m+k-j)!}{(m+r-j)!}(1-x)^{m+r-j}, & j-r \leq m, \\ 0, & j-r > m. \end{cases}$$

By Proposition 3.2.11,

$$(g_jh_j)^{(k)}(x) = \sum_{r=0}^{k}\binom{k}{r}g_j^{(r)}(x)h_j^{(k-r)}(x).$$

Also note that

$$\begin{aligned}\binom{m+k}{j}\frac{j!}{(j-r)!}\frac{(m+k-j)!}{(m+r-j)!} &= \frac{(m+k)!}{j!(m+k-j)!}\frac{j!}{(j-r)!}\frac{(m+k-j)!}{(m+r-j)!} \\ &= \frac{(m+k)!}{m!}\frac{m!}{(m-(j-r))!(j-r)!} = \frac{(m+k)!}{m!}\binom{m}{j-r}.\end{aligned}$$

Putting this all together we have

$$\begin{aligned}B_{m+k}^{(k)}(x) &= \sum_{j=0}^{m+k}\sum_{r=0}^{k} f(\tfrac{j}{m+k})\binom{m+k}{j}\binom{k}{r}g_j^{(r)}(x)h_j^{(k-r)}(x) \\ &= \sum_{r=0}^{k}\sum_{l=-r}^{m+k-r} f(\tfrac{l+r}{m+k})\binom{m+k}{l+r}\binom{k}{r}g_{l+r}^{(r)}(x)h_{l+r}^{(k-r)}(x) \\ &= \sum_{r=0}^{k}\sum_{l=0}^{m}(-1)^{k-r}\binom{k}{r}f(\tfrac{l+r}{m+k})\binom{m}{l}x^l(1-x)^{n-l},\end{aligned}$$

where we make the change of index $(l,r) = (j-r,r)$ in the second step and note that the derivatives of g_{l+r} and h_{l+r} vanish when $l < 0$ and $l > m$. Let $h = \frac{1}{m+k}$. Since

$$\Delta_h^k f(\tfrac{j}{m+k}) = \sum_{r=0}^{k}(-1)^{k-r}\binom{k}{r}f(\tfrac{j+r}{m+k})$$

this part of the result follows.

(vii) It follows from (3.21) that $B_mf_0(x) = 1$ for every $x \in [0,1]$. We also compute

$$\begin{aligned}B_mf_0(x) &= \sum_{k=0}^{m}\frac{k}{m}\frac{m!}{m!(m-k)!}x^k(1-x)^{m-k} \\ &= x\sum_{k=0}^{m-1}\frac{(m-1)!}{(k-1)!((m-1)-(k-1))!}x^k(1-x)^{m-1-k} \\ &= x(x+(1-x))^{m-1} = x,\end{aligned}$$

where we use the Binomial Theorem. To compute $B_m f_2$ we first compute

$$\begin{aligned}\frac{k^2}{m^2}\frac{m!}{k!(m-k)!} &= \frac{(k-1)+1}{m}\frac{(m-1)!}{(k-1)!(m-k)!}\\ &= \frac{(k-1)(n-1)}{n(n-1)}\frac{(m-1)!}{(k-1)!(m-k)!} + \frac{1}{m}\frac{(m-1)!}{(k-1)!(m-k)!}\\ &= \frac{m-1}{m}\binom{n-2}{k-2} + \frac{1}{m}\binom{n-1}{k-1},\end{aligned}$$

where we adopt the convention that $\binom{j}{l} = 0$ if either j or l are zero. We now compute

$$\begin{aligned}B_m f_2(x) &= \sum_{k=0}^{m} \frac{k^2}{m^2}\binom{m}{k} x^k(1-x)^{m-k}\\ &= \frac{m-1}{m}\sum_{k=2}^{m}\binom{m-2}{k-2}x^k(1-x)^{m-k} + \frac{1}{m}\sum_{k=1}^{m}\binom{m-1}{k-1}x^k(1-x)^{m-k}\\ &= \frac{m-1}{m}x^2(x+(1-x))^{m-2} + \frac{1}{m}x(x+(1-x))^{m-1} = \frac{m-1}{m}x^2 + \frac{1}{m}x,\end{aligned}$$

as desired. ■

Now, heuristics aside, we state the main result in this section, a consequence of which is that every continuously function on a compact interval can be approximated arbitrarily well (in the sense that the maximum difference can be made as small as desired) by a polynomial function.

3.6.21 Theorem (Weierstrass Approximation Theorem) *Consider a compact interval* $[a, b] \subseteq \mathbb{R}$ *and let* $f\colon [a, b] \to \mathbb{R}$ *be continuous. Then the sequence* $(B_m^{[a,b]} f)_{m\in\mathbb{Z}_{>0}}$ *converges uniformly to* f *on* $[a, b]$*.*

Proof It is evident (why?) that we can take $[a, b] = [0, 1]$ and then let us denote $B_m f = B_m^{[0,1]} f$ for simplicity.

Let $\epsilon \in \mathbb{R}_{>0}$. Since f is uniformly continuous by Theorem 3.1.24 there exists $\delta \in \mathbb{R}_{>0}$ such that $|f(x) - f(y)| \le \frac{\epsilon}{2}$ whenever $|x - y| \le \delta$. Let

$$M = \sup\{|f(x)| \mid x \in [0, 1]\},$$

noting that $M < \infty$ by Theorem 3.1.23. Note then that if $|x - y| \le \delta$ then

$$|f(x) - f(y)| \le \tfrac{\epsilon}{2} \le \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x-y)^2.$$

If $|x - y| > \delta$ then

$$|f(x) - f(y)| \le 2M \le 2M\left(\tfrac{x-y}{\delta}\right)^2 \le \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x-y)^2.$$

That is to say, for every $x, y \in [0, 1]$,

$$|f(x) - f(y)| \le \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x-y)^2. \tag{3.22}$$

Now, fix $x_0 \in [0,1]$ and compute, using the lemma above (along with the notation f_0, f_1, and f_2 introduced in the lemma) and (3.22),

$$\begin{aligned}|B_m f(x) - f(x_0)| = |B_m(f - f(x_0)f_0)(x)| &\le B_m\left(\tfrac{\epsilon}{2} f_0 + \tfrac{2M}{\delta^2}(f_1 - x_0 f_0)^2\right)(x)\\ &= \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x^2 + \tfrac{1}{m}(x - x^2) - 2x_0 x + x_0^2)\\ &= \tfrac{\epsilon}{2} + \tfrac{2M}{\delta^2}(x - x_0)^2 + \tfrac{2M}{m\delta^2}(x - x^2),\end{aligned}$$

this holding for every $m \in \mathbb{Z}_{\ge 0}$. Now evaluate at $x = x_0$ to get

$$|B_m f(x_0) - f(x_0)| \le \tfrac{\epsilon}{2} + \tfrac{2M}{m\delta^2}(x_0 - x_0^2) \le \tfrac{\epsilon}{2} + \tfrac{M}{2m\delta^2},$$

using the fact that $x_0 - x_0^2 \le \frac{1}{4}$ for $x_0 \in [0,1]$. Therefore, if $N \in \mathbb{Z}_{>0}$ is sufficiently large that $\frac{M}{2m\delta^2} < \frac{\epsilon}{2}$ for $m \ge N$ we have

$$|B_m f(x_0) - f(x_0)| < \epsilon,$$

and this holds for every $x_0 \in [0,1]$, giving us the desired uniform convergence. ■

For fun, let us illustrate the Bernstein approximations in an example.

3.6.22 Example (Bernstein approximation) Let us consider $f\colon [0,1] \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} x, & x \in [0, \frac{1}{2}], \\ 1 - x, & x \in (\frac{1}{2}, 1]. \end{cases}$$

In Figure 3.15 we show some Bernstein approximations to f. Note that the convergence is rather poor. One might wish to contrast the 100th approximation in Figure 3.15 with the 10 approximation of the same function using Fourier series depicted in Figure IV-5.11. (If you have no clue what a Fourier series is, that is fine. We will get there in time.) •

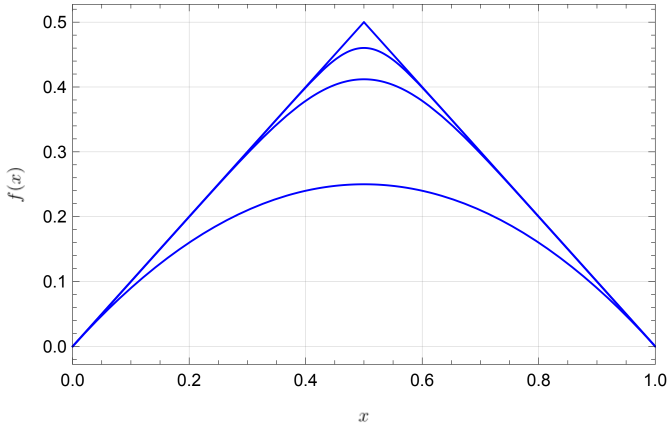


Figure 3.15 Bernstein approximations for $m \in \{2, 50, 100\}$

We shall revisit the Weierstrass Approximation Theorem in Sections II-1.7.2.

3.6.7 Swapping limits with other operations

In this section we give some basic result concerning the swapping of various function operations with limits. The first result we consider pertains to integration. When we consider Lebesgue integration in Chapter III-2 we shall see that there are more powerful limit theorems available. Indeed, the *raison d'etre* for the Lebesgue integral is just these limit theorems, as these are not true for the Riemann integral. However, for the moment these theorems have value in that they apply in at least some cases, and indicate what *is* true for the Riemann integral.

3.6.23 Theorem (Uniform limits commute with Riemann integration) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $(\mathrm{f}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of continuous* $\mathbb{R}$*-valued functions defined on* $[\mathrm{a}, \mathrm{b}]$ *that converge uniformly to* f*. Then*

$$\lim_{\mathrm{j}\to\infty} \int_\mathrm{a}^\mathrm{b} \mathrm{f}_\mathrm{j}(\mathrm{x})\,\mathrm{dx} = \int_\mathrm{a}^\mathrm{b} \mathrm{f}(\mathrm{x})\,\mathrm{dx}.$$

Proof As the functions $(f_j)_{j\in\mathbb{Z}_{>0}}$ are continuous and the convergence to f is uniform, f must be continuous by Theorem 3.6.8. Since the interval $[a, b]$ is compact, the functions f and f_j, $j \in \mathbb{Z}_{>0}$, are also bounded. Therefore, by Proposition 3.4.25,

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \le M(b-a)$$

where $M = \sup\{|f(x)| \mid x \in [a, b]\}$. Let $\epsilon \in \mathbb{R}_{>0}$ and select $N \in \mathbb{Z}_{>0}$ such that $|f_j(x) - f(x)| < \frac{\epsilon}{b-a}$ for all $x \in [a, b]$, provided that $j \ge N$. Then

$$\begin{aligned}\left|\int_a^b f_j(x)\,\mathrm{d}x - \int_a^b f(x)\,\mathrm{d}x\right| &= \left|\int_a^b (f_j(x) - f(x))\,\mathrm{d}x\right| \\ &\le \frac{\epsilon}{b-a}(b-a) = \epsilon.\end{aligned}$$

This is the desired result. ■

Next we state a result that tells us when we may switch limits and differentiation.

3.6.24 Theorem (Uniform limits commute with differentiation) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval and let* $(\mathrm{f}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence continuously differentiable* $\mathbb{R}$*-valued functions on* $[\mathrm{a}, \mathrm{b}]$*, and suppose that the sequence converges pointwise to* f*. Also suppose that the sequence* $(\mathrm{f}'_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *of derivatives converges uniformly to* g*. Then* f *is differentiable and* $\mathrm{f}' = \mathrm{g}$*.*

Proof Our hypotheses ensure that we may write, for each $j \in \mathbb{Z}_{>0}$,

$$f_j(x) = f_j(a) + \int_a^x f'_j(\xi)\,\mathrm{d}\xi.$$

for each $x \in [a, b]$. By Theorem 3.6.23, we may interchange the limit as $j \to \infty$ with the integral, and so we get

$$f(t) = f(a) + \int_a^x g(\xi)\,d\xi.$$

Since g is continuous, being the uniform limit of continuous functions (by Theorem 3.6.8), the Fundamental Theorem of Calculus ensures that $f' = g$. ■

The next result in this section has a somewhat different character than the rest. It actually says that it is possible to differentiate a sequence of monotonically increasing functions term-by-term, except on a set of measure zero. The interesting thing here is that only pointwise convergence is needed.

3.6.25 Theorem (Termwise differentiation of sequences of monotonic functions is a.e. valid) *Let* $\mathrm{I} = [\mathrm{a}, \mathrm{b}]$ *be a compact interval, let* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of monotonically increasing functions such that the series* $\mathrm{S} = \sum_{\mathrm{j}=1}^{\infty} \mathrm{f_j}(\mathrm{x})$ *converges pointwise to a function* f. *Then there exists a set* $\mathrm{Z} \subseteq \mathrm{I}$ *such that*

(i) Z *has measure zero and*

(ii) $\mathrm{f}'(\mathrm{x}) = \sum_{\mathrm{j}=1}^{\infty} \mathrm{f}'_{\mathrm{j}}(\mathrm{x})$ *for all* $\mathrm{x} \in \mathrm{I} \setminus \mathrm{Z}$.

Proof Note that the limit function f is monotonically increasing. Denote by $Z_1 \subseteq [a, b]$ the set of points for which all of the functions f and f_j, $j \in \mathbb{Z}_{>0}$, do not possess derivatives. Note that by Theorem 3.2.26 it follows that Z_1 is a countable union of sets of measure zero. Therefore, by Exercise 2.5.11, Z_1 has measure zero. Now let $x \in I \setminus Z_1$ and let $\epsilon \in \mathbb{R}_{>0}$ be sufficiently small that $x + \epsilon \in [a, b]$. Then

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} = \sum_{j=1}^{\infty} \frac{f_j(x+\epsilon) - f_j(x)}{\epsilon}.$$

Since $f_j(x + \epsilon) - f_j(x) \geq 0$, for any $k \in \mathbb{Z}_{>0}$ we have

$$\frac{f(x+\epsilon) - f(x)}{\epsilon} \geq \sum_{j=1}^{k} \frac{f_j(x+\epsilon) - f_j(x)}{\epsilon},$$

which then gives

$$f'(x) \geq \sum_{j=1}^{k} f'_j(x).$$

The sequence of partial sums for the series $\sum_{j=1}^{\infty} f'_j(x)$ is therefore bounded above. Moreover, by Theorem 3.2.26, it is increasing. Therefore, by Theorem 2.3.8 the series $\sum_{j=1}^{\infty} f'_j(x)$ converges for every $x \in I \setminus Z_1$.

Let us now suppose that $f(a) = 0$ and $f_j(a) = 0$, $j \in \mathbb{Z}_{>0}$. This can be done without loss of generality by replacing f with $f - f(a)$ and f_j with $f_j - f_j(a)$, $j \in \mathbb{Z}_{>0}$. With this assumption, for each $x \in [a, b]$ and $k \in \mathbb{Z}_{>0}$, we have $f(x) - S_k(x) \geq 0$ where $(S_k)_{k\in\mathbb{Z}_{>0}}$ is the sequence of partial sums for S. Choose a subsequence $(S_{k_l})_{l\in\mathbb{Z}_{>0}}$ of $(S_k)_{k\in\mathbb{Z}_{>0}}$

having the property that $0 \leq f(b) - S_{k_l}(b) \leq 2^{-l}$, this being possible since the sequence $(S_k(b))_{k\in\mathbb{Z}_{>0}}$ converges to $f(b)$. Note that

$$f(x) - S_{k_l}(x) = \sum_{j=k_l+1}^{\infty} f_j(x),$$

meaning that $f - S_{k_l}$ is a monotonically increasing function. Therefore, $0 \leq f(x) - S_{k_l}(x) \leq 2^{-l}$ for all $x \in [a, b]$. This shows that the series $\sum_{l=1}^{\infty}(f(x) - S_{k_l}(x))$ is a pointwise convergent sequence of monotonically increasing functions. Let g denote the limit function, and let $Z_2 \subseteq [a, b]$ be the set of points where all of the functions g and $f - S_{k_l}$, $l \in \mathbb{Z}_{>0}$, do not possess derivatives, noting that this set is, in the same manner as was Z_1, a set of measure zero. The argument above applies again to show that, for $x \in I \setminus Z_2$, the series $\sum_{l=1}^{\infty}(f'(x) - S'_{k_l}(x))$ converges. Thus, for $x \in I \setminus Z_2$, it follows that $\lim_{l\to\infty}(f'(x) - S'_{k_l}(x)) = 0$. Now, for $x \in I \setminus Z_1$, we know that $(S'_k(x))_{k\in\mathbb{Z}_{>0}}$ is a monotonically increasing sequence. Therefore, for $x \in I \setminus (Z_1 \cup Z_2)$, the sequence $(f'(x) - S'_k(x))_{k\in\mathbb{Z}_{>0}}$ must converge to zero. This gives the result by taking $Z = Z_1 \cup Z_2$. ■

As a final result, we indicate how convexity interacts with pointwise limits.

3.6.26 Theorem (The pointwise limit of convex functions is convex) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is convex and if* $(\mathrm{f_j})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *is a sequence of convex functions converging pointwise to* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$*, then* f *is convex.*

Proof Let $x_1, x_2 \in I$ and let $s \in [0, 1]$. Then

$$\begin{aligned} f((1-s)x_1 + sx_2) &= \lim_{j\to\infty} f_j((1-s)x_1 + sx_2) \leq \lim_{j\to\infty}((1-s)f_j(x_1) + sf_j(x_2)) \\ &= (1-s)\lim_{j\to\infty} f_j(x_1) + s\lim_{j\to\infty} f_j(x_2) \\ &= (1-s)f(x_1) + sf(x_2), \end{aligned}$$

where we have used Proposition 2.3.23. ■

3.6.8 Notes

There are many proofs available of the Weierstrass Approximation Theorem, and the rather explicit proof we give is due to Bernstein [1912].

Exercises

3.6.1 Consider the sequence of functions $\{f_j\}_{j\in\mathbb{Z}_{>0}}$ defined on the interval $[0, 1]$ by $f_j(x) = x^{1/2^j}$. Thus

$$f_1(x) = \sqrt{x}, \quad f_2(x) = \sqrt{f_1(x)} = \sqrt{\sqrt{x}}, \quad \ldots, \quad f_j(x) = \sqrt{f_{j-1}(x)} = x^{1/2^j}, \ldots$$

(a) Sketch the graph of f_j for $j \in \{1, 2, 3\}$.

(b) Does the sequence of functions $(f_j)_{j\in\mathbb{Z}_{>0}}$ converge pointwise? If so, what is the limit function?

(c) Is the convergence of the sequence of functions $(f_j)_{j\in\mathbb{Z}_{>0}}$ uniform?

(d) Is it true that

$$\lim_{j\to\infty}\int_0^1 f_j(x)\,\mathrm{d}x = \int_0^1 \lim_{j\to\infty} f_j(x)\,\mathrm{d}x?$$

3.6.2 In each of the following exercises, you will be given a sequence of functions defined on the interval $[0,1]$. In each case, answer the following questions.

1. Sketch the first few functions in the sequence.
2. Does the sequence converge pointwise? If so, what is the limit function?
3. Does the sequence converge uniformly?

The sequences are as follows:

(a) $(f_j(x) = (x - \frac{1}{j^2})^2)_{j\in\mathbb{Z}_{>0}}$;

(b) $(f_j(x) = x - x^j)_{j\in\mathbb{Z}_{>0}}$.

3.6.3 Let $I \subseteq \mathbb{R}$ be an interval and let $(f_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence of locally bounded functions on I converging pointwise to $f\colon I \to \mathbb{R}$. Show that there exists a function $g\colon I \to \mathbb{R}$ such that $(f_j)_{j\in\mathbb{Z}_{>0}}$ converges dominated by g.

Section 3.7

ℝ-power series

In Section 3.6.4 we considered the convergence of general series of functions. In this section we consider special series of functions where the functions in the series are given by $f_j(x) = a_j x^j$, $j \in \mathbb{Z}_{\geq 0}$. This class of series is important in a surprising number of ways. For example, as we shall see in Section 3.7.4, one can associate a power series to every function of class C^∞, and this power series sometimes approximates the function in some sense.

Do I need to read this section? The material in this section is of a somewhat technical character, and so can probably be skipped until it is needed. One of the main uses will occur in Section II-3.3 when we explore the intimate relationship between power series and analytic functions in complex analysis. There will also be occasions throughout these volumes when it is convenient to use Taylor's Theorem. •

3.7.1 ℝ-formal power series

We begin with a discussion that is less analytical, and more algebraic in flavour. This discussion serves to separate the simpler algebraic features of power series from the more technical analytical features. A purely logical presentation of this material would certain present the material Section 4.4 before our present discussion. However, we have decided to make a small sacrifice in logic for the sake of organisation. Readers wishing to preserve the logical structure may wish to look ahead at this point to Section 4.4.

Let us first give a formal definition of what we mean by a ℝ-formal power series, while at the same time defining the operations of addition and multiplication in this set.

3.7.1 Definition (ℝ-formal power series) A ℝ-***formal power series*** is a sequence $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ in ℝ. If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ are two ℝ-formal power series, then define ℝ-formal power series $A + B$ and $A \cdot B$ by

$$A + B = (a_j + b_j)_{j\in\mathbb{Z}_{\geq 0}}, \quad A \cdot B = \left(\sum_{j=0}^{k} a_j b_{k-j}\right)_{k\in\mathbb{Z}_{\geq 0}},$$

which are the ***sum*** and ***product*** of A and B, respectively. If $\alpha \in \mathbb{R}$ then αA denotes the ℝ-formal power series $(\alpha a_j)_{j\in\mathbb{Z}_{\geq 0}}$ which is the ***product*** of α and A. •

In order to distinguish between multiplication of two ℝ-formal power series and multiplication of a ℝ-formal power series by a real number, we shall call the latter ***scalar multiplication***. This is reflective of the idea of a vector space that we introduce in Section 4.5. Note that the product of $\mathbb{R}$-formal power series is very much related to the Cauchy product of series in Definition 2.4.29. As we shall see, this is not surprising given the natural manner of thinking about $\mathbb{R}$-formal power series.

Our definition of $\mathbb{R}$-formal power series is meant to be rigorous, but suffers from being at the same time obtuse. A less obtuse working definition is possible, and requires the following notion.

3.7.2 Definition (Indeterminate) The ***indeterminate*** in the set of $\mathbb{R}$-formal power series is the element $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_j = \begin{cases} 1, & j = 1, \\ 0, & \text{otherwise.} \end{cases}$$

If the indeterminate is denoted by the symbol ξ, then $\mathbb{R}[[\xi]]$ denotes the ***set of*** $\mathbb{R}$***-formal power series in indeterminate*** ξ. •

Now let us see what are the notational implications of introducing the indeterminate into the picture. A direct application of the definition of the product shows that, if the indeterminate is denoted by ξ and if $k \in \mathbb{Z}_{>0}$, then ξ^k (the k-fold product of ξ with itself) is the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ given by

$$a_j = \begin{cases} 1, & j = k, \\ 0, & \text{otherwise.} \end{cases}$$

Let us adopt the convention that ξ^0 denotes the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_j = \begin{cases} 1, & j = 0, \\ 0, & j \in \mathbb{Z}_{>0}. \end{cases}$$

Now let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be an *arbitrary* $\mathbb{R}$-formal power series and, for $k \in \mathbb{Z}_{\geq 0}$, let A_k denote the $\mathbb{R}$-formal power series $(a_{k,j})_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_{k,j} = \begin{cases} a_j, & j \leq k, \\ 0, & j > k. \end{cases}$$

Note that, using the definition of

$$\begin{aligned} A_k &= (a_0, a_1, \dots, a_k, 0, \dots) \\ &= (a_0, 0, \dots, 0, 0, \dots) + (0, a_1, \dots, 0, 0, \dots) + \cdots + (0, 0, \dots, a_k, 0, \dots) \\ &= a_0\xi^1 + a_a\xi^1 + \cdots + a_k\xi^k. \end{aligned}$$

We would now like to write $A = \lim_{k\to\infty} A_k$, but the problem is that we do not really know what the limit means in this case. It certainly does not mean the limit thinking of the sum as one of real numbers; this limit will generally not exist. Thus we define what the limit means as follows.

3.7.3 Definition (Limit of ℝ-formal power series) Let $(A_k = (a_{k,j})_{j\in\mathbb{Z}_{\geq 0}})_{k\in\mathbb{Z}_{\geq 0}}$ be a sequence of ℝ-formal power series and let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a ℝ-formal power series. The sequence $(A_k)_{k\in\mathbb{Z}_{\geq 0}}$ ***converges*** to A, and we write $A = \lim_{k\to\infty} A_k$, if, for each $j \in \mathbb{Z}_{\geq 0}$, there exists $N_j \in \mathbb{Z}_{\geq 0}$ such that $a_{k,j} = a_j$ for $k \geq N_j$. •

With this notion of convergence in the set of ℝ-formal power series we can prove what we want.

3.7.4 Proposition (ℝ-formal power series as limits of finite sums) *If* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* ℝ*-formal power series, then*

$$\mathrm{A} = \lim_{\mathrm{k}\to\infty} \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{a_j}\xi^{\mathrm{j}}.$$

Proof Let $A_k = \sum_{j=0}^{k} a_j\xi^j$ and denote $A_k = (a_{k,j})_{j\in\mathbb{Z}_{\geq 0}}$. For $j \in \mathbb{Z}_{\geq 0}$ note that $a_{k,j} = a_j$ for $k \geq j$, which gives the condition that $(A_k)_{k\in\mathbb{Z}_{\geq 0}}$ converge to A by taking $N_j = j$ in the definition. ■

The upshot of the preceding exceedingly ponderous discussion is that we can write the ℝ-formal power $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ as

$$\sum_{j=0}^{\infty} a_j\xi^j,$$

and all of the symbols in this expression make exact sense. Moreover, with this representation of a ℝ-formal power series, addition is merely the addition of the coefficients of like powers of the indeterminate. Multiplication is to be interpreted as follows. Suppose that one wishes to find the coefficient of ξ^k in the product $A \cdot B$. One does this by writing, in indeterminate form, the first $k + 1$ terms in A and B, and multiplying them using the usual rules for multiplication of finite sums in ℝ. Thus we write

$$A_k = \sum_{j=0}^{k} a_j\xi^j, \quad B_k = \sum_{j=0}^{k} b_j\xi^j,$$

and compute

$$A_k \cdot B_k = \sum_{l=0}^{2k} \sum_{j=0}^{l} a_j b_{l-j}\xi^j$$

(this formula is easily proved, cf. Theorem 4.4.2). One then can see that the coefficient of ξ^k in this expression is exactly the $(k+1)$st term in the sequence $A \cdot B$.

Let us present the basic properties of the operations of addition and multiplication of ℝ-formal power series. To do this, we let $0_{\mathbb{R}[[\xi]]}$ denote the ℝ-formal power series $(0)_{j\in\mathbb{Z}_{\geq 0}}$ and we let $1_{\mathbb{R}[[\xi]]}$ denote the ℝ-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ given by

$$a_j = \begin{cases} 1, & j = 0, \\ 0, & j \in \mathbb{Z}_{>0}. \end{cases}$$

If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a $\mathbb{R}$-formal power series, then we let $-A$ denote the $\mathbb{R}$-formal power series $(-a_j)_{j\in\mathbb{Z}_{\geq 0}}$. If $a_0 \neq 0$ then we define the $\mathbb{R}$-formal power series $A^{-1} = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ by inductively defining

$$\begin{aligned} b_0 &= \tfrac{1}{a_0}, \\ b_1 &= \tfrac{1}{a_0}(-a_1 b_0), \\ &\vdots \\ b_k &= -\frac{1}{a_0}\sum_{j=1}^{k} a_j b_{k-j}, \\ &\vdots \end{aligned}$$

With these definitions, the following result is straightforward to prove, and follows from our discussion of polynomials in Section 4.4.

3.7.5 Proposition (Properties of addition and multiplication of $\mathbb{R}$-formal power series) *Let $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$, $\mathrm{B} = (\mathrm{b_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$, and $\mathrm{C} = (\mathrm{c_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ be $\mathbb{R}$-formal power series. Then the following statements hold:*

(i) $\mathrm{A} + \mathrm{B} = \mathrm{B} + \mathrm{A}$ *(**commutativity** of addition);*

(ii) $(\mathrm{A} + \mathrm{B}) + \mathrm{C} = \mathrm{A} + (\mathrm{B} + \mathrm{C})$ *(**associativity** of addition);*

(iii) $\mathrm{A} + 0_{\mathbb{R}[[\xi]]} = \mathrm{A}$ *(**additive identity**);*

(iv) $\mathrm{A} + (-\mathrm{A}) = 0_{\mathbb{R}[[\xi]]}$ *(**additive inverse**);*

(v) $\mathrm{A} \cdot \mathrm{B} = \mathrm{B} \cdot \mathrm{A}$ *(**commutativity** of multiplication);*

(vi) $(\mathrm{A} \cdot \mathrm{B}) \cdot \mathrm{C} = \mathrm{A} \cdot (\mathrm{B} \cdot \mathrm{C})$ *(**associativity** of multiplication);*

(vii) $\mathrm{A} \cdot (\mathrm{B} + \mathrm{C}) = \mathrm{A} \cdot \mathrm{B} + \mathrm{A} \cdot \mathrm{C}$ *(**left distributivity**);*

(viii) $(\mathrm{A} + \mathrm{B}) \cdot \mathrm{C} = \mathrm{A} \cdot \mathrm{C} + \mathrm{B} \cdot \mathrm{C}$ *(**right distributivity**);*

(ix) $\mathrm{A} \cdot 1_{\mathbb{R}[[\xi]]} = \mathrm{A}$ *(**multiplicative identity**);*

(x) if $\mathrm{a}_0 \neq 0$ *then* $\mathrm{A} \cdot \mathrm{A}^{-1} = 1_{\mathbb{R}[[\xi]]}$ *(**multiplicative inverse**).*

Proof With the exception of the multiplicative inverse, these properties all follow in the same manner as for polynomials as proved in Theorem 4.4.2. The formula for the multiplicative inverse arises from writing down the elements in the equation $A \cdot A^{-1} = 1_{\mathbb{R}[[\xi]]}$, and solving recursively for the unknown elements of the sequence A^{-1}, starting with the zeroth term. ■

The preceding properties of addition and scalar multiplication can be summarised in the language of Section 4.2 by saying that $\mathbb{R}[[\xi]]$ is a ring. Note that the multiplicative inverse of a formal $\mathbb{R}$-power series does not always exist, even when $A \neq 0_{\mathbb{R}[[\xi]]}$.

For multiplication of a $\mathbb{R}$-formal power series by a real number, we have the following properties.

3.7.6 Proposition (Properties of scalar multiplication of ℝ-formal power series) *Let* $\mathrm{A} = (\mathrm{a}_j)_{j \in \mathbb{Z}_{\geq 0}}$ *and* $\mathrm{B} = (\mathrm{b}_j)_{j \in \mathbb{Z}_{\geq 0}}$ *be ℝ-formal power series and let* $\alpha, \beta \in \mathbb{R}$*. Then the following statements hold:*

(i) $\alpha(\beta\,\mathrm{A}) = (\alpha\beta)\mathrm{A}$ *(**associativity**);*

(ii) $1\,\mathrm{A} = \mathrm{A}$*;*

(iii) $\alpha(\mathrm{A} + \mathrm{B}) = \alpha\,\mathrm{A} + \alpha\,\mathrm{B}$ *(**distributivity**);*

(iv) $(\alpha + \beta)\mathrm{A} = \alpha\,\mathrm{A} + \beta\,\mathrm{B}$ *(**distributivity** again).*

Proof These all follow directly from the definition of scalar multiplication and the properties of addition and multiplication in ℝ as given in Proposition 2.2.4. ■

According to the terminology of Section 4.5, the preceding result, along with the properties of addition from Proposition 3.7.5, ensure that $\mathbb{R}[[\xi]]$ is a ℝ-vector space. With the additional structure given by the product, we further see that $\mathbb{R}[[\xi]]$ is, in fact, a commutative and associative ℝ-algebra.

In terms of our definition of convergence in $\mathbb{R}[[\xi]]$, one has the following properties of addition, multiplication, and scalar multiplication.

3.7.7 Proposition (Sums and products, and convergence in $\mathbb{R}[[\xi]]$) *Let* $(\mathrm{A}_k = (\mathrm{a}_{k,j})_{j \in \mathbb{Z}_{\geq 0}})_{k \in \mathbb{Z}_{>0}}$ *and* $(\mathrm{B}_k = (\mathrm{b}_{k,j})_{j \in \mathbb{Z}_{\geq 0}})_{k \in \mathbb{Z}_{>0}}$ *be sequences of ℝ-formal power series converging to the ℝ-formal power series* $\mathrm{A} = (\mathrm{a}_j)_{j \in \mathbb{Z}_{\geq 0}}$ *and* $\mathrm{B} = (\mathrm{b}_j)_{j \in \mathbb{Z}_{\geq 0}}$*, respectively, and let* $\alpha \in \mathbb{R}$*. Then the following statements hold:*

(i) $\lim_{k \to \infty}(\mathrm{A}_k + \mathrm{B}_k) = \mathrm{A} + \mathrm{B}$*;*

(ii) $\lim_{k \to \infty}(\mathrm{A}_k \cdot \mathrm{B}_k) = \mathrm{A} \cdot \mathrm{B}$*;*

(iii) $\lim_{k \to \infty}(\alpha\,\mathrm{A}_k) = \alpha\,\mathrm{A}$.

Proof The first two conclusions follow from the definition of convergence of ℝ-formal power series, noting that the operations of addition and multiplication have the property that, if two ℝ-formal power series agree for sufficiently large values of the index, then so too do their sum and product. We leave the elementary, albeit slightly tedious, details to the reader. The final assertion follows trivially from the definition of convergence. ■

The first two parts of the previous result say that addition and multiplication are continuous, where continuity is as defined according to the notion of convergence in Definition 3.7.3.

One can also perform calculus for ℝ-formal power series without having to worry about the analytical problems concerning limits in ℝ. To do so, we simply "pretend" that an element of $\mathbb{R}[[\xi]]$ can be differentiated and integrated term-by-term with respect to ξ. After one is finished pretending, then one makes the following definition.

3.7.8 Definition (Differentiation and integration of $\mathbb{R}$-formal power series) Let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a $\mathbb{R}$-formal power series.

(i) The ***derivative*** of A is the $\mathbb{R}$-formal power series $A' = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by $b_j = (j+1)a_{j+1}$, $j \in \mathbb{Z}_{\geq 0}$.

(ii) The ***integral*** of A is the $\mathbb{R}$-formal power series $\int A = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$b_j = \begin{cases} 0, & j = 0, \\ \frac{a_{j-1}}{j}, & j \in \mathbb{Z}_{>0}. \end{cases} \qquad \bullet$$

In terms of the indeterminate representation of a $\mathbb{R}$-formal power series, we have the following representation. If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a $\mathbb{R}$-formal power series, then

$$A' = \left(\sum_{j=0}^{\infty} a_j \xi^j\right)' = \sum_{j=1}^{\infty} j a_j \xi^{j-1} = \sum_{j=0}^{\infty} (j+1) a_{j+1} \xi^j.$$

This is simply termwise differentiation with respect to the indeterminate. Note that in this case we can ignore the matter of whether it is valid to switch the sum and the derivative since we are not actually talking about functions. Similar statements hold, of course, for the integral of a $\mathbb{R}$-formal power series.

For this derivative operation, one has the usual rules.

3.7.9 Proposition (Properties of differentiation and integration of $\mathbb{R}$-formal power series) *Let* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *and* $\mathrm{B} = (\mathrm{b_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *be $\mathbb{R}$-formal power series and let $\alpha \in \mathbb{R}$. Then the following statements hold:*

(i) $(\mathrm{A} + \mathrm{B})' = \mathrm{A}' + \mathrm{B}'$;

(ii) $(\mathrm{A} \cdot \mathrm{B})' = \mathrm{A}' \cdot \mathrm{B} + \mathrm{A} \cdot \mathrm{B}'$;

(iii) $(\alpha\,\mathrm{A})' = \alpha\,\mathrm{A}'$;

(iv) $\int(\mathrm{A} + \mathrm{B}) = \int \mathrm{A} + \int \mathrm{B}$;

(v) $\int(\alpha\,\mathrm{A}) = \alpha \int \mathrm{A}$.

Proof The second statement is the only possibly nontrivial one, so it is the only thing we will prove. We note that

$$A \cdot B = \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} a_j b_{k-j}\right) \xi^k,$$

so that

$$\begin{aligned} (A \cdot B)' &= \sum_{k=1}^{\infty} \left(\sum_{j=0}^{k} a_j b_{k-j}\right) k \xi^{k-1} \\ &= \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} (j+1) a_{j+1} b_{k-j}\right) \xi^k + \sum_{k=0}^{\infty} \left(\sum_{j=0}^{k} (j+1) a_{k-j} b_{j+1}\right) \xi^k \\ &= A' \cdot B + A \cdot B', \end{aligned}$$

as desired. ■

The derivative also commutes with limits, as one would hope to be the case.

3.7.10 Proposition (Differentiation and integration, and convergence in $\mathbb{R}[[\xi]]$) *If* $(\mathsf{A}_k = (\mathsf{a}_{k,j})_{j\in\mathbb{Z}_{\geq 0}})_{k\in\mathbb{Z}_{>0}}$ *is a sequence in* $\mathbb{R}[[\xi]]$ *converging to* A*, then* $\mathsf{A}' = \lim_{k\to\infty}\mathsf{A}_k'$ *and* $\int\mathsf{A} = \lim_{k\to\infty}\int\mathsf{A}_k$.

Proof This is a more or less obvious result, given the definition of convergence of ℝ-formal power series. ■

Now that we have finished playing algebraic games, we turn to the matter of when a formal power series actually represents a function.

3.7.2 ℝ-convergent power series

The one thing that we did not do in the preceding section is think of ℝ-formal power series as functions. This is because not all ℝ-formal power series *can* be thought of as functions. For example, if $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is the ℝ-formal power series defined by $a_j = j!$, $j \in \mathbb{Z}_{\geq 0}$, then the series $\sum_{j=1}^{\infty} a_j x^j$ diverges for any $x \in \mathbb{R}\setminus\{0\}$. In this section we address this matter by thinking of power series as being series of functions, just as we discussed in Section 3.6.4.

First we classify ℝ-formal power series according to the convergence properties possessed by the corresponding series of functions.

3.7.11 Proposition (Classification of ℝ-formal power series by convergence) *For each* ℝ*-formal power series* $(\mathsf{a}_j)_{j\in\mathbb{Z}_{\geq 0}}$*, exactly one of the following statements holds:*

(i) the series $\sum_{j=0}^{\infty}\mathsf{a}_j\mathsf{x}^j$ *converges absolutely for all* $\mathsf{x}\in\mathbb{R}$*;*

(ii) the series $\sum_{j=0}^{\infty}\mathsf{a}_j\mathsf{x}^j$ *diverges for all* $\mathsf{x}\in\mathbb{R}\setminus\{0\}$*;*

(iii) there exists $\mathsf{R}\in\mathbb{R}_{>0}$ *such that the series* $\sum_{j=0}^{\infty}\mathsf{a}_j\mathsf{x}^j$ *converges absolutely for all* $\mathsf{x}\in\mathsf{B}(\mathsf{R},0)$*, and diverges for all* $\mathsf{x}\in\mathbb{R}\setminus\overline{\mathsf{B}}(\mathsf{R},0)$.

Proof First let us prove a lemma.

1 Lemma *If the series* $\sum_{j=0}^{\infty}\mathsf{a}_j\mathsf{x}_0^j$ *converges for some* $\mathsf{x}_0\in\mathbb{R}$*, then the series* $\sum_{j=0}^{\infty}\mathsf{a}_j\mathsf{x}^j$ *converges absolutely for* $\mathsf{x}\in\mathsf{B}(|\mathsf{x}_0|,0)$.

Proof Note that the sequence $(a_jx_0^j)_{j\in\mathbb{Z}_{\geq 0}}$ converges to zero, and so is bounded by Proposition 2.3.4. Thus let $M\in\mathbb{R}_{>0}$ have the property that $|a_jx_0^j|\leq M$ for each $j\in\mathbb{Z}_{\geq 0}$. Then, for $x\in\mathsf{B}(|x_0|,0)$, we have

$$|a_jx^j| = |a_jx_0^j|\left|\frac{x}{x_0}\right|^j \leq M\left|\frac{x}{x_0}\right|^j, \qquad j\in\mathbb{Z}_{\geq 0}.$$

Since $\left|\frac{x}{x_0}\right| < 1$ the series $\sum_{j=0}^{\infty} M\left|\frac{x}{x_0}\right|$ converges as shown in Example 2.4.2–1. Therefore, by the Comparison Test, the series $\sum_{j=0}^{\infty}a_jx^j$ converges absolutely for $x\in\mathsf{B}(|x_0|,0)$. ▼

Now let

$$R = \sup\left\{x\in\mathbb{R}_{\geq 0}\ \middle|\ \sum_{j=0}^{\infty}a_jx^j \text{ converges}\right\}.$$

We have three cases.

1. $R = \infty$: For $x \in \mathbb{R}$ choose $x_0 > 0$ such that $|x| < x_0$. By the lemma, the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely. This is case (i) of the statement of the result.
2. $R = 0$: Let $x \in \mathbb{R} \setminus \{0\}$ and choose $x_0 > 0$ such that $|x| > x_0$. If $\sum_{j=0}^{\infty} a_j x^j$ converges, then by the lemma, the series $\sum_{j=0}^{\infty} a_j x_0^j$ converges absolutely, and so converges. But this contradicts the definition of R, so the series $\sum_{j=0}^{\infty} a_j x^j$ must diverge for every nonzero $x \in \mathbb{R}$. This is case (ii) of the statement of the result.
3. $R \in \mathbb{R}_{>0}$: If $x \in \mathsf{B}(R, 0)$ then, by the lemma, the series $\sum_{j=0}^{\infty} a_j x^j$ converges absolutely. If $x \in \mathbb{R} \setminus \overline{\mathsf{B}}(R, 0)$ then there exists $x_0 > R$ such that $|x| > x_0$. If the series $\sum_{j=0}^{\infty} a_j x^j$ converges, then by the lemma the series $\sum_{j=0}^{\infty} a_j x_0^j$ converges absolutely, and so converges. But this contradicts the definition of R. This is case (iii) of the statement of the result.

These three possibilities clearly are exhaustive and mutually exclusive. ■

Now we can sensibly define what we mean by a power series that converges.

3.7.12 Definition (ℝ-convergent power series) A ℝ-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ is a ℝ-***convergent power series*** if it falls into either case (i) or (iii) of Proposition 3.7.11. •

One can also say that a ℝ-formal power series that is not convergent has a zero radius of convergence, and sometimes it will be convenient to use this language.

Of course, one is interested in actually determining whether a given ℝ-formal power series is convergent or not. It turns out that this is actually possible, as the following result indicates.

3.7.13 Theorem (Cauchy–Hadamard[14] test for power series convergence) *Let* $(\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *be a* ℝ*-formal power series, and define* $\rho \in \overline{\mathbb{R}}_{\geq 0}$ *by* $\rho = \limsup_{\mathrm{j}\to\infty} |\mathrm{a_j}|^{1/\mathrm{j}}$. *Then define* $\mathrm{R} \in \overline{\mathbb{R}}_{\geq 0}$ *by*

$$\mathrm{R} = \begin{cases} \infty, & \rho = 0, \\ \frac{1}{\rho}, & \rho \in \mathbb{R}_{>0}, \\ 0, & \rho = \infty. \end{cases}$$

Then R *is the radius of convergence for* $(\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$.

Proof Let $x \in \mathbb{R}$. We have

$$\limsup_{j\to\infty} |a_j x^j|^{1/j} = \limsup_{j\to\infty} |x||a_j|^{1/j} = |x|\rho.$$

Now, by the Root Test, $\sum_{j=0} a_j x^j$ converges if $|x|\rho < 1$ and diverges if $|x|\rho > 1$. From these statements, the result follows. ■

[14] Jacques Salomon Hadamard (1865–1963) was a French mathematician. He made significant contributions to the fields of complex analysis, number theory, differential equations, geometry and linear algebra.

Note that in Proposition 3.7.11 we make no assertions about the convergence of power series for values of x whose magnitude us equal to the radius of convergence.

3.7.14 Definition (Region of (absolute) convergence) Let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a $\mathbb{R}$-formal power series and consider the classification of Proposition 3.7.11. In case (i) the ***radius of convergence*** is ∞, and in case (iii) the ***radius of convergence*** is the positive number R asserted in the statement of the proposition. The ***region of absolute convergence*** is $\mathscr{R}_{\mathrm{abs}}(A) = (-R, R)$, and the region of convergence is the largest interval $\mathscr{R}_{\mathrm{conv}}(A) \subseteq \mathbb{R}$ on which the series $\sum_{j=0}^{\infty} a_j x^j$ converges. •

Note that the region of convergence could be either $(-R, R)$, $[-R, R)$, $(-R, R]$, or $[-R, R]$. The following examples show that all possibilities are realised.

3.7.15 Examples (Region of (absolute) convergence)

1. Consider the $\mathbb{R}$-formal power series $A = (a_j = \frac{1}{2^j j^2})_{j\in\mathbb{Z}_{>0}}$ (take $a_0 = 0$). We compute
$$\lim_{j\to\infty}\left|\frac{a_{j+1}}{a_j}\right| = \lim_{j\to\infty}\left|\frac{2^j j^2}{2^{j+1}(j+1)^2}\right| = \frac{1}{2}.$$
By Proposition 2.4.15 we conclude that the radius of convergence of the power series $\sum_{j=1}^{\infty}\frac{x^j}{2^j j^2}$ is 2. When $x = 2$ the series becomes $\sum_{j=1}^{\infty}\frac{1}{j^2}$, which we know converges by Example 2.4.2–4. When $x = -2$ the series becomes $\sum_{j=1}^{\infty}\frac{(-1)^j}{j^2}$, which again is convergent, this time by the Alternating Test. Thus $\mathscr{R}_{\mathrm{abs}}(A) = (-2, 2)$, while $\mathscr{R}_{\mathrm{conv}}(A) = [-2, 2]$.
2. Now consider the $\mathbb{R}$-formal power series $A = (a_j = \frac{1}{2^j j})_{j\in\mathbb{Z}_{>0}}$ (take $a_0 = 0$). We again use Proposition 2.4.15 and the computation
$$\lim_{j\to\infty}\left|\frac{a_{j+1}}{a_j}\right| = \lim_{j\to\infty}\left|\frac{2^j j}{2^{j+1}(j+1)}\right| = \frac{1}{2}$$
to deduce that this power series has radius of convergence 2. For $x = 2$ the series becomes $\sum_{j=1}^{\infty}\frac{1}{j}$ which diverges by Example 2.4.2–4, and for $x = -2$ the series becomes $\sum_{j=1}^{\infty}\frac{(-1)^j}{j}$ which converges by Example 2.4.2–3. Thus $\mathscr{R}_{\mathrm{abs}}(A) = (-2, 2)$, while $\mathscr{R}_{\mathrm{conv}}(A) = [-2, 2)$.
3. Now we define the $\mathbb{R}$-formal power series $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ by
$$a_j = \begin{cases} 0, & j = 0, \\ 0, & j \text{ odd}, \\ \frac{2}{2^{-\frac{j}{2}} j}, & \text{otherwise}. \end{cases}$$
Thus the corresponding series is $\sum_{k=1}^{\infty}\frac{x^{2k}}{2^k k}$. We have
$$\limsup_{j\to\infty}|a_j|^{1/j} = \limsup_{k\to\infty}\left|\frac{1}{2^k k}\right|^{1/2k} = \frac{1}{\sqrt{2}}\lim_{k\to\infty}\left(\frac{1}{k}\right)^{1/2k} = \frac{1}{\sqrt{2}}.$$

Thus the radius of convergence is $\sqrt{2}$. For $x = \pm\sqrt{2}$ the series becomes $\sum_{k=1}^{\infty} \frac{1}{k}$ which diverges. Thus $\mathscr{R}_{\mathrm{abs}}(A) = \mathscr{R}_{\mathrm{conv}}(A) = (-\sqrt{2}, \sqrt{2})$. •

An important property of $\mathbb{R}$-convergent power series, is that, not only do they converge absolutely, they converge uniformly on any compact interval in the region of absolute convergence.

3.7.16 Theorem (Uniform convergence of $\mathbb{R}$-convergent power series) *If* $\mathrm{A} = (\mathrm{a}_{\mathrm{j}})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-convergent power series, then the series* $\sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}\mathrm{x}^{\mathrm{j}}$ *converges uniformly on any compact interval* $\mathrm{J} \subseteq \mathscr{R}_{\mathrm{abs}}(\mathrm{A})$.

Proof It suffices to consider the case where $J = [-R_0, R_0]$ since any compact interval will be contained in an interval of this form. Let $x \in [-R_0, R_0]$. Since $\sum_{j=0}^{\infty} a_j R_0^j$ converges absolutely and since $|a_j x^j| \leq a_j R_0^j$, uniform convergence follows from the Weierstrass M-test. ■

The next result gives the value of the limit function at points in the boundary of the region of convergence.

3.7.17 Theorem (Continuous extension to region of convergence) *Let* $(\mathrm{a}_{\mathrm{j}})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *be a* $\mathbb{R}$*-convergent power series with radius of convergence* R. *If the series* $\sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}\mathrm{R}^{\mathrm{j}}$ *(resp.* $\sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}(-\mathrm{R})^{\mathrm{j}}$*) converges, then*

$$\lim_{\mathrm{x}\uparrow\mathrm{R}} \sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}\mathrm{x}^{\mathrm{j}} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}\mathrm{R}^{\mathrm{j}} \qquad \left(\textit{resp.}\ \lim_{\mathrm{x}\downarrow-\mathrm{R}} \sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}\mathrm{x}^{\mathrm{j}} = \sum_{\mathrm{j}=0}^{\infty} \mathrm{a}_{\mathrm{j}}(-\mathrm{R}^{\mathrm{j}})\right).$$

Proof We shall only prove the theorem in the limit as x approaches R; the other case follows entirely similarly (or by a change of variable from x to $-x$). Denote by $f\colon \mathsf{B}(R, 0) \to \mathbb{R}$ the limit function for the power series. Let $S_{-1} = 0$ and for $k \in \mathbb{Z}_{\geq 0}$ define

$$S_k = \sum_{j=0}^{k} a_j R^j.$$

We then directly have

$$\sum_{j=0}^{k} a_j x^j = \sum_{j=0}^{k} (S_j - S_{j-1})(\tfrac{x}{R})^j = (1 - \tfrac{x}{R}) \sum_{j=0}^{k-1} S_j (\tfrac{x}{R})^j + S_k (\tfrac{x}{R})^k.$$

For $x \in \mathsf{B}(R, 0)$ we note that $\lim_{k\to\infty} S_k(\frac{x}{R})^k = 0$, and therefore

$$f(x) = \sum_{j=0}^{\infty} a_j x^j = (1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} S_j (\tfrac{x}{R})^j.$$

If $S = \lim_{j\to\infty} S_j$, for $\epsilon \in \mathbb{R}_{>0}$ take $N \in \mathbb{Z}_{>0}$ such that $|S - S_j| < \frac{\epsilon}{2}$ for $j \geq N$. Note that, from Example 2.4.2–1, we have

$$(1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} (\tfrac{x}{R})^j = 1$$

for $x \in \mathsf{B}(R, 0)$. It therefore follows that for $x \in (0, R)$ we have

$$(1 - \tfrac{x}{R}) \sum_{j=N+1}^{\infty} |S_j - S|(\tfrac{x}{R})^j \leq \tfrac{\epsilon}{2}(1 - \tfrac{x}{R}) \sum_{j=N+1}^{\infty} (\tfrac{x}{R})^j < \tfrac{\epsilon}{2}. \tag{3.23}$$

Now let $\delta \in \mathbb{R}_{>0}$ have the property that for $x \in (R - \delta, R)$

$$(1 - \tfrac{x}{R}) \sum_{j=0}^{N} |S_j - S| < \tfrac{\epsilon}{2}.$$

It therefore follows that for $x \in (R - \delta, R)$ we also have

$$(1 - \tfrac{x}{R}) \sum_{j=0}^{N} |S_j - S|(\tfrac{x}{R})^j < \tfrac{\epsilon}{2}. \tag{3.24}$$

We therefore obtain, for $x \in (R - \delta, R)$,

$$\begin{aligned} |f(x) - S| = \left|(1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} (S_j - S)(\tfrac{x}{R})^j\right| &\leq (1 - \tfrac{x}{R}) \sum_{j=0}^{\infty} |S_j - S|(\tfrac{x}{R})^j \\ &\leq (1 - \tfrac{x}{R}) \sum_{j=0}^{N} |S_j - S|(\tfrac{x}{R})^j + (1 - \tfrac{x}{R}) \sum_{j=N+1}^{\infty} |S_j - S|(\tfrac{x}{R})^j \\ &< \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon, \end{aligned}$$

using (3.23) and (3.24). It therefore follows that $\lim_{x \uparrow R} f(x) = S$, as desired. ■

The preceding two theorems have the following important corollary.

3.7.18 Corollary (ℝ-convergent power series have a continuous limit function) *If $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j} \in \mathbb{Z}_{\geq 0}}$ is a ℝ-convergent power series, then the limit function on $\mathscr{R}_{\mathrm{conv}}(\mathrm{A})$ is continuous.*

Proof This follows immediately from the previous two theorems along with Theorem 3.6.8. ■

3.7.3 ℝ-convergent power series and operations on functions

In this section we explore how various operations on functions interact with power series. The results in this section have the usual mundane character of other similar sections in this chapter. However, it is worth noting that there is one rather spectacular conclusion that emerges, namely that the limit function of a ℝ-convergent power series is infinitely differentiable. The significance of this is perhaps not to be fully appreciated until we realise that, when this conclusion is extended to power series for complex functions, it allows the correspondence between analytic functions and power series.

But first some mundane things.

3.7.19 Proposition (Addition and multiplication, and $\mathbb{R}$-convergent power series) *If* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *and* $\mathrm{B} = (\mathrm{b_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *are* $\mathbb{R}$*-convergent power series, then the following statements hold:*

(i) $\mathscr{R}_{\mathrm{conv}}(\mathrm{A}+\mathrm{B}) \subseteq \mathscr{R}_{\mathrm{conv}}(\mathrm{A}) \cap \mathscr{R}_{\mathrm{conv}}(\mathrm{B})$*, and so, in particular,* $\mathrm{A}+\mathrm{B}$ *is a* $\mathbb{R}$*-convergent power series;*

(ii) $\mathscr{R}_{\mathrm{conv}}(\mathrm{A}\cdot\mathrm{B}) \subseteq \mathscr{R}_{\mathrm{conv}}(\mathrm{A}) \cap \mathscr{R}_{\mathrm{conv}}(\mathrm{B})$*, and so, in particular,* $\mathrm{A}\cdot\mathrm{B}$ *is a* $\mathbb{R}$*-convergent power series.*

Proof This follows immediately from Proposition 2.4.30. ■

In the language of Section 4.2, the preceding result says that the set of $\mathbb{R}$-convergent power series is a subring of the set of $\mathbb{R}$-formal power series. This in and of itself is not hugely interesting. However, the exact properties of the ring of $\mathbb{R}$-convergent power series *is* of quite some importance in the study of analytic functions; we refer the reader to Section 3.7.5 for further discussion and references.

3.7.20 Proposition (Differentiation and integration of $\mathbb{R}$-convergent power series) *If* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-convergent power series, then the following statements hold:*

(i) $\mathscr{R}_{\mathrm{abs}}(\mathrm{A}') = \mathscr{R}_{\mathrm{abs}}(\mathrm{A})$*, and so, in particular,* A' *is a* $\mathbb{R}$*-convergent power series;*

(ii) $\mathscr{R}_{\mathrm{abs}}(\int \mathrm{A}) = \mathscr{R}_{\mathrm{abs}}(\mathrm{A})$*, and so, in particular,* $\int \mathrm{A}$ *is a* $\mathbb{R}$*-convergent power series.*

Furthermore, if the series defined by A *converges to* $\mathrm{f}\colon \mathscr{R}_{\mathrm{abs}}(\mathrm{A}) \to \mathbb{R}$*, then the series defined by* A' *converges to* f' *on* $\mathscr{R}_{\mathrm{abs}}(\mathrm{A})$ *and the series defined by* $\int \mathrm{A}$ *converges to the function* $\mathrm{x} \mapsto \int_0^{\mathrm{x}} \mathrm{f}(\xi)\,\mathrm{d}\xi$ *on* $\mathscr{R}_{\mathrm{abs}}(\mathrm{A})$*.*

Proof That $\mathscr{R}_{\mathrm{abs}}(A') = \mathscr{R}_{\mathrm{abs}}(A)$ and $\mathscr{R}_{\mathrm{abs}}(\int A) = \mathscr{R}_{\mathrm{abs}}(A)$ follows since $\lim_{j\to\infty} j^{1/j} = \lim_{j\to\infty}(\frac{1}{j})^{1/j} = 1$ by Proposition 3.8.12, allowing us to conclude that

$$\limsup_{j\to\infty} |ja_j|^{1/j} = \limsup_{j\to\infty} |a_j|^{1/j}, \quad \limsup_{j\to\infty} \left|\frac{a_j}{j}\right|^{1/j} = \limsup_{j\to\infty} |a_j|^{1/j}.$$

That the series defined by A' and $\int A$ have the properties stated follows from Theorems 3.6.23 and 3.6.24, along with the definitions of A' and $\int A$. ■

This gives the following remarkable corollary concerning the character of the limit function for $\mathbb{R}$-convergent power series.

3.7.21 Corollary (Limits of $\mathbb{R}$-convergent power series are infinitely differentiable) *If* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-convergent power series converging to* $\mathrm{f}\colon \mathscr{R}_{\mathrm{abs}}(\mathrm{A}) \to \mathbb{R}$*, then* f *is infinitely differentiable on* $\mathscr{R}_{\mathrm{abs}}(\mathrm{A})$*, and* $\mathrm{a_j} = \frac{\mathrm{f}^{(\mathrm{j})}(0)}{\mathrm{j}!}$*.*

Proof This follows simply by a repeated application of Proposition 3.7.20, and by performing term-by-term differentiation, and evaluating the resulting expressions at 0. ■

3.7.4 Taylor series

In the preceding section we indicated how, for the special class of ℝ-formal power series that are convergent, one can construct a limit function that is infinitely differentiable. In this section we consider the possibility of "reversing" this operation, and producing a ℝ-formal power series from an infinitely differentiable function. Even in cases when a function is not infinitely differentiable, we shall attempt to approximate it using a truncated power series. What we shall see in this section is that the correspondence between functions and the power series which purport to approximate them is a complicated one. Indeed, it is only for a special class of functions, those which we call "real analytic," that this correspondence as a useful one.

Let $I \subseteq \mathbb{R}$ be an interval and let $x_0 \in \mathrm{int}(I)$. Suppose that $f\colon I \to \mathbb{R}$ is infinitely differentiable. If one takes as the final objective the idea that we wish to approximate f near x_0. If $x_0 = 0$ then we might like to write

$$f(x) = \sum_{j=0}^{\infty} a_j x^j.$$

For $x_0 \neq 0$ it makes sense to write this approximation as

$$f(x) = \sum_{j=0}^{\infty} a_j (x - x_0)^j.$$

Indeed, if we write our approximation in this way, and then believe that differentiation can be performed term-by-term on the right, we obtain

$$f(x_0) = a_0,\ f^{(1)}(x_0) = a_1,\ f^{(2)}(x_0) = 2a_2, \dots, f^{(j)}(x_0) = j!a_j, \dots$$

With this as motivation, we make the following definition.

3.7.22 Definition (Taylor polynomial and Taylor series) Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in \mathrm{int}(I)$, and let $f\colon I \to \mathbb{R}$ be r-times differentiable for $r \in \mathbb{Z}_{>0} \cup \{\infty\}$.

(i) For $k \leq r$, the ***Taylor polynomial*** of degree k for f about x_0 is the polynomial function $\mathscr{T}_k(f, x_0)$ defined by

$$\mathscr{T}_k(f, x_0)(x) = \sum_{j=0}^{k} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j.$$

(ii) If $r = \infty$ then the ***Taylor series*** for f about x_0 is the ℝ-formal power series $\mathscr{T}_\infty(f, x_0) = (\frac{f^{(j)}(x_0)}{j!})_{j \in \mathbb{Z}_{\geq 0}}$. •

Sometimes it can be tedious to compute the derivatives needed to explicitly exhibit the Taylor polynomial or the Taylor series. In some cases, the following result is helpful.

3.7.23 Proposition (Property of Taylor polynomial) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{r} \in \mathbb{Z}_{>0}$*, and let* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *be a function that is* r*-times differentiable with* $\mathrm{f}^{(\mathrm{r})}$ *locally bounded. If* $\mathrm{x}_0 \in \mathrm{I}$ *and if* $\mathrm{P}\colon \mathrm{I} \to \mathbb{R}$ *is a polynomial function of degree* $\mathrm{r}-1$*, then* $\mathrm{P} = \mathscr{T}_{\mathrm{r}-1}(\mathrm{f}, \mathrm{x}_0)$ *if and only if*

$$\lim_{\mathrm{x}\to_{\mathrm{I}}\mathrm{x}_0} \frac{\mathrm{f}(\mathrm{x}) - \mathrm{P}(\mathrm{x})}{(\mathrm{x}-\mathrm{x}_0)^{\mathrm{r}-1}} = 0.$$

Proof We will use Taylor's Theorem stated below. Suppose that $P = \mathscr{T}_{r-1}(f, x_0)$. Then, by Taylor's Theorem, for x in a neighbourhood of x_0, we have

$$|f(x) - P(x)| \le M|x-x_0|^r \quad\implies\quad \lim_{x\to_I x_0} \left|\frac{f(x)-P(x)}{(x-x_0)^{r-1}}\right| \le \lim_{x\to_I x_0} M|x-x_0| = 0.$$

Now suppose that

$$\lim_{x\to_I x_0} \frac{f(x)-P(x)}{(x-x_0)^{r-1}} = 0.$$

By Taylor's Theorem, write

$$f(x) = \mathscr{T}_{r-1}(f, x_0)(x) + R_r(f, x_0)(x),$$

where $R_r(f, x_0)(x)$ is a function defined in a neighbourhood of x_0 satisfying $|R_r(f, x_0)(x)| \le M|x-x_0|^r$. Then, using Exercise 2.2.8,

$$\begin{aligned}
&\lim_{x\to_I x_0} \left|\frac{f(x)-P(x)}{(x-x_0)^{r-1}}\right| = 0,\\
\implies\quad &\lim_{x\to_I x_0} \left|\frac{\mathscr{T}_{r-1}(f,x_0)(x) + R_r(f,x_0)(x) - P(x)}{(x-x_0)^{r-1}}\right| = 0,\\
\implies\quad &\lim_{x\to_I x_0} \left|\left|\frac{\mathscr{T}_{r-1}(f,x_0)(x) - P(x)}{(x-x_0)^{r-1}}\right| - \left|\frac{R_r(f,x_0)(x)}{(x-x_0)^{r-1}}\right|\right| = 0.
\end{aligned}$$

Since

$$\lim_{x\to_I x_0} \left|\frac{R_r(f,x_0)(x)}{(x-x_0)^{r-1}}\right| = 0$$

by the properties of $R_r(f, x_0)$, we conclude that

$$\lim_{x\to_I x_0} \left|\frac{\mathscr{T}_{r-1}(f,x_0)(x) - P(x)}{(x-x_0)^{r-1}}\right| = 0.$$

If P and $\mathscr{T}_{r-1}(f, x_0)$ were distinct degree $r-1$ polynomials, then we would either have

$$\lim_{x\to_I x_0} \left|\frac{\mathscr{T}_{r-1}(f,x_0)(x) - P(x)}{(x-x_0)^{r-1}}\right| = \alpha > 0, \quad\text{or}\quad \lim_{x\to_I x_0} \left|\frac{\mathscr{T}_{r-1}(f,x_0)(x) - P(x)}{(x-x_0)^{r-1}}\right| = \infty.$$

Thus the result follows. ■

The way to interpret the result is that the Taylor polynomial of degree k about x_0 provides the best (in some sense) degree k polynomial approximation to f near x_0. In this sense, the Taylor polynomial can be thought of as the generalisation of the derivative, the derivative providing the best linear approximation of a function.

There are two fundamentally different sorts of questions arising from the notions of the Taylor polynomial and the Taylor series.

1. Is the Taylor series for an infinitely differentiable function a $\mathbb{R}$-convergent power series?
2. (a) Does the Taylor polynomial approximate f in some sense?
   (b) If f is infinitely differentiable and the Taylor series *is* a $\mathbb{R}$-convergent power series, does it approximate f in some sense?

Before we proceed to explore these questions in detail, let us give a definition which they immediately suggest.

3.7.24 Definition (Real analytic function) Let $I \subseteq \mathbb{R}$ be an interval, let $x_0 \in I$, and let $f\colon I \to \mathbb{R}$ be an infinitely differentiable function with Taylor series $\mathscr{T}_\infty(f, x_0) = (\frac{f^{(j)}(x_0)}{j!})_{j\in\mathbb{Z}_{\geq 0}}$. We say that f is ***real analytic*** at x_0 if $\mathscr{T}_\infty(f, x_0)$ is a $\mathbb{R}$-convergent power series, and if there exists a neighbourhood U of x_0 such that

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j$$

for all $x \in U$. •

Thus real analytic functions are exactly those that are perfectly approximated by their Taylor series. What is not clear at this time is whether "real analytic" is actually different than "infinitely differentiable." The following result addresses this in rather dramatic fashion.

3.7.25 Theorem (Borel's Theorem) *If* $(\mathrm{a}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *is a* $\mathbb{R}$*-formal power series, then there exists an interval* $\mathrm{I} \subseteq \mathbb{R}$ *with* $0 \in \mathrm{int}(\mathrm{I})$ *and a function* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *of class* C^∞ *such that* $\mathscr{T}_\infty(\mathrm{f}, 0) = (\mathrm{a}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$.

Proof Define $\wedge\colon [-1, 1] \to \mathbb{R}$ by

$$\wedge(x) = \begin{cases} 0, & x \in \{-1, 1\}, \\ \mathrm{e}^{-\frac{1}{1-x^2}}\mathrm{e}, & x \in (-1, 1), \end{cases}$$

and note that

1. $\wedge$ is infinitely differentiable,
2. $\wedge(\pm 1) = 0$,
3. $\wedge(0) = 1$, and
4. $\wedge(x) \in (0, 1)$ for $|x| \in (0, 1)$.

(We refer the reader to Example 3.7.28–2 for the details concerning this function.) We take $I = [-1, 1]$ and, for $\epsilon \in (0, 1)$, define $g_\epsilon\colon I \to \mathbb{R}$ by

$$g_\epsilon(x) = \begin{cases} 0, & |x| \in [\epsilon, 1], \\ \wedge(1 + \frac{2x}{\epsilon}), & x \in (-\epsilon, -\frac{\epsilon}{2}), \\ \wedge(-1 + \frac{2x}{\epsilon}), & x \in (\frac{\epsilon}{2}, \epsilon), \\ 1, & |x| \in [0, \frac{\epsilon}{2}]. \end{cases}$$

Then, for $k \in \mathbb{Z}_{\geq 0}$, define $f_{\epsilon,k}\colon I \to \mathbb{R}$ inductively by taking $f_{\epsilon,0} = g_\epsilon$ and

$$f_{\epsilon,k}(x) = \int_0^x f_{\epsilon,k-1}(\xi)\,\mathrm{d}\xi.$$

Note that

1. $f_{\epsilon,k}^{(j)}(0) = 0$, $j \in \{0, 1, \dots, k-1\}$,
2. $f_{\epsilon,k}^{(k)}(0) = 1$, and
3. $\left|f_{\epsilon,k}^{(j)}(x)\right| \leq \epsilon$ for $j \in \{0, 1, \dots, k-1\}$ and $x \in I$.

Now let $(\epsilon_j)_{j\in\mathbb{Z}_{>0}}$ be a sequence in $\mathbb{R}_{>0}$ for which the series $\sum_{j=0}^\infty |a_j|\epsilon_j$ converges. We claim that if

$$f(x) = \sum_{j=0}^\infty a_j f_{\epsilon_j,j}(x),$$

then f is well-defined and infinitely differentiable on $[-1, 1]$, and has the property that $\mathscr{T}_\infty(f, 0) = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$.

By our choice of the sequence $(\epsilon_j)_{j\in\mathbb{Z}_{\geq 0}}$, it follows from the Weierstrass M-test that f is well-defined by virtue of the absolute and uniform convergence of the series $\sum_{j=0}^\infty a_j f_{\epsilon_j,j}(x)$ for $x \in [-1, 1]$. Moreover, the hypotheses of Theorem 3.6.24 hold, and so the series can be differentiated term-by-term. One may then directly verify that the Weierstrass M-test again ensures that the resulting differentiated series is again uniformly convergent. This argument may be repeated to show that f is infinitely differentiable, and the series for the kth derivative is the kth derivative of the series taken term-by-term. One now uses the properties of the functions $f_{\epsilon,j}$, $j \in \mathbb{Z}_{\geq 0}$, to directly verify that $\mathscr{T}_\infty(f, 0) = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$. We leave the tedious, but direct, checking of the details of the assertions in this paragraph to the reader. ■

This result, therefore, rules out any sort of complete correspondence between a function and its Taylor series. Indeed, it even rules out the convergence of Taylor series.

It is clear, then, that a real analytic must have a rather specific character to its Taylor series. The following result precisely characterises this.

3.7.26 Theorem (Derivatives of real analytic functions) *If* $\mathrm{I} \subseteq \mathbb{R}$ *is an open interval and if* $\mathrm{f}\colon \mathrm{I} \to \mathbb{R}$ *is infinitely differentiable, then the following statements are equivalent:*

(i) f *is real analytic;*

(ii) for each $\mathrm{x}_0 \in \mathrm{I}$ *there exists a neighbourhood* $\mathrm{U} \subseteq \mathrm{I}$ *of* x_0 *and* $\mathrm{C}, \mathrm{r} \in \mathbb{R}_{>0}$ *such that*

$$|\mathrm{f}^{(\mathrm{m})}(\mathrm{x})| \leq \mathrm{C}\mathrm{m}!\mathrm{r}^{-\mathrm{m}}$$

for all $\mathrm{x} \in \mathrm{U}$ *and* $\mathrm{m} \in \mathbb{Z}_{\geq 0}$.

Proof First suppose that f is real analytic and let $x_0 \in I$. Let $\delta \in \mathbb{R}_{>0}$ be such that

$$f(x) = \sum_{k=0}^\infty a_k(x - x_0)^k, \qquad |x - x_0| < \delta.$$

This implies that, for each $\rho \in (0,\delta)$, the sequence $(a_k\rho^k)_{k\in\mathbb{Z}_{\geq 0}}$ is bounded, say by $C' \in \mathbb{R}_{>0}$. Therefore, by Corollary 3.7.21 we have

$$|f^{(m)}(x_0)| \leq C'm!\rho^{-m}.$$

Let us fix some $\rho \in (0,\delta)$.

By differentiating the power series for f term-by-term on $\mathsf{B}(x_0,\delta)$ we have

$$\frac{f^{(m)}(t)}{m!} = \frac{1}{m!}\sum_{k=0}^{\infty}(k+1)\cdots(k+m)a_{k+m}(x-x_0)^k = \sum_{k=0}^{\infty}\binom{k+m}{m}a_{k+m}(x-x_0)^k,$$

where

$$\binom{j}{l} = \frac{j!}{l!(j-l)!}$$

is the binomial coefficient defined for $j,l \in \mathbb{Z}_{\geq 0}$ with $j \geq l$. By Exercise 2.2.1 we have

$$2^j = (1+1)^j = \sum_{l=0}^{j}\binom{j}{l}.$$

Therefore,

$$\binom{j}{l} \leq 2^j, \qquad l \in \{0,1,\dots,j\}.$$

Therefore, if $|x-x_0| < \frac{\rho}{3}$,

$$\left|\frac{f^{(m)}(x)}{m!}\right| \leq C'\rho^{-m}\sum_{k=0}^{\infty}\binom{k+m}{m}\rho^{-k}|x-x_0|^k \leq C'\left(\frac{\rho}{2}\right)^{-m}\sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k = 3C'\left(\frac{\rho}{2}\right)^{-m},$$

using Example 2.4.2–1. This gives the desired estimate, taking $C = 3C'$ and $r = \frac{\rho}{2}$.

Conversely suppose that for $x_0 \in I$, $|f^{(m)}(x)| \leq Cm!r^{-m}$ for some $C,r \in \mathbb{R}_{>0}$ and for each $m \in \mathbb{Z}_{\geq 0}$. Then, for $|x-x_0| < r$ we have

$$\sum_{k=0}^{\infty}\frac{|f^{(k)}(x_0)|}{k!}|x-x_0|^k \leq C\sum_{k=0}^{\infty}\left(\frac{|x-x_0|}{r}\right)^k < \infty$$

by Example 2.4.2–1. Thus the series

$$\sum_{k=0}^{\infty}\frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

converges absolutely, and so converges, for each $x \in \mathsf{B}(x_0,r)$. Thus f is real analytic. ■

We now explore the question of how well a Taylor polynomial or Taylor series approximates the function generating it, under suitable hypotheses. We begin with the case where the function f is differentiable to finite order.

3.7.27 Theorem (Taylor's Theorem) *Let* $\mathrm{I} \subseteq \mathbb{R}$ *be an interval, let* $\mathrm{r} \in \mathbb{Z}_{>0}$*, and let* $\mathrm{f} \colon \mathrm{I} \to \mathbb{R}$ *be a function that is* r*-times differentiable with* $\mathrm{f}^{(\mathrm{r})}$ *locally bounded. Then, if* $[\mathrm{a}, \mathrm{b}] \subseteq \mathrm{I}$ *is a compact interval, there exists* $\mathrm{c} \in [\mathrm{a}, \mathrm{b}]$ *such that*

$$\mathrm{f}(\mathrm{b}) = \mathscr{T}_{\mathrm{r}-1}(\mathrm{f}, \mathrm{a})(\mathrm{b}) + \frac{\mathrm{f}^{(\mathrm{r})}(\mathrm{c})}{\mathrm{r}!}(\mathrm{b} - \mathrm{a})^{\mathrm{r}}.$$

In particular, if $\mathrm{J} \subseteq \mathrm{I}$ *is a compact interval containing* x_0 *then there exists* $\mathrm{M} \in \mathbb{R}_{>0}$ *such that*

$$|\mathrm{f}(\mathrm{x}) - \mathscr{T}_{\mathrm{r}-1}(\mathrm{f}, \mathrm{x}_0)(\mathrm{x})| \leq \mathrm{M}|\mathrm{x} - \mathrm{x}_0|^{\mathrm{r}}$$

for all $\mathrm{x} \in \mathrm{J}$.

Proof Define $\alpha \in \mathbb{R}$ by asking that $f(b) = \mathscr{T}(f, a)(b) + \alpha(b - a)^r$. Now, if for $x \in [a, b]$ we define

$$g(x) = f(x) - \mathscr{T}_{r-1}(f, a)(x) - \alpha(x - a)^r,$$

then we have $g^{(r)}(x) = f^{(r)}(x) - r!\alpha$ since $\mathscr{T}_{r-1}(f, a)$ is a polynomial of degree $r - 1$. We directly compute, using the definition of $\mathscr{T}(f, a)$, that $g^{(j)}(a) = 0$ for $j \in \{0, 1, \dots, r - 1\}$. We also directly have $g(b) = 0$. Therefore, there exists $c_1 \in [a, b]$ such that $g^{(1)}(c_1) = 0$ by the Mean Value Theorem applied to g. We similarly assert the existence of $c_2 \in [a, c_1]$ such that $g^{(2)}(c_2) = 0$, again by the Mean Value Theorem, but now applied to $g^{(1)}$. Continuing in this way we arrive at $c_r \in [a, c_{r-1}]$ such that $g^{(r)}(c_r) = 0$. Taking $c = c_r$, the result follows since $g^{(r)}(x) = f^{(r)}(x) - r!\alpha$. ■

One might be inclined to conjecture that, if f is of class C^∞, then increasing sequences of Taylor polynomials ought to better and better approximate a function. Of course, Theorem 3.7.25 immediately rules this out. The following examples serve to illustrate just how complicated is the correspondence between a function and its Taylor series.

3.7.28 Examples (Taylor series)

1. The first example we give is one of a function that is infinitely differentiable on $\mathbb{R}$, but whose Taylor series about 0 only converges in a bounded neighbourhood of 0.

   We define $f \colon \mathbb{R} \to \mathbb{R}$ by $f(x) = \frac{1}{1+x^2}$. This function, being the quotient of an infinitely differentiable function by a nonvanishing infinitely differentiable function is it self infinitely differentiable. To determine the Taylor series for f, let make an educated guess, and then check it using Proposition 3.7.23. By Example 2.4.2–1 we have, for $x^2 < 1$,

   $$\frac{1}{1 + x^2} = \sum_{j=0}^{\infty} (-1)^j x^{2j}.$$

   Let us verify that this is actually the series associated to the Taylor series for f about 0. As we saw during the course of Example 2.4.2–1,

   $$\sum_{j=0}^{k} (-1)^j x^{2j} = \frac{1 - (-x^2)^{k+1}}{1 + x^2}.$$

Therefore

$$\left|\frac{1}{1+x^2}-\sum_{j=0}^{k}(-1)^jx^{2j}\right|=\frac{x^{2k+2}}{1+x^2}.$$

Thus

$$\lim_{x\to 0}\left|\frac{\frac{1}{1+x^2}-\sum_{j=0}^{k}(-1)^jx^{2j}}{x^{2k+1}}\right|=0,$$

and we conclude from Proposition 3.7.23 that $\sum_{j=0}^{k}(-1)^jx^{2j}=\mathscr{T}_{2k+1}(f,0)$. Thus we do indeed have $\mathscr{T}_\infty(f,0)=(a_j)_{j\in\mathbb{Z}_{>0}}$ where

$$a_j=\begin{cases}0, & j \text{ odd},\\ (-1)^{j/2}, & j \text{ odd}.\end{cases}$$

By Example 2.4.2–1 the radius of convergence for the Taylor series is 1. Indeed, one easily sees that $\mathscr{R}_{\mathrm{abs}}(\mathscr{T}_\infty(f,0))=\mathscr{R}_{\mathrm{conv}}(\mathscr{T}_\infty(f,0))=(-1,1)$.

Thus we indeed have a function, infinitely differentiable on all of ℝ, whose Taylor series converges on a bounded interval. Note that this function *is* real analytic at 0. In fact, one can verify that the function is real analytic everywhere. But even this is not enough to ensure the global convergence of the Taylor series about a given point. In order to understand why the Taylor series for this function does not converge on all of ℝ, it is necessary to understand ℂ-power series, as we do in Section II-3.3.

2. The next function we construct is one with a Taylor series whose radius of convergence is infinite, but which converges to the function only at one point. We define $f\colon\mathbb{R}\to\mathbb{R}$ by

$$f(x)=\begin{cases}\mathrm{e}^{-\frac{1}{x^2}}, & x\neq 0,\\ 0, & x=0,\end{cases}$$

and in Figure 3.16 we show the graph of f. We claim that $\mathscr{T}_\infty(f,0)$ is the zero ℝ-formal power series. To prove this, we must compute the derivatives of f at $x=0$. The following lemma is helpful in this regard.

1 Lemma *For* $\mathrm{j}\in\mathbb{Z}_{\geq 0}$ *there exists a polynomial* $\mathrm{p_j}$ *of degree at most* $2\mathrm{j}$ *such that*

$$\mathrm{f^{(j)}(x)}=\frac{\mathrm{p_j(x)}}{\mathrm{x^{3j}}}\mathrm{e}^{-\frac{1}{\mathrm{x}^2}},\qquad \mathrm{x}\neq 0.$$

Proof We prove this by induction on j. Clearly the lemma holds for $j=0$ by taking $p_0(x)=1$. Now suppose the lemma holds for $j\in\{0,1,\dots,k\}$. Thus

$$f^{(k)}(x)=\frac{p_k(x)}{x^{3k}}\mathrm{e}^{-\frac{1}{x^2}}$$

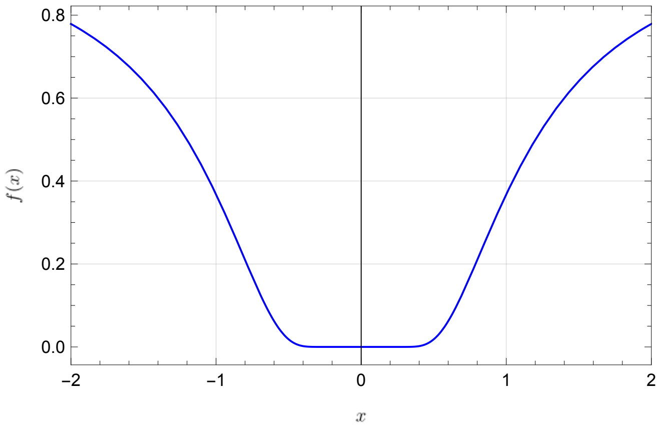


Figure 3.16 A function that is infinitely differentiable but not analytic

for a polynomial p_k of degree at most $2k$. Then we compute

$$f^{(k+1)}(x) = \frac{x^3 p_k'(x) - 3kx^2 p_k(x) - 2p_k(x)}{x^{3(k+1)}} \mathrm{e}^{-\frac{1}{x^2}}.$$

Using the rules for differentiation of polynomials, one easily checks that $x \mapsto x^3 p_k'(x) - 3kx^2 p_k(x) - 2p_k(x)$ is a polynomial whose degree is at most $2(k+1)$. ▼

From the lemma we infer the infinite differentiability of f on $\mathbb{R} \setminus \{0\}$. We now need to consider the derivatives at 0. For this we employ another lemma.

2 Lemma $\lim_{\mathrm{x}\to 0} \frac{\mathrm{e}^{-\frac{1}{\mathrm{x}^2}}}{\mathrm{x}^{\mathrm{k}}} = 0$ *for all* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$.

Proof We note that

$$\lim_{x\downarrow 0} \frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^k} = \lim_{y\to\infty} \frac{y^k}{\mathrm{e}^{y^2}}, \qquad \lim_{x\uparrow 0} \frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^k} = \lim_{y\to-\infty} \frac{y^k}{\mathrm{e}^{y^2}}.$$

Using the properties of the exponential function as given in Section 3.8.1, we have

$$\mathrm{e}^{y^2} = \sum_{j=0}^{\infty} \frac{y^{2j}}{j!}$$

In particular, $\mathrm{e}^{y^2} \geq \frac{y^{2k}}{k!}$, and so

$$\left|\frac{y^k}{\mathrm{e}^{y^2}}\right| \leq \left|\frac{k!}{y^k}\right|,$$

and so

$$\lim_{x\to 0} \frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^k} = 0,$$

as desired. ▼

Now, letting $p_k(x) = \sum_{j=0}^{2k} a_j x^j$, we may directly compute

$$\lim_{x\to 0} f^{(k)}(x) = \lim_{x\to 0} \sum_{j=0}^{2k} a_j x^{2j} \frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^{3k}} = \sum_{j=0}^{2k} a_j \lim_{x\to 0} \frac{\mathrm{e}^{-\frac{1}{x^2}}}{x^{3k-j}} = 0.$$

Thus we arrive at the conclusion that f is infinitely differentiable on ℝ, and that f and all of its derivatives are zero at $x = 0$. Thus $\mathscr{T}_\infty(f, 0) = (0)_{j\in\mathbb{Z}_{\geq 0}}$. This is clearly a ℝ-convergent power series; it converges everywhere to the zero function. However, $f(x) \neq 0$ except when $x = 0$. Thus the Taylor series about 0 for f, while convergent everywhere, converges to f only at $x = 0$. This is therefore an example of a function that is infinitely differentiable at a point, but not real analytic there. This function may seem rather useless, but in actuality it is quite an important one. For example, we used it in the construction for the proof of Theorem 3.7.25. •

These examples, along with Borel's Theorem, indicate the intricate nature of the correspondence between a function and its Taylor series. For the correspondence to have any real meaning, the function must be analytic, and even then the correspondence is only local.

3.7.5 Notes

As we shall see in Section II-3.3, there is, for ℂ-power series, a correspondence between convergent power series and holomorphic functions. This correspondence also applies to the real case, where "holomorphic" gets replaced with "real analytic." The ring-theoretic structure of the ℝ-convergent power series are of some importance. In particular, this ring possesses the property of being "Noetherian.[15]" Because of the correspondence between convergent power series and analytic functions, the ring theoretic structure gets transfered, at least locally, to the set of analytic functions. This leads to some rather remarkable features of analytic functions as compared to, say, merely infinitely differentiable functions. We refer to [Krantz and Parks 2002] for a discussion of this in the real analytic case, and to [Hörmander 1966] for the holomorphic case.

Exercises

3.7.1 State and prove a version of the Fundamental Theorem of Calculus for ℝ-formal power series.

3.7.2 State and prove an integration by parts formula for ℝ-formal power series.

3.7.3 Prove part (vi) of Proposition 2.4.30 using Proposition 3.7.17.

[15] Amalie Emmy Noether (1882-1935) was a German mathematician whose name is attached to important properties of rings in algebra and to conservation laws in physics

Section 3.8

Some $\mathbb{R}$-valued functions of interest

In this section we present, in a formal way, some of the special functions that will, and indeed already have, come up in these volumes.

Do I need to read this section? It is much more than likely the case that the reader has already encountered the functions we discuss in this section. However, it may be the case that the formal definitions and rigorous presentation of their properties will be new. This section, therefore, fits into the "read for pleasure" category. •

3.8.1 The exponential function

One of the most important functions in mathematics, particularly in applied mathematics, is the exponential function. This importance is nowhere to be found in the following definition, but hopefully at the end of their reading these volumes, the reader will have some appreciation for the exponential function.

3.8.1 Definition (Exponential function) The ***exponential function***, denoted by $\exp\colon \mathbb{R} \to \mathbb{R}$, is given by

$$\exp(x) = \sum_{j=0}^{\infty} \frac{x^j}{j!}. \qquad \bullet$$

In Figure 3.17 we show the graphs of exp and its inverse log that we will be discussing in the next section.

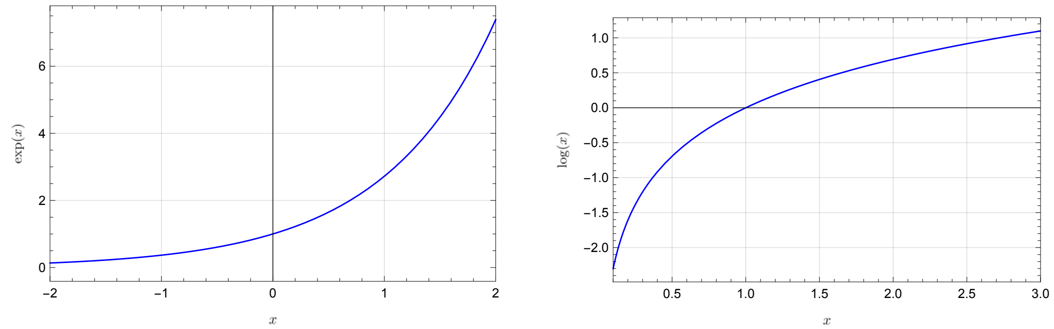


Figure 3.17 The function exp (left) and its inverse log (right)

One can use Theorem 3.7.13, along with Proposition 2.4.15, to easily show that the power series for exp has an infinite radius of convergence, and so indeed defines a function on $\mathbb{R}$. Let us record some of the more immediate and useful properties of exp.

3.8.2 Proposition (Properties of the exponential function) *The exponential function enjoys the following properties:*

(i) $\exp$ *is infinitely differentiable;*

(ii) $\exp$ *is strictly monotonically increasing;*

(iii) $\exp(x) > 0$ *for all* $x \in \mathbb{R}$*;*

(iv) $\lim_{x\to\infty} \exp(x) = \infty$*;*

(v) $\lim_{x\to-\infty} \exp(x) = 0$*;*

(vi) $\exp(x + y) = \exp(x)\exp(y)$ *for all* $x, y \in \mathbb{R}$*;*

(vii) $\exp' = \exp$*;*

(viii) $\lim_{x\to\infty} x^k \exp(-x) = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

Proof (i) This follows from Corollary 3.7.21, along with the fact that the radius of convergence of the power series for exp is infinite.

(vi) Using the Binomial Theorem and Proposition 2.4.30(iv) we compute

$$\exp(x)\exp(y) = \left(\sum_{j=0}^{\infty} \frac{x^j}{j!}\right)\left(\sum_{j=0}^{\infty} \frac{x^k}{k!}\right) = \sum_{k=0}^{\infty}\sum_{j=0}^{k} \frac{x^j}{j!}\frac{y^{k-j}}{(k-j)!}$$
$$= \sum_{k=0}^{\infty} \frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j} x^j y^{k-j} = \sum_{k=0}^{\infty} \frac{(x+y)^k}{k!}.$$

(viii) We have $\exp(-x) = \frac{1}{\exp(x)}$ by part (vi), and so we compute

$$\lim_{x\to\infty} x^k \exp(-x) = \lim_{x\to\infty} \frac{x^k}{\sum_{j=0}^{\infty} \frac{x^j}{j!}} \le \lim_{x\to\infty} \frac{(k+1)!x^k}{x^{k+1}} = 0.$$

(ii) From parts (i) and (viii) we know that exp has an everywhere positive derivative. Thus, from Proposition 3.2.23 we know that exp is strictly monotonically increasing.

(iii) Clearly $\exp(x) > 0$ for all $x \in \mathbb{R}_{\ge 0}$. From part (vi) we have

$$\exp(x)\exp(-x) = \exp(0) = 1.$$

Therefore, for $x \in \mathbb{R}_{<0}$ we have $\exp(x) = \frac{1}{\exp(-x)} > 0$.

(iv) We have

$$\lim_{x\to\infty} \exp(x) = \lim_{x\to\infty} \sum_{j=0}^{\infty} \frac{x^j}{j!} \ge \lim_{x\to\infty} x = \infty.$$

(v) By parts (vi) and (iv) we have

$$\lim_{x\to-\infty} \exp(x) = \lim_{x\to\infty} \frac{1}{\exp(-x)} = 0.$$

(vii) Using part (vi) and the power series representation for exp we compute

$$\exp'(x) = \lim_{h\to 0} \frac{\exp(x+h) - \exp(x)}{h} = \lim_{h\to 0} \frac{\exp(x)(\exp(h) - 1)}{h} = \exp(x). \qquad \blacksquare$$

One of the reasons for the importance of the function exp in applications can be directly seen from property (vii). From this one can see that exp is the solution to the "initial value problem"

$$y'(x) = y(x), \quad y(0) = 1. \tag{3.25}$$

Most readers will recognise this as the differential equation governing a scalar process which exhibits "exponential growth." It turns out that many physical processes can be modelled, or approximately modelled, by such an equation, or by a suitable generalisation of such an equation. Indeed, one could use the solution of (3.25) as the *definition* of the function exp. However, to be rigorous, one would then be required to show that this equation has a unique solution; this is not altogether difficult, but does take one off topic a little. Such are the constraints imposed by rigour.

In Section 2.4.3 we defined the constant e by

$$\mathrm{e} = \sum_{j=0}^{\infty} \frac{1}{j!}.$$

From this we see immediately that $\mathrm{e} = \exp(1)$. To explore the relationship between the exponential function exp and the constant e, we first prove the following result, which recalls from Proposition 2.2.3 and the discussion immediately following it, the definition of x^q for $x \in \mathbb{R}_{>0}$ and $q \in \mathbb{Q}$.

3.8.3 Proposition (exp(x) = e$^{\mathbf{x}}$) $\exp(x) = \sup\{\mathrm{e}^q \mid q \in \mathbb{Q},\ q < x\}$.

Proof First let us take the case where $x = q \in \mathbb{Q}$. Write $q = \frac{j}{k}$ for $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$. Then, by repeated application of part (vi) of Proposition 3.8.2 we have

$$\exp(q)^k = \exp(kq) = \exp(j) = \exp(j \cdot 1) = \exp(1)^j (\mathrm{e}^1)^j = \mathrm{e}^j.$$

By Proposition 2.2.3 this gives, by definition, $\exp(q) = \mathrm{e}^q$.

Now let $x \in \mathbb{R}$ and let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be a monotonically increasing sequence in $\mathbb{Q}$ such that $\lim_{j\to\infty} q_j = x$. By Theorem 3.1.3 we have $\exp(x) = \lim_{j\to\infty} \exp(q_j)$. By part (ii) of Proposition 3.8.2 the sequence $(\exp(q_j))_{j\in\mathbb{Z}_{>0}}$ is strictly monotonically increasing. Therefore, by Theorem 2.3.8,

$$\lim_{j\to\infty} \exp(q_j) = \lim_{j\to\infty} \mathrm{e}^{q_j} = \sup\{\mathrm{e}^q \mid q < x\},$$

as desired. ■

We shall from now on alternately use the notation e^x for $\exp(x)$, when this is more convenient.

3.8.2 The natural logarithmic function

From Proposition 3.8.2 we know that exp is a strictly monotonically increasing, continuous function. Therefore, by Theorem 3.1.30 we know that exp is an invertible function from $\mathbb{R}$ to image(exp). From parts (iii), (iv), and (v) of Proposition 3.8.2, as well as from Theorem 3.1.30 again, we know that image(exp) $= \mathbb{R}_{>0}$. This then leads to the following definition.

3.8.4 Definition (Natural logarithmic function) The ***natural logarithmic function***, denoted by $\log\colon \mathbb{R}_{>0} \to \mathbb{R}$, is the inverse of exp. •

We refer to Figure 3.17 for a depiction of the graph of log.

3.8.5 Notation (log versus ln) It is not uncommon to see the function that we denote by "log" written instead as "ln." In such cases, log is often used to refer to the base 10 logarithm (see Definition 3.8.13), since this convention actually sees much use in applications. However, we shall refer to the base 10 logarithm as $\log_{10}$. •

Now let us record the properties of log that follow immediately from its definition.

3.8.6 Proposition (Properties of the natural logarithmic function) *The natural logarithmic function enjoys the following properties:*

(i) $\log$ *is infinitely differentiable;*

(ii) $\log$ *is strictly monotonically increasing;*

(iii) $\log(x) = \int_1^x \frac{1}{\xi}\,d\xi$ *for all* $x \in \mathbb{R}_{>0}$*;*

(iv) $\lim_{x\to\infty} \log(x) = \infty$*;*

(v) $\lim_{x\downarrow 0} \log(x) = -\infty$*;*

(vi) $\log(xy) = \log(x) + \log(y)$ *for all* $x, y \in \mathbb{R}_{>0}$*;*

(vii) $\lim_{x\to\infty} x^{-k}\log(x) = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

Proof (iii) From the Chain Rule and using the fact that $\log \circ \exp(x) = x$ for all $x \in \mathbb{R}$ we have

$$\log'(\exp(x)) = \frac{1}{\exp(x)} \qquad \Longrightarrow \qquad \log'(y) = \frac{1}{y}$$

for all $y \in \mathbb{R}_{>0}$. Using the fact that $\log(1) = 0$ (which follows since $\exp(0) = 1$), we then apply the Fundamental Theorem of Calculus, this being valid since $y \mapsto \frac{1}{y}$ is Riemann integrable on any compact interval in $\mathbb{R}_{>0}$, we obtain $\log(x) = \int_1^y \frac{1}{\eta}\,d\eta$, as desired.

(i) This follows from part (iii) using the fact that the function $x \mapsto \frac{1}{x}$ is infinitely differentiable on $\mathbb{R}_{>0}$.

(ii) This follows from Theorem 3.1.30.

(iv) We have

$$\lim_{x\to\infty} \log(x) = \lim_{y\to\infty} \log(\exp(y)) = \lim_{y\to\infty} y = \infty.$$

(v) We have

$$\lim_{x\downarrow 0} \log x = \lim_{y\to-\infty} \log(\exp(y)) = \lim_{y\to-\infty} y = -\infty.$$

(vi) For $x, y \in \mathbb{R}_{>0}$ write $x = \exp(a)$ and $y = \exp(b)$. Then

$$\log(xy) = \log(\exp(a)\exp(b)) = \log(\exp(a+b)) = a + b = \log(x) + \log(y).$$

(vii) We compute

$$\lim_{x\to\infty} \frac{\log x}{x^k} = \lim_{y\to\infty} \frac{\log \exp(y)}{\exp(y)^k} = \lim_{y\to\infty} \frac{y}{\exp(y)^k} \le \lim_{y\to\infty} \frac{y}{(1+y+\frac{1}{2}y^2)^k} = 0. \qquad \blacksquare$$

3.8.3 Power functions and general logarithmic functions

For $x \in \mathbb{R}_{>0}$ and $q \in \mathbb{Q}$ we had defined, in and immediately following Proposition 2.2.3, x^q by $(x^{1/k})^j$ if $q = \frac{j}{k}$ for $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$. In this section we wish to extend this definition to x^y for $y \in \mathbb{R}$, and to explore the properties of the resulting function of both x and y.

3.8.7 Definition (Power function) If $a \in \mathbb{R}_{>0}$ then the function $\mathsf{P}_a \colon \mathbb{R} \to \mathbb{R}$ is defined by $\mathsf{P}_a(x) = \exp(x\log(a))$. If $a \in \mathbb{R}$ then the function $\mathsf{P}^a \colon \mathbb{R}_{>0} \to \mathbb{R}$ is defined by $\mathsf{P}^a(x) = \exp(a\log(x))$. •

Let us immediately connect this (when seen for the first time rather nonintuitive) definition to what we already know.

3.8.8 Proposition ($\mathsf{P}_a(x) = a^x$) $\mathsf{P}_a(x) = \sup\{a^q \mid q \in \mathbb{Q},\ q < x\}$.

Proof Let us first take $x = q \in \mathbb{Q}$ and write $q = \frac{j}{k}$ for $j \in \mathbb{Z}$ and $k \in \mathbb{Z}_{>0}$. We have

$$\exp(q\log(a))^k = \exp\left(\tfrac{j}{k}\log(a)\right)^k = \exp(j\log(a)) = \exp(\log(a))^j = a^j.$$

Therefore, by Proposition 2.2.3 we have

$$\exp(q\log(a)) = a^q.$$

Now let $x \in \mathbb{R}$ and let $(q_j)_{j\in\mathbb{Z}_{>0}}$ be a strictly monotonically increasing sequence in $\mathbb{Q}$ converging to x. Since exp and log are continuous, by Theorem 3.1.3 we have

$$\lim_{j\to\infty} \exp(q_j\log(a)) = \exp(x\log(a)).$$

As we shall see in Proposition 3.8.10, the function $x \mapsto \mathsf{P}_a(x)$ is strictly monotonically increasing. Therefore the sequence $(\exp(q_j\log(a)))_{j\in\mathbb{Z}_{>0}}$ is strictly monotonically increasing. Thus

$$\lim_{j\to\infty} \exp(q_j\log(a)) = \sup\{\mathsf{P}_a(q) \mid q \in \mathbb{Q},\ q < x\},$$

as desired. ■

Clearly we also have the following result.

3.8.9 Corollary ($\mathsf{P}^a(x) = x^a$) $\mathsf{P}^a(x) = \sup\{x^q \mid q \in \mathbb{Q},\ q < a\}$.

As with the exponential function, we will use the notation a^x for $\mathsf{P}_a(x)$ and x^a for $\mathsf{P}^a(x)$ when it is convenient to do so.

Let us now record some of the properties of the functions P_a and P^a that follow from their definition. When possible, we state the result using both the notation $\mathsf{P}_a(x)$ and a^x (or P^a and x^a).

3.8.10 Proposition (Properties of P_a) *For* $a \in \mathbb{R}_{>0}$, *the function* P_a *enjoys the following properties:*

(i) P_a *is infinitely differentiable;*

(ii) P_a *is strictly monotonically increasing when* $a > 1$, *is strictly monotonically decreasing when* $a < 1$, *and is constant when* $a = 1$*;*

(iii) $\mathsf{P}_a(x) = a^x > 0$ *for all* $x \in \mathbb{R}$*;*

(iv) $\lim_{x\to\infty} \mathsf{P}_a(x) = \lim_{x\to\infty} a^x = \begin{cases} \infty, & a > 1, \\ 0, & a < 1, \\ 1, & a = 1; \end{cases}$

(v) $\lim_{x\to-\infty} \mathsf{P}_a(x) == \lim_{x\to-\infty} a^x = \begin{cases} 0, & a > 1, \\ \infty, & a < 1, \\ 1, & a = 1; \end{cases}$

(vi) $\mathsf{P}_a(x + y) = a^{x+y} = a^x a^y = \mathsf{P}_a(x)\mathsf{P}_a(y)$*;*

(vii) $\mathsf{P}_a'(x) = \log(a)\mathsf{P}_a(x)$*;*

(viii) *if* $a > 1$ *then* $\lim_{x\to\infty} x^k\mathsf{P}_a(-x) = \lim_{x\to\infty} x^k a^{-x} = 0$ *for all* $k \in \mathbb{Z}_{>0}$*;*

(ix) *if* $a < 1$ *then* $\lim_{x\to\infty} x^k\mathsf{P}_a(x) = \lim_{x\to\infty} x^k a^x = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

Proof (i) Define $f, g\colon \mathbb{R} \to \mathbb{R}$ and $f(x) = x\log(a)$ and $g(x) = \exp(x)$. Then $\mathsf{P}_a = g \circ f$, and so is the composition of infinitely differentiable functions. This part of the result follows from Theorem 3.2.13.

(ii) Let $x_1 < x_2$. If $a > 1$ then $\log(a) > 0$ and so

$$x_1\log(a) < x_2\log(a) \quad\Longrightarrow\quad \exp(x_1\log(a)) < \exp(x_2\log(a))$$

since exp is strictly monotonically increasing. If $a < 1$ then $\log(a) < 0$ and so

$$x_1\log(a) > x_2\log(a) \quad\Longrightarrow\quad \exp(x_1\log(a)) > \exp(x_2\log(a)),$$

again since exp is strictly monotonically increasing. For $a = 1$ we have $\log(a) = 0$ so $\mathsf{P}_a(x) = 1$ for all $x \in \mathbb{R}$.

(iii) This follows since image(exp) $\subseteq \mathbb{R}_{>0}$.

(iv) For $a > 1$ we have

$$\lim_{x\to\infty} \mathsf{P}_a(x) = \lim_{x\to\infty} \exp(x\log(a)) = \lim_{y\to\infty} \exp(y) = \infty,$$

and for $a < 1$ we have

$$\lim_{x\to\infty} \mathsf{P}_a(x) = \lim_{x\to\infty} \exp(x\log(a)) = \lim_{y\to-\infty} \exp(y) = 0.$$

For $a = 1$ the result is clear since $\mathsf{P}_1(x) = 1$ for all $x \in \mathbb{R}$.

(v) For $a > 1$ we have

$$\lim_{x\to-\infty} \mathsf{P}_a(x) = \lim_{x\to-\infty} \exp(x\log(a)) = \lim_{y\to-\infty} \exp(y) = 0,$$

and for $a < 1$ we have

$$\lim_{x\to-\infty} \mathsf{P}_a(x) = \lim_{x\to-\infty} \exp(x\log(a)) = \lim_{y\to\infty} \exp(y) = \infty.$$

Again, for $a = 1$ the result is obvious.

(vi) We have

$$\mathsf{P}_a(x+y) = \exp((x+y)\log(a)) = \exp(x\log(a))\exp(y\log(a)) = \mathsf{P}_a(x)\mathsf{P}_a(y).$$

(vii) With f and g as in part (i), and using Theorem 3.2.13, we compute

$$\mathsf{P}_a'(x) = g'(f(x))f'(x) = \exp(x\log(a))\log(a) = \log(a)\mathsf{P}_a(x).$$

(viii) We compute

$$\lim_{x\to\infty} x^k\mathsf{P}_a(-x) = \lim_{x\to\infty} x^k\exp(-x\log(a)) = \lim_{y\to\infty}\left(\frac{y}{\log(a)}\right)^k \exp(-y) = 0,$$

using part (viii) of Proposition 3.8.2.

(ix) We have

$$\lim_{x\to\infty} x^k\mathsf{P}_a(x) = \lim_{x\to\infty} x^k\exp((-x)(-\log(a))) = 0$$

since $\log(a) < 0$. ■

3.8.11 Proposition (Properties of P^a) *For $a \in \mathbb{R}$, the function P^a enjoys the following properties:*

(i) P^a *is infinitely differentiable;*

(ii) P^a *is strictly monotonically increasing;*

(iii) $\mathsf{P}^a(x) = x^a > 0$ *for all* $x \in \mathbb{R}_{>0}$*;*

(iv) $\lim_{x\to\infty} \mathsf{P}^a(x) = \lim_{x\to\infty} x^a = \begin{cases} \infty, & a > 0, \\ 0, & a < 0, \\ 1, & a = 0; \end{cases}$

(v) $\lim_{x\downarrow 0} \mathsf{P}^a(x) = \lim_{x\downarrow 0} x^a = \begin{cases} 0, & a > 0, \\ \infty, & a < 0, \\ 1, & a = 0; \end{cases}$

(vi) $\mathsf{P}^a(xy) = (xy)^a = x^a y^a = \mathsf{P}^a(x)\mathsf{P}^a(y)$;

(vii) $(\mathsf{P}^a)'(x) = a\mathsf{P}^{a-1}(x)$.

Proof (i) Define $f\colon \mathbb{R}_{>0} \to \mathbb{R}$, $g\colon \mathbb{R} \to \mathbb{R}$, and $h\colon \mathbb{R} \to \mathbb{R}$ by $f(x) = \log(x)$, $g(x) = ax$, and $h(x) = \exp(x)$. Then $\mathsf{P}^a = h \circ g \circ f$. Since each of f, g, and h is infinitely differentiable, then so too is P^a by Theorem 3.2.13.

(ii) Let $x_1, x_2 \in \mathbb{R}_{>0}$ satisfy $x_1 < x_2$. Then

$$\mathsf{P}^a(x_1) = \exp(a\log(x_1)) < \exp(a\log(x_2)) = \mathsf{P}^a(x_2)$$

using the fact that both log and exp are strictly monotonically increasing.

(iii) This follows since image(exp) $\subseteq \mathbb{R}_{>0}$.

(iv) For $a > 0$ we have

$$\lim_{x\to\infty} \mathsf{P}^a(x) = \lim_{x\to\infty} \exp(a\log(x)) = \lim_{y\to\infty} \exp(y) = \infty,$$

and for $a < 0$ we have

$$\lim_{x\to\infty} \mathsf{P}^a(x) = \lim_{x\to\infty} \exp(a\log(x)) = \lim_{y\to-\infty} \exp(y) = 0.$$

For $a = 0$ we have $\mathsf{P}^a(x) = 1$ for all $x \in \mathbb{R}_{>0}$.

(v) For $a > 0$ we have

$$\lim_{x\downarrow 0} \mathsf{P}^a(x) = \lim_{x\downarrow 0} \exp(a\log(x)) = \lim_{y\to-\infty} \exp(y) = 0,$$

and for $a < 0$ we have

$$\lim_{x\downarrow 0} \mathsf{P}^a(x) = \lim_{x\downarrow 0} \exp(a\log(x)) = \lim_{y\to\infty} \exp(y) = \infty.$$

For $a = 1$, the result is trivial again.

(vi) We have

$$\mathsf{P}^a(xy) = \exp(a\log(xy)) = \exp(a(\log(x)+\log(y))) = \exp(a\log(x))\exp(a\log(y)) = \mathsf{P}^a(x)\mathsf{P}^a(y).$$

(vii) With f, g, and h as in part (i), and using the Chain Rule, we have

$$\begin{aligned}(\mathsf{P}^a)'(x) = h'(g(f(x)))g'(f(x))f'(x) &= a\exp(a\log(x))\tfrac{1}{x}\\ &= a\exp(a\log(x))\exp(-1\log(x)) = a\exp((a-1)\log(x)) = a\mathsf{P}^{a-1}(x),\end{aligned}$$

as desired, using part (vi) of Proposition 3.8.10. ■

The following result is also sometimes useful.

3.8.12 Proposition (Property of $\mathsf{P}_x(x^{-1})$) $\lim_{x\to\infty}\mathsf{P}_x(x^{-1}) = \lim_{x\to\infty} x^{1/x} = 1.$

Proof We have

$$\lim_{x\to\infty}\mathsf{P}_x(x^{-1}) = \lim_{x\to\infty}\exp(x^{-1}\log(x)) = \lim_{y\to 0}\exp(y) = 1,$$

using part (vii) of Proposition 3.8.6. ■

Now we turn to the process of inverting the power function. For the exponential function we required that $\log(e^x) = x$. Thus, if our inverse of P_a is denoted (for the moment) by f_a, then we expect that $f_a(a^x) = x$. This definition clearly has difficulties when $a = 1$, reflecting the fact that P_1 is not invertible. In all other case, since P_a is continuous, and either strictly monotonically increasing or strictly monotonically decreasing, we have the following definition, using Theorem 3.1.30.

3.8.13 Definition (Arbitrary base logarithm) For $a \in \mathbb{R}_{>0} \setminus \{1\}$, the function $\log_a\colon \mathbb{R}_{>0} \to \mathbb{R}$, called the ***base a logarithmic function***, is the inverse of P_a. When $a = 10$ we simply write $\log_{10} = \log$. •

The following result relates the logarithmic function for an arbitrary base to the natural logarithmic function.

3.8.14 Proposition (Characterisation of $\log_a$) $\log_a(x) = \dfrac{\log(x)}{\log(a)}.$

Proof Let $x \in \mathbb{R}_{>0}$ and write $x = a^y$ for some $y \in \mathbb{R}$. First suppose that $y \neq 0$. Then we have $\log(x) = y\log(a)$ and $\log_a(x) = y$, and the result follows by eliminating y from these two expressions. When $y = 0$ we have $x = a = a^1$. Therefore, $\log_a(x) = 1 = \frac{\log(x)}{\log(a)}$. ■

With this result we immediately have the following generalisation of Proposition 3.8.6. We leave the trivial checking of the details to the reader.

3.8.15 Proposition (Properties of $\log_a$) *For* $a \in \mathbb{R}_{>0} \setminus \{1\}$*, the function* $\log_a$ *enjoys the following properties:*

(i) $\log_a$ *is infinitely differentiable;*

(ii) $\log_a$ *is strictly monotonically increasing when* $a > 1$ *and is strictly monotonically decreasing when* $a < 1$*;*

(iii) $\log_a(x) = \frac{1}{\log(a)}\int_1^x \frac{1}{\xi}\,d\xi$ *for all* $x \in \mathbb{R}_{>0}$*;*

(iv) $\lim_{x\to\infty}\log_a(x) = \begin{cases}\infty, & a > 1,\\ -\infty, & a < 1;\end{cases}$

(v) $\lim_{x\downarrow 0}\log_a(x) = \begin{cases}-\infty, & a > 1,\\ \infty, & a < 1;\end{cases}$

(vi) $\log_a(xy) = \log_a(x) + \log_a(y)$ *for all* $x, y \in \mathbb{R}_{>0}$*;*

(vii) $\lim_{x\to\infty} x^{-k}\log_a(x) = 0$ *for all* $k \in \mathbb{Z}_{>0}$.

3.8.4 Trigonometric functions

Next we turn to describing the standard trigonometric functions. These functions are perhaps most intuitively introduced in terms of the concept of "angle" in plane geometry. However, to really do this properly would, at this juncture, require a significant expenditure of effort. Therefore, we define the trigonometric functions by their power series expansion, and then proceed to show that they have the expected properties. In the course of our treatment we will also see that the constant π introduced in Section 2.4.3 has the anticipated relationships to the trigonometric functions. Convenience in this section forces us to make a fairly serious logical jump in the presentation. While all constructions and theorems are stated in terms of real numbers, in the proofs we use complex numbers rather heavily.

3.8.16 Definition (sin and cos) The ***sine function***, denoted by $\sin\colon \mathbb{R} \to \mathbb{R}$, and the ***cosine function***, denoted by $\cos\colon \mathbb{R} \to \mathbb{R}$, are defined by

$$\sin(x) = \sum_{j=1}^{\infty} \frac{(-1)^{j+1} x^{2j-1}}{(2j-1)!}, \quad \cos(x) = \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j}}{(2j)!},$$

respectively. •

In Figure 3.18 we show the graphs of the functions sin and cos.

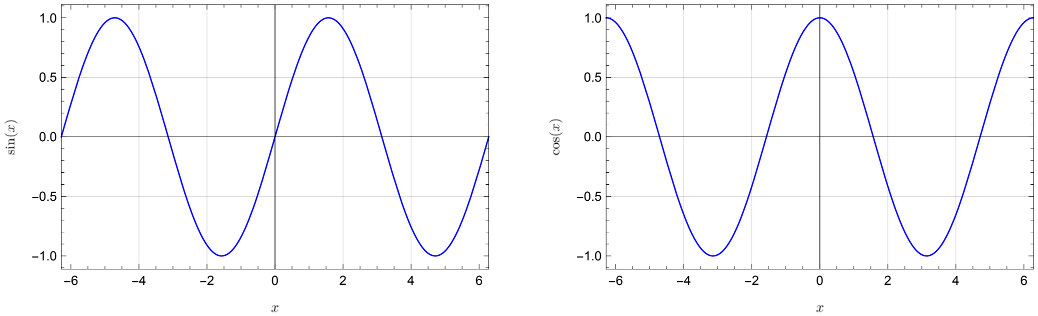


Figure 3.18 The functions sin (left) and cos (right)

3.8.17 Notation Following normal conventions, we shall frequently write $\sin x$ and $\cos x$ rather than the more correct $\sin(x)$ and $\cos(x)$. •

An application of Proposition 2.4.15 and Theorem 3.7.13 shows that the power series expansions for sin and cos are, in fact, convergent for all x, and so the functions are indeed defined with domain $\mathbb{R}$.

First we prove the existence of a number having the property that we know π to possess. In fact, we construct the number $\frac{\pi}{2}$, where π is as given in Section 2.4.3.

3.8.18 Theorem (Construction of π) *There exists a positive real number* p_0 *such that*

$$\mathrm{p}_0 = \inf\{\mathrm{x} \in \mathbb{R}_{>0} \mid \cos(\mathrm{x}) = 0\}.$$

Moreover, $\mathrm{p}_0 = \frac{\pi}{2}$.

Proof First we record the derivative properties for sin and cos.

1 Lemma *The functions* sin *and* cos *are infinitely differentiable and satisfy* $\sin' = \cos$ *and* $\cos' = -\sin$.

Proof This follows directly from Proposition 3.7.20 where it is shown that convergent power series can be differentiated term-by-term. ▼

Let us now perform some computations using complex variables that will be essential to many of the proofs in this section. We suppose the reader to be acquainted with the necessary elementary facts about complex numbers. The next observation is the most essential along these lines. We denote $\mathbb{S}^{\mathbb{C}}_1 = \{z \in \mathbb{C} \mid |z| = 1\}$, and recall that all points in $z \in \mathbb{S}^1_{\mathbb{C}}$ can be written as $z = \mathrm{e}^{\mathrm{i}x}$ for some $x \in \mathbb{R}$, and that, conversely, for any $x \in \mathbb{R}$ we have $\mathrm{e}^{\mathrm{i}x} \in \mathbb{S}^1_{\mathbb{C}}$.

2 Lemma $\mathrm{e}^{\mathrm{ix}} = \cos(\mathrm{x}) + \mathrm{i}\sin(\mathrm{x})$.

Proof This follows immediately from the $\mathbb{C}$-power series for the complex exponential function:

$$\mathrm{e}^z = \sum_{j=0}^{\infty} \frac{x^j}{j!}.$$

Substituting $z = \mathrm{i}x$, using the fact that $\mathrm{i}^{2j} = (-1)^j$ for all $j \in \mathbb{Z}_{>0}$, and using Proposition 2.4.30, we get the desired result. ▼

From the preceding lemma we then know that $\cos(x) = \mathrm{Re}(\mathrm{e}^{\mathrm{i}x})$ and that $\sin(x) = \mathrm{Im}(\mathrm{e}^{\mathrm{i}x})$. Therefore, since $\mathrm{e}^{\mathrm{i}x} \in \mathbb{S}^1_{\mathbb{C}}$, we have

$$\cos(x)^2 + \sin(x)^2 = 1. \tag{3.26}$$

Let us show that the set $\{x \in \mathbb{R}_{>0} \mid \cos(x) = 0\}$ is nonempty. Suppose that it is empty. Since $\cos(0) = 1$ and since cos is continuous, it must therefore be the case (by the Intermediate Value Theorem) that $\cos(x) > 0$ for all $x \in \mathbb{R}$. Therefore, by Lemma 1, $\sin'(x) > 0$ for all $x \in \mathbb{R}$, and so sin is strictly monotonically increasing by Proposition 3.2.23. Therefore, since $\sin(0) = 0$, $\sin(x) > 0$ for $x > 0$. Therefore, for $x_1, x_2 \in \mathbb{R}_{>0}$ satisfying $x_1 < x_2$, we have

$$\sin(x_1)(x_2 - x_1) < \int_{x_1}^{x_2} \sin(x)\,\mathrm{d}x = \cos(x_2) - \cos(x_1) \le 2,$$

where we have used the fact that sin is strictly monotonically increasing, Lemma 1, the Fundamental Theorem of Calculus, and (3.26). We thus have arrive at the contradiction that $\limsup_{x_2 \to \infty} \sin(x_1)(x_2 - x_1) \le 2$.

Since cos is continuous, the set $\{x \in \mathbb{R}_{>0} \mid \cos(x) = 0\}$ is closed. Therefore, $\inf\{x \in \mathbb{R}_{>0} \mid \cos(x) = 0\}$ is contained in this set, and this gives the existence of p_0. Note that, by (3.26), $\sin(p_0) \in \{-1, 1\}$. Since $\sin(0) = 0$ and since $\sin(x) = \cos(x) > 0$ for $x \in [0, p_0)$, we must have $\sin(p_0) = 1$.

The following property of p_0 will also be important.

3 Lemma $\cos(\frac{\mathrm{p}_0}{2}) = \sin(\frac{\mathrm{p}_0}{2}) = \frac{1}{\sqrt{2}}$.

Proof Let $x_0 = \cos(\frac{p_0}{2})$, $y_0 = \sin(\frac{p_0}{2})$, and $z_0 = x_0 + \mathrm{i}y_0$. Then, using Proposition II-3.4.1,

$$(\mathrm{e}^{\mathrm{i}\frac{p_0}{2}})^2 = \mathrm{e}^{\mathrm{i}p_0} = \mathrm{i}$$

since $\cos(p_0) = 0$ and $\sin(p_0) = 1$. Thus

$$(\mathrm{e}^{\mathrm{i}\frac{p_0}{2}})^4 = \mathrm{i}^2 = -1,$$

again using Proposition II-3.4.1. Using the definition of complex multiplication we also have

$$(\mathrm{e}^{\mathrm{i}\frac{p_0}{2}})^4 = (x_0 + \mathrm{i}y_0)^4 = x_0^4 - 6x_0^2y_0^2 + y_0^4 + 4\mathrm{i}x_0y_0(x_0^2 - y_0^2).$$

Thus, in particular, $x_0^2 - y_0^2 = 0$. Combining this with $x_0^2 + y_0^2 = 1$ we get $x_0^2 = y_0^2 = \frac{1}{2}$. Since both x_0 and y_0 are positive by virtue of $\frac{p_0}{2}$ lying in $(0, p_0)$, we must have $x_0 = y_0 = \frac{1}{\sqrt{2}}$, as claimed. ▼

Now we show, through a sequence of seemingly irrelevant computations, that $p_0 = \frac{\pi}{2}$. Define the function $\tan\colon (-p_0, p_0) \to \mathbb{R}$ by $\tan(x) = \frac{\sin(x)}{\cos(x)}$, noting that tan is well-defined since $\cos(-x) = \cos(x)$ and since $\cos(x) > 0$ for $x \in [0, p_0)$. We claim that tan is continuous and strictly monotonically increasing. We have, using the quotient rule,

$$\tan'(x) = \frac{\cos(x)^2 + \sin(x)^2}{\cos(x)^2} = \frac{1}{\cos(x)^2}.$$

Thus $\tan'(x) > 0$ for all $x \in (-p_0, p_0)$, and so tan is strictly monotonically increasing by Proposition 3.2.23. Since $\sin(p_0) = 1$ and (since $\sin(-x) = -\sin(x)$) since $\sin(-p_0) = -1$, we have

$$\lim_{x \uparrow p_0} \tan(x) = \infty, \quad \lim_{x \downarrow p_0} \tan(x) = -\infty.$$

This shows that tan is an invertible and differentiable mapping from $(-p_0, p_0)$ to $\mathbb{R}$. Moreover, since tan′ is nowhere zero, the inverse, denoted by $\tan^{-1}\colon \mathbb{R} \to (-p_0, p_0)$, is also differentiable and the derivative of its inverse is given by

$$(\tan^{-1})'(x) = \frac{1}{\tan'(\tan^{-1}(x))},$$

as per Theorem 3.2.24. We further claim that

$$(\tan^{-1})'(x) = \frac{1}{1 + x^2}.$$

Indeed, our above arguments show that $(\tan^{-1})'(x) = (\cos(\tan^{-1}(x)))^2$. If $y = \tan^{-1}(x)$ then

$$\frac{\sin(y)}{\cos(y)} = x.$$

Since $\sin(y) > 0$ for $y \in (0, p_0)$, we have $\sin(y) = \sqrt{1 - \cos(y)}$ by (3.26). Therefore,

$$\frac{1 - \cos(y)^2}{\cos(y)^2} = x^2 \quad \Longrightarrow \quad \cos(y)^2 = \frac{1}{1 + x^2}$$

as desired.

By the Fundamental Theorem of Calculus we then have

$$\int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x = \tan^{-1}(1) - \tan^{-1}(0).$$

Since $\tan^{-1}(1) = \frac{p_0}{2}$ by Lemma 3 above and since $\tan^{-1}(0) = 0$ (and using part (v) of Proposition 3.8.19 below), we have

$$\int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x = \frac{p_0}{2}. \tag{3.27}$$

Now recall from Example 3.7.28–1 that we have

$$\frac{1}{1+x^2} = \sum_{j=0}^{\infty} (-1)^j x^{2j},$$

with the series converging uniformly on any compact subinterval of $(-1,1)$. Therefore, by Proposition 3.7.20, for $\epsilon \in (0,1)$ we have

$$\begin{aligned}\int_0^{1-\epsilon} \frac{1}{1+x^2}\,\mathrm{d}x &= \int_0^{1-\epsilon} \sum_{j=0}^{\infty} (-1)^j x^{2j}\,\mathrm{d}x\\ &= \sum_{j=0}^{\infty} (-1)^j \int_0^{1-\epsilon} x^{2j}\,\mathrm{d}x\\ &= \sum_{j=0}^{\infty} (-1)^j \frac{(1-\epsilon)^{2j+1}}{2j+1}.\end{aligned}$$

The following technical lemma will allow us to conclude the proof.

4 Lemma $\lim_{\epsilon\downarrow 0} \sum_{j=0}^{\infty} (-1)^j \frac{(1-\epsilon)^{2j+1}}{2j+1} = \sum_{j=0} \frac{(-1)^j}{2j+1}.$

Proof By the Alternating Test, the series $\sum_{j=0}^{\infty} (-1)^j \frac{(1-\epsilon)^{2j+1}}{2j+1}$ converges for $\epsilon \in [0,2]$. Define $f\colon [0,2] \to \mathbb{R}$ by

$$f(x) = \sum_{j=0}^{\infty} (-1)^{j+1} \frac{(x-1)^{2j+1}}{2j+1}$$

and define $g\colon [-1,1] \to \mathbb{R}$ by

$$g(x) = \sum_{j=0}^{\infty} (-1)^{j+1} \frac{x^{2j+1}}{2j+1}$$

so that $f(x) = g(x-1)$. Since g is defined by a $\mathbb{R}$-convergent power series, by Corollary 3.7.18 g is continuous. In particular,

$$g(-1) = \lim_{x\downarrow -1} \sum_{j=0}^{\infty} (-1)^{j+1} \frac{x^{2j+1}}{2j+1}.$$

From this it follows that

$$f(0) = \lim_{x\downarrow 0} \sum_{j=0}^{\infty} (-1)^{j+1} \frac{(x-1)^{2j+1}}{2j+1},$$

which is the result. ▼

Combining this with (3.27) we have

$$\frac{p_0}{2} = \lim_{\epsilon\downarrow 0} \int_0^{1-\epsilon} \frac{1}{1+x^2}\,\mathrm{d}x = \lim_{\epsilon\downarrow 0} \sum_{j=0}^{\infty} (-1)^j \frac{(1-\epsilon)^{2j+1}}{2j+1} = \sum_{j=0} \frac{(-1)^j}{2j+1} = \frac{\pi}{4},$$

using the definition of π in Definition 2.4.20. ■

Now that we have on hand a reasonable characterisation of π, we can proceed to state the familiar properties of sin and cos.

3.8.19 Proposition (Properties of sin and cos) *The functions* sin *and* cos *enjoy the following properties:*

(i) sin *and* cos *are infinitely differentiable, and furthermore satisfy* $\sin' = \cos$ *and* $\cos' = -\sin$*;*

(ii) $\sin(-x) = \sin(x)$ *and* $\cos(-x) = \cos(x)$ *for all* $x \in \mathbb{R}$*;*

(iii) $\sin(x)^2 + \cos(x)^2 = 1$ *for all* $x \in \mathbb{R}$*;*

(iv) $\sin(x + 2\pi) = \sin(x)$ *and* $\cos(x + 2\pi) = \cos(x)$ *for all* $x \in \mathbb{R}$*;*

(v) the map

$$[0, 2\pi) \ni x \mapsto (\cos(x), \sin(x)) \in \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$$

is a bijection.

Proof (i) This was proved as Lemma 1 in the proof of Theorem 3.8.18.

(ii) This follows immediately from the ℝ-power series for sin and cos.

(iii) This was proved as (3.26) in the course of the proof of Theorem 3.8.18.

(iv) Since $e^{i\frac{\pi}{2}} = i$ by Theorem 3.8.18, we use Proposition II-3.4.1 to deduce

$$e^{2\pi i} = (e^{i\frac{\pi}{2}})^4 = i^4 = 1.$$

Again using Proposition II-3.4.1 we then have

$$e^{z+2\pi i} = e^z e^{2\pi i} = e^z$$

for all $z \in \mathbb{C}$. Therefore, for $x \in \mathbb{R}$, we have

$$\cos(x + 2\pi) + i\sin(x + 2\pi) = e^{i(x+2\pi)} = e^{ix} = \cos(x) + i\sin(x),$$

which gives the result.

(v) Denote $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$, and note that, if we make the standard identification of $\mathbb{C}$ with $\mathbb{R}^2$ (as we do), then $\mathbb{S}^1_{\mathbb{C}}$ (see the proof of Theorem 3.8.18) becomes identified with $\mathbb{S}^1$, with the identification explicitly being $x + \mathrm{i}y \mapsto (x, y)$. Thus the result we are proving is equivalent to the assertion that the map

$$f\colon [0, 2\pi) \ni x \mapsto \mathrm{e}^{\mathrm{i}x} \in \mathbb{S}^1_{\mathbb{C}}$$

is a bijection. This is what we will prove. By part (iii), this map is well-defined in the sense that it actually does take values in $\mathbb{S}^1_{\mathbb{C}}$. Suppose that $\mathrm{e}^{\mathrm{i}x_1} = \mathrm{e}^{\mathrm{i}x_2}$ for distinct points $x_1, x_2 \in [0, 2\pi)$, and suppose for concreteness that $x_1 < x_2$. Then $x_2 - x_1 \in (0, 2\pi)$, and $\frac{1}{4}(x_2 - x_1) \in (0, \frac{\pi}{2})$. We then have

$$\mathrm{e}^{\mathrm{i}x_1} = \mathrm{e}^{\mathrm{i}x_2} \quad \Longrightarrow \quad \mathrm{e}^{\mathrm{i}(x_2 - x_1)} = 1 \quad \Longrightarrow \quad (\mathrm{e}^{\mathrm{i}\frac{1}{4}(x_2 - x_1)})^4 = 1.$$

Let $\mathrm{e}^{\mathrm{i}\frac{1}{4}(x_2 - x_1)} = \xi + \mathrm{i}\eta$. Since $\frac{1}{4}(x_2 - x_1) \in (0, \frac{\pi}{2})$, we saw during the course of the proof of Theorem 3.8.18 that $\xi, \eta \in (0, 1)$. We then use the definition of complex multiplication to compute

$$(\mathrm{e}^{\mathrm{i}\frac{1}{4}(x_2 - x_1)})^4 = \xi^4 - 6\xi^2\eta^2 + \eta^4 + 4\mathrm{i}\xi\eta(\xi^2 - \eta^2).$$

Since $(\mathrm{e}^{\mathrm{i}\frac{1}{4}(x_2 - x_1)})^4 = 1$ is real, we conclude that $\xi^2 - \eta^2 = 0$. Combining this with $\xi^2 + \eta^2 = 1$ gives $\xi^2 = \eta^2 = \frac{1}{2}$. Since both ξ and η are positive we have $\xi = \eta = \frac{1}{\sqrt{2}}$. Substituting this into the above expression for $(\mathrm{e}^{\mathrm{i}\frac{1}{4}(x_2 - x_1)})^4$ gives $(\mathrm{e}^{\mathrm{i}\frac{1}{4}(x_2 - x_1)})^4 = -1$. Thus we arrive at a contradiction, and it cannot be the case that $\mathrm{e}^{\mathrm{i}x_1} = \mathrm{e}^{\mathrm{i}x_2}$ for distinct $x_1, x_2 \in [0, 2\pi)$. Thus f is injective.

To show that f is surjective, we let $z = x + \mathrm{i}y \in \mathbb{S}^1_{\mathbb{C}}$, and consider four cases.

1. $x, y \geq 0$: Since cos is monotonically decreasing from 1 to 0 on $[0, \frac{\pi}{2}]$, there exists $\theta \in [0, \frac{\pi}{2}]$ such that $\cos(\theta) = x$. Since $\sin(\theta)^2 = 1 - \cos(\theta)^2 = 1 - x^2 = y^2$, and since $\sin(\theta) \geq 0$ for $\theta \in [0, \frac{\pi}{2}]$, we conclude that $\sin(\theta) = y$. Thus $z = \mathrm{e}^{\mathrm{i}\theta}$.
2. $x \geq 0$ and $y \leq 0$: Let $\xi = x$ and $\eta = -y$ so that $\xi, \eta \geq 0$. From the preceding case we deduce the existence of $\phi \in [0, \frac{\pi}{2}]$ such that $\mathrm{e}^{\mathrm{i}\phi} = \xi + \mathrm{i}\eta$. Thus $\cos(\phi) = x$ and $\sin(\phi) = -y$. By part (ii) we then have $\cos(-\phi) = x$ and $\sin(-\phi) = y$, and we note that $-\phi \in [-\frac{\pi}{2}, 0]$. Define

   $$\theta = \begin{cases} 2\pi - \phi, & \phi \in (0, \frac{\pi}{2}], \\ 0, & \phi = 0. \end{cases}$$

   By part (iv) we then have $\cos(\theta) = x$ and $\sin(\theta) = y$, and that $\theta \in [\frac{3\pi}{2}, 2\pi)$ if $\phi \in (0, \frac{\pi}{2}]$.
3. $x \leq 0$ and $y \geq 0$: Let $\xi = -x$ and $\eta = y$ si that $\xi, \eta \geq 0$. As in the first case we have $\phi \in [0, \frac{\pi}{2}]$ such that $\cos(\phi) = \xi$ and $\sin(\phi) = \eta$. We then have $-\cos(\phi) = x$ and $\sin(\phi) = y$. Next define $\theta = \pi - \phi$ and note that

   $$\mathrm{e}^{\mathrm{i}\theta} = \mathrm{e}^{\mathrm{i}\pi}\mathrm{e}^{-\mathrm{i}\phi} = -(\cos(\phi) - \mathrm{i}\sin(\phi)) = -\cos(\phi) + \mathrm{i}\sin(\phi) = x + \mathrm{i}y,$$

   as desired.
4. $x \leq 0$ and $y \leq 0$: Take $\xi = -x$ and $\eta = -y$ so that $\xi, \eta \geq 0$. As in the first case, we have $\phi \in [0, \frac{\pi}{2}]$ such that $\cos(\phi) = \xi = -x$ and $\sin(\phi) = \eta = -y$. Then, taking $\theta = \pi + \phi$, we have

   $$\mathrm{e}^{\mathrm{i}\theta} = \mathrm{e}^{\mathrm{i}\pi}\mathrm{e}^{\mathrm{i}\phi} = -(\cos(\phi) + \mathrm{i}\sin(\phi)) = x + \mathrm{i}y,$$

   as desired. ■

From the basic construction of sin and cos that we give, and the properties that follow directly from this construction, there is of course a great deal that one can proceed to do; the resulting subject is broadly called “trigonometry.” Rigorous proofs of many of the facts of basic trigonometry follow easily from our constructions here, particularly since we give the necessary properties, along with a rigorous definition, of π. We do assume that the reader has an acquaintance with trigonometry, as we shall use certain of these facts without much ado.

The reciprocals of sin and cos are sometimes used. Thus we define $\csc\colon (0, 2\pi) \to \mathbb{R}$ and $\sec\colon (-\pi, \pi) \to \mathbb{R}$ by $\csc(x) = \frac{1}{\sin(x)}$ and $\sec(x) = \frac{1}{\cos(x)}$. These are the ***cosecant*** and ***secant*** functions, respectively. One can verify that the restrictions of csc and sec to $(0, \frac{\pi}{2})$ are bijective. In Figure 3.19

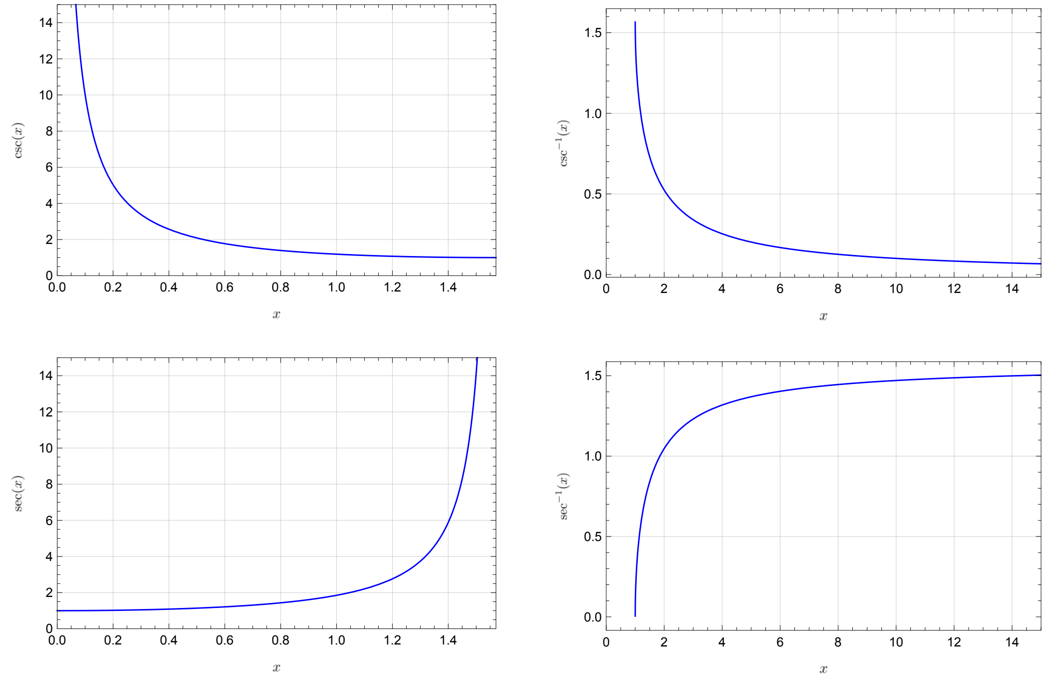


Figure 3.19 Cosecant and its inverse (top) and secant and its inverse (bottom) on $(0, \frac{\pi}{2})$

One useful and not perfectly standard construction is the following. Define $\tan\colon (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$ by $\tan(x) = \frac{\sin(x)}{\cos(x)}$, noting that the definition makes sense since $\cos(x) > 0$ for $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$. In Figure 3.20 we depict the graph of tan and its inverse $\tan^{-1}$. During the course of the proof of Theorem 3.8.18 we showed that the function tan had the following properties.

3.8.20 Proposition (Properties of tan) *The function* tan *enjoys the following properties:*

(i) tan *is infinitely differentiable;*

(ii) tan *is strictly monotonically increasing;*

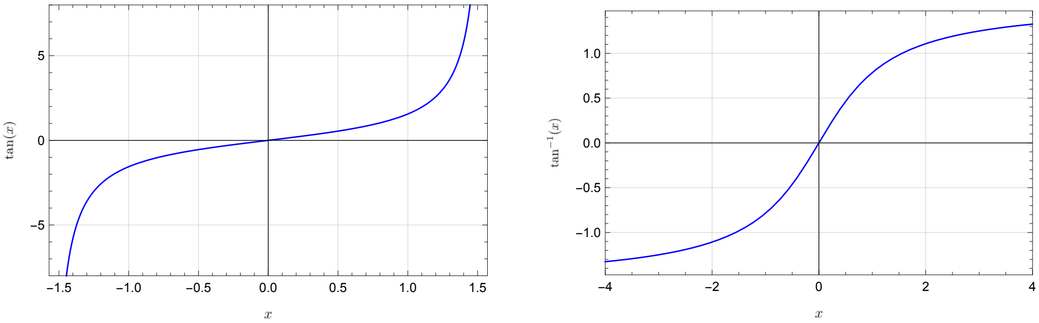


Figure 3.20 The function tan (left) and its inverse $\tan^{-1}$ (right)

(iii) the inverse of tan*, denoted by* $\tan^{-1}\colon \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$ *is infinitely differentiable.*

It turns out to be useful to extend the definition of $\tan^{-1}$ to $(-\pi, \pi]$ by defining the function $\operatorname{atan}\colon \mathbb{R}^2 \setminus \{(0,0)\} \to (-\pi, \pi]$ by

$$\operatorname{atan}(x,y) = \begin{cases} \tan^{-1}(\frac{y}{x}), & x > 0, \\ \pi - \tan^{-1}(\frac{y}{x}), & x < 0, \\ \frac{\pi}{2}, & x = 0,\ y > 0, \\ -\frac{\pi}{2}, & x = 0,\ y < 0. \end{cases}$$

As we shall see in Section II-3.1 when we discuss the geometry of the complex plane, this function returns that angle of a point (x, y) measured from the positive x-axis.

3.8.5 Hyperbolic trigonometric functions

In this section we shall quickly introduce the hyperbolic trigonometric functions. Just why these functions are called "trigonometric" is only best seen in the setting of $\mathbb{C}$-valued functions in Section II-3.4.1.

3.8.21 Definition (sinh and cosh) The ***hyperbolic sine function***, denoted by $\sinh\colon \mathbb{R} \to \mathbb{R}$, and the ***hyperbolic cosine function***m denoted by $\cosh\colon \mathbb{R} \to \mathbb{R}$, are defined by

$$\sinh(x) = \sum_{j=1}^{\infty} \frac{x^{2j-1}}{(2j-1)!}, \quad \cosh(x) = \sum_{j=0}^{\infty} \frac{x^{2j}}{(2j)!},$$

respectively. •

In Figure 3.21 we depict the graphs of sinh and cosh.

As with sin and cos, an application of Proposition 2.4.15 and Theorem 3.7.13 shows that the power series expansions for sinh and cosh are convergent for all x.

The following result gives some of the easily determined properties of sinh and cosh.

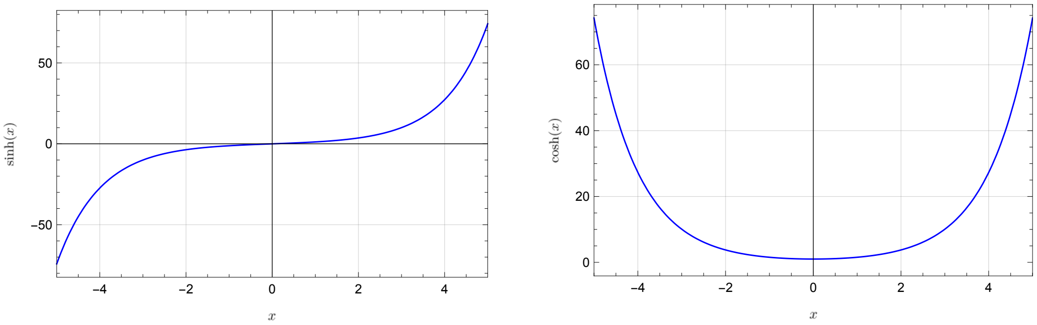


Figure 3.21 The functions sinh (left) and cosh (right)

3.8.22 Proposition (Properties of sinh and cosh) *The functions* sinh *and* cosh *enjoy the following properties:*

(i) $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$ *and* $\cosh(x) = \frac{1}{2}(e^x + e^{-x})$*;*

(ii) sinh *and* cosh *are infinitely differentiable, and furthermore satisfy* $\sinh' = \cosh$ *and* $\cosh' = \sinh$*;*

(iii) $\sinh(-x) = \sinh(x)$ *and* $\cosh(-x) = \cosh(x)$ *for all* $x \in \mathbb{R}$*;*

(iv) $\cosh(x)^2 - \sinh(x)^2 = 1$ *for all* $x \in \mathbb{R}$*.*

Proof (i) These follows directly from the $\mathbb{R}$-power series definitions for exp, sinh, and cosh.

(ii) This follows from Corollary 3.7.21 and the fact that $\mathbb{R}$-convergent power series can be differentiated term-by-term.

(iii) These follow directly from the $\mathbb{R}$-power series for sinh and cosh.

(iv) This can be proved directly using part (i). ■

Also sometimes useful is the ***hyperbolic tangent function*** $\tanh\colon \mathbb{R} \to \mathbb{R}$ defined by $\tanh(x) = \frac{\sinh(x)}{\cosh(x)}$.

Exercises

3.8.1 For representative values of $a \in \mathbb{R}_{>0}$, give the graph of P_a, showing the features outlined in Proposition 3.8.10.

3.8.2 For representative values of $a \in \mathbb{R}$, give the graph of P^a, showing the features outlined in Proposition 3.8.11.

3.8.3 Prove the following trigonometric identities:

(a) $\cos a \cos b = \frac{1}{2}(\cos(a+b) + \cos(a-b))$;

(b) $\cos a \sin b = \frac{1}{2}(\sin(a+b) - \sin(a-b))$;

(c) $\sin a \sin b = \frac{1}{2}(\cos(a-b) - \cos(a+b))$.

3.8.4 Prove the following trigonometric identities:

(a)

3.8.5 Show that tanh is injective.

Chapter 4

Algebraic structures

During the course of these volumes, we shall occasionally, sometimes in essential ways, make use of certain ideas from abstract algebra, particular abstract linear algebra. In this chapter we provide the necessary background in abstract algebra, saving the subject of linear algebra for Chapter 5. Our idea is to provide sufficient detail to give some context to the instances when we make use of algebra.

Do I need to read this chapter? Provided that the reader is comfortable with the very basic arithmetic ideas concerning integers, real numbers, complex numbers, and polynomials, the material in Sections 4.1–4.7 can probably be skipped until it is needed in the course of the text. When it is needed, however, a reader with little exposure to abstract algebra can expect to expend some effort even for the basic material we present here. The material in Section 4.5 appears immediately in Chapter IV-1 in our initial consideration of the concept of spaces of signals. For this reason, the material should be considered essential. However, it is possible that certain parts of the chapter can be skimmed at a first reading, since the most essential concept is that of a vector space as defined and discussed in Section 4.5. The preparatory material of Sections 4.1–4.7 in not essential for understanding what a vector space is, particularly if one is comfortable with the algebraic structure of the set $\mathbb{R}$ of real numbers and the set $\mathbb{C}$ of complex numbers. Section 4.8 will not be important for significant portions of the text, so can easily be skipped until needed or wanted. •

Contents

Section 4.1

Groups

One of the basic structures in mathematics is that of a group. A group structure often forms the building block for more particular algebraic structures.

Do I need to read this section? Since the material in this section is not difficult, although it is abstract, it may be useful reading for those who feel as if they need to get some familiarity with simple abstract constructions and proofs. The content of the section itself is necessary reading for those who want to understand the material in Sections 4.2–4.4. •

4.1.1 Definitions and basic properties

There are a few structures possessing less structure than a group, so we first define these. Many of our definitions of algebraic structure involve the notion of a "binary operation," so let us make this precise.

4.1.1 Definition (Binary operation) A ***binary operation*** on a set S is a map $B\colon S \times S \to S$. A pair (S, B) where B is a binary operation on S is a ***magma***. •

We begin with one of the most basic of algebraic structures, even more basic than a group.

4.1.2 Definition (Semigroup) A ***semigroup*** is a nonempty set S with a binary operation on S, denoted by $(s_1, s_2) \mapsto s_1 \cdot s_2$, having the property that

(i) $(s_1 \cdot s_2) \cdot s_3 = s_1 \cdot (s_2 \cdot s_3)$ for all $s_1, s_2, s_3 \in \mathsf{S}$ (***associativity***). •

Slightly more structured than a semigroup is the idea of a monoid.

4.1.3 Definition (Monoid) A ***monoid*** is a nonempty set M with a binary operation on M, denoted by $(m_1, m_2) \mapsto m_1 \cdot m_2$, having the following properties:

(i) $m_1 \cdot (m_2 \cdot m_3) = (m_1 \cdot m_2) \cdot m_3$ for all $m_1, m_2, m_3 \in \mathsf{M}$ (***associativity***);

(ii) there exists $e \in \mathsf{M}$ such that $m \cdot e = e \cdot m = m$ for all $m \in \mathsf{M}$ (***identity element***). •

Now we define what we mean by a group.

4.1.4 Definition (Group) A ***group*** is a nonempty set G endowed with a binary operation, denoted by $(g_1, g_2) \mapsto g_1 \cdot g_2$, having the following properties:

(i) $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$ for all $g_1, g_2, g_3 \in \mathsf{G}$ (***associativity***);

(ii) there exists $e \in \mathsf{G}$ such that $g \cdot e = e \cdot g = g$ for all $g \in \mathsf{G}$ (***identity element***);

(iii) for each $g \in \mathsf{G}$ there exists $g^{-1} \in \mathsf{G}$ such that $g \cdot g^{-1} = g^{-1} \cdot g = e$ (***inverse element***).

A group is ***Abelian*** if $g_1 \cdot g_2 = g_2 \cdot g_1$ for all $g_1, g_2 \in \mathsf{G}$. •

As we did when we defined the operation of multiplication in $\mathbb{R}$, we will often omit the symbol "·" for the binary operation in a group (or semigroup or monoid), and simply write $g_1 g_2$ in place of $g_1 \cdot g_2$. When talking simultaneously about more than one group, it is sometimes advantageous to denote the identity element of a group G by e_{G}.

Clearly the following inclusions hold:

$$\text{Semigroups} \subseteq \text{Monoids} \subseteq \text{Groups}.$$

Throughout these volumes, we shall encounter many examples of groups. For the moment, let us give some very simple examples that illustrate the difference between the ideas of a semigroup, monoid, and group.

4.1.5 Examples (Semigroups, monoids, and groups)

1. A singleton $\{x\}$ with the (only possible) binary operation $x \cdot x = x$ is a group with identity element x and with inverse element defined by $x^{-1} = x$.
2. The set $\mathbb{Z}_{>0}$ with the binary operation of addition is a semigroup. However, it is not a monoid since it has no identity element, and it is not a group, because it has no identity element and so there are also no inverse elements.
3. The set $\mathbb{Z}_{>0}$ with the binary operation of multiplication is a monoid with identity element $e = 1$. It is not a group.
4. The set $\mathbb{Z}_{\geq 0}$ with the binary operation of addition is a monoid with identity element 0, but not a group.
5. The set $\mathbb{Z}_{\geq 0}$ with the binary operation of multiplication is a monoid with identity element 1. It is not a group.
6. The set $\mathbb{Z}$ with the binary operation of addition is a group with identity element 0, and with inverse defined by $k^{-1} = -k$.
7. The set $\mathbb{Z}$ with the binary operation of multiplication is a monoid with identity 1, but it is not a group.
8. The sets $\mathbb{Q}$ and $\mathbb{R}$ with the binary operations of addition are groups with identity element 0 and with inverse defined by $x^{-1} = -x$.
9. The sets $\mathbb{Q}$ and $\mathbb{R}$ with the binary operations of multiplication are monoids with identity element 1. They are not groups.
10. The sets $\mathbb{Q}^* \triangleq \mathbb{Q} \setminus \{0\}$ and $\mathbb{R}^* \triangleq \mathbb{R} \setminus \{0\}$ with the binary operation of multiplication are groups with identity element 1 and with inverse given by $x^{-1} = \frac{1}{x}$.
11. Let $\mathfrak{S}_k$, $k \in \mathbb{Z}_{>0}$, denote the set of bijections of the set $\{1, \dots, k\}$, and equip $\mathfrak{S}_k$ with the binary operation $(\sigma_1, \sigma_2) \mapsto \sigma_1 \circ \sigma_2$. One can easily verify that $\mathfrak{S}_k$ is a group with identity given by the identity map, and with inverse given by the inverse map. This group is called the ***permutation group*** or the ***symmetric***

group on k symbols. It is conventional to represent a permutation $\sigma \in \mathfrak{S}_k$ using the following matrix-type representation:

$$\begin{pmatrix} 1 & 2 & \cdots & k \\ \sigma(1) & \sigma(2) & \cdots & \sigma(k) \end{pmatrix}.$$

Thus the first row contains the elements $\{1, \dots, k\}$ in order, and the second row contains the images of these elements under σ.

We claim that $\mathfrak{S}_k$ is Abelian when $k \in \{1, 2\}$, and otherwise is not Abelian. We leave it to the reader to check directly that $\mathfrak{S}_1$ and $\mathfrak{S}_2$ are Abelian. Let us show that $\mathfrak{S}_3$ is not Abelian. Define $\sigma_1, \sigma_2 \in \mathfrak{S}_3$ by

$$\begin{aligned} \sigma_1(1) = 2, \quad \sigma_1(2) = 1, \quad \sigma_1(3) = 3, \\ \sigma_2(1) = 1, \quad \sigma_2(2) = 3, \quad \sigma_2(3) = 2. \end{aligned}$$

One can then verify that

$$\begin{aligned} \sigma_1 \circ \sigma_2(1) = 2, \quad \sigma_1 \circ \sigma_2(2) = 3, \quad \sigma_1 \circ \sigma_2(3) = 1, \\ \sigma_2 \circ \sigma_1(1) = 3, \quad \sigma_2 \circ \sigma_1(2) = 1, \quad \sigma_2 \circ \sigma_1(3) = 2. \end{aligned}$$

Thus $\mathfrak{S}_3$ in indeed not Abelian.

That $\mathfrak{S}_k$ is not Abelian for $k > 3$ follows since in Example 4.1.12–7 we will show that $\mathfrak{S}_3$ is a isomorphic to a subgroup of $\mathfrak{S}_k$ (asking the readers forgiveness that the terms "isomorphic" and "subgroup" have yet to be defined; they will be shortly).

We shall have more to say about the symmetric group in Section 4.1.6.

All groups in the above list may be verified to be Abelian, with the exception of the permutation group on k symbols for $k \geq 2$. •

Having introduced the notions of a semigroup and monoid, we shall not make much use of them. They are, however, useful in illustrating what a group is and is not.

The following properties of groups are more or less easily verified, and we leave the verifications to the reader as Exercise 4.1.1.

4.1.6 Proposition (Elementary properties of groups) *If* G *is a group, then the following statements hold:*

(i) *there is exactly one element* $\mathrm{e} \in \mathsf{G}$ *that satisfies* $\mathrm{g} \cdot \mathrm{e} = \mathrm{e} \cdot \mathrm{g} = \mathrm{g}$ *for all* $\mathrm{g} \in \mathsf{G}$*, i.e., the identity element in a group is unique;*

(ii) *for* $\mathrm{g} \in \mathsf{G}$*, there exists exactly one element* $\mathrm{g}' \in \mathsf{G}$ *such that* $\mathrm{g}' \cdot \mathrm{g} = \mathrm{g} \cdot \mathrm{g}' = \mathrm{e}$*, i.e., inverse elements are unique;*

(iii) *for* $\mathrm{g} \in \mathsf{G}$*,* $(\mathrm{g}^{-1})^{-1} = \mathrm{g}$*;*

(iv) *for* $\mathrm{g}_1, \mathrm{g}_2 \in \mathsf{G}$*,* $(\mathrm{g}_1 \cdot \mathrm{g}_2)^{-1} = \mathrm{g}_2^{-1} \cdot \mathrm{g}_1^{-1}$*;*

(v) *if* $g_1, g_2, h \in G$ *satisfy* $h \cdot g_1 = h \cdot g_2$*, then* $g_1 = g_2$*;*

(vi) *if* $g_1, g_2, h \in G$ *satisfy* $g_1 \cdot h = g_2 \cdot h$*, then* $g_1 = g_2$*;*

(vii) *if* $g_1, g_2 \in G$*, then there exists a unique* $h \in G$ *such that* $g_1 \cdot h = g_2$*;*

(viii) *if* $g_1, g_2 \in G$*, then there exists a unique* $h \in G$ *such that* $h \cdot g_1 = g_2$*.*

There is some useful notation associated with iterated group multiplication. Namely, if G is a semigroup, if $g \in G$, and if $k \in \mathbb{Z}_{>0}$, then we define $g^k \in G$ iteratively by $g^1 = g$ and $g^k = g \cdot g^{k-1}$. The following result records the fact that this notation behaves as we expect.

4.1.7 Proposition (Properties of g^k) *If* G *is a semigroup, if* $g \in G$*, and if* $k_1, k_2 \in \mathbb{Z}_{>0}$*, then the following statements hold:*

(i) $g^{k_1} \cdot g^{k_2} = g^{k_1+k_2}$*;*

(ii) $(g^{k_1})^{k_2} = g^{k_1k_2}$*.*

Proof (i) Let $g \in G$ and $k_1 \in \mathbb{Z}_{>0}$. If $k_2 = 1$ then, by definition,

$$g^{k_1} \cdot g^{k_2} = g^{k_1} \cdot g = g^{k_1+1} = g^{k_1+k_2},$$

so the result holds for $k_2 = 1$. Now suppose that the result holds for $k_2 \in \{1, \dots, k\}$. Then, if $k_2 = k + 1$,

$$g^{k_1} g^{k_2} = g^{k_1} \cdot g^{k+1} = g^{k_1} \cdot g^k \cdot g = g^{k_1+k} \cdot g = g^{k_1+k+1} = g^{k_1+k_2},$$

giving the result by induction on k_2.

(ii) Let $g \in G$ and $k_1 \in \mathbb{Z}_{>0}$. If $k_2 = 1$ then clearly $(g^{k_1})^{k_2} = g^{k_1k_2}$. Now suppose that the result holds for $k_2 \in \{1, \dots, k\}$, and for $k_2 = k + 1$ compute

$$(g^{k_1})^{k_2} = (g^{k_1})^{k+1} = (g^{k_1})^k \cdot g^{k_1} = g^{k_1k} \cdot g^{k_1} = g^{k_1k+k_1} = g^{k_1(k+1)} = g^{k_1k_2},$$

giving the result by induction on k_2. ■

4.1.8 Notation (g^k for Abelian groups) When a group is Abelian, then the group operation is sometimes thought of as addition, since it shares the property of commutativity possessed by addition. In such cases, one often write "kg" in place of "g^k" to reflect the idea that the group operation is "additive." •

4.1.2 Subgroups

It is often useful to consider subsets of groups that respect the group operation.

4.1.9 Definition (Subgroup) A nonempty subset H of a group G is a ***subgroup*** if

(i) $h_1 \cdot h_2 \in H$ for all $h_1, h_2 \in H$ and

(ii) $h^{-1} \in H$ for all $h \in H$. •

The following property of subgroups are easily verified, as the reader can see by doing Exercise 4.1.5.

4.1.10 Proposition (A subgroup is a group) *A nonempty subset* $\mathsf{H} \subseteq \mathsf{G}$ *of a group* G *is a subgroup if and only if* H *is a group using the binary operation of multiplication in* G*, restricted to* H*.*

4.1.11 Remark (On sub"objects") Mathematics can be perhaps thought of as the study of sets having some prescribed structure. It is frequent that one is interested in subsets which inherit this structure from the superset. Such subsets are almost always named with the prefix "sub." The above notion of a subgroup is our first encounter with this idea, although it will come up frequently in these volumes. •

Let us give some examples of subgroups.

4.1.12 Examples (Subgroups)

1. For any group G, $\{e\}$ is a subgroup, often called the ***trivial subgroup***.
2. Let $k \in \mathbb{Z}_{>0}$. The subset $k\mathbb{Z}$ of $\mathbb{Z}$ defined by
$$k\mathbb{Z} = \{kj \mid j \in \mathbb{Z}\}$$
(i.e., $k\mathbb{Z}$ consists of multiples of k) is a subgroup of $\mathbb{Z}$ if $\mathbb{Z}$ possesses the binary operation of addition.
3. $\mathbb{Z}$ and $\mathbb{Q}$ are subgroups of $\mathbb{R}$ if $\mathbb{R}$ possesses the binary operation of addition.
4. $\mathbb{Q}^*$ is a subgroup of $\mathbb{R}^*$ if $\mathbb{R}$ possesses the binary operation of multiplication.
5. $\mathbb{Z}$ is not a subgroup of $\mathbb{Q}$ if $\mathbb{Q}$ possesses the binary operation of multiplication.
6. Neither $\mathbb{Z}_{>0}$ nor $\mathbb{Z}_{\geq 0}$ are subgroups of $\mathbb{Z}$ if $\mathbb{Z}$ possesses the binary operation of addition.
7. Let $l, k \in \mathbb{Z}_{>0}$ with $l < k$. Let $\mathfrak{S}_{l,k}$ be the subset of $\mathfrak{S}_k$ defined by
$$\mathfrak{S}_{l,k} = \{\sigma \in \mathfrak{S}_k \mid \sigma(j) = j,\ j > l\}.$$
We claim that $\mathfrak{S}_{l,k}$ is a subgroup of $\mathfrak{S}_k$. It is clear by definition that, if $\sigma_1, \sigma_2 \in \mathfrak{S}_{l,k}$, then $h_1 \circ h_2 \in \mathfrak{S}_{l,k}$. If $\sigma \in \mathfrak{S}_{l,k}$ then let us write $\psi(j) = \sigma(j)$ for $j \in \{1, \dots, l\}$. This then defines $\psi \in \mathfrak{S}_l$. One can then directly verify that σ^{-1} is defined by
$$\sigma^{-1}(j) = \begin{cases} \psi^{-1}(j), & j \in \{1, \dots, l\}, \\ j, & j > l. \end{cases}$$
Thus $\sigma^{-1} \in \mathfrak{S}_{l,k}$, as desired.
Note that our above computations show that essentially $\mathfrak{S}_{l,k}$ consists of a copy of $\mathfrak{S}_l$ sitting inside $\mathfrak{S}_k$. In the language we are about to introduce in Definition 4.1.22, $\mathfrak{S}_{l,k}$ is isomorphic to $\mathfrak{S}_l$ (see Example 4.1.25–2). •

An important idea in many algebraic settings is that of the smallest subobject containing some subset. For groups this construction rests on the following result.

4.1.13 Proposition (Existence of subgroup generated by a subset) *Let* G *be a group and let* $\mathsf{S} \subseteq \mathsf{G}$*. Then there exists a subgroup* $\mathsf{H}_\mathsf{S} \subseteq \mathsf{G}$ *such that*

(i) $\mathsf{S} \subseteq \mathsf{H}_\mathsf{S}$ *and*

(ii) if $\mathsf{H} \subseteq \mathsf{G}$ *is a subgroup for which* $\mathsf{S} \subseteq \mathsf{H}$ *then* $\mathsf{H}_\mathsf{S} \subseteq \mathsf{H}$*.*

Moreover,

$$\mathsf{H}_\mathsf{S} = \{\mathsf{g}_1 \cdots \mathsf{g}_\mathsf{k} \mid \mathsf{k} \in \mathbb{Z}_{>0},\ \mathsf{g}_\mathsf{j} \in \mathsf{S} \textit{ or } \mathsf{g}_\mathsf{j}^{-1} \in \mathsf{S},\ \mathsf{j} \in \{1,\dots,\mathsf{k}\}\}$$

is the unique subgroup having the above two properties.

Proof Let

$$\mathscr{H}_S = \{\mathsf{H} \subseteq \mathsf{G} \mid \mathsf{H} \text{ is a subgroup with } S \subseteq \mathsf{H}\}.$$

Since $\mathsf{G} \in \mathscr{H}_S$ it follows that $\mathscr{H}_S$ is nonempty. We claim that $\mathsf{H}_S \triangleq \cap_{\mathsf{H}\in\mathscr{H}_S}\mathsf{H}$ has the required properties. First let $g \in S$. Then $g \in \mathsf{H}$ for every $\mathsf{H} \in \mathscr{H}_S$. Thus $g \in \mathsf{H}_S$ and so $S \subseteq \mathsf{H}_S$. Now let $g_1, g_2 \in \mathsf{H}_S$. Then $g_1, g_2 \in \mathsf{H}$ for every $\mathsf{H} \in \mathscr{H}_S$ and so $g_1 \cdot g_2 \in \mathsf{H}$ for every $\mathsf{H} \in \mathscr{H}_S$. Similarly, if $g \in \mathsf{H}$ for every $\mathsf{H} \in \mathscr{H}_S$ then $g^{-1} \in \mathsf{H}$ for every $\mathsf{H} \in \mathscr{H}_S$. Thus H_S is a subgroup containing S. Furthermore, if H is a subgroup containing S and if $g \in \mathsf{H}_S$ then clearly $g \in \mathsf{H}$ since $\mathsf{H} \in \mathscr{H}_S$. Thus $\mathsf{H}_S \subseteq \mathsf{H}$. We, moreover, claim that there is only one subgroup having the two stated properties. Indeed, suppose that $\mathsf{H}'_S \subseteq \mathsf{G}$ is a subgroup containing S and if H'_S is contained in any subgroup containing S. Then $\mathsf{H}'_S \subseteq \mathsf{H}_S$. Moreover, since $\mathsf{H}'_S \in \mathscr{H}_S$ we have $\mathsf{H}_S \subseteq \mathsf{H}'_S$. Thus $\mathsf{H}'_S = \mathsf{H}_S$.

To prove the final assertion it now suffices to show that

$$\mathsf{H}'_S = \{g_1 \cdots g_k \mid k \in \mathbb{Z}_{>0},\ g_j \in S \text{ or } g_j^{-1} \in S,\ j \in \{1,\dots,k\}\}$$

is a subgroup containing S and has the property that $\mathsf{H}'_S \subseteq \mathsf{H}$ for any subgroup H containing S. Clearly $S \subseteq \mathsf{H}'_S$. Now let

$$g_1 \cdots g_k, g'_1, \dots, g'_{k'} \in \mathsf{H}'_S.$$

Then clearly

$$g_1 \cdots g_k \cdot g'_1, \dots, g'_{k'} \in \mathsf{H}'_S.$$

Moreover,

$$(g_1 \cdots g_k)^{-1} = g_k^{-1} \cdots g_1^{-1} \in \mathsf{H}'_S$$

and so H'_S is a subgroup. Now let H be a subgroup containing S. Then $g_1 \cdot g_2 \in \mathsf{H}$ and $g^{-1} \in \mathsf{H}$ for every $g, g_1, g_2 \in S$. This means that $g_1 \cdots g_k \in \mathsf{H}$ for every $g_1, \dots, g_k \in \mathsf{G}$ such that either g_j or g_j^{-1} are in S, $j \in \{1,\dots,k\}$. Thus $\mathsf{H}'_S \subseteq \mathsf{H}$ and so we conclude that $\mathsf{H}'_S = \mathsf{H}_S$. ■

The following notion is one that we shall occasionally make reference to.

4.1.14 Definition (Subgroup generated by a subset) If G is a group and if $S \subseteq \mathsf{G}$, the subgroup H_S of Proposition 4.1.13 is the ***subgroup generated by* S**, denoted $\langle S \rangle$. •

4.1.15 Remark (On "generated by") The previous definition is an instance of an important idea in mathematics, particularly with algebraic structures. The idea is that one has some subset of a set with structure, and one wants to talk about the smallest subset containing the given subset, and which possesses the given structure. For groups, this is the clear content of Proposition 4.1.13, and we shall see it repeatedly in other settings. •

4.1.3 Quotients and products

Let us now turn to some important ideas connected with subgroups.

4.1.16 Definition (Left and right cosets) Let G be a group with H a subgroup.

(i) The ***left coset*** of H through $g \in \mathsf{G}$ is the set $g\mathsf{H} = \{gh \mid h \in \mathsf{H}\}$.

(ii) The ***right coset*** of H through $g \in \mathsf{G}$ is the set $\mathsf{H}g = \{hg \mid h \in \mathsf{H}\}$.

The set of left (resp. right) cosets is denoted by G/H (resp. $\mathsf{H}\backslash\mathsf{G}$), and the map assigning to $g \in \mathsf{G}$ the coset $g\mathsf{H} \in \mathsf{G}/\mathsf{H}$ (resp. $\mathsf{H}g \in \mathsf{H}\backslash\mathsf{G}$) is denoted by π_{H} (resp. ${}_{\mathsf{H}}\pi$), and is called the ***canonical projection***. •

Of course, if G is Abelian, then $g\mathsf{H} = \mathsf{H}g$ for each $g \in \mathsf{G}$, and, as a consequence, the sets G/H and $\mathsf{H}\backslash\mathsf{G}$ are the same. It is common to refer to G/H or $\mathsf{H}\backslash\mathsf{G}$ as the ***quotient*** of G by H.

An alternative description of cosets is given by the following result.

4.1.17 Proposition (Cosets as equivalence classes) *The set G/H (resp. $\mathsf{H}\backslash\mathsf{G}$) is the same as the set of equivalence classes in G associated to the equivalence relation $\mathsf{g}_1 \sim \mathsf{g}_2$ if $\mathsf{g}_2^{-1}\mathsf{g}_1 \in \mathsf{H}$ (resp. $\mathsf{g}_2\mathsf{g}_1^{-1} \in \mathsf{H}$).*

Proof We prove the proposition only for left cosets, and the proof for right cosets follows, *mutatis mutandis*. First let us prove that the relation defined by $g_1 \sim g_2$ if $g_2^{-1}g_1 \in \mathsf{H}$ is an equivalence relation.

1. Note that $g^{-1}g = e \in \mathsf{H}$, so the relation is reflexive.
2. If $g_1 \sim g_2$ then $g_2^{-1}g_1 \in \mathsf{H}$, which implies that $(g_2^{-1}g_1)^{-1} \in \mathsf{H}$ since H is a subgroup. By Proposition 4.1.6 this means that $g_1^{-1}g_2 \in \mathsf{H}$; i.e., that $g_2 \sim g_1$. Thus the relation is symmetric.
3. If $g_1 \sim g_2$ and $g_2 \sim g_3$, or equivalently that $g_2 \sim g_1$ and $g_3 \sim g_2$, then $g_1^{-1}g_2, g_2^{-1}g_3 \in \mathsf{H}$. Then, since H is a subgroup,

$$(g_1^{-1}g_2)(g_2^{-1}g_3) \in \mathsf{H} \quad \Longrightarrow \quad g_1^{-1}g_3 \in \mathsf{H}.$$

   Thus $g_3 \sim g_1$, or $g_1 \sim g_3$, and the relation is transitive.

Now let $g \in \mathsf{G}$ and let $g' \in g\mathsf{H}$. Then $g' = gh$ for some $h \in \mathsf{H}$, so $g^{-1}g' \in \mathsf{H}$, so $g' \sim g$. Conversely, suppose that $g' \sim g$ so that $g^{-1}g' = h$ for some $h \in \mathsf{H}$. Then $g' = gh$, so $g' \in g\mathsf{H}$. This gives the result. ■

Let us give some examples of cosets and collections of cosets.

4.1.18 Examples (Cosets)

1. Let $k \in \mathbb{Z}_{>0}$. Consider the group $\mathbb{Z}$ with the binary operation of addition, and also consider the subgroup $k\mathbb{Z}$ consisting of multiples of k. We claim that $\mathbb{Z}/k\mathbb{Z}$ is a set with k elements. Using the Theorem 4.2.45 below, we see that every element of $\mathbb{Z}$ lies in the coset of exactly one of the elements from the set $\{0, 1, \dots, k-1\}$, which gives our claim. For reasons which will become clear in Example 4.2.2–4 it is convenient to denote the coset through $j \in \mathbb{Z}$ by $j + k\mathbb{Z}$. We will frequently encounter the group $\mathbb{Z}/k\mathbb{Z}$, and so give it the shorthand $\mathbb{Z}_k$.
2. Consider the group $\mathbb{R}$ equipped with the binary operation of addition, and consider the subgroup $\mathbb{Q}$. We claim that the set $\mathbb{R}/\mathbb{Q}$ is uncountable. Indeed, if it were not, then this would imply that $\mathbb{R}$ is the countable union of cosets, and each coset itself must be countable. That is to say, if $\mathbb{R}/\mathbb{Q}$ is countable, then $\mathbb{R}$ is a countable union of countable sets. But, by Proposition 1.7.16, this means that $\mathbb{R}$ is countable. However, in Exercise 2.1.4 the reader is asked to show $\mathbb{R}$ is actually not countable. The contradiction proves that $\mathbb{R}/\mathbb{Q}$ is uncountable. Further investigation of $\mathbb{R}/\mathbb{Q}$ takes one into the topic of field extensions, which we consider very briefly in Section 4.3.3, and then into Galois theory, which is somewhat beyond our focus here.
3. Consider the permutation group $\mathfrak{S}_3$ in 3 symbols and consider the subgroup $\mathfrak{S}_{2,3}$, which is isomorphic to $\mathfrak{S}_2$ as we showed in Example 4.1.25–2. Let us describe the cosets of $\mathfrak{S}_3/\mathfrak{S}_{2,3}$. Suppose that $\sigma_1, \sigma_2 \in \mathfrak{S}_3$ lie in the same coset of $\mathfrak{S}_{2,3}$. Then it must hold that $\sigma_1 \circ \sigma_2^{-1}(3) = 3$, or equivalently that $\sigma_1^{-1}(3) = \sigma_2^{-1}(3)$. Thus cosets are identified by their having in common the fact that the same elements in $\{1, 2, 3\}$ are images of the element 3. The cosets are then easily seen to be
   (a) $\left\{\begin{pmatrix}1 & 2 & 3\\1 & 2 & 3\end{pmatrix}, \begin{pmatrix}1 & 2 & 3\\2 & 1 & 3\end{pmatrix}\right\}$,
   (b) $\left\{\begin{pmatrix}1 & 2 & 3\\2 & 3 & 1\end{pmatrix}, \begin{pmatrix}1 & 2 & 3\\3 & 2 & 1\end{pmatrix}\right\}$, and
   (c) $\left\{\begin{pmatrix}1 & 2 & 3\\3 & 1 & 2\end{pmatrix}, \begin{pmatrix}1 & 2 & 3\\1 & 3 & 2\end{pmatrix}\right\}$. •

Next we discuss a particular sort of subgroup that, as we shall see, is distinguished by the structure of its set of cosets.

4.1.19 Definition (Normal subgroup) A subgroup H of a group G is a ***normal subgroup*** if $g\mathsf{H} = \mathsf{H}g$ for all $g \in \mathsf{G}$. •

The following result explains why normal subgroups are interesting.

4.1.20 Proposition (Quotients by normal subgroups are groups) *Let N be a normal subgroup of G and define a binary operation on G/N by*

$$(g_1\mathsf{N}, g_2\mathsf{N}) \mapsto (g_1g_2)\mathsf{N}.$$

Then this binary operation satisfies the conditions for group multiplication.

Proof First let us show that this binary operation is well-defined. Let $g_1, g_2, h_1, h_2 \in \mathsf{G}$ satisfy $g_1\mathsf{N} = h_1\mathsf{N}$ and $g_2\mathsf{N} = h_2\mathsf{N}$. Then we must have $g_1^{-1}h_1 = n_1$ and $g_2^{-1}h_2 = n_2$ for $n_1, n_2 \in \mathsf{N}$, and then we compute

$$\begin{aligned}(h_1h_2\mathsf{N}) &= \{h_1h_2n \mid n \in \mathsf{N}\} = \{g_1n_1g_2n_2n \mid n \in \mathsf{N}\}\\ &= \{g_1g_2n_3n_2n \mid n \in \mathsf{N}\} = \{g_1g_2n \mid n \in \mathsf{N}\} = (g_1g_2)\mathsf{N},\end{aligned}$$

where $n_3 \in \mathsf{N}$ is defined so that $n_1g_2 = g_2n_3$, this being possible by Exercise 4.1.8 since N is normal.

To then verify that the (now) well-defined binary operation satisfies the conditions for group multiplication is trivial. ■

Another useful construction for groups is the product of two groups.

4.1.21 Definition (Direct product of groups) If G and H are groups, their ***direct product*** is the group $\mathsf{G} \times \mathsf{H}$ with the group operation $(g_1, h_1) \cdot (g_2, h_2) = (g_1g_2, h_1h_2)$. •

If H, K are subgroups of G, then we denote

$$\mathsf{HK} = \{hk \mid h \in \mathsf{H},\ k \in \mathsf{K}\}.$$

We note that it need not be the case that HK is itself a subgroup.

4.1.4 Group homomorphisms

Another important concept for groups, and for many other structures in mathematics, is that of a map that preserves the structure.

4.1.22 Definition (Group homomorphism, epimorphism, monomorphism, and isomorphism) For semigroups (resp. monoids, groups) G and H, a map $\phi\colon \mathsf{G} \to \mathsf{H}$ is a:

(i) ***semigroup*** (resp. ***monoid, group***) ***homomorphism***, or simply a ***homomorphism***, if $\phi(g_1 \cdot g_2) = \phi(g_1) \cdot \phi(g_2)$ for all $g_1, g_2 \in \mathsf{G}$;

(ii) ***epimorphism*** if it is a surjective homomorphism;

(iii) ***monomorphism*** if it is an injective homomorphism;

(iv) ***isomorphism*** if it a bijective homomorphism. •

We shall mainly be concerned with group homomorphisms, although homomorphisms of semigroups and monoids will arise at times.

4.1.23 Remark (On morphisms of various sorts) As with the idea of a sub"object" as discussed in Remark 4.1.11, the idea of a map between sets that preserves the structure of those sets, e.g., the group structure in the case of a group homomorphism, is of fundamental importance. The expression "morphosis" comes from Greek for "form," whereas the prefixes "homo," "epi," "mono," and "isos" are from the Greek for roughly "alike," "on," "one," and "equal," respectively. •

The following result gives a couple of basic properties of homomorphisms.

4.1.24 Proposition (Properties of group homomorphisms) *If* G *and* H *are monoids and if* $\phi\colon \mathsf{G} \to \mathsf{H}$ *is a monoid homomorphism, then*

(i) $\phi(\mathrm{e}_{\mathsf{G}}) = \mathrm{e}_{\mathsf{H}}$*, and*

(ii) if G *and* H *are additionally groups, then* $\phi(\mathrm{g}^{-1}) = (\phi(\mathrm{g}))^{-1}$*.*

Proof (i) Let $g \in \mathsf{G}$ and note that

$$\phi(e_{\mathsf{G}}g) = \phi(ge_{\mathsf{G}}) = \phi(e_{\mathsf{G}})\phi(g) = \phi(g)\phi(e_{\mathsf{G}}) = \phi(g).$$

In particular, $\phi(g)\phi(e_{\mathsf{G}}) = \phi(g)e_{\mathsf{H}}$, and the result follows by multiplication by $\phi(g)^{-1}$.

(ii) Now, if $g \in \mathsf{G}$ then $\phi(g)\phi(g^{-1}) = \phi(gg^{-1}) = \phi(e_{\mathsf{G}}) = e_{\mathsf{H}}$, which shows that $\phi(g^{-1}) = (\phi(g))^{-1}$. ■

4.1.25 Examples (Group homomorphisms)

1. If G and H are groups with identity elements e_{G} and e_{H}, respectively, then the map $\phi\colon \mathsf{G} \to \mathsf{H}$ defined by $\phi(g) = e_{\mathsf{H}}$ for all $g \in \mathsf{G}$ is readily verified to be a homomorphism. It is an epimorphism if and only if $\mathsf{H} = \{e_{\mathsf{H}}\}$ and a monomorphism if and only if $\mathsf{G} = \{e_{\mathsf{G}}\}$.
2. Let $l, k \in \mathbb{Z}_{>0}$ with $l < k$. The map $\phi\colon \mathfrak{S}_l \to \mathfrak{S}_k$ defined by

$$\phi(\sigma)(j) = \begin{cases} \sigma(j), & j \in \{1, \dots, l\}, \\ j, & j > l \end{cases}$$

is verified to be a monomorphism. In fact, it is easily verified to be an isomorphism from $\mathfrak{S}_l$ to $\mathfrak{S}_{l,k} \subseteq \mathfrak{S}_k$. •

Associated to every homomorphism of groups are two important subsets, one of the domain and one of the codomain of the homomorphism.

4.1.26 Definition (Image and kernel of group homomorphism) Let G and H be groups and let $\phi\colon \mathsf{G} \to \mathsf{H}$ be a homomorphism.

(i) The ***image*** of ϕ is $\mathrm{image}(\phi) = \{\phi(g) \mid g \in \mathsf{G}\}$.

(ii) The ***kernel*** of ϕ is $\ker(\phi) = \{g \in \mathsf{G} \mid \phi(g) = e_{\mathsf{H}}\}$. •

The image and the kernel have useful properties relative to the group structure.

4.1.27 Proposition (Image and kernel are subgroups) *If* G *and* H *are groups and if* $\phi\colon \mathsf{G} \to \mathsf{H}$ *is a homomorphism, then*

(i) $\mathrm{image}(\phi)$ *is a subgroup of* H *and*

(ii) $\ker(\phi)$ *is a normal subgroup of* G*.*

Proof (i) If $g_1, g_2 \in \mathsf{G}$ then $\phi(g_1)\phi(g_2) = \phi(g_1g_2) \in \mathrm{image}(\phi)$. From part (ii) of Proposition 4.1.24 we have $(\phi(g))^{-1} \in \mathrm{image}(\phi)$ for every $g \in \mathsf{G}$.

(ii) Let $g_1, g_2 \in \ker(\phi)$. Then $\phi(g_1g_2) = \phi(g_1)\phi(g_2) = e_{\mathsf{H}}$ so that $g_1g_2 \in \ker(\phi)$. If $g \in \ker(\phi)$ then

$$e_{\mathsf{H}} = \phi(e_{\mathsf{G}}) = \phi(gg^{-1}) = \phi(g)\phi(g^{-1}) = \phi(g^{-1}).$$

Thus $g^{-1} \in \ker(\phi)$, and so $\ker(\phi)$ is a subgroup. To show that $\ker(\phi)$ is normal, let $g \in \mathsf{G}$ and let $h \in \ker(\phi)$. Then

$$\phi(ghg^{-1}) = \phi(g)\phi(h)\phi(g^{-1}) = \phi(g)\phi(g^{-1}) = e_{\mathsf{H}}.$$

Thus $ghg^{-1} \in \ker(\phi)$ for every $g \in \mathsf{G}$ and $h \in \ker(\phi)$. The result now follows by Exercise 4.1.8. ■

The following result characterising group monomorphisms is simple, but is one that we use continually, so it is worth recording.

4.1.28 Proposition (Characterisation of monomorphisms) *A group homomorphism* $\phi\colon \mathsf{G} \to \mathsf{H}$ *is a monomorphism if and only if* $\ker(\phi) = \mathrm{e}_{\mathsf{G}}$.

Proof Suppose that $\ker(\phi) = \{e_{\mathsf{G}}\}$ and that $\phi(g_1) = \phi(g_2)$. Then

$$e_{\mathsf{H}} = \phi(g_1)(\phi(g_2))^{-1} = \phi(g_1)\phi(g_2^{-1}) = \phi(g_1g_2^{-1}),$$

implying that $g_1g_2^{-1} \in \ker(\phi)$ whence $g_1 = g_2$, and so ϕ is injective.

Conversely, suppose ϕ is a monomorphism and let $g \in \ker(\phi)$. Thus $\phi(g) = e_{\mathsf{H}}$. However, since ϕ is a monomorphism and since $\phi(e_{\mathsf{G}}) = e_{\mathsf{H}}$, we must have $g = e_{\mathsf{G}}$. ■

4.1.5 The isomorphism theorems

One of the central themes in mathematics is classification, and this means that one wants to understand equivalence classes, where equivalence of two objects means that there exists an isomorphism between them. For sets, this gives rise to the notion of cardinality. For groups, this notion of equivalence is determined by the notion of group isomorphism. In this section we give three fundamental and quite elementary results that might be thought of as simple examples of the classification project for groups. These go under the unimaginative names “First Isomorphism Theorem,” “Second Isomorphism Theorem,” and “Third Isomorphism Theorem.”

All of the isomorphism theorems can be formulated as special instances of the following result.

4.1.29 Theorem (Interconnection of normal subgroups and homomorphisms) *Let* G *and* H *be groups, let* $\phi\colon \mathsf{G} \to \mathsf{H}$ *be an homomorphism, and let* N *be a normal subgroup of* G *contained in* $\ker(\phi)$. *Then there exists a unique homomorphism* $\overline{\phi}\colon \mathsf{G}/\mathsf{N} \to \mathsf{H}$ *satisfying* $\overline{\phi}(g\mathsf{N}) = \phi(g)$ *for every* $g \in \mathsf{G}$. *Moreover,*

(i) $\mathrm{image}(\overline{\phi}) = \mathrm{image}(\phi)$,

(ii) $\ker(\overline{\phi}) = \ker(\phi)/\mathsf{N}$, *and*

(iii) the following three statements are equivalent:

(a) $\overline{\phi}$ *is an isomorphism;*

(b) ϕ *is an epimorphism and* $\mathsf{N} = \ker(\phi)$.

Proof First let us define $\overline{\phi}(g\mathsf{N}) = \phi(g)$. To show that this definition makes sense, we show that it does not depend on the choice of representative in $g\mathsf{N}$. Indeed, if $a = gn$ for $n \in \mathsf{N}$, then we have

$$\phi(g) = \phi(an) = \phi(a)\phi(n) = \phi(a),$$

which shows that $\overline{\phi}(g\mathsf{N})$ is well defined. Moreover, the very definition of $\overline{\phi}$ ensures that it is the unique homomorphism satisfying the stated condition.

(i) This is immediate.

(ii) We have

$$g\mathsf{N} \in \ker(\overline{\phi}) \iff \overline{\phi}(g\mathsf{N}) = e_{\mathsf{H}} \iff \phi(g) = e_{\mathsf{H}} \iff g \in \ker(\phi).$$

(iii)(a) $\implies$ (iii)(b) Since $\overline{\phi}$ is an isomorphism, for $h \in \mathsf{H}$ there exists $g\mathsf{N} \in \mathsf{G}/\mathsf{N}$ such that $\overline{\phi}(g\mathsf{N}) = h$. But then $\phi(g) = h$ and so ϕ is an epimorphism. Also, if $g \in \ker(\phi)$, then $g\mathsf{N} \in \ker(\phi)/\mathsf{N} = \ker(\overline{\phi})$ and so $g\mathsf{N} = e_{\mathsf{G}/\mathsf{H}}$, whence $g\mathsf{N} = \mathsf{N}$ and so $g \in \mathsf{N}$.

(iii)(b) $\implies$ (iii)(a) Since $\operatorname{image}(\overline{\phi}) = \operatorname{image}(\phi)$, we have that $\overline{\phi}$ is an epimorphism. If $g\mathsf{N} \in \ker(\overline{\phi})$ then $g\mathsf{N} \in \ker(\phi)/\mathsf{N} = e_{\mathsf{G}/\mathsf{N}}$ and so $\overline{\phi}$ is a monomorphism by Proposition 4.1.28. ■

First is the first.

4.1.30 Corollary (First Isomorphism Theorem) *If* G *and* H *are groups and if* $\phi\colon \mathsf{G} \to \mathsf{H}$ *is an homomorphism, then there is an isomorphism* $\overline{\phi}\colon \mathsf{G}/\ker(\phi) \to \operatorname{image}(\phi)$ *for which the diagram*

$$\begin{tikzcd}
\mathsf{G} \arrow[r, "\phi"] \arrow[d] & \mathsf{H} \\
\mathsf{G}/\ker(\phi) \arrow[r, "\overline{\phi}"'] & \operatorname{image}(\phi) \arrow[u]
\end{tikzcd}$$

commutes.

Proof This follows from Theorem 4.1.29 by considering "$\mathsf{N} = \ker(\phi)$" and "$\mathsf{H} = \operatorname{image}(\phi)$." ■

Second is the second.

4.1.31 Corollary (Second Isomorphism Theorem) *Let* G *be a group, let* $\mathsf{N} \subseteq \mathsf{G}$ *be a normal subgroup, and let* $\mathsf{H} \subseteq \mathsf{G}$ *be a subgroup. Then there is an isomorphism of* $\mathsf{H}/(\mathsf{N} \cap \mathsf{H})$ *and* $(\mathsf{NH})/\mathsf{N}$.

Proof By Exercise 4.1.10, $\mathsf{NH} = \langle \mathsf{N} \cup \mathsf{H} \rangle$ and N is a normal subgroup of NH. Thus we have the sequence of homomorphisms

$$\begin{tikzcd}
\mathsf{H} \arrow[r] & \mathsf{NH} \arrow[r] & (\mathsf{NH})/\mathsf{N}
\end{tikzcd}$$

which, by composition, give an homomorphism $\phi\colon \mathsf{H} \to (\mathsf{NH})/\mathsf{N}$. Note that $\phi(h) = h\mathsf{N}$, thinking of $h\mathsf{N}$ as a coset in NH. We claim that $\ker(\phi) = \mathsf{N} \cap \mathsf{H}$. Indeed, if $\phi(h) = e_{(\mathsf{NH})/\mathsf{N}}$, then $h \in \mathsf{N}$ and so this gives $\ker(\phi) \subseteq \mathsf{N} \cap \mathsf{H}$. On the other hand, if $h \in \mathsf{N} \cap \mathsf{H}$, then $\phi(h) = e_{(\mathsf{NH})/\mathsf{N}}$, because $h \in \mathsf{N}$.

Now we apply the First Isomorphism Theorem to give an isomorphism $\overline{\phi}\colon \mathsf{H}/(\mathsf{N} \cap \mathsf{H}) \to \mathrm{image}(\phi)$. Now, if $nh\mathsf{N} \in (\mathsf{NH})/\mathsf{N}$ for $n \in \mathsf{N}$ and $h \in \mathsf{H}$, then $nh = hn'$ for some $n' \in \mathsf{N}$ since N is normal in NH. Thus $nh\mathsf{N} = h\mathsf{N} = \phi(h)$, and so ϕ is an epimorphism. This gives the result. ■

Third is the third.

4.1.32 Corollary (Third Isomorphism Theorem) *If* G *is a group and if* N *and* K *are normal subgroups of* G*, with* $\mathsf{N} \subseteq \mathsf{K}$*, then* K/N *is a normal subgroup of* G/N *and there is an isomorphism of* $(\mathsf{G}/\mathsf{N})/(\mathsf{K}/\mathsf{N})$ *with* G/K.

Proof We define $\phi\colon \mathsf{G}/\mathsf{N} \to \mathsf{G}/\mathsf{K}$ by $\phi(g\mathsf{N}) = g\mathsf{K}$. To see that ϕ is well defined, suppose that $g_1\mathsf{N} = g_2\mathsf{N}$ so that $g_1 = g_2 n$ with $n \in \mathsf{N}$. Since $\mathsf{N} \subseteq \mathsf{K}$, $g_1\mathsf{K} = g_2\mathsf{K}$, giving well-definedness of ϕ. Also, it is clear that ϕ is an epimorphism.

Moreover,

$$\ker(\phi) = \{g\mathsf{N} \mid g \in \mathsf{K}\} = \mathsf{K}/\mathsf{N}.$$

By Exercise 4.1.11 we conclude that K/N is a normal subgroup of G/N. By the First Isomorphism Theorem, we have an isomorphism of $(\mathsf{G}/\mathsf{N})/\ker(\phi)$ with G/K. Since $\ker(\phi) = \mathsf{K}/\mathsf{N}$, the result follows. ■

4.1.6 The symmetric group

In Example 4.1.5–11 we introduced the symmetric group. We shall have occasion to use some of the structure of the symmetric group, and in this section we collect the pertinent facts.

First of all let us define a simple collection of elements of the symmetric group and some notions associated with them.

4.1.33 Definition (Cycle, transposition, even permutation, odd permutation) Let $k \in \mathbb{Z}_{>0}$.

(i) An element $\sigma \in \mathfrak{S}_k$ is a ***cycle*** if there exists distinct $j_1, \dots, j_m \in \{1, \dots, k\}$ such that

$$\sigma(j_1) = j_2,\ \sigma(j_2) = j_3,\ \cdots,\ \sigma(j_{m-1}) = j_m,\ \sigma(j_m) = j_1,$$

and such that $\sigma(j) = j$ for $j \notin \{j_1, \dots, j_m\}$. The number m is the ***length*** of the cycle. We denote the above cycle by $(j_1\ j_2\ \cdots\ j_m)$.

(ii) An element $\sigma \in \mathfrak{S}_k$ is a ***transposition*** if it is a cycle of length 2. Thus $\sigma = (j_1\ j_2)$ for distinct $j_1, j_2 \in \{1, \dots, k\}$.

(iii) An element $\sigma \in \mathfrak{S}_k$ is ***even*** (resp. ***odd***) if it is a finite product of an even (resp. odd) number of transpositions. •

Let us illustrate the notion of a cycle with an elementary example.

4.1.34 Example (Cycle) The permutation

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

is a cycle using the elements 2, 4, and 5, and is written as (2 5 4), representing the fact that $\sigma(2) = 5$, $\sigma(5) = 4$, and $\sigma(4) = 2$. It is clear that one could also write the cycle as (5 4 2) or (4 2 5), and, therefore, the notation we use to represent a cycle is not unique. •

It turns out that every permutation is a product of cycles. If we ask that the cycles have an additional property, then the product is unique. This property is the following.

4.1.35 Definition (Disjoint permutations) Let $k \in \mathbb{Z}_{>0}$. Permutations $\sigma_1, \sigma_2 \in \mathfrak{S}_k$ are ***disjoint*** if, for every $j \in \{1, \dots, k\}$, $\sigma_1(j) \neq j$ implies that $\sigma_2(j) = j$ and $\sigma_2(j) \neq j$ implies that $\sigma_1(j) = j$. •

The idea is that the set of elements of $\{1, \dots, k\}$ not fixed by disjoint permutations are distinct. It is easy to show that disjoint permutations commute; this is Exercise 4.1.14.

We now have the following important structural result describing a typical permutation.

4.1.36 Theorem (Permutations are products of cycles) *Let $\mathrm{k} \in \mathbb{Z}_{>0}$. If $\sigma \in \mathfrak{S}_{\mathrm{k}}$ then there exist disjoint cycles $\sigma_1, \dots, \sigma_{\mathrm{r}} \in \mathfrak{S}_{\mathrm{k}}$ such that $\sigma = \sigma_1 \circ \cdots \circ \sigma_{\mathrm{r}}$. Moreover, if $\sigma'_1, \dots, \sigma'_{\mathrm{r}'} \in \mathfrak{S}_{\mathrm{k}}$ are disjoint permutations such that $\sigma'_1 \circ \cdots \circ \sigma'_{\mathrm{r}'}$, then $\mathrm{r} = \mathrm{r}'$ and there exists a bijection $\phi \colon \{1, \dots, \mathrm{r}\} \to \{1, \dots, \mathrm{r}\}$ such that $\sigma'_{\mathrm{j}} = \sigma_{\phi(\mathrm{j})}$, $\mathrm{j} \in \{1, \dots, \mathrm{r}\}$.*

Proof For $\sigma \in \mathfrak{S}_k$ and $j \in \{1, \dots, k\}$ let us denote

$$O(\sigma, j) = \{\sigma^m(j) \mid m \in \mathbb{Z}_{\geq 0}\}$$

and suppose that $\mathrm{card}(O(\sigma, j)) = N_{\sigma,j}$.

1 Lemma *With the above notation the following statements hold:*

(i) $\mathrm{j}, \sigma(\mathrm{j}), \dots, \sigma^{\mathrm{N}_{\sigma,\mathrm{j}}-1}(\mathrm{j})$ *are distinct;*

(ii) $\sigma^{\mathrm{N}_{\sigma,\mathrm{j}}}(\mathrm{j}') = \mathrm{j}'$ *for each* $\mathrm{j}' \in \mathrm{O}(\sigma, \mathrm{j})$;

(iii) $\mathrm{O}(\sigma, \mathrm{j}) = \{\mathrm{j}, \sigma(\mathrm{j}), \dots, \sigma^{\mathrm{N}_{\sigma,\mathrm{j}}-1}(\mathrm{j})\}$;

(iv) $\mathrm{O}(\sigma, \mathrm{j}') = \mathrm{O}(\sigma, \mathrm{j})$ *for every* $\mathrm{j}' \in \mathrm{O}(\sigma, \mathrm{j})$.

Proof (i) Suppose that $\sigma^{m_1}(j) = \sigma^{m_2}(j)$ for distinct $m_1, m_2 \in \{0, 1, \dots, N_{\sigma,j} - 1\}$. Suppose that $m_2 > m_1$ so that $\sigma^{m_2 - m_1}(j) = j$ with $m_2 - m_1 \in \{1, \dots, N_{\sigma,j} - 1\}$. For $m \in \mathbb{Z}_{>0}$ let us use the division algorithm for $\mathbb{Z}$ (Theorem 4.2.45) to write $m = q(m_2 - m_1) + r$ for $r \in \{0, 1, \dots, m_2 - m_1 - 1\}$. Then $\sigma^m(j) = \sigma^r(j)$ and so it follows that

$$O(\sigma, j) \subseteq \{j, \sigma(j), \dots, \sigma^{m_2 - m_1 - 1}(j)\}.$$

This, however, contradicts the definition of $N_{\sigma,j}$ since $m_2 - m_1 < N_{\sigma,j}$.

(ii) Since $\mathrm{card}(O(\sigma, j)) = N_{\sigma,j}$ and by the previous part of the lemma we must have $\sigma^{N_{\sigma,j}}(j) = \sigma^m(j)$ for some $m \in \{0, 1, \dots, N_{\sigma,j} - 1\}$. Thus $\sigma^{N_{\sigma,j}-m}(j) = j$ and so, by the previous part of the lemma we must have $m = 0$. Thus $\sigma^{N_{\sigma,j}}(j) = j$. Now, if $m \in \{1, \dots, N_{\sigma,j} - 1\}$, then

$$\sigma^{N_{\sigma,j}} \circ \sigma^m(j) = \sigma^m \circ \sigma^{N_{\sigma,j}}(j) = \sigma^m(j),$$

giving this part of the lemma.

(iii) Clearly

$$\{j, \sigma(j), \dots, \sigma^{N_{\sigma,j}-1}(j)\} \subseteq O(\sigma, j).$$

By definition of $N_{\sigma,j}$ and by part (i) equality follows.

(iv) Let $m' \in \{1, \dots, N_{\sigma,j} - 1\}$ and let $j' = \sigma^{m'}(j)$.

$$O(\sigma, j') = \{\sigma^m(j') \mid m \in \mathbb{Z}_{\geq 0}\} = \{\sigma^{m+m'}(j) \mid m \in \mathbb{Z}_{\geq 0}\} \subseteq O(\sigma, j).$$

On the other hand, if $m \in \mathbb{Z}_{>0}$ we can write $m - m' = qN_{\sigma,j} + r$ for $r \in \{0, 1, \dots, N_{\sigma,j} - 1\}$ using the division algorithm. Then

$$\sigma^m(j) = \sigma^{m-m'} \circ \sigma^{m'}(j) = \sigma^r \circ \sigma^{m'}(j) = \sigma^r(j'),$$

and so $O(\sigma, j) \subseteq O(\sigma, j')$. ▼

From the lemma and since the set $\{1, \dots, k\}$ is finite it follows that there exist $j_1, \dots, j_r \in \{1, \dots, k\}$ such that

1. $\{1, \dots, k\} = \cup_{l=1}^r O(\sigma, j_l)$ and
2. $O(\sigma, j_l) \cap O(\sigma, j_m) = \varnothing$ for $l \neq m$.

Let $N_l = \mathrm{card}(O(\sigma, j_l))$ for $l \in \{1, \dots, r\}$. For $l \in \{1, \dots, r\}$ define $\sigma_l \in \mathfrak{S}_k$ by

$$\sigma_l(j) = \begin{cases} \sigma(j), & j \in O(\sigma, j_l), \\ j, & \text{otherwise.} \end{cases}$$

By the lemma we have $\sigma_l = (j_l\ \sigma(j_l)\ \cdots\ \sigma^{N_l-1}(j_l))$. Moreover, for distinct $l, m \in \{1, \dots, r\}$ the permutations σ_l and σ_m are clearly disjoint. Therefore, by Exercise 4.1.14, the permutations $\sigma_1, \dots, \sigma_l$ commute with one another. We claim that $\sigma = \sigma_1 \circ \dots \circ \sigma_r$. Indeed, let $j \in \{1, \dots, k\}$ and let $l_j \in \{1, \dots, r\}$ satisfy $j \in O(\sigma, l_j)$. Then, by construction, $\sigma_l(j) = j$ for $l \neq l_j$. We thus have

$$\sigma_1 \circ \dots \circ \sigma_{l_j} \circ \dots \circ \sigma_r(j) = \sigma_{l_j} \circ \sigma_1 \circ \dots \circ \sigma_{l_j - 1} \circ \sigma_{l_j + 1} \circ \dots \circ \sigma_r(j) = \sigma_{l_j}(j) = \sigma(j),$$

giving the theorem. ■

It is not clear that a permutation cannot be both even and odd, so let us establish this in an illuminating way. In the statement of the result we consider the set $\{-1, 1\}$ to be a group with the product being multiplication in the usual way.

4.1.37 Theorem (The sign homomorphism from the symmetric group) *Let* $\mathrm{k} \in \mathbb{Z}_{>0}$*. If* $\sigma \in \mathfrak{S}_{\mathrm{k}}$ *then* σ *is the product of a finite number of transpositions. Moreover, the map* $\mathrm{sign}\colon \mathfrak{S}_{\mathrm{k}} \to \{-1, 1\}$ *given by*

$$\mathrm{sign}(\sigma) = \begin{cases} 1, & \textit{σ is a product of an even number of transpositions,} \\ -1, & \textit{σ is a product of an odd number of transpositions} \end{cases}$$

is a well-defined group homomorphism.

Proof By Theorem 4.1.36 it suffices to show that a cycle is a finite product of adjacent transpositions. However, for a cycle $(j_1 \ \cdots \ j_m)$ we can write

$$(j_1 \ \cdots \ j_m) = (j_1 \ j_2) \cdot (j_1 \ j_3) \cdot \cdots \cdot (j_1 \ j_m),$$

which can be verified directly.

Now we prove that sign is well-defined. Let $\sigma \in \mathfrak{S}_k$. By Theorem 4.1.36 there exist unique (up to order) disjoint cycles $\sigma_1, \dots, \sigma_r$ such that $\sigma = \sigma_1 \circ \cdots \circ \sigma_r$. Let us define $C(\sigma) = r$. In the following lemma we recall the notation $O(\sigma, j)$ introduced in the proof of Theorem 4.1.36.

1 Lemma *Let* $\sigma \in \mathfrak{S}_{\mathrm{k}}$ *and let* $\tau = (\mathrm{j}_1, \mathrm{j}_2)$*. Then*

(i) $\mathrm{C}(\sigma \circ \tau) = \mathrm{C}(\sigma) + 1$ *if* $\mathrm{O}(\sigma, \mathrm{j}_1) = \mathrm{O}(\sigma, \mathrm{j}_2)$ *and*

(ii) $\mathrm{C}(\sigma \circ \tau) = \mathrm{C}(\sigma) - 1$ *if* $\mathrm{O}(\sigma, \mathrm{j}_1) \neq \mathrm{O}(\sigma, \mathrm{j}_2)$*.*

Proof Suppose that $O(\sigma, j_1) = O(\sigma, j_2)$ and, using the lemma from the proof of Theorem 4.1.36, write

$$O(\sigma, j_1) = \{l_1 = j_1, \dots, l_s = j_2, \dots, l_m\}$$

with $l_p = \sigma^p(l_1)$ for $p \in \{1, \dots, m\}$. Let $\sigma' = (l_1 \ \cdots \ l_p)$. Then we can directly verify that

$$\sigma' \circ \tau = (l_1 \ \cdots \ l_p) \cdot (l_1 \ l_s) = (l_1 \ \cdots \ l_{s-1}) \cdot (l_s \ \cdots \ l_p),$$

giving $\sigma' \circ \tau$ as a product of two cycles. Now note that if j has the property that $O(\sigma, j) \neq O(\sigma, j_1)$ then, using the lemma from the proof of Theorem 4.1.36, $\sigma \circ \tau(j) = \sigma(j)$. Thus $O(\sigma \circ \tau, j) = O(\sigma, j)$ if $j \notin O(\sigma, j_1)$. For $j \in O(\sigma, j_1)$ we have $\sigma(j) = \sigma'(j)$ and also $\sigma \circ \tau(j) = \sigma' \circ \tau(j)$ since $\tau(j) \in O(\sigma, j_1)$. Thus

$$O(\sigma, j_1) = O(\sigma \circ \tau, j_1) \cup O(\sigma \circ \tau, j_2),$$

giving $C(\sigma \circ \tau) = C(\sigma) + 1$.

Now suppose that $O(\sigma, j_1) \neq O(\sigma, j_2)$. Let us write

$$O(\sigma, j_1) = \{j_1, \sigma(j_1), \dots, \sigma^{p_1 - 1}(j_1)\}, \quad O(\sigma, j_2) = \{j_2, \sigma(j_2), \dots, \sigma^{p_2 - 1}(j_2)\}.$$

Let us also define

$$\sigma_1' = (j_1 \ \sigma(j_1) \ \cdots \ \sigma^{p_1 - 1}(j_1)), \quad \sigma_2' = (j_2 \ \sigma(j_2) \ \cdots \ \sigma^{p_2 - 1}(j_2)).$$

One can then directly see that

$$\begin{aligned}\sigma_1' \circ \sigma_2' \circ \tau &= (j_1\ \sigma(j_1)\ \cdots\ \sigma^{p_1-1}(j_1)) \cdot (j_2\ \sigma(j_2)\ \cdots\ \sigma^{p_2-1}(j_2)) \cdot (j_1, j_2)\\ &= (j_1\ \sigma(j_1)\ \cdots\ \sigma^{p_1-1}(j_1)\ j_2\ \sigma(j_2)\ \cdots\ \sigma^{p_2-1}(j_2)).\end{aligned}$$

Now note that if $j \in O(\sigma, j_1) \cup O(\sigma, j_2)$ then $\sigma(j) = \sigma_1' \circ \sigma_2'(j)$ whence $\sigma \circ \tau(j) = \sigma_1' \circ \sigma_2' \circ \tau(j)$ since $\tau(j) \in O(\sigma, j_1) \cup O(\sigma, j_2)$. Therefore, $O(\sigma, j_1) \cup O(\sigma, j_2) = O(\sigma \circ \tau, j_1)$. Moreover, if $j \notin O(\sigma, j_1) \cup O(\sigma, j_2)$ then obviously $\sigma(j) = \sigma \circ \tau(j)$. Therefore, $O(\sigma \circ \tau, j) = O(\sigma, j)$ in this case. Summarising, $C(\sigma \circ \tau) = C(\sigma) - 1$. ▼

Let $\pi_2 \colon \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ be the canonical projection. Since $\pi_2(m+1) = \pi_2(m-1)$, the lemma shows that $\pi_2(C(\sigma)) = \pi_2(C(\sigma \circ \tau) + 1)$ for every $\sigma \in \mathfrak{S}_k$ and for every transposition τ.

To complete the proof note that $C(e) = k$ if e denotes the identity element of $\mathfrak{S}_k$. Now write $\sigma \in \mathfrak{S}_k$ as a finite product of transpositions: $\sigma = \tau_1 \circ \cdots \circ \tau_p$. Thus

$$\pi_2(C(\sigma)) = \pi_2(C(\tau_1 \circ \cdots \circ \tau_p)) = \pi_2(C(e) + p) = \pi_2(k + p).$$

Note that $\pi_2(C(\sigma))$ is defined independently of the choice of transpositions $\tau_1, \dots, \tau_p$. Thus, if $\sigma = \tau_1' \circ \cdots \circ \tau_{p'}'$ for transpositions $\tau_1', \dots, \tau_{p'}'$ then we must have $\pi_2(k + p) = \pi_2(k + p')$ meaning that $\pi_2(p) = \pi_2(p')$. But this means exactly that p and p' are either both even or both odd.

That sign is a homomorphism is a consequence of the obvious fact that the product of even permutations is even, the product of two odd permutations is even, and the product of an even and an odd permutation is odd. ■

Let us give some additional properties of the symmetric group that will be useful to us in our discussions of multilinear maps in Section 5.6, derivatives of such maps in Section II-1.4.2 and Theorem II-1.4.50.

Let $k_1, \dots, k_m \in \mathbb{Z}_{\geq 0}$ be such that $\sum_{j=1}^{m} k_m = k$. Let $\mathfrak{S}_{k_1|\cdots|k_m}$ be the subgroup of $\mathfrak{S}_k$ with the property that elements σ of $\mathfrak{S}_{k_1|\cdots|k_m}$ take the form

$$\begin{pmatrix} 1 & \cdots & k_1 & \cdots & k_1 + \cdots + k_{m-1} + 1 & \cdots & k_1 + \cdots + k_m \\ \sigma_1(1) & \cdots & \sigma_1(k_1) & \cdots & k_1 + \cdots + k_{m-1} + \sigma_m(1) & \cdots & k_1 + \cdots + k_{m-1} + \sigma_m(k_m) \end{pmatrix},$$

where $\sigma_j \in \mathfrak{S}_{k_j}$, $j \in \{1, \dots, m\}$. The assignment $(\sigma_1, \dots, \sigma_m) \mapsto \sigma$ with σ as above is an isomorphism of $\mathfrak{S}_{k_1} \times \cdots \times \mathfrak{S}_{k_m}$ with $\mathfrak{S}_{k_1|\cdots|k_m}$. Also denote by $\mathfrak{S}_{k_1,\dots,k_m}$ the subset of $\mathfrak{S}_k$ having the property that $\sigma \in \mathfrak{S}_{k_1,\dots,k_m}$ satisfies

$$\sigma(k_1 + \cdots + k_j + 1) < \cdots < \sigma(k_1 + \cdots + k_j + k_{j+1}), \qquad j \in \{0, 1, \dots, m-1\}.$$

Now we have the following result.

4.1.38 Proposition (Decompositions of the symmetric group) *With the above notation, the map $(\sigma_1, \cdots \sigma_{\mathrm{m}}) \mapsto \sigma_1 \circ \cdots \circ \sigma_{\mathrm{m}}$ from $\mathfrak{S}_{\mathrm{k}_1,\dots,\mathrm{k}_\mathrm{m}} \times \mathfrak{S}_{\mathrm{k}_1|\cdots|\mathrm{k}_\mathrm{m}}$ to $\mathfrak{S}_\mathrm{k}$ is a bijection.*

Proof Let P be the set of partitions $(S_1, \dots, S_m)$ of $\{1, \dots, k\}$ (i.e., $\{1, \dots, k\} = \overset{\circ}{\cup}{}_{j=1}^{m} S_j$) such that $\operatorname{card}(S_j) = k_j$, $j \in \{1, \dots, m\}$. Note that $\mathfrak{S}_k$ acts in a natural way on P. That is, if $(S_1, \dots, S_m) \in P$ and if $\sigma \in \mathfrak{S}_k$ then we can define $\sigma(S_1, \dots, S_m)$ to be the partition $(S'_1, \dots, S'_m) \in P$ for which $\sigma(S_j) = S'_j$ for each $j \in \{1, \dots, m\}$. Now specifically choose $S = (S_1, \dots, S_m) \in P$ by

$$S_j = \{k_0 + \dots + k_{j-1} + 1, \dots, k_1 + \dots + k_j\}, \qquad j \in \{1, \dots, m\},$$

taking $k_0 = 0$. Note that $\sigma \in \mathfrak{S}_k$ has the property that $\sigma(S) = S$ if and only if $\sigma \in \mathfrak{S}_{k_1|\cdots|k_m}$. For a general $T = (T_1, \dots, T_m) \in P$ let $\mathfrak{S}_{S \to T}$ be the set of $\sigma \in \mathfrak{S}_k$ that map S to T. Note that for a given $T \in P$ there exists a unique element of $\mathfrak{S}_{k_1, \dots, k_m} \cap \mathfrak{S}_{S \to T}$ (why?). Let us denote this unique permutation by $\sigma_T \in \mathfrak{S}_{k_1, \dots, k_m} \cap \mathfrak{S}_{S \to T}$. We claim that

$$\mathfrak{S}_{S \to T} = \{\sigma_T \circ \sigma' \mid \sigma' \in \mathfrak{S}_{k_1|\cdots|k_m}\}.$$

Indeed, if $\sigma \in \mathfrak{S}_{S \to T}$ then $\sigma_T^{-1} \circ \sigma(S) = S$ and so $\sigma_T^{-1} \circ \sigma = \sigma'$ for some $\sigma' \mathfrak{S}_{k_1|\cdots|k_m}$. Thus $\sigma = \sigma_T \circ \sigma'$ and so

$$\mathfrak{S}_{S \to T} \subseteq \{\sigma_T \circ \sigma' \mid \sigma' \in \mathfrak{S}_{k_1|\cdots|k_m}\}.$$

Conversely, if $\sigma' \in \mathfrak{S}_{k_1|\cdots|k_m}$ then $\sigma_T \circ \sigma' \in \mathfrak{S}_{S \to T}$ since $\sigma'(S) = S$. This gives $\mathfrak{S}_{S \to T} = \sigma_T \mathfrak{S}_{k_1|\cdots|k_m}$. Since σ_T is the unique element of $\mathfrak{S}_{k_1, \dots, k_m}$ for which this holds, it follows that if $\sigma \in \mathfrak{S}_{S \to T}$ for some $T \in P$ we have $\sigma = \sigma_1 \circ \sigma_2$ for unique $\sigma_1 \in \mathfrak{S}_{k_1, \dots, k_m}$ and $\sigma_2 \in \mathfrak{S}_{k_1|\cdots|k_m}$. Now, if $\sigma \in \mathfrak{S}_k$ then $\sigma \in \mathfrak{S}_{S \to T}$ for $T = \sigma^{-1}(S)$, and so the result holds. ■

Exercises

4.1.1 Do the following;

(a) prove Proposition 4.1.6;

(b) state which of the statements in Proposition 4.1.6 holds for semigroups;

(c) state which of the statements in Proposition 4.1.6 holds for monoids.

4.1.2 Let M be a monoid for which $ab = ba$ for all $a, b \in \mathsf{M}$, and let $m_1, \dots, m_k \in \mathsf{M}$ be elements for which there exists no inverse. Show that there is also no inverse for $m_1 \cdots m_k$.

4.1.3 Let G be a group and let $a, b, c \in \mathsf{G}$.

(a) Show that if $ab = ac$ then $b = c$.

(b) Show that if $ac = bc$ then $a = b$.

4.1.4 Let G and H be groups. Show that, if $\phi \colon \mathsf{G} \to \mathsf{H}$ is an isomorphism, then ϕ^{-1} is a homomorphism, and so also an isomorphism.

4.1.5 Prove Proposition 4.1.10.

4.1.6 Show that the following sets are subgroups of $\mathbb{R}$ with the group operation of addition:

(a) $\mathbb{Z}$;

(b) $\mathbb{Z}(\Delta) = \{j\delta \mid j \in \mathbb{Z}\}$;

(c) $\mathbb{Q}$.

The next two parts of this problem suppose that you know something about polynomials; we consider these in detail in Section 4.4. In any case, you should also show that the following sets are subgroups of $\mathbb{R}$ with the group operation of addition.

(d) the set $\bar{\mathbb{Q}} \cap \mathbb{R}$ of real algebraic numbers (recall that $z \in \mathbb{C}$ is an ***algebraic number*** if there exists a polynomial $P \in \mathbb{Z}[\xi]$ (i.e., one with integer coefficients) for which $P(z) = 0$, and we denote the set of algebraic numbers by $\bar{\mathbb{Q}}$);

(e) the set $\mathbb{K} \cap \mathbb{R}$ of real algebraic integers (recall that $z \in \mathbb{C}$ is an ***algebraic integer*** if there exists a monic polynomial $P \in \mathbb{Z}[\xi]$ (i.e., one with integer coefficients, and with the highest degree coefficient being 1) for which $P(z) = 0$, and we denote the set of algebraic integers by $\bar{\mathbb{K}}$).

4.1.7 Show that the subsets

(a) $x_0 + \mathbb{Q} = \{x_0 + q \mid q \in \mathbb{Q}\}$ for $x_0 \in \mathbb{R}$ and

(b) $\mathbb{Z}(x_0, \Delta) = \{x_0 + k\Delta \mid k \in \mathbb{Z}\}$ for $x_0 \in \mathbb{R}$

of $\mathbb{R}$ are semigroups with the binary operation

$$(x_0 + y_1) + (x_0 + y_2) = x_0 + y_1 + y_2.$$

Answer the following questions.

(c) Show that $x_0 + \mathbb{Q} = \mathbb{Q}$ if and only if $x_0 \in \mathbb{Q}$ and that $\mathbb{Z}(x_0, \Delta) = \mathbb{Z}(\Delta)$ if and only if $x_0 \in \mathbb{Z}(\Delta)$.

(d) Suppose that the binary operations on the semigroups $x_0+\mathbb{Q}$ and $\mathbb{Z}(x_0, \Delta)$ are as defined above. Show that the semigroup is a subgroup of $\mathbb{R}$ if and only if $x_0 = 0$.

4.1.8 Show that N is a normal subgroup of G if and only if $gng^{-1} \in \mathsf{N}$ for all $g \in \mathsf{G}$ and $n \in \mathsf{N}$.

4.1.9 Let G and H be groups and let $\phi\colon \mathsf{G} \to \mathsf{H}$ be an epimorphism. Show that the map $\phi_0\colon \mathsf{G}/\ker(\phi) \to \mathsf{H}$ defined by $\phi_0(g\ker(\phi)) = \phi(g)$ is a well-defined isomorphism.

4.1.10 Let G be a group and let H and N be subgroups of G with N normal. Prove that

(a) $\mathsf{N} \cap \mathsf{H}$ is a normal subgroup of H,

(b) N is a normal subgroup of $\langle \mathsf{N} \cup \mathsf{H} \rangle$, and

(c) $\mathsf{NH} = \langle \mathsf{N} \cup \mathsf{H} \rangle = \mathsf{HK}$.

4.1.11 Show that, if N is a normal subgroup of G, then there is a group H and an epimorphism $\phi\colon \mathsf{G} \to \mathsf{H}$ for which $\mathsf{N} = \ker(\phi)$.

In the following exercise you will use the definition that a transposition $\sigma \in \mathfrak{S}_k$ is ***adjacent*** if it has the form $\sigma = (j, j+1)$ for some $j \in \{1, \dots, k-1\}$. •

4.1.12 Show that any permutation $\sigma \in \mathfrak{S}_k$ is a finite product of adjacent transpositions.

4.1.13 Show that the only permutation that is a cycle of length 1 is the identity map.

4.1.14 Show that if $\sigma_1, \sigma_2 \in \mathfrak{S}_k$ are disjoint then $\sigma_1 \circ \sigma_2 = \sigma_2 \circ \sigma_1$.

Section 4.2

Rings

The number systems we have thus far considered in the text, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, are each endowed with two natural binary operations, those of addition and multiplication. Moreover, in each case the two binary operations fit together in a nice way. This combination of additive and multiplicative structures has the name "ring." In this section we study the basic structure of rings, including an introduction to some special properties of rings that will be possessed by some of the rings we will encounter in these volumes.

Do I need to read this section? The material in this section can perhaps be passed up unless (1) the reader is interested in understanding well the material from Sections 4.3 and 4.4, or (2) the reader is at one of those points in the text where ring theory comes up (and there are such points, e.g., Sections 5.8.2 and 5.8.3). •

4.2.1 Definitions and basic properties

4.2.1 Definition (Ring) A ***ring*** is a set R with two binary operations, $(r_1, r_2) \mapsto r_1 + r_2$ and $(r_1, r_2) \mapsto r_1 \cdot r_2$, called ***addition*** and ***multiplication***, respectively, and which together satisfy the following rules:

(i) $(r_1 + r_2) + r_3 = r_1 + (r_2 + r_3)$, $r_1, r_2, r_3 \in \mathsf{R}$ (***associativity*** of addition);

(ii) $r_1 + r_2 = r_2 + r_1$, $r_1, r_2 \in \mathsf{R}$ (***commutativity*** of addition);

(iii) there exists $0_{\mathsf{R}} \in \mathsf{R}$ such that $r + 0_{\mathsf{R}} = r$, $r \in \mathsf{R}$ (***additive identity***);

(iv) for $r \in \mathsf{R}$, there exists $-r \in \mathsf{R}$ such that $r + (-r) = 0_{\mathsf{R}}$ (***additive inverse***);

(v) $(r_1 \cdot r_2) \cdot r_3 = r_1 \cdot (r_2 \cdot r_3)$, $r_1, r_2, r_3 \in \mathsf{R}$ (***associativity*** of multiplication);

(vi) $r_1 \cdot (r_2 + r_3) = (r_1 \cdot r_2) + (r_1 \cdot r_3)$, $r_1, r_2, r_3 \in \mathsf{R}$ (***left distributivity***);

(vii) $(r_1 + r_2) \cdot r_3 = (r_1 \cdot r_3) + (r_2 \cdot r_3)$, $r_1, r_2, r_3 \in \mathsf{R}$ (***right distributivity***).

If there exists $1_{\mathsf{R}} \in \mathsf{R}$ such that $1_{\mathsf{R}} \cdot r = r \cdot 1_{\mathsf{R}}$ for all $r \in \mathsf{R}$, then the ring is a ***unit ring***, or a ***ring with unit***. The element 1_{R} in a unit ring is called the ***unity element***. If $r_1 \cdot r_2 = r_2 \cdot r_1$ for all $r_1, r_2 \in \mathsf{R}$, then the ring is ***commutative***. •

As usual, we may omit specific reference to "·" when writing the operation of multiplication, and will write $r_1 r_2$ in place of $r_1 \cdot r_2$.

Let us give some examples of rings.

4.2.2 Examples (Rings)

1. The set $\mathbb{Z}$ with its usual operations of addition and multiplication is a commutative unit ring.
2. The set $\mathbb{Q}$ with its usual operations of addition and multiplication is a commutative unit ring.

3. The set $\mathbb{R}$ with its usual operations of addition and multiplication is a commutative unit ring.
4. Let $k \in \mathbb{Z}_{>0}$ and on the set $\mathbb{Z}_k = \mathbb{Z}/k\mathbb{Z}$ of cosets in $\mathbb{Z}$ (see Example 4.1.18–1) define operations of addition and multiplication by

$$(j_1 + k\mathbb{Z}) + (j_2 + k\mathbb{Z}) = (j_1 + j_2) + k\mathbb{Z}, \qquad (j_1 + k\mathbb{Z}) \cdot (j_2 + k\mathbb{Z}) = (j_1 \cdot j_2) + k\mathbb{Z}.$$

We should show that these operations are well defined, in that they do not depend on the choice of representative in an equivalence class. If

$$j_a + k\mathbb{Z} = \tilde{j}_a + k\mathbb{Z}, \qquad a \in \{1, 2\},$$

then this implies that there exists $l_1, l_2 \in \mathbb{Z}$ such that $\tilde{j}_a = j_a + kl$, $a \in \{1, 2\}$. Therefore,

$$(\tilde{j}_1 + \tilde{j}_2) + k\mathbb{Z} = (j_1 + j_2 + k(l_1 + l_2)) + k\mathbb{Z} = (j_1 + j_2) + k\mathbb{Z}$$

and

$$(\tilde{j}_1 \cdot \tilde{j}_2) + k\mathbb{Z} = (j_1 \cdot j_2 + k(j_1 l_2 + j_2 l_1 + k l_1 l_2)) + k\mathbb{Z} = (j_1 \cdot j_2) + k\mathbb{Z}.$$

It is then a straightforward matter to check that these definitions of addition and multiplication satisfy the ring axioms, and so $\mathbb{Z}_k$ is indeed a ring.
5. Let S be a nonempty set, let R be a ring, and recall from Definition 1.3.1 the notation R^S for the set of maps from S to R. Define operations of addition and multiplication on R^S by

$$(f + g)(x) = f(x) + g(x), \quad (f \cdot g)(x) = f(x) \cdot g(x),$$

respectively. It is then trivial to verify that R^S is a ring when equipped with these operations. Indeed, if one sits down to check this, one sees that all one is doing is checking that the ring operations on R satisfy the necessary conditions. Also, if R is commutative, then so too is R^S, and if R is a unit ring with unit element 1_{R}, then the map $x \mapsto 1_{\mathsf{R}}$ is easily seen to be a unity element in R^S.
6. The set $\mathbb{R}[[\xi]]$ of formal $\mathbb{R}$-power series considered in Section 3.7 is a commutative unit ring.
7. If R is a ring with addition denoted by $(r_1, r_2) \mapsto r_1 + r_2$ and multiplication denoted by $(r_1, r_2) \mapsto r_1 \cdot r_2$, we define its ***opposite ring*** to be the set R with the operations of addition and multiplication, denoted by $+_{\text{op}}$ and $\cdot_{\text{op}}$, given by

$$r_1 +_{\text{op}} r_2 = r_1 + r_2, \quad r_1 \cdot_{\text{op}} r_2 = r_2 \cdot r_1.$$

One can easily verify that these operations do indeed define a ring structure, and we denote the resulting ring by R_{op}. If R is commutative, then clearly $R_{=}\mathsf{R}_{\text{op}}$ (where equality is not just as sets, but as rings). •

The following characterisation of the binary operations in a ring is useful to keep in mind.

4.2.3 Proposition (Properties of ring operations) *If* R *is a ring, then the following statements hold:*

(i) the set R*, with the binary operation of addition, is an Abelian group;*

(ii) the set R*, with the binary operation of multiplication, is a semigroup;*

(iii) if additionally R *is a unit ring, then the set* R*, with the binary operation of multiplication, is a monoid.*

The following notation concerning sums and products of ring elements is useful.

4.2.4 Notation (Sums and products in rings) Let R be a ring and let $(r_i)_{i\in I}$ be a family of elements of R, only finitely many of which are nonzero; denote the nonzero elements by $r_{i_1},\dots,r_{i_k}$. We then denote

$$\sum_{i\in I} r_i = r_{i_1} + \cdots + r_{i_k}.$$

Similarly, if R is a unit ring and if only finitely many terms in the family $(r_i)_{i\in I}$ are not equal to 1_{R} (say the nonzero elements are $r_{i_1},\dots,r_{i_k}$), then we denote

$$\prod_{i\in I} r_i = r_{i_1}\cdots r_{i_k}.$$ •

Aided by this observation, and by other fairly simple arguments that we leave to the reader as Exercise 4.2.2, we have the following properties of a ring.

4.2.5 Proposition (Elementary properties of rings) *If* R *is a ring, then the following statements hold:*

(i) there is exactly one element $0_{\mathsf{R}} \in \mathsf{R}$ *that satisfies* $0_{\mathsf{R}} + \mathrm{r} = \mathrm{r} + 0_{\mathsf{R}} = \mathrm{r}$ *for all* $\mathrm{r} \in \mathsf{R}$*, i.e., the additive identity is unique;*

(ii) for $\mathrm{r} \in \mathsf{R}$*, there exists exactly one element* $\mathrm{r}' \in \mathsf{R}$ *satisfying* $\mathrm{r} + \mathrm{r}' = 0_{\mathsf{R}}$*, i.e., additive inverses are unique;*

(iii) for $\mathrm{r} \in \mathsf{R}$, $-(-\mathrm{r}) = \mathrm{r}$*;*

(iv) for $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$, $-(\mathrm{r}_1 + \mathrm{r}_2) = -\mathrm{r}_1 - \mathrm{r}_2$*;*

(v) if $\mathrm{r}_1, \mathrm{r}_2, \in \mathsf{R}$ *satisfy* $\mathrm{r}_1 + \mathrm{s} = \mathrm{r}_2 + \mathrm{s}$*, then* $\mathrm{r}_1 = \mathrm{r}_2$*;*

(vi) if $\mathrm{r}_1, \mathrm{r}_2, \mathrm{s} \in \mathsf{R}$*, then there exists a unique* $\mathrm{s} \in \mathsf{S}$ *such that* $\mathrm{r}_1 + \mathrm{s} = \mathrm{r}_2$*;*

(vii) there is at most one element $1_{\mathsf{R}} \in \mathsf{R}$ *such that* $\mathrm{r} \cdot 1_{\mathsf{R}} = 1_{\mathsf{R}} \cdot \mathrm{r} = \mathrm{r}$ *for all* $\mathrm{r} \in \mathsf{R}$*, i.e., the unity element, if it exists, is unique;*

(viii) for $\mathrm{r} \in \mathsf{R}$, $\mathrm{r} \cdot 0_{\mathsf{R}} = 0_{\mathsf{R}} \cdot \mathrm{r} = 0_{\mathsf{R}}$*;*

(ix) for $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$, $-(\mathrm{r}_1 \cdot \mathrm{r}_2) = (-\mathrm{r}_1) \cdot \mathrm{r}_2 = \mathrm{r}_1 \cdot (-\mathrm{r}_2)$*;*

(x) for $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$, $(-\mathrm{r}_1) \cdot (-\mathrm{r}_2) = \mathrm{r}_1 \cdot \mathrm{r}_2$*;*

(xi) for $\mathrm{r}_1, \mathrm{r}_2, \mathrm{s} \in \mathsf{R}$, $(\mathrm{r}_1 - \mathrm{r}_2) \cdot \mathrm{s} = \mathrm{r}_1 \cdot \mathrm{s} - \mathrm{r}_2 \cdot \mathrm{s}$*;*

(xii) for $\mathrm{r}_1, \mathrm{r}_2, \mathrm{s} \in \mathsf{R}$, $\mathrm{s} \cdot (\mathrm{r}_1 - \mathrm{r}_2) = \mathrm{s} \cdot \mathrm{r}_1 - \mathrm{s} \cdot \mathrm{r}_2$.

Note that, in a ring R, the equations $r_1 \cdot s = r_2$ and $s \cdot r_1 = r_2$ may not have solutions for s given $r_1, r_2 \in \mathsf{R}$. This, of course, is because a ring with multiplication is only a semigroup, and not necessarily a group. However, it can happen that *some* of the elements of a ring may be invertible.

4.2.6 Definition (Unit) Let R be a ring with unity element 1_{R} and let $r \in \mathsf{R}$.

(i) The element r is ***left invertible*** if there exists $a \in \mathsf{R}$ such that $ar = 1_{\mathsf{R}}$. The ring element a is a ***left-inverse*** of r.

(ii) The element r is ***right invertible*** if there exists $b \in \mathsf{R}$ such that $rb = 1_{\mathsf{R}}$. The ring element b is a ***right-inverse*** of r.

(iii) The element r is a ***unit*** if it is both left and right invertible. •

Let us give some examples of unit elements in some of the rings we have encountered. We leave the straightforward verification of the assertions we make to the reader.

4.2.7 Examples (Unit elements)

1. The set of units in the ring $\mathbb{Z}$ is $\{1, -1\}$.
2. The set of units in the ring $\mathbb{Q}$ is $\mathbb{Q}^*$.
3. The set of units in the ring $\mathbb{R}$ is $\mathbb{R}^*$.
4. Let us consider the ring $\mathbb{Z}_k = \mathbb{Z}/k\mathbb{Z}$, and describe its units. We shall rely on some ideas that will only receive a full treatment in Section 4.2.7, although the reader may well have no difficulty believing the facts we assert based on their previous experience.

   An element $j + k\mathbb{Z} \in \mathbb{Z}_k$, $j \in \{0, 1, \dots, k-1\}$, is a unit if there exists $j' \in \mathbb{Z}$ such that $(j + k\mathbb{Z}) \cdot (j' + k\mathbb{Z}) = 1 + k\mathbb{Z}$. This means that $jj' = 1 + kl$ for some $l \in \mathbb{Z}$. Equivalently, $j + k\mathbb{Z}$ is a unit if and only if the equation

$$jj' + kl = 1$$

   has a solution for $j'l \in \mathbb{Z}$. As we shall see in Corollary 4.2.78, this is equivalent to the assertion that j and k have no common prime factors. Thus the set of units in $\mathbb{Z}_k$ is exactly

$$\{j + k\mathbb{Z} \mid \ j \in \{0, 1, \dots, k-1\},\ j \text{ and } k \text{ have no common prime factors}\}.$$

5. Let S be a nonempty set and R be a unit ring, and recall that R^S denotes the set of R-valued maps with domain S. The set of units for the ring R^S is then easily seen to be the set of maps which take values in the units of R.
6. The set of units in the ring $\mathbb{R}[[\xi]]$ consists of those formal $\mathbb{R}$-power series whose zeroth coefficient is nonzero. Indeed, in Proposition 3.7.5 we showed that such $\mathbb{R}$-formal power series possess a multiplicative inverse. One can easily

see conversely that if the zeroth coefficient of a $\mathbb{R}$-formal power series is zero, then it can have no multiplicative inverse, simply by using the definition of multiplication of $\mathbb{R}$-formal power series.

7. Let us give an example of a ring with an element that is left invertible but not right invertible. Our construction relies on some concepts we have not yet introduced, but with which the reader may well be familiar.

   We denote by $\mathbb{R}[\xi]$ the set of polynomials with real coefficients with indeterminate ξ (see Section 4.4). We think of $\mathbb{R}[\xi]$ as a $\mathbb{R}$-vector space (see Example 4.5.2–6). By $\mathsf{R} = \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}[\xi];\mathbb{R}[\xi])$ we denote the set of linear maps on $\mathbb{R}[\xi]$ (see Definition 4.5.4), noting that R is a ring with multiplication given by composition (see Corollary 5.4.18). Now define $r_d, r_i \in \mathsf{R}$ by

$$r_d\left(\sum_{j=0}^{k} a_j\xi^j\right) = \sum_{j=1}^{k} ja_j\xi^{j-1}, \qquad r_i\left(\sum_{j=0}^{k} a_j\xi^j\right) = \sum_{j=0}^{k} \frac{a_j}{j+1}\xi^{j+1}.$$

   Note that r_d returns the derivative of the polynomial and r_i returns the integral of the polynomial with the condition that the constant term of the integrated polynomial be zero. With these interpretations of r_d and r_i (or by direct computation) one can check that

$$r_d \circ r_i\left(\sum_{j=0}^{k} a_j\xi^k\right) = \sum_{j=0}^{k} a_j\xi^k, \qquad r_i \circ r_d\left(\sum_{j=0}^{k} a_j\xi^k\right) = \sum_{j=1}^{k} a_j\xi^j.$$

   Thus $r_d \circ r_i$ is the identity map (i.e., the unity element in R) and $r_i \circ r_d(P) = 0$ for any constant polynomial P. Thus r_i is left invertible with r_d as a left-inverse. This is reflected by the fact that, as a linear map, r_i is injective (cf. Proposition 5.4.46). In the same vein, since r_i is not surjective (no nonzero constant polynomial is contained in its image), it cannot be right invertible. Note that r_d has the opposite feature: it is right invertible but not left invertible. •

Let us give a description of some of the properties of the set of units of a ring. The verification of these is an easy application of Proposition 4.1.6

4.2.8 Proposition (Elementary properties of units) *If* R *is a ring with unit. then the following statements hold:*

(i) the set of units in R *is a group when equipped with the binary operation of multiplication;*

(ii) for a unit $\mathrm{r} \in \mathsf{R}$*, there exists exactly one element* $\mathrm{r}' \in \mathsf{R}$ *such that* $\mathrm{r}' \cdot \mathrm{r} = \mathrm{r} \cdot \mathrm{r}' = 1_{\mathsf{R}}$*;*

(iii) for a unit $\mathrm{r} \in \mathsf{R}$*,* r^{-1} *is a unit, and* $(\mathrm{r}^{-1})^{-1} = \mathrm{r}$*;*

(iv) if $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$*, if* $\mathrm{s} \in \mathsf{R}$ *is a unit, and if* $\mathrm{s} \cdot \mathrm{r}_1 = \mathrm{s} \cdot \mathrm{r}_2$*, then* $\mathrm{r}_1 = \mathrm{r}_2$*;*

(v) if $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$*, if* $\mathrm{s} \in \mathsf{R}$ *is a unit, and if* $\mathrm{r}_1 \cdot \mathrm{s} = \mathrm{r}_2 \cdot \mathrm{s}$*, then* $\mathrm{r}_1 = \mathrm{r}_2$*;*

(vi) if $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$ *with* r_1 *a unit, then there exists a unique* $\mathrm{s} \in \mathsf{R}$ *such that* $\mathrm{r}_1 \cdot \mathrm{s} = \mathrm{r}_2$*;*

(vii) *if* $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$ *with* r_2 *a unit, then there exists a unique* $\mathrm{s} \in \mathsf{R}$ *such that* $\mathrm{s} \cdot \mathrm{r}_1 = \mathrm{r}_2$.

Ring elements which differ by multiplication by a unit arise often in discussion of rings. We give these a name.

4.2.9 Definition (Associate) If R is a commutative unit ring, two elements $r_1, r_2 \in \mathsf{R}$ are ***associates*** if there exists a unit u such that $r_2 = ur_1$. Two ring elements that are not associates are ***nonassociate***. •

Note that the relation of being associates is an equivalence relation in R.

When discussing groups, we introduced the notation g^k where g is an element of a semigroup and k is a natural number. This notation can be profitably applied to rings. However, we must take more care with the notation since we have two operations to which the notation can be applied. Thus, for a ring R, for $r \in \mathsf{R}$, and $k \in \mathbb{Z}_{>0}$ we define elements $r^k, kr \in \mathsf{R}$ as follows. For $k = 1$ we take $r^k = r$ and $kr = r$. For general $k \in \mathbb{Z}_{>0}$ we define $r^k = r^{k-1} \cdot r$ and $kr = (k-1)r + r$. We also take $0r = 0_{\mathsf{R}}$ and $kr = -(-k)r$ for $k < 0$. In the case when R has a unity element, we define $r^0 = 1_{\mathsf{R}}$. It is then a direct application of Proposition 4.1.7 to get the following assertions.

4.2.10 Proposition (Properties of $\mathbf{r^k}$ and kr) *If* R *is a ring, if* $\mathrm{r} \in \mathsf{R}$, *and if* $\mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}_{>0}$, *then*

(i) $\mathrm{r}^{\mathrm{k}_1} \cdot \mathrm{r}^{\mathrm{k}_2} = \mathrm{r}^{\mathrm{k}_1+\mathrm{k}_2}$ *and*

(ii) $(\mathrm{r}^{\mathrm{k}_1})^{\mathrm{k}_2} = \mathrm{r}^{\mathrm{k}_1\mathrm{k}_2}$.

If $\mathrm{k}, \mathrm{k}_1, \mathrm{k}_2 \in \mathbb{Z}$ *and* $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$ *then*

(iii) $(\mathrm{k}_1\mathrm{r}) + (\mathrm{k}_2\mathrm{r}) = (\mathrm{k}_1 + \mathrm{k}_2)\mathrm{r}$,

(iv) $\mathrm{k}_1(\mathrm{k}_2\mathrm{r}) = (\mathrm{k}_1\mathrm{k}_2)\mathrm{r}$,

(v) $\mathrm{k}(\mathrm{r}_1 \cdot \mathrm{r}_2) = (\mathrm{k}\mathrm{r}_1) \cdot \mathrm{r}_2 = \mathrm{r}_1 \cdot (\mathrm{k}\mathrm{r}_2)$, *and*

(vi) *if* R *is commutative,* $(\mathrm{r}_1 \cdot \mathrm{r}_2)^{\mathrm{k}} = \mathrm{r}_1^{\mathrm{k}} \cdot \mathrm{r}_2^{\mathrm{k}}$.

Using this notation one can then state and prove a general version of the Binomial Theorem. The proof goes like the usual case (see Exercise 2.2.1).

4.2.11 Proposition (Binomial Theorem for commutative rings) *Let* R *be a commutative ring and let* $\mathrm{r}, \mathrm{s} \in \mathsf{R}$. *Then, for any* $\mathrm{k} \in \mathbb{Z}_{>0}$, *we have*

$$(\mathrm{r} + \mathrm{s})^{\mathrm{k}} = \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{B}_{\mathrm{k},\mathrm{j}} \mathrm{r}^{\mathrm{j}} \mathrm{s}^{\mathrm{k}-\mathrm{j}},$$

where

$$\mathrm{B}_{\mathrm{k},\mathrm{j}} = \binom{\mathrm{k}}{\mathrm{j}} \triangleq \frac{\mathrm{k}!}{\mathrm{j}!(\mathrm{k}-\mathrm{j})!}, \qquad \mathrm{j}, \mathrm{k} \in \mathbb{Z}_{>0},\ \mathrm{j} \leq \mathrm{k}.$$

4.2.2 Subrings and ideals

The study of rings is one that takes on a certain amount of depth. We shall need some topics from ring theory, although not a great many. The first construction to consider is the usual association of a subset of a ring which inherits the ring structure. This is the notion of a subring. Rings also have associated to them the important (for reasons we shall see as we proceed) concept of an ideal, which is more general than the idea of a subring.

4.2.12 Definition (Subring and ideal) Let R be a ring. A nonempty subset S of R is a:

(i) ***subsemiring*** if

(a) $r_1 + r_2 \in \mathsf{S}$ for all $r_1, r_2 \in \mathsf{S}$ and

(b) $r_1 \cdot r_2 \in \mathsf{S}$ for all $r_1, r_2 \in \mathsf{S}$;

(ii) ***subring*** if

(a) $r_1 + r_2 \in \mathsf{S}$ for all $r_1, r_2 \in \mathsf{S}$,

(b) $r_1 \cdot r_2 \in \mathsf{S}$ for all $r_1, r_2 \in \mathsf{S}$, and

(c) $-r \in \mathsf{S}$ for all $r \in \mathsf{S}$;

(iii) ***right ideal*** if

(a) it is a subring and

(b) $r_1 \cdot r_2 \in \mathsf{S}$ for all $r_1 \in \mathsf{S}$ and $r_2 \in \mathsf{R}$;

(iv) ***left ideal*** if

(a) it is a subring and

(b) $r_1 \cdot r_2 \in \mathsf{S}$ for all $r_1 \in \mathsf{R}$ and $r_2 \in \mathsf{S}$;

(v) ***two-sided ideal*** if it is both a right ideal and a left ideal. •

Clearly, for commutative rings, the notions of right ideal, left ideal, and two-sided ideal coincide. In such cases we may simply say ***ideal*** to denote any one of these equivalent concepts.

As with groups, subrings are themselves rings, as the reader can prove in Exercise 4.2.5.

4.2.13 Proposition (A subring is a ring) *If* S *is a subring of a ring* R*, then* $0_{\mathsf{R}} \in \mathsf{S}$*. In particular,* S *is a ring using the binary operations inherited from* R*.*

The following characterisation of ideals makes it somewhat easier to check whether a given subset is an ideal than is the case from simply looking at the definition.

4.2.14 Proposition (Characterisation of ideals) *A nonempty subset* I *of a ring* R *is a left (resp. right) ideal if and only if*

(i) $\mathrm{s}_1 - \mathrm{s}_2 \in \mathsf{I}$ *for all* $\mathrm{s}_1, \mathrm{s}_2 \in \mathsf{I}$ *and*

(ii) $\mathrm{r} \cdot \mathrm{s} \in \mathsf{I}$ *(resp.* $\mathrm{s} \cdot \mathrm{r} \in \mathsf{I}$*) for all* $\mathrm{r} \in \mathsf{R}$ *and* $\mathrm{s} \in \mathsf{I}$*.*

Proof First suppose that I is a left ideal (the case for a right ideal is proved similarly). It is clear that $r \cdot s \in \mathsf{I}$ for all $r \in \mathrm{r}$ and $s \in \mathsf{I}$. Now let $s_1, s_2 \in \mathsf{I}$. Since I is a subring, $-s_2 \in \mathsf{I}$ and so $s_1 + (-s_2) = s_1 - s_2 \in \mathsf{I}$.

Now suppose that I satisfies the two conditions in the statement of the proposition. Since the second of the conditions for a left ideal obviously holds, it remains to show that I is a subring. Note that $0_{\mathsf{R}} \in \mathsf{I}$ since $0_{\mathsf{R}} \cdot s = 0_{\mathsf{R}} \in \mathsf{I}$ for all $s \in \mathsf{I}$. Let $s \in \mathsf{I}$. Since $0_{\mathsf{R}} \in \mathsf{I}$ we have $0_{\mathsf{R}} - s = -s \in \mathsf{I}$. Now let $s_1, s_2 \in \mathsf{I}$. Then $-s_2 \in \mathsf{I}$ and so $s_1 - (-s_2) = s_1 + s_2 \in \mathsf{I}$. It is clear that $s_1 \cdot s_2 \in \mathsf{I}$ for all $s_1, s_2 \in \mathsf{I}$, and this shows that I is indeed a subring. ■

Let us give some examples of subrings and ideals. The most trivial of the assertions we make concerning these examples we leave to the reader to verify.

4.2.15 Examples (Subrings and ideals)

1. The subset $k\mathbb{Z}$ of $\mathbb{Z}$ is a subring, as is easily seen. Although $\mathbb{Z}$ is a ring with unity, $k\mathbb{Z}$ is a ring with unity if and only if $k = 1$. Note that $k\mathbb{Z}$ is also an ideal. Indeed, for $kj_1, kj_2 \in k\mathbb{Z}$ we have $kj_1 - kj_2 = k(j_1 - j_2) \in k\mathbb{Z}$, and for $kj_1 \in k\mathbb{Z}$ and $j_2 \in \mathbb{Z}$ we have $(kj_1)j_2 = k(j_1j_2) \in k\mathbb{Z}$.
2. $\mathbb{Z}$ is a subring of $\mathbb{Q}$, but it is fairly obviously not an ideal (the product of an integer with a rational number need not be an integer).
3. $\mathbb{Q}$ is a subring of $\mathbb{R}$, and it too is fairly obviously not an ideal (the product of a rational number with a real number need not be a rational number).
4. Let R be a ring and let $r_0 \in \mathsf{R}$. We claim that the set

$$(r_0) = \{r_0 r \mid r \in \mathsf{R}\}$$

is a right ideal. Indeed, if $r_0r_1, r_0r_2 \in (r_0)$ then $r_0r_1 - r_0r_2 = r_0(r_1 - r_2) \in (r_0)$, and if $r_0r_1 \in (r_0)$ and if $r_2 \in \mathsf{R}$, then $(r_0r_1)r_2 = r_0(r_1r_2) \in (r_0)$.

We shall explore ideals of this sort more fully in Section 4.2.8; in particular see Theorem 4.2.54. •

The importance of the notion of an ideal is explained by the following result.

4.2.16 Proposition (Quotients by ideals are rings) *Let* I *be a two-sided ideal of a ring* R *and let*

$$\mathrm{r} + \mathsf{I} = \{\mathrm{r} + \mathrm{s} \mid \mathrm{s} \in \mathsf{I}\} \quad \textit{and} \quad \mathsf{R}/\mathsf{I} = \{\mathrm{r} + \mathsf{I} \mid \mathrm{r} \in \mathsf{R}\}.$$

Then the binary operations on R/I *defined by*

$$(\mathrm{r}_1 + \mathsf{I}, \mathrm{r}_2 + \mathsf{I}) \mapsto (\mathrm{r}_1 + \mathrm{r}_2) + \mathsf{I} \quad \textit{and} \quad (\mathrm{r}_1 + \mathsf{I}, \mathrm{r}_2 + \mathsf{I}) \mapsto (\mathrm{r}_1\mathrm{r}_2) + \mathsf{I}$$

satisfy the conditions for addition and multiplication for a ring.

Proof Let us first verify that the binary operations defined make sense, in that they are independent of choice of representative. Let $r_1, r_2 \in \mathsf{R}$ and $s_1, s_2 \in \mathsf{I}$. We easily see that

$$((r_1 + s_1) + (r_2 + s_2)) + \mathsf{I} = (r_1 + r_2) + \mathsf{I}$$

and

$$((r_1 + s_1)(r_2 + s_2)) + \mathsf{I} = (r_1 r_2 + r_1 s_2 + r_2 s_2 + s_1 s_2) + \mathsf{I} = (r_1 r_2) + \mathsf{I},$$

which shows that the binary operations are indeed well-defined.

The matter of verifying that these binary operations satisfy the conditions for a ring is then a straightforward manipulation of symbols, the details of which we happily leave to the reader. ■

4.2.3 Prime and maximal ideals

Now we consider ideals having special properties that will be important to us later in this section, primarily in Section 4.2.9. It is convenient to use the following notation. Let R be a ring and let $\mathsf{I}, \mathsf{J} \subseteq \mathsf{R}$ be ideals. We then denote

$$\mathsf{IJ} = \{r_1 s_1 + \cdots + r_k s_k \mid r_j \in \mathsf{I},\ s_j \in \mathsf{J},\ j \in \{1, \dots, k\},\ k \in \mathbb{Z}_{>0}\}.$$

With this notation, we have the following definition.

4.2.17 Definition (Prime ideal, maximal ideal) Let R be a ring and let $\mathsf{I} \subseteq \mathsf{R}$ be a two-sided ideal. The ideal I is:

(i) ***prime*** if A and B are ideals for which $\mathsf{AB} \subseteq \mathsf{I}$, then either $\mathsf{A} \subseteq \mathsf{I}$ or $\mathsf{B} \subseteq \mathsf{I}$;

(ii) ***maximal*** if $\mathsf{I} \neq \mathsf{R}$ and if $\mathsf{J} \subseteq \mathsf{R}$ is an ideal for which $\mathsf{I} \subseteq \mathsf{J}$, then either $\mathsf{J} = \mathsf{I}$ or $\mathsf{J} = \mathsf{R}$. •

For prime ideals, there is an alternative characterisation, equivalent to the one we give in the case that the ring is commutative, that often makes it easier to check whether an ideal is prime.

4.2.18 Proposition (Characterisation of prime ideals) *If* R *is a ring, the following statements hold:*

(i) if $\mathsf{I} \subset \mathsf{R}$ *is an ideal such that* $\mathrm{rs} \in \mathsf{I}$ *implies that either* $\mathrm{r} \in \mathsf{I}$ *or* $\mathrm{s} \in \mathsf{I}$*, then* I *is prime;*

(ii) when R *is additionally commutative, then it holds that* $\mathrm{rs} \in \mathsf{I}$ *implies that either* $\mathrm{r} \in \mathsf{I}$ *or* $\mathrm{s} \in \mathsf{I}$ *if* I *is a prime ideal.*

Proof (i) Let $\mathsf{A}, \mathsf{B} \subseteq \mathsf{I}$ be ideals for which $\mathsf{AB} \subseteq \mathsf{I}$ and suppose that $\mathsf{A} \not\subseteq \mathsf{I}$. If $a \in \mathsf{A} - \mathsf{I}$ then $ab \in \mathsf{I}$ for every $b \in \mathsf{B}$. Therefore, $b \in \mathsf{I}$, and so $\mathsf{B} \subseteq \mathsf{I}$.

(ii) Let $ab \in \mathsf{I}$. Let $(ab) = \{rab \mid r \in \mathsf{R}\}$ and recall from Example 4.2.15–4 that (ab) is an ideal. We shall see in Theorem 4.2.52 that $(ab) \subseteq \mathsf{I}$ since (ab) is the smallest ideal containing ab. Moreover, in Theorem 4.2.54 we shall see that $(a)(b) \subseteq (ab)$, so that $(a)(b) \subseteq \mathsf{I}$, which implies that either $(a) \subseteq \mathsf{I}$ or $(b) \subseteq \mathsf{I}$, or that $a \in \mathsf{I}$ or $b \in \mathsf{I}$. ■

For maximal ideals, one of the interesting and nonobvious properties they possess is merely existing. Indeed, one has the following result.

4.2.19 Theorem (Maximal ideals exist) *If* R *is a unit ring with more than one element and if* $\mathsf{I} \subset \mathsf{R}$ *is a two-sided ideal, then there exists a maximal ideal* M *containing* I*.*

Proof Let $S(\mathsf{I})$ be the collection of ideals $\mathsf{J} \subset \mathsf{R}$ such that $\mathsf{I} \subseteq \mathsf{J}$. We partially order $S(\mathsf{I})$ by set inclusion; thus $\mathsf{J}_1 \preceq \mathsf{J}_2$ if $\mathsf{J}_1 \subseteq \mathsf{J}_2$. Let $\{\mathsf{I}_a \mid a \in A\}$ be a totally ordered subset of $S(\mathsf{I})$. We claim that $\bar{\mathsf{I}} = \cup_{a\in A}\mathsf{I}_a$ is an ideal. Let $r_1, r_2 \in \bar{\mathsf{I}}$ so that $r_1 \in \mathsf{I}_{a_1}$ and $r_2 \in \mathsf{I}_{a_2}$ for some $a_1, a_2 \in A$. Since $\{\mathsf{I}_a \mid a \in A\}$ is totally ordered we may suppose, without loss of generality, that $r_1 \in \mathsf{I}_{a_2}$. Then it follows that $r_1 - r_2 \in \mathsf{I}_{a_2} \subseteq \bar{\mathsf{I}}$. It also holds that $rr_1, r_1 r \in \mathsf{I}_{a_1} \subseteq \bar{\mathsf{I}}$ for each $r \in \mathsf{R}$. Thus $\bar{\mathsf{I}}$ is indeed an ideal. It is clear that $\mathsf{I} \subseteq \bar{\mathsf{I}}$. We claim that $\bar{\mathsf{I}} \neq \mathsf{R}$. This follows since $1_{\mathsf{R}} \notin \mathsf{I}_a$ for each $a \in A$ (or else we would have $\mathsf{I}_a = \mathsf{R}$ for some $a \in A$). Therefore, $1_{\mathsf{R}} \notin \bar{\mathsf{I}}$. Thus $\bar{\mathsf{I}}$ is an upper bound for $\{\mathsf{I}_a \mid a \in A\}$ in $S(\mathsf{I})$. The result now follows by Zorn's Lemma. ■

The theorem has the following corollary that will be useful in Section 4.2.9 in describing irreducible elements of a ring.

4.2.20 Corollary (Characterisation of maximal ideals) *Let* R *be a unit ring and let* $\mathsf{S}(\mathsf{R})$ *denote the collection of two-sided ideals that are strict subsets of* R*, and partially order* $\mathsf{S}(\mathsf{R})$ *by set inclusion. Then an ideal* I *is maximal if and only if it is a maximal element of the partially ordered set* $\mathsf{S}(\mathsf{R})$*.*

For many rings, maximal ideals are prime.

4.2.21 Proposition (Maximal ideals are prime in commutative unit rings) *If* R *is a commutative unit ring and if* M *is a maximal ideal in* R*, then* M *is a prime ideal.*

Proof Our proof relies on some constructions from Section 4.2.9.

Suppose that M is not prime so that, by Proposition 4.2.18, there exists $r, s \in \mathsf{R}$ such that $rs \in \mathsf{M}$ but $r, s \notin \mathsf{M}$. Let I_r and I_s be the ideals

$$\mathsf{I}_r = \{r_1 + r_2 r \mid r_1 \in \mathsf{M},\ r_2 \in \mathsf{R}\}, \quad \mathsf{I}_s = \{r_1 + r_2 s \mid r_1 \in \mathsf{M},\ r_2 \in \mathsf{R}\} \tag{4.1}$$

(one can check that these are indeed ideals, cf. Exercise 4.2.16). Since M is maximal we must have $\mathsf{I}_r = \mathsf{I}_s = \mathsf{R}$. By Theorem 4.2.54 we have $(r)(s) \subseteq (rs) \subseteq \mathsf{M}$. Now let $r' \in \mathsf{R}$. Since R is a unit ring, it is easy to see that $\mathsf{R} = \mathsf{R}\mathsf{R}$ which implies that $\mathsf{R} = \mathsf{I}_r\mathsf{I}_s$. Thus every element in R is a finite sum of products of elements from I_r and I_s. But such sums, by (4.1), are necessarily of the form

$$r_1 r_2 + a_1 r_3 + b_1 r_4 + a_2 b_2,$$

where $r_1, r_2, r_3, r_4 \in \mathsf{M}$, $a_1, a_2 \in (r)$, and $b_1, b_2 \in (s)$. But we then have $r_1 r_2 \in \mathsf{M}$ since M is a subring, $a_1 r_3, b_1 r_4 \in \mathsf{M}$ since M is an ideal, and $a_2 b_2 \in \mathsf{M}$ since $(rs) \subseteq \mathsf{M}$. This shows that $\mathsf{R} \subseteq \mathsf{M}$, contradicting the fact that M is a maximal ideal. Thus we conclude that either $r \in \mathsf{M}$ or $s \in \mathsf{M}$, and so M is prime. ■

We shall not really see the importance of prime and maximal ideals until we study prime and irreducible elements of a ring in Section 4.2.9. (Important properties of prime and maximal ideals are also given in Theorems 4.2.37 and 4.3.9, respectively.) For now, let us give some elementary examples of prime and maximal ideals.

4.2.22 Examples (Prime and maximal ideals) For the first two examples we assume the reader knows what a prime number is, and knows the property that every integer can be written as a product of primes. We shall prove these facts in Section 4.2.10.

1. Let p be a prime integer. We claim that $(p) = \{jp \mid j \in \mathbb{Z}\}$ is a prime ideal. By Example 4.2.15–4 we know that (p) is an ideal. To see that it is prime, let $\mathsf{A}, \mathsf{B} \subseteq (p)$ and let $a \in \mathsf{A}$ and $b \in \mathsf{B}$. Then, if $\mathsf{AB} \subseteq (p)$ we must have $ab \in (p)$, which means that ab is a multiple of p. Since ab can be factored as a product of prime numbers, and since p is a factor in ab, it must be a factor for either a or b. Thus either $a \in (p)$ or $b \in (p)$, which gives our claim by Proposition 4.2.18.
2. In the ring $\mathbb{Z}$ we consider the ideals
$$(3) = \{3j \mid j \in \mathbb{Z}\}, \quad (4) = \{4j \mid j \in \mathbb{Z}\}.$$
we claim that (3) is maximal but that (4) is not. To see that (3) is maximal, let $(3) \subseteq \mathsf{J}$ for an ideal J. If $\mathsf{J} \neq (3)$ then there exists $k \in \mathsf{J}$ that is not a multiple of 3. Since $\{jk \mid j \in \mathbb{Z}\} \subseteq \mathsf{J}$ by virtue of J being an ideal, this means that every integer not containing a 3 in its prime factorisation lies in J. But since $(3) \subseteq \mathsf{J}$, this means that $\mathsf{J} = \mathbb{Z}$. To see that (4) is not a maximal ideal, Note that $(4) \subseteq (2) = \{2j \mid j \in \mathbb{Z}\}$, but that $(2) \neq \mathbb{Z}$.
3. Next we consider the ring $\mathbb{R}[[\xi]]$ of $\mathbb{R}$-formal power series. We claim that the ideal
$$(\xi) = \{\xi \cdot A \mid A \in \mathbb{R}[[\xi]]\},$$
where ξ denotes the indeterminate in $\mathbb{R}[[\xi]]$ (see Definition 3.7.2), is maximal. Suppose that I is an ideal in $\mathbb{R}[[\xi]]$ containing (ξ). Define $A_0 = (a_j)_{j \in \mathbb{Z}_{\geq 0}} \in \mathbb{R}[[\xi]]$ by
$$a_j = \begin{cases} 1, & j = 0, \\ 0, & j \neq 0. \end{cases}$$
Then we have two cases.
   (a) $A_0 \in \mathsf{I}$: In this case, since for any $A \in \mathbb{R}[[\xi]]$ we have $A = A \cdot A_0$, we must have $\mathsf{I} = \mathbb{R}[[\xi]]$.
   (b) $A_0 \notin \mathsf{I}$: In this case every member A of I can be written as
$$A = \sum_{j=1}^{\infty} a_j \xi^j = \xi \sum_{j=1}^{\infty} a_j \xi^{j-1}.$$
   In particular, $A \in (\xi)$.

   This shows that either $\mathsf{I} = \mathbb{R}[[\xi]]$ or that $\mathsf{I} \subseteq (\xi)$. •

As we shall see during the course of Section 4.2.10, for the ring $\mathbb{Z}$, an ideal is prime if and only if it is maximal. We shall also see that there is a general relationship that holds between prime ideals and ring elements that we shall call “prime.” There is also a corresponding concept for ring elements, “irreducibility,” that corresponds to maximal ideals.

4.2.4 Ring homomorphisms

As expected, with rings one can talk about the maps between rings that preserve their structure.

4.2.23 Definition (Ring homomorphism, epimorphism, monomorphism, and isomorphism) For rings R and S, a map $\phi\colon \mathsf{R} \to \mathsf{S}$ is a:

(i) ***ring homomorphism***, or simply a ***homomorphism***, if $\phi(r_1 + r_2) = \phi(r_1) + \phi(r_2)$ and $\phi(r_1 \cdot r_2) = \phi(r_1) \cdot \phi(r_2)$ for all $r_1, r_2 \in \mathsf{G}$;

(ii) ***epimorphism*** if it is a surjective homomorphism;

(iii) ***monomorphism*** if it is an injective homomorphism;

(iv) ***isomorphism*** if it a bijective homomorphism.

If there exists an isomorphism from R to S then R and S are ***isomorphic***. •

The following result follows directly from Proposition 4.1.24.

4.2.24 Proposition (Properties of ring homomorphisms) *If* R *and* S *are rings and if* $\phi\colon \mathsf{R} \to \mathsf{S}$ *is a homomorphism, then*

(i) $\phi(0_{\mathsf{R}}) = 0_{\mathsf{S}}$*,*

(ii) $\phi(-\mathrm{r}) = -\phi(\mathrm{r})$ *for every* $\mathrm{r} \in \mathsf{R}$*, and*

(iii) if R *and* S *are additionally unit rings, then* $\phi(1_{\mathsf{R}}) = 1_{\mathsf{S}}$*.*

As with groups, one can certain subsets associated with a homomorphism of rings.

4.2.25 Definition (Image and kernel of ring homomorphism) Let R and S be rings and let $\phi\colon \mathsf{R} \to \mathsf{S}$ be a ring homomorphism.

(i) The ***image*** of ϕ is $\operatorname{image}(\phi) = \{\phi(r) \mid r \in \mathsf{R}\}$.

(ii) The ***kernel*** of ϕ is $\ker(\phi) = \{r \in \mathsf{R} \mid \phi(r) = 0_{\mathsf{S}}\}$. •

4.2.26 Proposition (Image and kernel are subring and ideal, respectively) *If* R *and* S *are rings and if* $\phi\colon \mathsf{R} \to \mathsf{S}$ *is a ring homomorphism, then*

(i) $\operatorname{image}(\phi)$ *is a subring of* S *and*

(ii) $\ker(\phi)$ *is a two-sided ideal of* R*.*

Proof It is easy to see that both $\operatorname{image}(\phi)$ and $\ker(\phi)$ are subrings. The only possibly non-obvious thing to prove is that $\ker(\phi)$ is a two-sided ideal. To see this, let $r_1 \in \mathsf{R}$ and $r_2 \in \ker(\phi)$. Then $\phi(r_1 r_2) = \phi(r_1)\phi(r_2) = 0_{\mathsf{S}}$ since $r_2 \in \ker(\phi)$ and by Proposition 4.2.5. Thus $r_1 r_2 \in \ker(\phi)$. Similarly one shows that $r_1 r_2 \in \ker(\phi)$ if $r_1 \in \ker(\phi)$ and $r_2 \in \mathsf{R}$. ■

4.2.27 Examples (Ring homomorphisms)

1. The natural maps $i_{\mathbb{Z}}\colon \mathbb{Z} \to \mathbb{Q}$ and $i_{\mathbb{Q}}\colon \mathbb{Q} \to \mathbb{R}$ are ring homomorphisms. Each of these ring homomorphisms is injective, and so has kernel equal to zero (cf. Exercise 4.2.7).
2. The map $\phi_k\colon \mathbb{Z} \to \mathbb{Z}_k$ defined by $j \mapsto j + k\mathbb{Z}$ is easily verified to be a ring homomorphism. One can readily see that $\ker(\phi_k) = k\mathbb{Z}$. •

4.2.5 Characteristic

In this section we consider briefly the notion of the characteristic of a ring. In the text we shall only be concerned with rings that have characteristic zero. Our principle intent, therefore, is to consider how these rings are special among the class of rings with general characteristic.

4.2.28 Definition (Characteristic) Let R be a ring and define

$$C(\mathsf{R}) = \{k \in \mathbb{Z}_{>0} \mid kr = 0_{\mathsf{R}} \text{ for all } r \in \mathsf{R}\}.$$

The ***characteristic*** of R is

(i) zero if $C(\mathsf{R}) = \varnothing$ and is

(ii) $\inf C(\mathsf{R})$ if $C(\mathsf{R}) \neq \varnothing$. •

Let us give some examples of the characteristic for some of the rings we have considered.

4.2.29 Examples (Characteristic) We leave to the reader the verification of the statements we make here.

1. The characteristic of the rings $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ is zero.
2. The characteristic of $\mathbb{Z}_k$ is k. •

The next result shows that the two proceeding examples are, in some sense, the building blocks for general rings with given characteristic.

4.2.30 Proposition (Property of rings with given characteristic) *If R is a commutative unit ring then the following statements hold:*

(i) if R has characteristic zero then there exists a subring S of R that is isomorphic to $\mathbb{Z}$;

(ii) if R has characteristic $\mathrm{k} \in \mathbb{Z}_{>0}$ then there exists a subring S of R that is isomorphic to $\mathbb{Z}_{\mathrm{k}}$.

Proof Define $\phi\colon \mathbb{Z} \to \mathsf{R}$ by $\phi(j) = j1_{\mathsf{R}}$. It is a trivial matter to check that ϕ is a homomorphism. Therefore, $\operatorname{image}(\phi)$ is a subring by Proposition 4.2.26.

First suppose that R has characteristic zero. We claim that $\phi\colon \mathbb{Z} \to \operatorname{image}(\phi)$ is a monomorphism. Suppose that $\phi(j_1) = \phi(j_2)$. Then $\phi(j_1 - j_2) = (j_1 - j_2)1_{\mathsf{R}} = 0_{\mathsf{R}}$. Since R has characteristic zero this implies that $j_1 - j_2 = 0$, and so ϕ is a monomorphism. Therefore, the map $j \mapsto \phi(j)$ is an isomorphism from $\mathbb{Z}$ to the subring $\operatorname{image}(\phi)$ of R.

Now suppose that R has characteristic $k \in \mathbb{Z}_{>0}$ and consider the ring homomorphism $\phi_k\colon \mathbb{Z} \to \mathbb{Z}_k$ defined by $\phi_k(j) = j + k\mathbb{Z}$. We claim that the map $\psi\colon \mathbb{Z}_k \to \mathsf{R}$ defined by $\psi(j + k\mathbb{Z}) = \phi(j)$ is an isomorphism onto image(ϕ). First we should show that ψ is well defined. To see this note that, for any $l \in \mathbb{Z}$,

$$\psi(j + lk + k\mathbb{Z}) = \phi(j + lk) = (j + lk)1_{\mathsf{R}} = j1_{\mathsf{R}} + lk1_{\mathsf{R}} = j1_{\mathsf{R}} + l(k1_{\mathsf{R}}) = j1_{\mathsf{R}} = \psi(j + k\mathbb{Z}).$$

Next we show that ψ is injective. Suppose that $\psi(j_1+k\mathbb{Z}) = \psi(j_2+k\mathbb{Z})$. Then $j_11_{\mathsf{R}} = j_21_{\mathsf{R}}$ which directly implies that $(j_1 - j_2)1_{\mathsf{R}} = 0_{\mathsf{R}}$. Therefore $j_1 = j_2 + lk$ for some $l \in \mathbb{Z}$. Thus $j_1 + k\mathbb{Z} = j_2 + k\mathbb{Z}$. Thus ψ is injective. Now we show that image(ψ) = image(ϕ). Let $j1_{\mathsf{R}} \in$ image(ϕ) and then note that $\psi(j + k\mathbb{Z}) = j1_{\mathsf{R}}$. Thus image($\phi$) ⊆ image($\psi$). Since obviously image(ψ) ⊆ image(ϕ) we have image(ψ) = image(ϕ) as desired. Therefore the map $j+k\mathbb{Z} \mapsto \phi(j)$ is an isomorphism from $\mathbb{Z}_k$ to the subring image(ϕ) as desired. ■

4.2.6 Integral domains

We now begin a brief excursion into rings with special structure. As we have seen, elements of a group do not necessarily possess multiplicative inverses. The following definition gives a special collection of elements that are not units.

4.2.31 Definition (Zerodivisor) An element r in a ring R is a ***zerodivisor*** if either there exists $s_1 \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ such that $rs_1 = 0_{\mathsf{R}}$ or there exists $s_2 \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ such that $s_2r = 0_{\mathsf{R}}$. An element $r \in \mathsf{R}$ that is not a zerodivisor is a ***nonzerodivisor***. •

We have already seen certain elements of rings that are *not* zerodivisors.

4.2.32 Proposition (Units are not zerodivisors) *If* R *is a unit ring and if* r *is a unit, then* r *is not a zerodivisor.*

Proof Let r be a unit and suppose that $rs_1 = 0_{\mathsf{R}}$ for $s_1 \in \mathsf{R}$. Then $r^{-1}(rs_1) = r^{-1}0_{\mathsf{R}} = 0_{\mathsf{R}}$ which gives $s_1 = 0_{\mathsf{R}}$. Similarly we can show that if $s_2r = 0_{\mathsf{R}}$ then $s_2 = 0_{\mathsf{R}}$, which shows that r is not a zerodivisor. ■

The following properties of elements that are not zerodivisors is useful.

4.2.33 Proposition (Cancellation law for elements that are not zerodivisors) *Let* R *be a ring and suppose that* $r \in \mathsf{R}$ *is not a zerodivisor. Then the following statements hold:*

(i) if $r \cdot s_1 = r \cdot s_2$ *then* $s_1 = s_2$*;*

(ii) if $s_1 \cdot r = s_2 \cdot r$ *then* $s_1 = s_2$*.*

Proof We shall only prove the first assertion, since the second follows in a similar manner. If $r \cdot s_1 = r \cdot s_2$ then $r \cdot (s_1 - s_2) = 0_{\mathsf{R}}$. Therefore, since r is not a zerodivisor, it follows that $s_1 - s_2 = 0_{\mathsf{R}}$. ■

Let us consider the zerodivisors in some of the rings we have encountered.

4.2.34 Examples (Zerodivisors)

1. The only zerodivisor in the ring $\mathbb{Z}$ is 0.
2. The only zerodivisor in the ring $\mathbb{Q}$ is 0.
3. The only zerodivisor in the ring $\mathbb{R}$ is 0.
4. Let us construct the zerodivisors in $\mathbb{Z}_k$. Again we rely on some ideas from Section 4.2.7. If $j + k\mathbb{Z}$, $j \in \{0, 1, \dots, k\}$, is a zerodivisor, then there exists $j' \in \{0, 1, \dots, k-1\}$ such that $jj' + k\mathbb{Z} = 0 + k\mathbb{Z}$. Thus jj' must be a multiple of k.

   We claim that there are no nonzero zerodivisors in $\mathbb{Z}_k$ if and only if k is prime. To see this, first let k be prime and suppose that $jj' + k\mathbb{Z} = 0 + k\mathbb{Z}$ for $j, j' \in \mathbb{Z}$. Then $jj' = lk$ for some $l \in \mathbb{Z}$. Thus k appears in the prime factorisation of jj', which means that k must appear in the prime factorisation of at least one of j or j'. But this means that either $j + k\mathbb{Z} = 0 + k\mathbb{Z}$ or that $j' + k\mathbb{Z} = 0 + k\mathbb{Z}$. Thus there are no nonzero zerodivisors in $\mathbb{Z}_k$ when k is prime.

   If k is not prime then there exists $j, j' \in \{1, \dots, k-1\}$ such that $jj' = k$. Thus $jj' + k\mathbb{Z} = 0 + k\mathbb{Z}$, and so $j + k\mathbb{Z}$ is a nonzero zerodivisor.
5. Let S be a nonempty set and let R be a ring, and let R^S be the ring of R-valued functions on S. We claim that $f \in \mathsf{R}^S$ is a zerodivisor if and only if there exists $x_0 \in S$ such that $f(x_0)$ is a zerodivisor.

   To see this, suppose first that f is a zerodivisor. Then there exists a nonzero function g such that either $(fg)(x) = 0_{\mathsf{R}}$ for all $x \in S$ or $(gf)(x) = 0_{\mathsf{R}}$ for all $x \in S$. Since g is nonzero, this means that there exists $x_0 \in S$ such that $g(x_0) \neq 0_{\mathsf{R}}$. Therefore, either $f(x_0)g(x_0) = 0_{\mathsf{R}}$ or $g(x_0)f(x_0) = 0_{\mathsf{R}}$. Thus $f(x_0)$ is a zerodivisor.

   Conversely, suppose that there exists $x_0 \in S$ such that $f(x_0)$ is a zerodivisor. Then either there exists $s_1 \in \mathsf{R}$ such that $f(x_0)s_1 = 0_{\mathsf{R}}$ or there exists $s_2 \in \mathsf{R}$ such that $s_2 f(x_0) = 0_{\mathsf{R}}$. Then define $g_1, g_2 \colon S \to \mathsf{R}$ by

   $$g_1(x) = \begin{cases} s_1, & x = x_0, \\ 0_{\mathsf{R}}, & x \neq x_0, \end{cases} \qquad g_2(x) = \begin{cases} s_2, & x = x_0, \\ 0_{\mathsf{R}}, & x \neq x_0. \end{cases}$$

   Then it holds that either $(fg_1)(x) = 0_{\mathsf{R}}$ for all $x \in S$ or that $(g_2 f)(x) = 0_{\mathsf{R}}$ for all $x \in S$. Thus f is a zerodivisor in R^S. •

Based on the notion of zerodivisors, we single out a special class of rings as follows.

4.2.35 Definition (Integral domain) An ***integral domain*** is a commutative unit ring for which the only zerodivisor is 0_{R}. •

We indicate which of our examples of rings are integral domains.

4.2.36 Examples (Integral domains)

1. $\mathbb{Z}$ is an integral domain (indeed, “integral” in “integral domain” comes from the fact that the set of integers is an integral domain).

2. $\mathbb{Q}$ is an integral domain.
3. $\mathbb{R}$ is an integral domain.
4. $\mathbb{Z}_k$ is an integral domain if and only if k is a prime number.
5. If S is a set for which $\mathrm{card}(S) = 1$, then R^S is an integral domain if and only if R is an integral domain. If $\mathrm{card}(S) > 1$, then one can easily verify that R^S is not an integral domain (why?).
6. The ring $\mathbb{R}[[\xi]]$ of $\mathbb{R}$-formal power series is an integral domain. To see this we shall show that if $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ are nonzero $\mathbb{R}$-formal power series, then $C = A \cdot B$ is also nonzero. Denote

$$n_A = \inf\{j \in \mathbb{Z}_{\geq 0} \mid a_j \neq 0\}, \quad n_B = \inf\{j \in \mathbb{Z}_{\geq 0} \mid b_j \neq 0\}$$

and write $C = (c_j)_{j\in\mathbb{Z}_{\geq 0}}$. Then, using the definition of multiplication of $\mathbb{R}$-formal power series, we may compute $c_{n_A+n_B} = a_{n_A} b_{n_B}$. In particular it follows that $c_{n_A+n_B} \neq 0$ so that $A \cdot B$ is nonzero. Thus $\mathbb{R}[[\xi]]$ is indeed an integral domain. •

The following result relates integral domains to prime ideals, hopefully to the better understanding of both concepts.

4.2.37 Theorem (Quotients by prime ideals are integral domains, and vice versa) *If R is a commutative unit ring and I is an ideal of R, then the following two statements are equivalent:*

(i) I is a prime ideal;

(ii) R/I is an integral domain.

Proof Suppose that I is prime and that $(r + \mathsf{I})(s + \mathsf{I}) = 0_{\mathsf{R}} + \mathsf{I}$. Then $rs \in \mathsf{I}$, and so by Proposition 4.2.18, either $r \in \mathsf{I}$ or $s \in \mathsf{I}$. Thus either $r + \mathsf{I} = 0_{\mathsf{R}} + \mathsf{I}$ or $s + \mathsf{I} = 0_{\mathsf{R}} + \mathsf{I}$. By Exercise 4.2.11 this implies that R/I is an integral domain.

Now suppose that R/I is an integral domain and let $r, s \in \mathsf{R}$ have the property that $rs \in \mathsf{I}$. Then $(r + \mathsf{I})(s + \mathsf{I}) = 0_{\mathsf{R}} + \mathsf{I}$, and so either $r + \mathsf{I} = 0_{\mathsf{R}} + \mathsf{I}$ or $s + \mathsf{I} = 0_{\mathsf{R}} + \mathsf{I}$ by Exercise 4.2.11. Therefore, either $r \in \mathsf{I}$ or $s \in \mathsf{I}$, and so I is prime by Proposition 4.2.18. ■

4.2.7 Euclidean rings and domains

In this section we consider a special class of rings within which the notion of long division, as learned in school for integers and for polynomials, makes sense. Let us recall how long division for integers works. If $j, k \in \mathbb{Z}$ with $k \neq 0$, then long division tells us that there exists $q, r \in \mathbb{Z}$ with $|r| < |k|$ and such that $j = qk + r$. The key ingredients are that j is a sum of a multiple qk of k and a remainder r, and that the remainder is less in magnitude than k.

We shall prove this form of long division shortly, but for now let us give the general form of a ring within which long division will turn out to be possible.

4.2.38 Definition (Euclidean[1] ring and domain) A ***Euclidean ring*** is a pair (R, δ) where R is a commutative ring and where $\delta\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ has the following properties:

(i) if $a, b \in \mathsf{R}$ and if $ab \neq 0_{\mathsf{R}}$, then $\delta(ab) \geq \delta(a)$;

(ii) if $a, b \in \mathsf{R}$ with $b \neq 0_{\mathsf{R}}$, then there exists $q, r \in \mathsf{R}$ such that

(a) $a = qb + r$ and such that

(b) $\delta(r) < \delta(b)$.

A ***Euclidean domain*** is a Euclidean ring that is also an integral domain. •

As is the case with many of the structures we talk about that are defined by ordered pairs, we shall often refer to a Euclidean ring (R, δ) simply as R, understanding that δ is in the background. The function δ is sometimes called the ***degree function***.

Let us give a somewhat uninteresting example of a Euclidean ring, just to get some idea of how all the pieces tie together.

4.2.39 Example ($\mathbb{R}$ is a (boring) Euclidean domain) On the integral domain $\mathbb{R}$ define $\delta\colon \mathbb{R} \to \mathbb{Z}_{\geq 0}$ by

$$\delta(x) = \begin{cases} 1, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Then clearly $\delta(xy) \geq \delta(x)$ (indeed, equality holds) for all $x, y \in \mathbb{R}^*$. Also, if $x, y \in \mathbb{R}$ with $y \neq 0$ then we can write $x = qy + r$ where $q = xy^{-1}$ and $r = 0$, and we note that $\delta(r) = 0 < 1 = \delta(y)$. Thus $(\mathbb{R}, \delta)$ is a Euclidean domain. After learning about fields in Section 4.3, the reader will readily see that this example can be extended to show that any field is a Euclidean domain for an appropriately defined map δ. •

The following language is commonly attached to the second part of the properties of the map δ in a Euclidean ring.

4.2.40 Notation (Division Algorithm) Let (R, δ) be a Euclidean ring. The fact that, given $a, b \in \mathsf{R}$ with $b \neq 0_{\mathsf{R}}$, we can write $a = qb + r$ with $\delta(r) < \delta(b)$ is often referenced by saying that the ring R possesses a ***division algorithm***. The element q is called the ***quotient*** and the element r is called the ***remainder***. •

Let us prove some basic facts about Euclidean rings before we prove that the integers form a Euclidean ring.

[1] Named for the Greek mathematician Euclid, whose extraordinarily influential treatise *The Elements* dates to around 300BC. Not much is known about Euclid, and indeed there is legitimate debate over whether he existed, or whether his work is really the work of a group of mathematicians. Whether or not he existed, the importation of his body of work, along with other things Greek, to the west by the Islamic culture played a significant rôle in the emergence of western civilisation from the Dark Ages.

4.2.41 Proposition (Properties of Euclidean rings) *If* (R,δ) *is a Euclidean ring, then the following statements hold:*

(i) $\delta(0_{\mathsf{R}}) < \delta(1_{\mathsf{R}})$*;*

(ii) if $\mathrm{a} \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ *then* $\delta(\mathrm{a}) \geq \delta(1_{\mathsf{R}})$*;*

(iii) $\delta(\mathrm{a}) = \delta(0_{\mathsf{R}})$ *if and only if* $\mathrm{a} = 0_{\mathsf{R}}$.

If R *is additionally an integral domain then the following statements also hold:*

(iv) for $\mathrm{a} \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$*,* $\delta(\mathrm{ab}) = \delta(\mathrm{a})$ *if and only if* b *is a unit;*

(v) $\mathrm{a} \in \mathsf{R}$ *is a unit if and only if* $\delta(\mathrm{a}) = \delta(1_{\mathsf{R}})$.

Proof (ii) Let $a \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ so that $\delta(a) = \delta(1_{\mathsf{R}}a) \geq \delta(1_{\mathsf{R}}) = \delta(1_{\mathsf{R}})$.

(i) Let $a \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ and use the definition of Euclidean ring to assert that there exists $q, r \in \mathsf{R}$ such that $a = q1_{\mathsf{R}} + r$ where $\delta(r) < \delta(1_{\mathsf{R}}) = \delta(1_{\mathsf{R}})$. By part (ii) we must therefore have $r = 0_{\mathsf{R}}$, and consequently also we have $\delta(0_{\mathsf{R}}) < \delta(1_{\mathsf{R}})$.

(iii) This was proved en route to proving part (i).

(iv) Let a and b have the properties that $a \neq 0_{\mathsf{R}}$ and $\delta(ab) = \delta(a)$. First we claim that $b \neq 0_{\mathsf{R}}$. Indeed, if $b = 0_{\mathsf{R}}$ then we have $\delta(a) = \delta(ab) = \delta(0_{\mathsf{R}})$, meaning that $ab = 0_{\mathsf{R}}$. Since R is an integral domain we must have either $a = 0_{\mathsf{R}}$ or $b = 0_{\mathsf{R}}$, and so we arrive at a contradiction. Now, using the definition of a Euclidean domain there exists $q, r \in \mathsf{R}$ such that $a = q(ab) + r$ and $\delta(r) < \delta(ab) = \delta(a)$. Rearranging things gives $r = a(1_{\mathsf{R}} - qb)$ so that $\delta(r) = \delta(a(1_{\mathsf{R}} - qb)) \geq \delta(a)$, unless $a(1_{\mathsf{R}} - qb) = 0_{\mathsf{R}}$. However, since $\delta(r) < \delta(a)$ we must indeed have $a(1_{\mathsf{R}} - qb) = 0_{\mathsf{R}}$. Thus, since R is an integral domain, $1_{\mathsf{R}} = qb$ showing that b is a unit.

Now suppose that b is a unit and that $a \neq 0_{\mathsf{R}}$. Then $\delta(ab) \geq \delta(a)$ and $\delta(a) = \delta(a1_{\mathsf{R}}) = \delta(abb^{-1}) \geq \delta(ab)$. Therefore, $\delta(a) = \delta(ab)$.

(v) From part (iv) we know that a is a unit if and only if $\delta(a) = \delta(a1_{\mathsf{R}}) = \delta(1_{\mathsf{R}}) = \delta(1_{\mathsf{R}})$. ■

Note that a consequence of the previous result is that the set image(δ) is ordered as $\{\delta_0, \delta_1, \delta_2, \dots\}$, where $\delta_0 = \delta(0_{\mathsf{R}})$, $\delta_1 = \delta(1_{\mathsf{R}})$, and where δ_k, $k \geq 2$, is the image of a nonzero nonunit.

We have made no assertions concerning the uniqueness of the quotient and remainder that are produced by the Division Algorithm. Indeed, generally the quotient and remainder are *not* unique (see Exercise 4.2.14). However, in some Euclidean domains, or more generally in some subsets of Euclidean domains, a unique quotient and remainder can be guaranteed. It will save us having to prove some things twice (once for the Euclidean ring $\mathbb{Z}$ of integers of Theorem 4.2.45, and once for the Euclidean ring $\mathsf{F}[\xi]$ of polynomials over a field F of Corollary 4.4.14) if we introduce some terminology to cover the matter of uniqueness of the quotient and remainder, as well as some other ring constructions that we will encounter.

4.2.42 Definition (δ-closed subset, δ-positive subset) Let (R, δ) be a Euclidean domain.

(i) A subset $A \subseteq \mathsf{R}$ is ***trivial*** if $C = \{0_{\mathsf{R}}\}$, and is ***nontrivial*** otherwise.

(ii) A nonempty subset $C \subseteq \mathsf{R}$ is ***δ-closed*** if, for each $a, b \in C$ with $b \neq 0_{\mathsf{R}}$, there exists $q, r \in C$ such that $a = qb + r$ and such that $\delta(r) < \delta(b)$.

(iii) A subset $C \subseteq \mathsf{R}$ ***admits a unique Division Algorithm*** if, for each $a, b \in C$ with $b \neq 0_\mathsf{R}$, there exists unique $q, r \in C$ such that $a = qb + r$ and such that $\delta(r) < \delta(b)$.

(iv) A nonempty subset $P \subseteq \mathsf{R}$ is ***δ-positive*** if, for each $a, b \in P$, we have $\delta(a - b) \leq \max\{\delta(a), \delta(b)\}$. •

Our primary interest will be in nontrivial, δ-closed, and δ-positive subsemirings. The following elementary properties of δ-closed subsets will be useful.

4.2.43 Proposition (δ-closed sets contain 0_R and 1_R) *If* (R, δ) *is a Euclidean domain and if* $\mathsf{C} \subseteq \mathsf{R}$ *is a nontrivial δ-closed subset, then* $0_\mathsf{R}, 1_\mathsf{R} \in \mathsf{C}$.

Proof Let $b \in C - \{0_\mathsf{R}\}$. Since C is δ-closed there exists $q, r \in C$ such that $b = qb + r$ with $\delta(r) < \delta(b)$. We claim that this implies that $q = 1_\mathsf{R}$ and $r = 0_\mathsf{R}$. Suppose that $q \neq 1_\mathsf{R}$. Then

$$\delta(b) \leq \delta((1_\mathsf{R} - q)b) = \delta(r) < \delta(b)$$

which is a contradiction. Thus $q = 1_\mathsf{R}$, and it then follows that $r = 0_\mathsf{R}$. ■

The next result is the first of our results that indicates why δ-closed and δ-positive subsets are important; others are Theorems 4.2.48 and 4.2.84.

4.2.44 Proposition (Uniqueness of quotient and remainder in δ-closed and δ-positive subsemirings) *If* (R, δ) *is a Euclidean domain and if* S *is a nontrivial, δ-closed, and δ-positive subsemiring of* R*, then* S *admits a unique Division Algorithm.*

Proof Suppose that $a = q_1 b + r_1 = q_2 b + r_2$ for $q_1, q_2, r_1, r_2 \in \mathsf{S}$ with $\delta(r_1), \delta(r_2) < \delta(b)$. Then $(q_1 - q_2)b = r_2 - r_1$, and so

$$\delta((q_1 - q_2)b) = \delta(r_1 - r_2) \leq \max\{\delta(r_1), \delta(r_2)\} < \delta(b),$$

using δ-closedness of S. This implies that $(q_1 - q_2)b = 0_\mathsf{R}$. Since $b \neq 0_\mathsf{R}$ this implies that $q_1 - q_2 = 0_\mathsf{R}$ and so $q_1 = q_2$. We then immediately have $r_1 = r_2$. ■

Now let us turn to the first of our primary examples of a Euclidean domain.

4.2.45 Theorem ($\mathbb{Z}$ is a Euclidean domain) *The pair* $(\mathbb{Z}, \delta)$ *is a Euclidean domain if we define* $\delta\colon \mathbb{Z} \to \mathbb{Z}_{\geq 0}$ *by* $\delta(\mathrm{j}) = |\mathrm{j}|$. *That is to say, if* $\mathrm{j}, \mathrm{k} \in \mathbb{Z}$ *with* $\mathrm{k} \neq 0$*, then there exists* $\mathrm{q}, \mathrm{r} \in \mathbb{Z}$ *such that* $\mathrm{j} = \mathrm{qk} + \mathrm{r}$ *and such that* $|\mathrm{r}| < |\mathrm{k}|$. *Moreover,* $\mathbb{Z}_{\geq 0}$ *is a nontrivial, δ-closed, and δ-positive subsemiring of* $\mathbb{Z}$*; therefore, if* $\mathrm{j} \in \mathbb{Z}_{\geq 0}$ *and* $\mathrm{k} \in \mathbb{Z}_{>0}$*, then there exists unique* $\mathrm{q}, \mathrm{r} \in \mathbb{Z}_{\geq 0}$ *such that* $\mathrm{j} = \mathrm{qk} + \mathrm{r}$.

Proof It is clear that if $j, k \in \mathbb{Z} \setminus 0$ then $|jk| \geq |j|$ since $|k| \geq 1$ for when $k \neq 0$.

Let us suppose that $j \geq 0$ and $k > 0$. Define

$$S(j, k) = \{j - mk \mid m \in \mathbb{Z},\ j - mk \geq 0\}.$$

We claim that $S(j, k)$ is not empty. Indeed, note that $j + jk \geq 0$ so that $-j \in S(j, k)$. Thus $S(j, k)$ is a nonempty subset of the well ordered set $\mathbb{Z}_{\geq 0}$, and so possesses a least element, which we denote by r. Suppose that $q \in \mathbb{Z}$ is the number for which $r = j - qk$. Clearly we have $j = qk + r$. Now we show that $r \in \{0, 1, \dots, k - 1\}$. Suppose that $r \geq k$. Then $r - k \geq 0$ and, since $r - k = j - (q + 1)k$ we have $r - k \in S(j, k)$. Since $b > 0$ we also have $r - b < r$ which contradicts the fact that r is the least element of $S(j,k)$. This proves the theorem for $j, k > 0$.

Suppose that $j > 0$ and $k < 0$ so that $j, -k > 0$. Then there exists $\tilde{q}, \tilde{r} \in \mathbb{Z}$ such that $j = \tilde{q}(-k) + \tilde{r}$ with $\tilde{r} \in \{0, 1, \dots, k-1\}$. Therefore, the theorem follows by taking $q = -\tilde{q}$ and $r = \tilde{r}$. The other cases where $j < 0$ and $k > 0$, and $j, k < 0$ follow similarly from the case where $j, k > 0$.

The final assertion of the theorem follows since (1) in the first part of the proof, the $q, r \in \mathbb{Z}_{\geq 0}$ we constructed for $j \in \mathbb{Z}_{\geq 0}$ and $k \in \mathbb{Z}_{\geq 0}$ were members of $\mathbb{Z}_{\geq 0}$, and (2) if $j, k \in \mathbb{Z}_{\geq 0}$ then $|j - k| \leq \max\{|j|, |k|\}$, as can easily be verified directly. ■

Note that $\mathbb{Z}$ is not a δ-positive subset of itself (cf. Exercise 4.2.14). We shall encounter in Corollary 4.4.14 a Euclidean ring which *is* a δ-positive subset of itself, and therefore which has the property that the quotient and remainder of the Division Algorithm are always unique.

As a corollary to Proposition 4.2.41 we have the following more or less obvious properties of the integers.

4.2.46 Corollary (Properties of $\mathbb{Z}$) *For $\mathbb{Z}$, the following statements hold:*

(i) $0 < 1$;

(ii) *if* $j \in \mathbb{Z} \setminus \{0\}$ *then* $|j| \geq 1$;

(iii) $|j| = 0$ *if and only if* $|j| = 0$;

(iv) *if* $j \in \mathbb{Z} \setminus \{0\}$ *then* $|jk| = |j|$ *if and only if* $k \in \{-1, 1\}$;

(v) *the only units in $\mathbb{Z}$ are* 1 *and* -1.

Of course, if the only Euclidean ring we encounter were to be the ring of integers, it would not be worth introducing the concept. However, we shall see in Section 4.4 that the ring of polynomials over a field is also a Euclidean domain, as we shall see in Corollary 4.4.14.

By Proposition 4.2.44 follows the uniqueness of the Division Algorithm in δ-closed and δ-positive subsemirings. The converse of this assertion is also true in some case.

4.2.47 Proposition (δ-positivity is sometimes implied by uniqueness of the Division Algorithm) *Let (R, δ) be a Euclidean domain and let $\mathsf{S} \subseteq \mathsf{R}$ be a nontrivial, δ-closed subsemiring with the following properties:*

(i) S *generates* R *as a ring;*

(ii) S *admits a unique Division Algorithm.*

Then S is δ-positive.

Proof Note that since S is a subsemiring, S generates R as a ring if and only if, for every $r \in \mathsf{R}$, it holds that either $r \in \mathsf{S}$ or $-r \in \mathsf{S}$. Suppose that S is not δ-positive so that $\delta(a - b) > \max\{\delta(a), \delta(b)\}$ for some $a, b \in \mathsf{S}$. Suppose that $b - a \in \mathsf{S}$. Then

$$b = 0_{\mathsf{R}} \cdot (b - a) + b, \quad \delta(b) < \delta(b - a),$$
$$b = 1_{\mathsf{R}} \cdot (b - a) + a, \quad \delta(a) < \delta(b - a),$$

which shows that S does not admit a unique Division Algorithm. An entirely similar argument gives the same conclusion when $a - b \in \mathsf{S}$. ■

Let us see how the structure of a Euclidean domain lends itself to the a generalisation of the notion of writing integers with respect to a given base (base 10 being the one we use in everyday life).

4.2.48 Theorem (Base expansion in δ-closed, δ-positive subsemirings) *Let* (R, δ) *be a Euclidean domain, let* $\mathsf{S} \subseteq \mathsf{R}$ *be a nontrivial, δ-closed, and δ-positive subsemiring, and let* $\mathrm{b} \in \mathsf{S}$ *be a nonzero nonunit. Then, given* $\mathrm{a} \in \mathsf{S} \setminus \{0_{\mathsf{R}}\}$, *there exists a unique* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$ *and unique* $\mathrm{r}_0, \mathrm{r}_1, \dots, \mathrm{r}_\mathrm{k} \in \mathsf{S}$ *such that*

(i) $\mathrm{r}_\mathrm{k} \neq 0_{\mathsf{R}}$,

(ii) $\delta(\mathrm{r}_0), \delta(\mathrm{r}_1), \dots, \delta(\mathrm{r}_\mathrm{k}) < \delta(\mathrm{b})$ *and*

(iii) $\mathrm{a} = \mathrm{r}_0 + \mathrm{r}_1\mathrm{b} + \mathrm{r}_2\mathrm{b}^2 + \cdots + \mathrm{r}_\mathrm{k}\mathrm{b}^\mathrm{k}$.

Proof We prove the result by induction on $\delta(a)$. By Proposition 4.2.43 we have

$$\inf\{\delta(a) \mid a \in \mathsf{S}\} = 0.$$

Since we do not consider the case $\delta(a) = \delta(0_{\mathsf{R}})$, first consider $a \in \mathsf{R}$ such that $\delta(a) = \delta(1_{\mathsf{R}})$. Then a is a unit by Proposition 4.2.41. Thus, since b is a nonzero nonunit, we have $\delta(a) < \delta(b)$, and the existence part of the result follows by taking $k = 0$ and $r_0 = a$. Now suppose that the result holds for all $a \in \mathsf{S}$ such that $\delta(a) \in \{\delta(1_{\mathsf{R}}, \dots, m\}$. Let a be such that

$$\delta(a) = \inf\{\delta(r) \mid r \in \mathsf{S},\ \delta(r) > m\}.$$

If $\delta(a) < \delta(b)$ then take $k = 0$ and $r_0 = a$ to give existence in this case. Otherwise, apply the Division Algorithm to give $a = qb + r$ with $\delta(r) < \delta(b)$. Since S is δ-closed, we can moreover suppose that $q, r \in \mathsf{S}$. Now, since b is a nonzero nonunit, since we are supposing that $\delta(a) \geq \delta(b) > \delta(r)$, and since S is δ-positive,

$$\delta(q) < \delta(qb) = \delta(a - r) \leq \max\{\delta(a), \delta(r)\} = \delta(a).$$

Therefore, we may apply the induction hypothesis to q to give

$$q = r'_0 + r'_1 b + r'_2 b^2 + \cdots + r'_k b^k$$

for some $k \in \mathbb{Z}_{\geq 0}$ and for $r'_0, r'_1, \dots, r'_k \in \mathsf{S}$. Then

$$a = (r'_0 + r'_1 b + r'_2 b^2 + \cdots + r'_k b^k)b + r = r + r'_0 b + r'_1 b^2 + \cdots + r'_k b^{k+1},$$

showing that the existence part of the result holds for $\delta(a) = \inf\{\delta(r) \mid r \in \mathsf{S},\ \delta(r) > m\}$. This proves the existence part of the result for all $a \in \mathsf{S}$ by induction.

We also prove the uniqueness assertion by induction on $\delta(a)$. First we use a technical lemma concerning the general expansion of 0_{R} in the base b.

1 **Lemma** *Let* (R,δ) *be a Euclidean domain with* $\mathrm{b} \in \mathsf{R}$ *a nonzero nonunit. If* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$ *and* $\mathrm{r}_0, \mathrm{r}_1, \dots, \mathrm{r}_\mathrm{k} \in \mathsf{R}$ *satisfy*

(i) $\mathrm{r}_0 + \mathrm{r}_1\mathrm{b} + \mathrm{r}_2\mathrm{b}^2 + \cdots + \mathrm{r}_\mathrm{k}\mathrm{b}^\mathrm{k} = 0_\mathsf{R}$ *and*

(ii) $\delta(\mathrm{r}_0), \delta(\mathrm{r}_1), \dots, \delta(\mathrm{r}_\mathrm{k}) < \delta(\mathrm{b})$,

then $\mathrm{r}_0 = \mathrm{r}_1 = \cdots = \mathrm{r}_\mathrm{k} = 0_\mathsf{R}$.

Proof We prove this by induction on k. For $k = 0$ the result is trivial. For $k = 1$ we have $r_0 + r_1 b = 0_\mathsf{R}$, and we claim that $r_0 = r_1 = 0_\mathsf{R}$. Suppose that $r_1 \neq 0_\mathsf{R}$. Then

$$\delta(b) \leq \delta(r_1 b) = \delta(-r_0) = \delta(r_0) < \delta(b),$$

which is a contradiction. Thus $r_1 = 0_\mathsf{R}$, and then also $r_0 = 0_\mathsf{R}$. Now suppose the result holds for $k \in \{0, 1, \dots, m\}$ and consider the expression

$$0_\mathsf{R} = r_0 + r_1 b + r_2 b^2 + \cdots + r_{m+1} b^{m+1} = (r_1 + r_2 b + \cdots + r_{m+1} b^m) b + r_0.$$

Since the result holds for $k = 1$, it follows that

$$r_1 + r_2 b + \cdots + r_{m+1} b^m = 0_\mathsf{R}, \quad r_0 = 0_\mathsf{R}.$$

By the induction hypothesis, $r_1 = r_2 = \cdots = r_{m+1} = 0_\mathsf{R}$, and so the result follows. ▼

Now we carry on with the uniqueness part of the proof. First consider the case when $\delta(a) = \delta(1_\mathsf{R})$. Then, since b is a nonzero nonunit, $\delta(a) < \delta(b)$ by Proposition 4.2.41. Suppose that

$$a = r_0 + r_1 b + r_2 b^2 + \cdots + r_k b^k = (r_1 + r_2 b + \cdots + r_k b^{k-1}) b + r_0 \tag{4.2}$$

for $r_0, r_1, \dots, r_k \in \mathsf{S}$ with $\delta(r_0), \delta(r_1), \dots, \delta(r_k) < \delta(b)$. By Proposition 4.2.44 there is only one way to express a as $qb + r$ with $\delta(r) < \delta(b)$ and with $q, r \in \mathsf{S}$, and from the existence part of the proof we know that this implies that

$$r_1 + r_2 b + \cdots + r_k b^{k-1} = 0_\mathsf{R}, \quad r_0 = a.$$

By the lemma we can then assert that $r_1 = \cdots = r_k = 0_\mathsf{R}$, and so we must have $k = 0$ and $r_0 = a$ as the unique solution to (4.2). Thus the result holds when $\delta(a) = \delta(1_\mathsf{R})$. Next suppose the result true for $\delta(a) \in \{\delta(1_\mathsf{R}, \dots, m\}$, and suppose that $a \in \mathsf{S}$ satisfies

$$\delta(a) = \inf\{\delta(r) \mid r \in \mathsf{S},\ \delta(r) > m\}.$$

Then suppose that

$$a = r_0 + r_1 b + \cdots + r_k b^k = r'_0 + r'_1 b + \cdots + r'_{k'} b^{k'}$$

for $k, k' \in \mathbb{Z}_{\geq 0}$, $r_0, r_1, \dots, r_k \in \mathsf{S}$, and $r'_0, r'_1, \dots, r'_{k'} \in \mathsf{S}$ satisfying $\delta(r_j), \delta(r'_{j'}) < \delta(b)$ for $j \in \{0, 1, \dots, k\}$ and $j' \in \{0, 1, \dots, k'\}$. Also suppose that $r_k, r'_{k'} \neq 0_\mathsf{R}$. Then

$$\underbrace{(r_1 + r_2 b + \cdots + r_k b^{k-1})}_{q} b + r_0 = \underbrace{(r'_1 + r'_2 b + \cdots + r'_{k'} b^{k'-1})}_{q'} b + r'_0.$$

By Proposition 4.2.44 we have $q = q'$ and $r_0 = r'_0$. First suppose that $\delta(a) < \delta(b)$. Then, by Proposition 4.2.44, we have $q = q' = 0_R$ and $r_0 = r'_0 = a$. By the lemma it follows that $r_1 = \cdots = r_k = 0_R$ and $r'_1 = \cdots = r'_{k'} = 0_R$, and so we have $k = k' = 0$ and $r_0 = r'_0 = a$. Next suppose that $\delta(a) \geq \delta(b)$. Then it follows that $q, q' \neq 0_R$, since otherwise we have $a = r_0 = r'_0$, contradicting the fact that $\delta(r_0), \delta(r'_0) < \delta(b)$. Then we have

$$\delta(q) < \delta(qb) = \delta(a - r_0) \leq \max\{\delta(a), \delta(r_0)\} = \delta(a)$$

since b is a nonzero nonunit and since $\delta(a) \geq \delta(b) > \delta(r_0)$. Similarly, $\delta(q') < \delta(a)$. Therefore, the induction hypothesis applies to q and q' and we conclude that $k-1 = k'-1$ and $r_j = r'_j$ for $j \in \{1, \dots, k\}$, so proving the uniqueness part of the result by induction on $\delta(a)$. ■

The proof of the theorem is constructive, and we ask the reader to verify the procedure for determining the coefficients $r_0, r_1, \dots, r_k$ in Exercise 4.2.15.

Applying Theorem 4.2.48 to the Euclidean domain $\mathbb{Z}$ and the δ-closed and δ-positive subsemiring $\mathbb{Z}_{\geq 0}$, we have the following conclusion, probably known to the reader. The case of $k = 2$ corresponds to the binary expansion of a positive integer.

4.2.49 Corollary (Base expansion of positive integers) *If* $a, b \in \mathbb{Z}_{>0}$ *with* $b \geq 2$*, then there exists unique* $k \in \mathbb{Z}_{\geq 0}$ *and* $r_0, r_1, \dots, r_k \in \mathbb{Z}_{\geq 0}$ *such that*

(i) $r_k \neq 0$,

(ii) $r_0, r_1, \dots, r_k < b$, *and*

(iii) $a = r_0 + r_1 b + r_2 b^2 + \cdots + r_k b^k$.

The base expansion also has the following useful consequence.

4.2.50 Theorem (Base expansion in the smallest base) *Let* (R, δ) *be a Euclidean domain, let* $S \subseteq R$ *be a nontrivial,* δ*-closed, and* δ*-positive subsemiring of* R*, and let*

$$U = \{r \in S \mid r \text{ is a unit}\} \cup \{0_R\}.$$

If $U \subset S$ *and if* $x \in S$ *satisfies*

$$\delta(x) = \inf\{\delta(r) \mid r \in S,\ \delta(r) > \delta(1_R)\},$$

then, for $a \in S \setminus \{0_R\}$*, there exists a unique* $k \in \mathbb{Z}_{\geq 0}$ *and* $c_0, c_1, \dots, c_k \in U$ *such that*

(i) $c_k \neq 0_R$ *and*

(ii) $a = c_0 + c_1 x + \cdots + c_k x^k$.

Moreover, if $U \subset S$ *and if* $a, b \in S \setminus \{0_R\}$ *are written as*

$$a = c_0 + c_1 x + \cdots + c_k x^k, \quad b = d_0 + d_1 x + \cdots + d_l x^l$$

for $c_0, c_1, \dots, c_k, d_0, d_1, \dots, d_l \in U$ *such that* $c_k, d_l \neq 0_R$*, then* $\delta(a) > \delta(b)$ *if and only if* $k > l$.

Proof Since x is a nonzero nonunit, from Theorem 4.2.48 we can write $a = c_0 + c_1x + \cdots + c_kx^k$ for unique $c_0, c_1, \dots, c_k \in \mathsf{S}$ with $c_k \neq 0_{\mathsf{R}}$ and $\delta(c_0), \delta(c_1), \dots, \delta(c_k) < \delta(x)$. The hypotheses on x immediately give $c_0, c_1, \dots, c_k \in U$.

Now let a and b be as stated in the second assertion and write $a = qb + r$ for $q, r \in \mathsf{S}$ with $\delta(r) < \delta(b)$, this being possible by δ-closedness of S.

Let us assume that $\delta(a) > \delta(b)$. We will show by induction on $\delta(b)$ that $k > l$. First suppose that $\delta(b) = \delta(1_{\mathsf{R}})$ so that $b \in U$. Since $\delta(a) > \delta(b)$ it follows that a is a nonzero nonunit and so, by the first part of the result, $k > 1$, giving the result in this case. Assume the result holds for $\delta(b) \in \{\delta(1_{\mathsf{R}}), \dots, n\}$ and suppose that

$$\delta(b) = \inf\{\delta(r) \mid r \in c, \ \delta(r) > n\}.$$

We claim that the hypothesis that $\delta(a) > \delta(b)$ implies that q is a nonzero nonunit. If $q = 0_{\mathsf{R}}$ then $a = r$ and so $\delta(b) > \delta(r) = \delta(a)$, in contradiction with our assumption. If q is a unit then

$$\delta(b) = \delta(qb) = \delta(a - r) = \delta(a),$$

the last equality holding since $\delta(r) < \delta(b) < \delta(a)$ and since $\delta(a - r) \leq \max\{\delta(a), \delta(r)\}$ by δ-positivity of S. Thus q being a unit leads to the contradiction $\delta(b) = \delta(a)$. Since q is a nonzero nonunit, by the first conclusion of the proposition we have $q = u_0 + u_1x + \cdots + u_mx^m$ for $m \in \mathbb{Z}_{>0}$ with $u_0, u_1, \dots, u_m \in U$ and $u_m \neq 0_{\mathsf{R}}$. Since $\delta(r) < \delta(b)$ the induction hypotheses imply that $r = v_0 + v_1x + \cdots + v_px^p$ for $p < l$ with $v_0, v_1, \dots, v_p \in U$ and $v_p \neq 0_{\mathsf{R}}$. Therefore,

$$\begin{aligned} a &= c_0 + c_1x + \cdots + c_kx^k \\ &= (u_0 + u_1x + \cdots + u_mx^m)(d_0 + d_1x + \cdots + d_lx^l) + v_0 + v_1x + \cdots + v_px^p, \end{aligned}$$

from which we deduce that $k > l$ since R is a domain and since $p < m + l$.

Now assume that $k > l$. Let us write

$$q = u_0 + u_1x + \cdots + u_mx^m, \quad r = v_0 + v_1x + \cdots + v_px^p$$

with $u_0, u_1, \dots, u_m, v_0, v_1, \dots, v_p \in U$ and $u_m, v_p \neq 0_{\mathsf{R}}$. Since $\delta(r) < \delta(b)$ the previous part of the proof gives $p < l$. By the uniqueness part of Theorem 4.2.48 we must have $m = k - l > 0$. Therefore, again by the uniqueness part of Theorem 4.2.48, we conclude that q is not a unit and so $\delta(q) > \delta(1_{\mathsf{R}})$. Therefore,

$$\delta(b) < \delta(qb) = \delta(a - r) \leq \max\{\delta(a), \delta(r)\} = \delta(a),$$

the last equality holding since $\delta(r) < \delta(b)$. This gives the result. ■

An interesting result regarding Euclidean domains admitting a unique Division Algorithm now almost immediately follows, and relies on the notion of a polynomial which we only introduce in Section 4.4.

4.2.51 Theorem (Characterisation of Euclidean domains admitting a unique Division Algorithm) *If* (R, δ) *is a Euclidean domain that admits a unique Division Algorithm, then*

(i) the set of units in R *forms a field which we denote by* F_{R} *and*

(ii) if $\mathsf{F}_{\mathsf{R}} \subset \mathsf{R}$ *then* R *is isomorphic to* $\mathsf{F}_{\mathsf{R}}[\xi]$*.*

Proof We claim that R admits a unique Division Algorithm if and only if $\delta(a + b) \leq \max\{\delta(a), \delta(b)\}$ for every $a, b \in \mathsf{R}$. Certainly, if $\delta(a + b) \leq \max\{\delta(a), \delta(b)\}$ for every $a, b \in \mathsf{R}$, then R is a δ-closed and δ-positive subsemiring of itself, and then uniqueness of quotient and remainder follows from Proposition 4.2.44. Conversely, suppose that $a, b \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ satisfy $\delta(a + b) > \max\{\delta(a), \delta(b)\}$. Then we can write $a = 0_{\mathsf{R}} \cdot (a + b) + a$ with $\delta(a) < \delta(a + b)$ and also $a = 1_{\mathsf{R}} \cdot (a + b) + (-b)$ with $\delta(-b) < \delta(a + b)$. Thus R does not admit a unique Division Algorithm.

That the units in R form a field will follow if we can show that, if units $a, b \in \mathsf{R}$ satisfy $a + b \neq 0_{\mathsf{R}}$, then $a + b$ is a unit. This, however, follows since

$$\delta(1_{\mathsf{R}}) \leq \delta(a + b) \leq \max\{\delta(a), \delta(b)\} = \delta(1_{\mathsf{R}}),$$

and so $\delta(a + b) = \delta(1_{\mathsf{R}})$, implying that $a + b$ is a unit.

The final assertion of the corollary follows from Theorem 4.2.50 since every $r \in \mathsf{R}$ can be written as

$$r = a_0 + a_1 x + \cdots + a_k x^k$$

for unique $a_0, a_1, \ldots, a_k \in \mathsf{F}_{\mathsf{R}}$ with $a_k \neq 0_{\mathsf{R}}$ and with x as defined in the statement of Theorem 4.2.50. We then easily see that the map

$$\mathsf{R} \ni a_0 + a_1 x + \cdots + a_k x^k \mapsto a_0 + a_1 \xi + \cdots + a_k \xi^k \in \mathsf{F}_{\mathsf{R}}[\xi]$$

is the desired isomorphism. ■

As a final comment in this section, we note that we have focussed here on the situation when the map δ is given as part of the data of the Euclidean ring. This sidesteps the question, "Given a ring R, does there exists a map $\delta\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ such that (R, δ) is a Euclidean ring?" We will not expound on this in any length, but we do make the following remark. In subsequent sections we will show that Euclidean rings possess certain properties (e.g., being a principal ideal ring) that can be formulated without using the map δ. Therefore, any ring *not* possessing these properties cannot be a Euclidean ring for *any* map δ. We shall also give in Example 4.2.56 an example of an integral domain that cannot be made into a Euclidean domain for any map δ.

4.2.8 Principal ideal rings and domains

We now turn to a class of rings which, as we shall see, generalises the notion of a Euclidean ring. This class of rings is described by the structure of its ideals.

We next restrict our attention to ideals that are of a certain sort.

4.2.52 Theorem (Intersections of ideal are ideals) *Let* R *be a ring, let* $\mathsf{S} \subseteq \mathsf{R}$*, and let* $\mathrm{I}_l(\mathrm{S})$ *(resp.* $\mathrm{I}_r(\mathrm{S})$*,* $\mathrm{I}(\mathrm{S})$*) be the collection of left (resp. right,two-sided) ideals of* R *for which if* $\mathsf{I} \in \mathrm{I}_l(\mathrm{S})$*, (resp.* $\mathsf{I} \in \mathrm{I}_r(\mathrm{S})$*,* $\mathsf{I} \in \mathrm{I}(\mathrm{S})$*) then* $\mathsf{S} \subseteq \mathsf{I}$*. Then* $\cap_{\mathsf{I}\in\mathrm{I}_l(\mathrm{S})}\mathsf{I}$ *(resp.* $\cap_{\mathsf{I}\in\mathrm{I}_r(\mathrm{S})}\mathsf{I}$*,* $\cap_{\mathsf{I}\in\mathrm{I}(\mathrm{S})}\mathsf{I}$*) is a left (resp. right, two-sided) ideal which contains* S*. Moreover,* $\cap_{\mathsf{I}\in\mathrm{I}_l(\mathrm{S})}\mathsf{I}$ *(resp.* $\cap_{\mathsf{I}\in\mathrm{I}_r(\mathrm{S})}\mathsf{I}$*,* $\cap_{\mathsf{I}\in\mathrm{I}(\mathrm{S})}\mathsf{I}$*) is contained in any left (resp. right, two-sided) ideal that contains* S*.*

Proof We shall prove the theorem for left ideals. The result for right and two-sided ideals follows in a similar manner.

First let us show that $S \subseteq \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$. This is clear: let $r \in S$. Then $r \in \mathsf{I}$ for each $\mathsf{I} \in I_l(S)$, and so $r \in \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$ and so $S \subseteq \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$. Now let $s_1, s_2 \in \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$. Let $\mathsf{I} \in I_l(S)$. Since $s_1, s_2 \in \mathsf{I}$ and since I is an ideal, $s_1 - s_2 \in \mathsf{I}$. Thus $s_1 - s_2 \in \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$. In an entirely similar manner one shows that if $r \in \mathsf{R}$ and $s \in \mathsf{I}$ for each $\mathsf{I} \in \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$, then $rs \in \cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$. This shows that $\cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$ is a left ideal by Proposition 4.2.14.

That $\cap_{\mathsf{I}\in I_l(S)}\mathsf{I}$ is contained in any left ideal containing S is obvious by definition. ■

With this result, the following definition makes sense.

4.2.53 Definition (Ideal generated by a set, principal ideal, principal ideal ring) Let R be a ring and let $S \subseteq \mathsf{R}$. Then, using the notation of Theorem 4.2.52, $\cap_{\mathsf{I}\in I(S)}\mathsf{I}$ is the ***ideal generated by* S**, and is denoted by (S). If $S = \{a_1, \dots, a_k\}$ then we denote $(S) = (a_1, \dots, a_k)$. An ideal I of R is ***principal*** if $\mathsf{I} = (a)$ for some $a \in \mathsf{R}$, and R is a ***principal ideal ring*** if every ideal is principal. A ***principal ideal domain*** is a principal ideal ring that is also an integral domain. •

The notation $(a_1, \dots, a_k)$ for the ideal generated by $\{a_1, \dots, a_k\}$ is imperfect because it suggests that $(a_1, \dots, a_k)$ is an element of the k-fold Cartesian product of R with itself. However, the notation is so entrenched that our changing it would only delude the reader.

Nothing we have said so far gives us much of an idea of what the ideal (S) generated by S *looks like*. The next result addresses this rather important matter for principal ideals.

4.2.54 Theorem (Characterisation of principal ideals) *Let* R *be a ring and let* $\mathrm{a} \in \mathsf{R}$*. Then the following statements hold:*

(i) $$(\mathrm{a}) = \left\{\mathrm{r}_0 \cdot \mathrm{a} + \mathrm{a} \cdot \mathrm{s}_0 + \mathrm{la} + \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{r}_\mathrm{j} \cdot \mathrm{a} \cdot \mathrm{s}_\mathrm{j} \;\middle|\; \mathrm{r}_0, \mathrm{s}_0, \mathrm{r}_\mathrm{j}, \mathrm{s}_\mathrm{j} \in \mathsf{R},\ \mathrm{k} \in \mathbb{Z}_{>0},\ \mathrm{l} \in \mathbb{Z}\right\};$$

(ii) if R *is a unit ring then* $$(\mathrm{a}) = \left\{\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{r}_\mathrm{j} \cdot \mathrm{a} \cdot \mathrm{s}_\mathrm{j} \;\middle|\; \mathrm{r}_\mathrm{j}, \mathrm{s}_\mathrm{j} \in \mathsf{R},\ \mathrm{k} \in \mathbb{Z}_{>0}\right\};$$

(iii) if R *is commutative then* $(\mathrm{a}) = \{\mathrm{r} \cdot \mathrm{a} + \mathrm{ka} \mid \mathrm{r} \in \mathsf{R},\ \mathrm{k} \in \mathbb{Z}\}$*;*

(iv) if R *is a commutative unit ring then* $(\mathrm{a}) = \{\mathrm{r} \cdot \mathrm{a} \mid \mathrm{r} \in \mathsf{R}\}$*.*

Proof (i) Let $\mathsf{I}_a = \left\{r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j \;\middle|\; r_0, s_0, r_j, s_j \in \mathsf{R},\ k \in \mathbb{Z}_{>0},\ l \in \mathbb{Z}\right\}$.

Note that $a \in \mathsf{I}_a$. We also claim that I_a is an ideal. To see this, let

$$s = r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j \in \mathsf{I}_a$$

and let $r \in \mathsf{R}$. Then

$$r \cdot s = (r \cdot r_0 + r \cdot ((l-1)a)) \cdot a + a \cdot 0_{\mathsf{R}} + 0_{\mathsf{R}}a + \left(r \cdot a \cdot s_0 + \sum_{j=1}^{k} (r \cdot r_j) \cdot a \cdot s_j \right) \in \mathsf{I}_a$$

and

$$s \cdot r = 0_{\mathsf{R}} \cdot a + a \cdot (s_0 \cdot r) + 0_{\mathsf{R}}a + \left(r_0 \cdot a \cdot r + ((l-1)a) \cdot a \cdot r + \sum_{j=1}^{k} r_j a s_j \right) \in \mathsf{I}_a.$$

Also, if

$$s_1 = r_{10} \cdot a + a \cdot s_{10} + l_1 a + \sum_{j=1}^{k_1} r_{1j} \cdot a \cdot s_{1j},$$
$$s_2 = r_{20} \cdot a + a \cdot s_{20} + l_2 a + \sum_{j=1}^{k_2} r_{2j} \cdot a \cdot s_{2j}$$

are elements of I_a, then

$$s_1 - s_2 = (r_{10} - r_{20}) \cdot a + a \cdot (s_{10} - s_{20}) + (l_1 - l_2) \cdot a + \left(\sum_{j=1}^{k_1} r_{1j} \cdot a \cdot s_{1j} + \sum_{j=1}^{k_2} (-r_{2j}) \cdot a \cdot s_{2j} \right) \in \mathsf{I}_a.$$

Thus I_a is an ideal by Proposition 4.2.14.

Now suppose that I is an ideal containing a. Then, since I is a subring, $la \in \mathsf{I}$ for all $l \in \mathbb{Z}_{\geq 0}$. Also since I is a subring, $-a \in \mathsf{I}$, so $-la \in \mathsf{I}$ for all $l \in \mathbb{Z}_{\geq 0}$. Thus $la \in \mathsf{I}$ for all $l \in \mathbb{Z}$. Also, since I is a left ideal, $r \cdot a \in \mathsf{I}$ for all $r \in \mathsf{R}$, and since I is a right ideal, $a \cdot s \in \mathsf{I}$ for all $s \in \mathsf{R}$. Since I is a left and right ideal, for each $k \in \mathbb{Z}_{>0}$ and for each collection $r_1, \dots, r_k, s_1, \dots, s_k \in \mathsf{R}$, $r_j \cdot a \cdot s_j \in \mathsf{I}$, $j \in \{1, \dots, k\}$. Since I is a subring,

$$\sum_{j=1}^{k} r_j \cdot a \cdot s_j \in \mathsf{I}$$

This all shows that $\mathsf{I}_a \subseteq \mathsf{I}$. Thus I_a is contained in any ideal that contains a. This gives the result.

(ii) We know from part (i) that

$$(a) = \left\{ r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j \;\middle|\; r_0, s_0, r_j, s_j \in \mathsf{R},\ k \in \mathbb{Z}_{>0},\ l \in \mathbb{Z} \right\}.$$

The result follows from the observation that if

$$r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j \in (a),$$

then we can write

$$r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j$$
$$= r_0 \cdot a \cdot 1_{\mathsf{R}} + 1_{\mathsf{R}} \cdot a \cdot s_0 + ((l-1)a) \cdot a \cdot 1_{\mathsf{R}} + \sum_{j=1}^{k} r_j \cdot a \cdot s_j.$$

(iii) We know from part (i) that

$$(a) = \left\{ r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j \;\middle|\; r_0, s_0, r_j, s_j \in \mathsf{R},\ k \in \mathbb{Z}_{>0},\ l \in \mathbb{Z} \right\}.$$

The result follows from the observation that if

$$r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j \in (a),$$

then we can write

$$r_0 \cdot a + a \cdot s_0 + la + \sum_{j=1}^{k} r_j \cdot a \cdot s_j = \left(r_0 + s_0 + \sum_{j=1}^{k} r_j \cdot s_j \right) \cdot a + la.$$

(iv) This follows from combining parts (ii) and (iii). ■

From the theorem, we know that if R is a commutative principal ideal ring with unit, then every ideal in R can be written as $\{ra \mid r \in \mathsf{R}\}$ for some $a \in \mathsf{R}$. Our main result concerning principal ideal rings is the following, which asserts that principal ideal rings generalise Euclidean rings.

4.2.55 Theorem (Euclidean rings are principal ideal rings) *If* (R, δ) *is a Euclidean ring, then* R *is a principal ideal ring. Moreover, if* (a) *is an ideal in* R*, then*

$$\delta(\mathrm{a}) = \inf\{\delta(\mathrm{b}) \mid \mathrm{b} \in (\mathrm{a}) \setminus \{0_{\mathsf{R}}\}\}.$$

Proof Let $\mathsf{I} \subseteq \mathsf{R}$ be an ideal. If $\mathsf{I} = \{0_{\mathsf{R}}\}$ then $\mathsf{I} = (0_{\mathsf{R}})$ and so the result follows. So suppose that $\mathsf{I} \neq \{0_{\mathsf{R}}\}$ and define $a \in \mathsf{I}$ such that

$$\delta(a) = \inf\{\delta(b) \mid b \in \mathsf{I} \setminus \{0_{\mathsf{R}}\}\},$$

this being possible since $\mathbb{Z}_{\geq 0}$ is well ordered. For $b \in \mathsf{I}$ we write $b = qa + r$ for $q, r \in \mathsf{R}$ with $\delta(r) < \delta(a)$. Then, since $b, qa \in \mathsf{I}$, $r \in \mathsf{I}$. Now, either $r = 0_{\mathsf{R}}$ or $r \neq 0_{\mathsf{R}}$ and $\delta(r) < \delta(a)$. In the latter case we contradict the definition of a, so we must have $r = 0_{\mathsf{R}}$. Thus, if $b \in \mathsf{I}$ then $b = qa$ for some $q \in \mathsf{R}$. Thus $\mathsf{I} = (a)$. ■

There exist principal ideal domains that are not Euclidean domains. The following not entirely trivial example exhibits a principal ideal domain that is not a Euclidean domain.

4.2.56 Example (A principal ideal domain that is not a Euclidean domain) Our example relies on knowing about complex numbers. The reader who does not know about complex numbers can refer to Section 4.7, and can also be expected to have to work to know what we are doing here.

We let R denote the subset of the ring $\mathbb{C}$ given by

$$\mathsf{R} = \left\{ j + \tfrac{1}{2}k(1 + \mathrm{i}\sqrt{19}) \;\middle|\; j, k \in \mathbb{Z} \right\}.$$

We denote $\alpha = \frac{1}{2}(1 + \mathrm{i}\sqrt{19})$, and we note the following easily verified facts about α:

1. $\bar{\alpha} = 1 - \alpha$;
2. $\alpha\bar{\alpha} = 5$;
3. $\alpha^2 = \alpha - 5$;
4. if $j + k\alpha \in \mathsf{R}$ then $\alpha(j + k\alpha) = -5k + (j + k)\alpha$.

Using these facts it is a simple matter, that we leave to the reader, to verify that R is a subring with unity of $\mathbb{C}$. Since $\mathbb{C}$ is an integral domain, so too is R.

As a first step in understanding R, let us determine the units in R. To do so, we introduce the map $N\colon \mathsf{R} \to \mathbb{Z}$ given by

$$N(j + k\alpha) = (j + k\alpha)(j + k\bar{\alpha}) = j^2 + jk + 5k^2. \tag{4.3}$$

One can verify directly that $N((j_1 + k_1\alpha)(j_2 + k_2\alpha)) = N(j_1 + k_1\alpha)N(j_2 + k_2\alpha)$. It is also clear that $N(j + k\alpha) \geq 0$ and $N(j + k\alpha) = 0$ if and only if $j = k = 0$. We then have the following lemma characterising the units of R.

1 Lemma *The set of units of* R *is* $\{-1, 1\}$.

Proof Let r be a unit in R so that there exists $r^{-1} \in \mathsf{R}$ for which $rr^{-1} = 1$. Then

$$1 = N(1) = N(rr^{-1}) = N(r)N(r^{-1}),$$

which shows that $N(r)$ is a unit in $\mathbb{Z}$, and hence equal to either 1 or -1. Since N takes values in $\mathbb{Z}_{\geq 0}$ we must have $N(r) = 1$. Write $r = j + k\alpha$ so that $j^2 + jk + 5k^2 = 1$. First suppose that $jk \geq 0$. Then we must have $k = 0$ from which it follows that $j \in \{-1, 1\}$. Thus the lemma follows in this case. Next suppose that $jk < 0$. Then, since $j + k\bar{\alpha} = j + k - k\alpha$,

$$1 = N(j + k\alpha) = N(j + k\bar{\alpha}) = (j + k)^2 - jk + 4k^2.$$

Therefore we again conclude that $k = 0$ and so $j \in \{-1, 1\}$. Again, the lemma follows. ▼

The next lemma will also be useful, and relies on the notion of prime elements of a ring that we will not introduce until Section 4.2.9.

2 Lemma 2 *and* 3 *are prime in* R.

Proof As will be seen below, R is a principal ideal domain. Therefore, by Lemma 1 of Theorem 4.2.71, it suffices to show that 2 and 3 are irreducible. Thus suppose, for example, that $2 = (j_1 + k_1\alpha)(j_2 + k_2\alpha)$. Then

$$4 = N(2) = N(j_1 + k_1\alpha)N(j_2 + k_2\alpha).$$

If we suppose that neither of $j_1+k_1\alpha$ or $j_2+k_2\alpha$ is a unit, this implies that $N(j_1+k_1\alpha) = N(j_2 + k_2\alpha) = 2$. Therefore, in particular, following the computations of Lemma 1, we have

$$2 = N(j_1 + k_1\alpha) = N(j_1 + k_1\bar{\alpha}) = j_1^2 + j_1k_2 + 5k_1^2 = (j_1 + k_1)^2 - j_1k_2 + 4k_1^2.$$

Considering separately the case $j_1k_1 \geq 0$ and $j_1k_1 < 0$ we deduce that $k_1 = 0$. In like manner we conclude that $k_2 = 0$. Therefore, we arrive at $2 = j_1j_2$, meaning that $(j_1 + 0\alpha)(j_2 + 0\alpha)$ is a prime factorisation of 2 *in* $\mathbb{Z}$. Thus one of $j_1 + 0\alpha$ and $j_2 + 0\alpha$ is a unit in $\mathbb{Z}$, and hence in R by Lemma 1. Thus 2 is irreducible. An entirely similar computation shows that 3 is irreducible. ▼

We now show that there exists no map $\delta\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ for which (R, δ) is a Euclidean domain. Suppose that δ is such a map, and let $b \in \mathsf{R}$ have the property that

$$\delta(b) = \inf\{\delta(r) \mid r \in \mathsf{R} \setminus \{-1, 0, 1\}\}.$$

Since (R, δ) is a Euclidean domain there exists $q, r \in \mathsf{R}$ such that $2 = qb + r$ where $\delta(r) < \delta(b)$. Thus we must have $r \in \{-1, 0, 1\}$, using Proposition 4.2.41. Thus we have either $qb = 3$, $qb = 2$, or $qb = 1$. We cannot have the last instance since b is not a unit in R. Thus either $b|2$ or $b|3$. Since 2 and 3 are prime this implies that $b \in \{-3, -2, 2, 3\}$. Now write $q = j + k\alpha$ so that we have either

1. $-3j - 3k\alpha = -1$, $-3j - 3k\alpha = 0$, or $-3j - 3k\alpha = 1$,
2. $-2j - 2k\alpha = -1$, $-2j - 2k\alpha = 0$, or $-2j - 2k\alpha = 1$,
3. $2j + 2k\alpha = -1$, $2j + 2k\alpha = 0$, or $2j + 2k\alpha = 1$, or
4. $3j + 3k\alpha = -1$, $3j + 3k\alpha = 0$, or $3j + 3k\alpha = 1$.

These in turn imply

1. $N(3)(j^2 + jk + 5k^2) = 1$ or $N(3)(j^2 + jk + 5k^2) = 0$, or
2. $N(2)(j^2 + jk + 5k^2) = 1$ or $N(2)(j^2 + jk + 5k^2) = 0$

and, using the equality $j + k\alpha = (j + k) - k\alpha$,

1. $N(3)((j + k)^2 - jk + 4k^2) = 1$ or $N(3)((j + k)^2 - jk + 4k^2) = 0$, or
2. $N(2)((j + k)^2 - jk + 4k^2) = 1$ or $N(2)((j + k)^2 - jk + 4k^2) = 0$.

Taking separately the cases $jk \geq 0$ and $jk < 0$, we see that none of these equalities can be satisfied for $j, k \in \mathbb{Z}$, and so we conclude that the map $\delta\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ having the property that (R, δ) is a Euclidean domain does not exist.

Next we show that R is a principal ideal domain. We do this employing a general strategy suggested by the following lemma.

3 **Lemma** *Let* R *be an integral domain and suppose that there exists a map* $\sigma\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ *having the following properties:*

(i) *if* $\mathsf{a}, \mathsf{b} \in \mathsf{R}$ *with* $\mathsf{ab} \neq 0$, *then* $\sigma(\mathsf{ab}) \geq \sigma(\mathsf{a})$*;*

(ii) *if* $\mathsf{a}, \mathsf{b} \in \mathsf{R}$ *with* $\mathsf{b} \neq 0$ *and if* $\sigma(\mathsf{a}) \geq \sigma(\mathsf{b})$, *then either* $\mathsf{b}|\mathsf{a}$ *or there exists* $\mathsf{r}, \mathsf{s} \in \mathsf{R}$ *such that* $0 < \sigma(\mathsf{ra} - \mathsf{sb}) < \sigma(\mathsf{b})$.

Then R *is a principal ideal domain.*

Proof Let $\mathsf{I} \subseteq \mathsf{R}$ be an ideal. If $\mathsf{I} = \{0\}$ then $\mathsf{I} = (0)$ and so I is principal. If $\mathsf{I} \neq \{0\}$, then define $b \in \mathsf{I}$ such that

$$\sigma(b) = \inf\{\sigma(a) \mid a \in \mathsf{I} \setminus \{0\}\}.$$

Suppose that there exists $a \in \mathsf{I}$ such that, for any $r \in \mathsf{R}$, $a \neq rb$. Since $a \neq 0$ we have $\sigma(a) \geq \sigma(b)$. Then there exists $r, s \in \mathsf{R}$ such that $\sigma(ra - sb) < \sigma(b)$. Since I is an ideal, $ra - sb \in \mathsf{I}$, and we have thus contradicted the definition of b. ▼

We shall now show that the map $N\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ defined in (4.3) has the properties of the map σ in the lemma. It is clear that $N(ab) \geq N(a)$ if $ab \neq 0$.

4 **Lemma** *Let* R *be the ring we are using in this example, and let* $\mathsf{a}, \mathsf{b} \in \mathsf{R}$ *with* $\mathsf{b} \neq 0$, *with* $\mathrm{N}(\mathsf{a}) \geq \mathrm{N}(\mathsf{b})$, *and such that* $\mathsf{b} \nmid \mathsf{a}$. *Then there exists* $\mathsf{r}, \mathsf{s} \in \mathsf{R}$ *such that* $0 < \mathrm{N}(\mathsf{ra} - \mathsf{sb}) < \mathrm{N}(\mathsf{b})$.

Proof Let us first say some things about $\frac{a}{b}$. We have

$$\frac{a}{b} = \frac{a\bar{b}}{b\bar{b}}.$$

By the properties of the number α given above we can then write

$$\frac{a}{b} = \beta + \gamma\alpha,$$

where $\beta, \gamma \in \mathbb{Q}$. Moreover, since $b \nmid a$, it must be that at least one of β and γ does not lie in $\mathbb{Z}$. We shall find $r, s \in \mathsf{R}$ such that

$$0 < N\left(\tfrac{a}{b}r - s\right) < 1,$$

and the lemma follows from this using the properties of N. We consider various cases. For $x \in \mathbb{R}$ we denote by $\{x\}$ the integer nearest x, taking the convention that $\{j + \frac{1}{2}\} = j$ for $j \in \mathbb{Z}$.

1. $\gamma \in \mathbb{Z}$: Here we have $\beta \notin \mathbb{Z}$. Define $r = 1$ and $s = \{\beta\} + \gamma\alpha$. Then

$$0 < N\left(\tfrac{a}{b}r - s\right) = N(\beta + \gamma\alpha - \{\beta\} - \gamma\alpha) \leq \tfrac{1}{4} < 1.$$

2. $\beta \in \mathbb{Z}$ and $5\gamma \notin \mathbb{Z}$: Take $r = \bar{\alpha}$ and $s = \{\beta + 5\gamma\} - \beta\alpha$. Then, using the properties of α given above,

$$\begin{aligned} 0 < N\left(\tfrac{a}{b}r - s\right) &= N\left(\tfrac{a}{b}\bar{\alpha} - \{\beta + 5\gamma\} + \beta\alpha\right) \\ &= N(\beta + 5\gamma - \beta\alpha - \{\beta + 5\gamma\} + \beta\alpha) \leq \tfrac{1}{4} < 1. \end{aligned}$$

3. $\beta \in \mathbb{Z}$ and $5\gamma \in \mathbb{Z}$: Note that $\gamma = \frac{j}{5}$ for $j \in \mathbb{Z}$. Then a simple induction on j shows that either $|\gamma - \{\gamma\}| = \frac{1}{5}$ or $|\gamma - \{\gamma\}| = \frac{2}{5}$. Take $r = 1$ and $s = \beta + \{\gamma\}\alpha$. Then
$$0 < N\left(\tfrac{a}{b}r - s\right) = N(\beta + \gamma\alpha - \beta - \{\gamma\}\alpha) = N(\gamma - \{\gamma\})N(\alpha) \leq \tfrac{4}{5} < 1.$$
4. $\beta, \gamma \notin \mathbb{Z}$ and $2\beta, 2\gamma \notin \mathbb{Z}$: In this case we either have $|\gamma - \{\gamma\}| \leq \frac{1}{3}$ or $|2\gamma - 2\{\gamma\}| < \frac{1}{3}$. In the first case we take $r = 1$ and $s = \{\beta\} + \{\gamma\}\alpha$ and compute
$$\begin{aligned} 0 < N\left(\tfrac{a}{b}r - s\right) &= N(\beta - \{\beta\} + (\gamma - \{\gamma\})\alpha) \\ &= (\beta - \{\beta\})^2 + (\beta - \{\beta\})(\gamma - \{\gamma\}) + 5(\gamma - \{\gamma\})^2 \leq \tfrac{1}{4} + \tfrac{1}{2}\tfrac{1}{3} + \tfrac{5}{9} = \tfrac{35}{36} < 1. \end{aligned}$$
In the second case we take $r = 2$ and $s = 2\{\beta\} + 2\{\gamma\}\alpha$, and the same computation as the first case gives the same conclusion.
5. $\beta, \gamma \notin \mathbb{Z}$ and $2\beta, 2\gamma \in \mathbb{Z}$: Note that for $\beta, \gamma \in \mathbb{Q}$ we have
$$(\beta + \gamma\alpha)\alpha = -5\gamma + (\beta + \gamma)\alpha,$$
as may be verified by direct computation. Moreover, in this case we have $\beta = \frac{j}{2}$ and $\gamma = \frac{k}{2}$ for $j, k \in \mathbb{Z}$. Therefore, $\beta + \gamma \in \mathbb{Z}$. Now take $r = \alpha$ and $s = \{-5\gamma\} + \{\beta + \gamma\}\alpha$ so that
$$\begin{aligned} 0 < N\left(\tfrac{a}{b}r - s\right) &= N\left(\tfrac{a}{b}\alpha - \{-5\gamma\} - \{\beta + \gamma\}\alpha\right) \\ &= N(-5\gamma + (\beta + \gamma)\alpha - \{-5\gamma\} - \{\beta + \gamma\}\alpha) = N(-5\gamma - \{-5\gamma\}) \leq \tfrac{1}{4} < 1. \end{aligned}$$
6. $\beta, \gamma \notin \mathbb{Z}$, $2\beta \in \mathbb{Z}$, $2\gamma \notin \mathbb{Z}$, and $5\gamma \in \mathbb{Z}$: Take $r = 5$ and $s = \{5\beta\} + 5\gamma\alpha$ and compute
$$0 < N\left(\tfrac{a}{b}r - s\right) = N(5\beta + 5\gamma\alpha - \{5\beta\} - 5\gamma\alpha) = N(5\beta - \{5\beta\}) \leq \tfrac{1}{4} < 1.$$
7. $\beta, \gamma \notin \mathbb{Z}$, $2\beta \in \mathbb{Z}$, $2\gamma \notin \mathbb{Z}$, and $5\gamma \notin \mathbb{Z}$: Here take $r = 2\bar{\alpha}$ and $s = \{2\beta + 10\gamma\} - 2\beta\alpha$ and compute
$$\begin{aligned} 0 < N\left(\tfrac{a}{b}r - s\right) &= N\left(2\tfrac{a}{b}\bar{\alpha} - \{2\beta + 10\gamma\} + 2\beta\alpha\right) \\ &= N(2\beta + 10\gamma - 2\beta\alpha - \{2\beta + 10\gamma\} + 2\beta\alpha) = N(2\beta + 10\gamma - \{2\beta + 10\gamma\}) \leq \tfrac{1}{4} < 1. \end{aligned}$$
8. $\beta, \gamma \notin \mathbb{Z}$, $2\beta \notin \mathbb{Z}$, and $2\gamma \in \mathbb{Z}$: Let $r = 2$ and $s = \{2\beta\} + 2\gamma\alpha$ and compute
$$0 < N\left(\tfrac{a}{b}r - s\right) = N(2\beta + 2\gamma\alpha - \{2\beta\} - 2\gamma\alpha) = N(2\beta - \{2\beta\}) \leq \tfrac{1}{4} < 1.$$

These cases may be easily seen to cover all possibilities, and so the lemma follows.
▼

The lemma, combined with Lemma 3, shows that R is a principal ideal domain.
●

An important property of principal ideal rings is the following which will be essential in the proof of Theorem 4.2.71. There we will see that the finiteness assertion in the theorem allows us to conclude a prime factorisation theorem for principal ideal domains.

4.2.57 Theorem (Nested sequences of ideals are finite in principal ideal rings) *Let* R *be a principal ideal ring and let* $(\mathsf{I}_{\mathrm{j}})_{\mathrm{j}\in\mathbb{Z}_{>0}}$ *be a sequence of ideals having the property that* $\mathsf{I}_{\mathrm{j}} \subseteq \mathsf{I}_{\mathrm{j}+1}$ *for* $\mathrm{j} \in \mathbb{Z}_{>0}$*. Then there exists* $\mathrm{N} \in \mathbb{Z}_{>0}$ *such that* $\mathsf{I}_{\mathrm{j}} = \mathsf{I}_{\mathrm{N}}$ *for* $\mathrm{j} \geq \mathrm{N}$*.*

Proof The sequence $(\mathsf{I}_j)_{j\in\mathbb{Z}_{>0}}$ is totally ordered in the set of ideals of R using the partial order of set inclusion. Therefore, as we saw during the course of the proof of Theorem 4.2.19, $\bar{\mathsf{I}} = \cup_{j\in\mathbb{Z}_{>0}}\mathsf{I}_j$ is an ideal. Since R is a principal ideal ring, $\bar{\mathsf{I}} = (r)$ for some $r \in \bar{\mathsf{R}}$. Thus $r \in \mathsf{I}_N$ for some $N \in \mathbb{Z}_{>0}$. For $j \geq N$ it therefore follows that $r \in \mathsf{I}_j$. Thus $(r) \subseteq \mathsf{I}_j \subseteq \bar{\mathsf{I}} = (r)$, and so $\mathsf{I}_j = (r)\mathsf{I}_N$ for $j \geq N$. ■

4.2.9 Divisors, primes, and irreducibles

The reader has more than likely been exposed to the idea of a prime number. The notion of a prime number is an example of a more general idea of a prime element in a ring. In this section we study this, along with the sometimes related notion of an irreducible element of a ring.

To begin the discussion we first discuss the notion of a divisor of an element of a ring.

4.2.58 Definition (Divisor) Let R be a commutative ring and let $r \in \mathsf{R}$. An element $d \in \mathsf{R} \setminus \{0\}$ is a ***divisor***, or a ***factor***, of e if there exists $s \in \mathsf{R}$ such that $r = sd$. If d is a divisor of r, then d ***divides*** r, r is ***divisible*** by d, and we write $d|r$. If d does not divide r then we write $d \nmid r$. •

4.2.59 Examples (Divisors)

1. Let $\mathsf{R} = \mathbb{Z}$. Then 2 divides every even integer, but does not divide any odd integer.
2. In the ring $\mathbb{Z}_4$, $3 + 4\mathbb{Z}$ divides $2 + 4\mathbb{Z}$ since $(3 + 4\mathbb{Z})(2 + 4\mathbb{Z}) = 6 + 4\mathbb{Z} = 2 + 4\mathbb{Z}$.
3. In the ring $\mathbb{R}$, every element nonzero element divides every other element. Indeed, if $x \in \mathbb{R}$ and if $y \in \mathbb{R}^*$, then $x = (xy^{-1})y$. Readers who have read Section 4.3, or who know about fields, know that this is a property of every field. •

Let us first record some elementary properties of divisors. These all follow in a straightforward way from the definition, so we leave the proofs as Exercise 4.2.17 for the reader.

4.2.60 Proposition (Elementary properties of divisors) *In a commutative ring* R *with* $\mathrm{d}, \mathrm{r}, \mathrm{s} \in \mathsf{R}$*, the following statements hold:*

(i) *if* $\mathrm{d}|\mathrm{r}$ *and* $\mathrm{r}|\mathrm{s}$ *then* $\mathrm{d}|\mathrm{s}$;

(ii) $\mathrm{d}|\mathrm{r}$ *if and only if* $\mathrm{d}|(\mathrm{ur})$ *for every unit* u;

(iii) $\mathrm{d}|\mathrm{r}$ *if and only if* $(\mathrm{ud})|\mathrm{r}$ *for every unit* u;

(iv) $\mathrm{u}|\mathrm{a}$ *for every* $\mathrm{a} \in \mathsf{R}$ *if and only if* u *is a unit;*

(v) *if* $\mathrm{d} \neq 0$ *then* $\mathrm{d}|0$;

(vi) *if* u *is a unit and if* $\mathrm{d}|\mathrm{u}$ *then* d *is a unit;*

(vii) *if* $\mathrm{r} = \mathrm{us}$ *for some unit* u*, then* $\mathrm{r}|\mathrm{s}$ *and* $\mathrm{s}|\mathrm{r}$*;*

(viii) *if* R *is an integral domain and if* $\mathrm{r}|\mathrm{s}$ *and* $\mathrm{s}|\mathrm{r}$*, then* $\mathrm{r} = \mathrm{us}$ *for a unit* u*;*

(ix) *if* $\mathrm{d}|\mathrm{r}$ *and* $\mathrm{d}|\mathrm{s}$*, then* $\mathrm{d}|(\mathrm{ar} + \mathrm{bs})$ *for every* $\mathrm{a}, \mathrm{b} \in \mathsf{R}$*.*

Divisors also have relationships to principal ideals as follows.

4.2.61 Proposition (Divisors and principal ideals) *If* R *is a commutative ring with identity, then the following statements hold for* $\mathrm{r}, \mathrm{s} \in \mathsf{R}$*:*

(i) $\mathrm{r}|\mathrm{s}$ *if and only if* $(\mathrm{s}) \subseteq (\mathrm{r})$*;*

(ii) $\mathrm{r}|\mathrm{s}$ *and* $\mathrm{s}|\mathrm{r}$ *if and only if* $(\mathrm{r}) = (\mathrm{s})$*;*

(iii) r *is a unit if and only if* $(\mathrm{r}) = \mathsf{R}$*.*

Proof (i) Suppose that $r|s$ and, by Theorem 4.2.54, let $r_1 s \in (s)$ for some $r_1 \in \mathsf{R}$. Then $r_1 s = r_1 r_2 r$ for some $r_2 \in \mathsf{R}$, meaning, again by Theorem 4.2.54, that $r_1 s \in (r)$. Now suppose that $(s) \subseteq (r)$. Then, by Theorem 4.2.54, $s = r'r$ for some $r' \in \mathsf{R}$, showing that $r|s$.

(ii) This follows from part (i).

(iii) Suppose that r is a unit and let $r' \in \mathsf{R}$. Then $r' = (r'r^{-1})r$, showing that $\mathsf{R} = (r)$. Now suppose that $(r) = \mathsf{R}$. Then $1_{\mathsf{R}} = r'r$ for some $r' \in \mathsf{R}$, which shows that r is a unit. ■

In any commutative ring R with identity and for any $r \in \mathsf{R}$, we always have $1_{\mathsf{R}}|r$ and $r|r$, and so every element always possesses the divisors 1_{R} and itself. For the ring $\mathbb{Z}$, we use the terminology that a positive number is “prime” when its only divisors are 1_{R} and itself. For general rings, this idea can be stated as the following definition. At this point, it is perhaps not obvious that the concepts we are talking about amount to our usual notion of a prime integer. However, this will be proved in Section 4.2.10.

4.2.62 Definition (Irreducible, prime) For a commutative ring R with identity, an element $r \in \mathsf{R}$ is:

(i) ***irreducible*** if r is nonzero, not a unit, and has the property that, if $r = a \cdot b$, then either a or b is a unit;

(ii) ***reducible*** if it is not irreducible;

(iii) ***prime*** if r is nonzero, not a unit, and has the property that $r|(a \cdot b)$ implies that $r|a$ or $r|b$. •

Let us give some examples that show that this really does agree with what one already knows.

4.2.63 Examples (Irreducibles and primes)

1. Consider the integral domain $\mathbb{Z}$. We claim that 2 is both irreducible and prime, but that 4 is neither irreducible nor prime. Note that neither 2 nor 4 are units.

   Indeed, suppose that $2 = jk$ for $j, k \in \mathbb{Z}$. Then $j, k \in \{1, 2\}$ or $j, k \in \{-1, -1\}$, and then one can directly see that either (1) $j = 1$ and $k = 2$ or that (2) $j = 2$ and $k = 1$ or that (3) $j = -1$ and $k = -2$ or that (4) $j = -2$ and $k = -1$. Thus 2 is irreducible. Next suppose that $2|(jk)$. This means that jk is even. One can directly see that this implies that either j or k must be even, so that $2|j$ or $2|k$. Thus 2 is prime.

   Now note that $4 = 2 \cdot 2$, but that 2 is not a unit. Therefore, 4 is not irreducible. Also note that $4|(2 \cdot 6)$, but that 4 divides neither 2 nor 6. Thus 4 is not prime either.

   We shall see below that "prime" and "irreducible" agree for $\mathbb{Z}$, and for a whole class of integral domains.

2. Consider the ring $\mathbb{Z}_6$. We claim that $2 + 6\mathbb{Z}$ is prime but not irreducible. First note that $2 + 6\mathbb{Z}$ is not a unit in $\mathbb{Z}_6$.

   Indeed, if $(2 + 6\mathbb{Z})|((j + 6\mathbb{Z})(k + 6\mathbb{Z}))$ then $jk = 2l + 6m$ for some $l, m \in \mathbb{Z}$. In particular, one can easily see that this implies that jk must be even. Therefore, either j or k is even, and so either $2 + 6\mathbb{Z}$ divides $j + 6\mathbb{Z}$ or $2 + 6\mathbb{Z}$ divides $k + 6\mathbb{Z}$. Thus $2 + 6\mathbb{Z}$ is prime.

   Now note that $2 + 6\mathbb{Z} = (2 + 6\mathbb{Z})(4 + 6\mathbb{Z})$, but that neither $2 + 6\mathbb{Z}$ nor $4 + 6\mathbb{Z}$ are units in $\mathbb{Z}_6$. Indeed, both $2 + 6\mathbb{Z}$ and $4 + 6\mathbb{Z}$ are zerodivisors.

3. Next take the ring $\mathbb{R}$. Since all nonzero elements of $\mathbb{R}$ are units, there are neither any primes nor any irreducibles in $\mathbb{R}$.

4. Our next example is one that gives elements of a ring that are irreducible but not prime. The ring we need to exhibit this is a little more complicated than our preceding two examples. We define a subset of the ring $\mathbb{R}$ by

   $$\mathsf{R} = \left\{ j + k\sqrt{10} \;\middle|\; j, k \in \mathbb{Z} \right\},$$

   and we leave to the reader the straightforward verification that R is a subring of $\mathbb{R}$. We claim that the elements $2, 3, 4 + \sqrt{10}, 4 - \sqrt{10} \in \mathsf{R}$ are irreducible, but not prime. As a first step, let us characterise the units of R. To do so, introduce the map $N \colon \mathsf{R} \to \mathbb{Z}$ defined by

   $$N(j + k\sqrt{10}) = j^2 - 10k^2.$$

   This map will be useful generally, but for now we use it to state the following lemma.

   1 Lemma *The set of units in* R *is* $\{\mathrm{j} + \mathrm{k}\sqrt{10} \mid \mathrm{j}^2 - 10\mathrm{k}^2 \in \{-1, 1\}\}$.

   Proof It is a straightforward calculation to check that $N(r_1 r_2) = N(r_1)N(r_2)$ for all $r_1, r_2 \in \mathsf{R}$. Also, suppose that $N(j + k\sqrt{10}) = j^2 - 10k^2 = 0$ and that $k \neq 0$.

Then $\frac{j^2}{k^2} = 10$ or $\left|\frac{j}{k}\right| = \sqrt{10}$. Since $\sqrt{10}$ is irrational (why?) this equation has no integer solutions for j and k. Thus we must have $k = 0$ whence also $j = 0$. Thus $N(j + k\sqrt{10}) = 0$ if and only if $j = k = 0$.

Now suppose that $j+k\sqrt{10}$ is a unit with multiplicative inverse $j' + k'\sqrt{10}$. Then

$$1 = N(1) = N((j + k\sqrt{10})(j' + k'\sqrt{10})) = N(j + k\sqrt{10})N(j' + k'\sqrt{10}).$$

Since N takes values in $\mathbb{Z}$ it follows that $N(j + k\sqrt{10})$ is a unit in $\mathbb{Z}$, or that $N(j + k\sqrt{10}) = j^2 - 10k^2 \in \{-1, 1\}$. ▼

One can then check directly that none of the four elements $2, 3, 4+\sqrt{10}, 4-\sqrt{10} \in$ R is a unit.

To see that 2 and 3 are irreducible, let $p \in \{-3, -2, 2, 3\}$, suppose that

$$p = (j_1 + k_1\sqrt{10})(j_2 + k_2\sqrt{10}).$$

Then

$$p^2 = N(p) = N(j_1 + k_1\sqrt{10})N(j_2 + k_2\sqrt{10}).$$

Since p is prime in $\mathbb{Z}$, in order that $j_1 + k_1\sqrt{10}$ and $j_2 + k_2\sqrt{10}$ not be units in R, by the lemma we must have $N(j_1 + k_1\sqrt{10}), N(j_2 + k_2\sqrt{10}) \in \{-p, p\}$. We now use a lemma.

2 Lemma *For* $\mathrm{p} \in \{-3, -2, 2, 3\}$, *there do not exist* $\mathrm{j}, \mathrm{k} \in \mathbb{Z}$ *such that* $\mathrm{j}^2 - 10\mathrm{k}^2 = \mathrm{p}$.

Proof Let $\pi_5 \colon \mathbb{Z} \to \mathbb{Z}_5$ be given by $\pi_5(j) = j + 5\mathbb{Z}$. If $j, k \in \mathbb{Z}$ satisfy $j^2 - 10k^2 = p$, then we have $j^2 = p + 5(2k^2)$, and so $\pi_5(j^2) = \pi_5(p)$. Therefore, it suffices to show that the equation $(j + 5\mathbb{Z})^2 = p + 5\mathbb{Z}$ has no solutions in $\mathbb{Z}_5$. Note that $-3 + 5\mathbb{Z} = 2 + 5\mathbb{Z}$ and $-2 + 5\mathbb{Z} = 3 + 5\mathbb{Z}$, so it suffices to consider only $p \in \{2, 3\}$. For this, we simply compute

$$\begin{aligned} (1 + 5\mathbb{Z})^2 &= 1 + 5\mathbb{Z}, & (2 + 5\mathbb{Z})^2 &= 4 + 5\mathbb{Z}, \\ (3 + 5\mathbb{Z})^2 &= 4 + 5\mathbb{Z}, & (4 + 5\mathbb{Z})^2 &= 1 + 5\mathbb{Z}. \end{aligned}$$

Therefore, $(j + 5\mathbb{Z})^2$ cannot take the value $2 + 5\mathbb{Z}$ or $3 + 5\mathbb{Z}$, and the lemma follows. ▼

The lemma immediately implies that if $p = (j_1 + k_1\sqrt{10})(j_2 + k_2\sqrt{10})$ then either $j_1 + k_1\sqrt{10}$ or $j_2 + k_2\sqrt{10}$ must be a unit. Thus either k_1 or $k_2 = 0$. In the first case we must have $k_2 = 0$ and in the second case we must have $k_1 = 0$. Thus we are left with $p = j_1 j_2$, which means that either j_1 or j_2 must be a unit in $\mathbb{Z}$. Thus either $j_1 + k_1\sqrt{10}$ or $j_2 + k_2\sqrt{10}$ must be a unit in R. Thus p is irreducible for $p \in \{-3, -2, 2, 3\}$.

To see that $4+\sqrt{10}$ is irreducible, write $4+\sqrt{10}=(j_1+k_1\sqrt{10})(j_2+k_2\sqrt{10})$ for $j_1+k_1\sqrt{10}$ and $j_2+k_2\sqrt{10}$ nonunits. Then

$$6=N(4+\sqrt{10})=N(j_1+k_1\sqrt{10})N(j_1+k_1\sqrt{10}).$$

Since $j_1+k_1\sqrt{10}$ and $j_2+k_2\sqrt{10}$ are nonunits, by Lemma 1 we must have $N(j_1+k_2\sqrt{10}), N(j_2+k_2\sqrt{10})\in\{-3,-2,2,3\}$. By Lemma 2, however, we know that this is not possible if $j_1+k_1\sqrt{10}$ and $j_2+k_2\sqrt{10}$ are nonunits. Thus $4+\sqrt{10}$ is irreducible. An entirely similar computation also shows that $4-\sqrt{10}$ is irreducible.

Now let us show that 2 is not prime. Note that $2|6$, and that $6=(4+\sqrt{10})(4-\sqrt{10})$. It is clear that $2\nmid(4+\sqrt{10})$ and $2\nmid(4-\sqrt{10})$, which means that 2 is not prime.

The matter of showing that $3, 4+\sqrt{10}$, and $4-\sqrt{10}$ are irreducible but not prime can be carried out in an entirely analogous manner, and we leave the details of this to the reader. •

Next we indicate the relationship between primes and irreducibles, and special sorts of ideals. The notion of an ideal came about exactly for the reason of a need to generalise objects such as primes. Therefore, in some sense, the following result plays a key rôle in comprehending why ideals are important. To state the result, it is convenient to recall from Corollary 4.2.20 the characterisation of maximal ideals as maximal elements of the set, partially ordered by set inclusion, $S(\mathsf{R})$ of all proper ideals. Here we modify this slightly by defining $P(\mathsf{R})$ to be the set of principal ideals, again ordered by set inclusion.

4.2.64 Theorem (Prime and irreducibles, and ideals) *For* R *a commutative unit ring the following statements hold:*

(i) $\mathrm{p}\in\mathsf{R}$ *is prime if and only if* (p) *is a prime ideal;*

(ii) *if* r *is irreducible then* (r) *is a maximal element of* $\mathrm{P}(\mathsf{R})$*;*

(iii) *if* R *is an integral domain and if* $(\mathrm{s})\subseteq(\mathrm{r})$ *for every* $(\mathrm{s})\in\mathrm{P}(\mathsf{R})$*, then* r *is irreducible.*

Proof (i) Suppose that p is prime and let $rs\in(p)$. Then $p|(rs)$ by Theorem 4.2.54 so that either $p|r$ or $p|s$. By Proposition 4.2.18 it follows that (p) is prime. Now suppose that (p) is prime and suppose that $p|(rs)$. Therefore, $rs\in(p)$ by Theorem 4.2.54 and so either $r\in(p)$ or $s\in(p)$. Again using Theorem 4.2.54 we conclude that $p|r$ or $p|s$, showing that p is prime.

(ii) Note that if r is irreducible it is not a unit, and so $(r)\subset\mathsf{R}$ by Proposition 4.2.61. Therefore, $(r)\in P(\mathsf{R})$. Now let $(s)\in P(\mathsf{R})$ and suppose that $(r)\subseteq(s)$. Then, by Proposition 4.2.61, $r=ds$ for some $d\in\mathsf{R}$. It must then hold that either d or s is a unit. By Proposition 4.2.61, if s is a unit then $(s)=\mathsf{R}$ and if d is a unit then $(r)=(s)$. Thus (r) is maximal.

(iii) Suppose that (r) is maximal in $P(\mathsf{R})$. Then r must be a nonzero nonunit by Proposition 4.2.61. Now suppose that $r=ab$ from which we deduce that $(r)\subseteq(a)$, again by Proposition 4.2.61. Since (r) is maximal, either $(a)=(r)$ or $(a)=\mathsf{R}$. In the first case $a = dr$ for some $d \in \mathsf{R}$ by Proposition 4.2.61, and so $r = ab = rdb$. By Proposition 4.2.33, $db = 1_{\mathsf{R}}$, whence b is a unit. In the second case, a is a unit by Proposition 4.2.61, and so we conclude that r is irreducible. ■

The following result gives some general relationships that one can infer about primes and irreducibles when R is an integral domain.

4.2.65 Proposition (Primes and irreducibles in integral domains) *If* R *is be an integral domain, and consider the following three statements concerning* $\mathrm{p} \in \mathsf{R}$*:*

(i) p *is prime;*

(ii) p *is irreducible;*

(iii) if $\mathrm{d}|\mathrm{p}$ *then either* d *is a unit or* p *and* d *are associates.*

Then (i) $\Longrightarrow$ *(ii)* $\Longrightarrow$ *(iii).*

Proof Suppose that p is prime and that $p = ab$. Then $p|(ab)$ and since p is prime, without loss of generality we can assert that $p|a$. Thus $a = qp$ for some $q \in \mathsf{R}$. Then $p = ab = aqp$ which implies that $aq = 1_{\mathsf{R}}$ by Proposition 4.2.33. Thus a is a unit, and so p is irreducible. This shows that (i) $\Longrightarrow$ (ii).

Let p be irreducible and suppose that $d|p$ so that $(p) \subseteq (d)$ by Proposition 4.2.61. By Theorem 4.2.64(ii) we have $(p) = (d)$ or $(d) = \mathsf{R}$. Now, by Proposition 4.2.61, in the first case we have $p|d$, and so, by Proposition 4.2.60, $p = ud$ for some unit u. In the second case, d is a unit by Proposition 4.2.61. This gives the implication (ii) $\Longrightarrow$ (iii). ■

4.2.10 Unique factorisation domains

In this section, as in our earlier sections dealing with special classes of rings, we will generalise a property of the ring of integers. The property of the integers that we will generalise is the prime factorisation where every positive integer can be written as a unique (up to ordering) product of prime numbers.

We can now state the main definition in this section.

4.2.66 Definition (Unique factorisation domain) A ***unique factorisation domain*** is an integral domain R such that:

(i) if $r \in \mathsf{R}$ is nonzero and not a unit, then there exists irreducible elements $f_1, \dots, f_k \in \mathsf{R}$ such that $r = f_1 \cdots f_k$;

(ii) if, for irreducible elements $f_1, \dots, f_k, g_1, \dots, g_l \in \mathsf{R}$, we have $f_1 \cdots f_k = g_1 \cdots g_l$, then $k = l$ and there exists $\sigma \in \mathfrak{S}_k$ such that $f_j | g_{\sigma(j)}$ and $g_{\sigma(j)} | f_j$ for each $j \in \{1, \dots, k\}$. •

The first part of the definition tells us that every nonzero element of R that is not a unit is expressible as a product of irreducibles. The second part of the definition, along with Proposition 4.2.60(viii), tells us that the expression as a product of irreducibles is unique up to order and the factors differing by a unit. It is often convenient to eliminate the ambiguity of knowing the irreducible factors only up to multiplication by units. Let us denote by $\mathscr{I}_{\mathsf{R}}$ the set of irreducible elements of a commutative unit ring. On $\mathscr{I}_{\mathsf{R}}$ define a relation by $p_1 \sim p_2$ if $p_2 = up_1$ for some unit u. In Exercise 4.2.19 the reader can show that this relation is an equivalence relation. The following definition develops some terminology associated to this.

4.2.67 Definition (Selection of irreducibles) Let R be a commutative unit ring and let $\mathscr{I}_{\mathsf{R}}$ be the set of irreducible elements in R, with $\sim$ the equivalence relation described above. A ***selection of irreducibles*** is a map $P\colon (\mathscr{I}_{\mathsf{R}}/\sim) \to \mathscr{I}_{\mathsf{R}}$ such that $P([p]) \in [p]$. We shall denote a selection of irreducibles P by $(p_a)_{a\in A_{\mathsf{R}}}$ where $A_{\mathsf{R}} = \mathscr{I}_{\mathsf{R}}/\sim$ and where $p_a = P(a)$. •

The following result encapsulates why the notion of a selection of irreducibles is valuable.

4.2.68 Proposition (Unique factorisation determined by selection of irreducibles) *If* R *is a unique factorisation domain and if* $(\mathrm{p}_{\mathrm{a}})_{\mathrm{a}\in\mathrm{A}_{\mathsf{R}}}$ *is a selection of irreducibles, then, given a nonzero nonunit* $\mathrm{r} \in \mathsf{R}$*, there exists unique* $\mathrm{p}_{\mathrm{a}_1}, \dots, \mathrm{p}_{\mathrm{a}_k} \in (\mathrm{p}_{\mathrm{a}})_{\mathrm{a}\in\mathrm{A}_{\mathsf{R}}}$ *and a unique unit* $\mathrm{u} \in \mathsf{R}$ *such that* $\mathrm{r} = \mathrm{u}\mathrm{p}_{\mathrm{a}_1}\cdots\mathrm{p}_{\mathrm{a}_k}$.

Proof Let $f_1, \dots, f_k$ be irreducibles such that $r = f_1 \cdots f_k$. For $j \in \{1, \dots, k\}$, define $p_{a_j} \in \{p_a \mid a \in A_{\mathsf{R}}\}$ so that $p_{a_j} \in [f_j]$. Then $f_j = u_j p_{a_j}$ for some unit $u_j \in \mathsf{R}$, $j \in \{1, \dots, k\}$. Then we have

$$r = u_1 \cdots u_k p_{a_1} \cdots p_{a_k},$$

giving the existence part of the result. Now suppose that

$$r = up_{a_1} \cdots p_{a_k} = u' p_{a'_1} \cdots p_{a'_{k'}} \tag{4.4}$$

are two representations of the desired form. Since $(up_{a_1})p_{a_2} \cdots p_{a_k}$ and $(u'p_{a'_1})p_{a_2} \cdots p_{a'_{k'}}$ are two factorisations by irreducibles, we immediately conclude that $k' = k$. For convenience, let us define $f_1 = up_{a_1}$, $f_j = p_{a_j}$, $j \in \{2, \dots, k\}$, and $f'_1 = u'p_{a'_1}$, $f'_j = p_{a'_j}$, $j \in \{2, \dots, k\}$. We can then assert the existence of $\sigma \in \mathfrak{S}_k$ such that $f'_j = v_j f_{\sigma(j)}$, $j \in \{1, \dots, k\}$, for units $v_1, \dots, v_k$. By the definition of a selection of irreducibles, for each $j \in \{1, \dots, k\}$ we have $f'_j = v_j f_{\sigma(j)} = w_j p_{b_j}$ for a unit w_j and $p_{b_j} \in (p_a)_{a\in A_{\mathsf{R}}}$. Now we have a few cases.

1. $f'_1 = u'p_{a'_1} = v_1 f_1 = v_1 u p_{a_1}$: In this case we have

$$u'p_{a'_1} = uv_1 p_{a_1} = w_1 p_{b_1}.$$

We conclude that $p_{a'_1} \sim p_{a_1} \sim p_{b_1}$, implying that $p_{a'_1} = p_{a_1} = p_{b_1}$. By Proposition 4.2.33 we also conclude that $u' = uv_1 = w_1$.

2. $f'_1 = u'p_{a'_1} = f_{\sigma(1)} = p_{a_{\sigma(1)}}$ for $\sigma(1) \neq 1$: Here we have

$$u'p_{a'_1} = v_1 p_{a_{\sigma(1)}} = w_1 p_{b_1},$$

and as above we conclude that $p_{a'_1} = p_{a_{\sigma(1)}} = p_{b_1}$ and $u' = v_1 = w_1$.

3. $f'_j = p_{a'_j} = v_j f_1 = v_j u p_1 = w_j p_{b_j}$ for $j \neq 1$: Here we conclude that $p_{a'_j} = p_1 = p_{b_j}$ and $1_{\mathsf{R}} = v_j u = w_j$.

4. $f'_j = p_{a'_j} = v_j f_{\sigma(j)} = v_j p_{a_j} = w_j p_{b_j}$ for $j \neq 1$ and $\sigma(j) \neq 1$: In this case we conclude that $p_{a'_j} = p_{a_j} = p_{b_j}$ and $1_{\mathsf{R}} = v_j = w_j$.

We then see that $p_{a'_1} \cdots p_{a'_k} = p_{a_{\sigma(1)}} \cdots p_{a_{\sigma(k)}}$, and we immediately conclude from (4.4) and Proposition 4.2.33 that $u' = u$, and this completes the proof of uniqueness. ■

Although we have not yet proven that $\mathbb{Z}$ is a unique factorisation domain, the reader has probably at least been told this at some point, so let us use this as an example to illustrate the definition. That $\mathbb{Z}$ is, in fact, a unique factorisation domain will follow from Corollary 4.2.73 below.

4.2.69 Example ($\mathbb{Z}$ as a unique factorisation domain) Let us assume for the moment the following fact: every positive integer not equal to 1 can be written as a product of positive prime integers. It then follows that every integer $j \in \mathbb{Z} \setminus \{-1, 0, 1\}$ can be written as

$$j = (\pm p_1) \cdots (\pm p_k)$$

for positive prime integers $p_1, \dots, p_k$. This expression of j as a product of primes is now unique up to the order, and up to the use of "+" or "−" in each of the factors. When one recalls that the only units in $\mathbb{Z}$ are 1 and −1, then this ambiguity of the signs in each of the factors corresponds to the fact that, in the definition of a unique factorisation domain, one only knows the terms in the factorisation up to multiplication by a unit. This ambiguity is typically resolved by the standard selection of irreducibles given by $P \colon \mathscr{I}_{\mathbb{Z}}/\sim \to \mathscr{I}_{\mathbb{Z}}$ having the property that $P([p]) = |p|$; thus we select the positive of the two primes in the same equivalence class. If we denote this selection of primes by $(p_a)_{a \in A_{\mathsf{R}}}$, then we can write any integer j as

$$j = \pm p_1 \cdots p_k$$

where $p_1, \dots, p_k$ are positive primes. This is the prime factorisation that we learn in school. •

For unique factorisation domains we have the following valuable characterisation of primes and irreducibles. This sharpens the conclusions of Proposition 4.2.65 in the case of a unique factorisation domain.

4.2.70 Proposition (Primes and irreducibles in unique factorisation domains) *If* R *is be a unique factorisation domain, then the following three statements concerning* $\mathsf{p} \in \mathsf{R}$ *are equivalent:*

(i) p *is prime;*

(ii) p *is irreducible;*

(iii) if $\mathsf{d}|\mathsf{p}$ *then either* d *is a unit or* p *and* d *are associates.*

Proof From Proposition 4.2.65 it only remains to show that (iii) $\implies$ (i). Suppose that $p|(ab)$ and that p satisfies (iii). Using the properties of a unique factorisation domain, write $p = f_1 \dots f_k$ for irreducibles $f_1, \dots, f_k$. It follows from (iii) that $k = 1$ since irreducibles are not units. Now write

$$a = g_1 \dots g_l, \quad b = h_1 \dots h_m$$

for irreducibles $g_1, \dots, g_l, h_1, \dots, h_m$. Then there exists $r \in \mathsf{R}$ such that

$$rf_1 = g_1 \dots g_l h_1 \dots h_m.$$

Now write r as a product of irreducibles: $r = s_1 \cdots s_q$. Then

$$s_1 \cdots s_q f_1 = g_1 \dots g_l h_1 \dots h_m.$$

Using the definition of a unique factorisation domain we conclude that, for some $a \in \{g_1, \dots, g_l, h_1, \dots, h_m\}$, we have $f_1 | a$ and $a | f_1$. This allows us to conclude that either $p|a$ or $p|b$. Thus p is prime. ■

The main result of this section is now the following.

4.2.71 Theorem (Principal ideal domains are unique factorisation domains) *If* R *is a principal ideal domain, then it is a unique factorisation domain.*

Proof Denote by $B(\mathsf{R}) \subseteq \mathsf{R}$ those nonzero nonunits of R that cannot be factored as in part (i) of Definition 4.2.66 and suppose that $B(\mathsf{R})$ is nonempty. Let $r \in B(\mathsf{R})$. By Proposition 4.2.61, $(r) \subset \mathsf{R}$. Let (s_r) be the maximal ideal containing (r) which exists by Theorem 4.2.19. By Theorem 4.2.64(ii), s_r is irreducible and, by Proposition 4.2.61, $s_r | r$. Thus we can write $r = s_r a_r$ for some $a_r \in \mathsf{R}$.

We claim that $a_r \in B(\mathsf{R})$. First of all, a_r is nonzero, and moreover it is a nonunit, since if it were a unit then the relationship $r = s_r a_r$ implies that r is irreducible by Exercise 4.2.18. If $a_r \notin B(\mathsf{R})$ then a_r has a factorisation by irreducibles, and since s_r is irreducible, so too does r. Thus $a_r \in B(\mathsf{R})$.

Now for any $r \in B(\mathsf{R})$ use the Axiom of Choice to select an irreducible s_r and an element $a_r \in B(\mathsf{R})$ such that $r = s_r a_r$. Now we define a map $f \colon B(\mathsf{R}) \to B(\mathsf{R})$ by $f(r) = a_r$.

Next we claim that $(r) \subset (a_r)$. By Proposition 4.2.61 we have $(r) \subseteq (a_r)$. If $(r) = (a_r)$ then $a_r = u_r r$ for some unit u_r by Propositions 4.2.60 and 4.2.61. Then $r = s_r a_r = s_r u_r r$ which gives $s_r u_r = 1_{\mathsf{R}}$ by Proposition 4.2.33. This would imply that s_r is a unit, and so not in $B(\mathsf{R})$. Thus we conclude that $(r) \subset (a_r)$.

Now, taking some $r \in B(\mathsf{R})$, recursively define $g \colon \mathbb{Z}_{>0} \to B(\mathsf{R})$ by $g(1) = r$ and $g(k+1) = f(g(k))$. Denote $r_j = g(j)$ for $j \in \mathbb{Z}_{>0}$. Our above constructions then give a sequence $((r_j))_{j \in \mathbb{Z}_{>0}}$ of ideals having the property that

$$(r_1) \subset (r_2) \subset \cdots \subset (r_k) \subset \cdots.$$

This contradicts Theorem 4.2.57, and so we conclude that $B(\mathsf{R})$ is finite. Therefore part (i) of Definition 4.2.66 holds for a principal ideal domain.

Now we show the uniqueness of the factorisation into irreducibles. First we prove a lemma.

1 Lemma *If* R *is a principal ideal domain and if* $\mathrm{p} \in \mathsf{R}$ *is irreducible, then* p *is prime.*

Proof If p is irreducible, then (p) is maximal by Theorem 4.2.64 and since R is a principal ideal domain. Therefore, by Proposition 4.2.21, (p) is prime. Another application of Theorem 4.2.64 gives p as prime. ▼

Now let r be a nonzero nonunit and write

$$r = f_1 \cdots f_k = g_1 \cdots g_l$$

for irreducibles $f_1, \dots, f_k, g_1, \dots, g_l$. By the lemma, f_1 is prime. Therefore, $f_1 | (g_1 \dots g_k)$, and we conclude that $f_1 | g_{j_1}$ for some $j_1 \in \{1, \dots, l\}$. By Proposition 4.2.70 it must therefore be the case that $f_1 = u_1 g_{j_1}$ for some unit u_1. We now have

$$g_{j_1} u_1 f_2 \dots f_k = g_{j_1} g_1 \dots g_{j_1 - 1} g_{j_1 + 1} g_l,$$

which gives, by Proposition 4.2.33,

$$u_1 f_2 \dots f_k = g_1 \dots g_{j_1 - 1} g_{j_1 + 1} g_l.$$

We can now proceed as above to deduce that $u_1 f_2$ is a prime which divides g_{j_2} for some $j_2 \in \{1, \dots, l\} \setminus \{j_1\}$. Repeating this k times gives part (ii) of Definition 4.2.66. ■

Combining this with Theorem 4.2.55 gives the following useful result.

4.2.72 Corollary (Euclidean domains are unique factorisation domains) *If* R *is a Euclidean domain, then it is a unique factorisation domain.*

Of course, this now also gives the prime factorisation for integers that we have already used many times.

4.2.73 Corollary (Prime factorisation in $\mathbb{Z}$) *If* $\mathrm{j} \in \mathbb{Z} \setminus \{-1, 0, 1\}$, *then there exists positive prime integers* $\mathrm{p}_1, \dots, \mathrm{p}_k$ *and* $\mathrm{l}_1, \dots, \mathrm{l}_k \in \mathbb{Z}_{>0}$ *such that* $\mathrm{j} = \mathrm{u}\mathrm{p}_1^{\mathrm{l}_1} \dots \mathrm{p}_k^{\mathrm{l}_k}$, *where* $\mathrm{u} \in \{-1, 1\}$.

The converse of Theorem 4.2.71 is not generally true, as the following example illustrates.

4.2.74 Example (A unique factorisation domain that is not a principal ideal domain) For this example we shall require some concepts from Section 4.4.

We claim that the polynomial ring $\mathbb{Z}[\xi]$ is a unique factorisation domain, but not a principal ideal domain. That $\mathbb{Z}[\xi]$ is a unique factorisation domain is stated below as Corollary 4.4.21.

To see that $\mathbb{Z}[\xi]$ is not a principal ideal domain, consider the ideal $(2, \xi)$ generated by two elements. We claim that this ideal is not principal. Indeed, suppose that $(2, \xi) = (P)$ for some polynomial P with integer coefficients. Then, by Exercise 4.2.16, given any $P_1, P_2 \in \mathbb{Z}[\xi]$, we must have $Q \in \mathbb{Z}[\xi]$ such that

$$2P_1(\xi) + P_2(\xi)\xi = Q(\xi)P(\xi).$$

In particular, taking $P_1(\xi) = 0$ and $P_2(\xi) = 1$ there exists $Q_1 \in \mathbb{Z}[\xi]$ such that $\xi = Q_1(\xi)P(\xi)$. This means that we must have either

1. $Q_1(\xi) \in \{1, -1\}$ and $P(\xi) \in \{\xi, -\xi\}$ or
2. $Q_1(\xi) \in \{\xi, -\xi\}$ and $P(\xi) \in \{1, -1\}$.

We cannot have $P(\xi) \in \{1, -1\}$ since, in this case, this would imply that $(P) = \mathbb{Z}[\xi]$ by Proposition 4.2.61(iii). Thus we must have $P(\xi) \in \{\xi, -\xi\}$. Also, taking $P_1(\xi) = 1$ and $P_2(\xi) = 0$, there exists $Q_2 \in \mathbb{Z}[\xi]$ such that $2 = Q_2(\xi)P(\xi)$. Thus we must have either

1. $Q_2(\xi) \in \{2, -2\}$ and $P(\xi) = \{1, -1\}$ or
2. $Q_2(\xi) \in \{1, -1\}$ and $P(\xi) \in \{2, -2\}$.

We still cannot have $P(\xi) \in \{1, -1\}$, and so we must have $P(\xi) \in \{2, -1\}$, giving a contradiction. Thus $\mathbb{Z}[\xi]$ is not a principal ideal domain. •

4.2.11 Greatest common divisors, least common multiples, and the Euclidean Algorithm

The notion of a greatest common divisor for two integers is probably known to the reader from their school studies. In this section we explore this concept in a more general setting that will prove useful to us at various points in the text.

4.2.75 Definition (Greatest common divisor) Let R be a commutative ring and let $S \subseteq \mathsf{R}$. A ***greatest common divisor*** for S is an element $d \in \mathsf{R}$ such that

(i) $d|a$ for every $a \in S$ and

(ii) if $d'|a$ for every $a \in S$ then $d'|d$.

If R is additionally a unit ring $S = \{a_1, \dots, a_k\}$ and if 1_{R} is a greatest common divisor for S then the elements $a_1, \dots, a_k$ are ***relatively prime*** or ***coprime***. •

While in the rings that one encounters early in life it is the case that greatest common divisors exist, generally this is not the case. And when a greatest common divisor exists, it is not typically not unique.

4.2.76 Examples (Greatest common divisors)

1. Consider the ring $2\mathbb{Z}$. Note that if S is any subset for which $2 \in S$, then S has no greatest common divisor since 2 has no divisors.
2. In the ring $\mathbb{Z}$ consider the set $S = \{-8, 36\}$. The divisors shared by -8 and 36 are 2, 4, -2, and -4. Note that neither 2 nor -2 are a greatest common divisor for S since 4 divides all elements of S but 4 divides neither 2 nor -2. However, both 4 and -4 are greatest common divisors. Note that these two greatest common divisors differ only by multiplication by a unit in $\mathbb{Z}$. This is a feature of greatest common divisors in general; see Exercise 4.2.21. In some cases one has a means of distinguishing a member of the set of greatest common divisors, and this distinguished member is called *the* greatest common divisor. For example, in $\mathbb{Z}$ it suffices to ask that the greatest common divisor be positive to uniquely distinguish it.
3. Consider the ring $\mathbb{R}$ and let $S \subseteq \mathbb{R}$ be any set of real numbers. If $d \in \mathbb{R}^*$ and $x \in S$ then we have $x = (xd^{-1})d$, so d is a common divisor for all elements of S. Now let $d_1, d_2 \in \mathbb{R}^*$. Then we have $d_1 = (d_1d_2^{-1})d_2$. Thus it turns out that every

element of $\mathbb{R}^*$ is a greatest common divisor for S. In this case there is such an abundance of greatest common divisors that the concept loses relevance. •

Since greatest common divisors do not generally exist, it becomes useful to assert conditions under which they do.

4.2.77 Proposition (Existence of greatest common divisors) *For* R *a commutative unit ring and for* $\mathsf{S} = \{\mathsf{a}_1, \dots, \mathsf{a}_k\} \subseteq \mathsf{R}$*, the following statements hold:*

(i) the following statements for $\mathsf{d} \in \mathsf{R}$ *are equivalent:*

(a) $(\mathsf{d}) = (\mathsf{a}_1, \dots, \mathsf{a}_k)$*;*

(b) d *is a greatest common divisor for* S *of the form* $\mathsf{d} = \mathsf{r}_1\mathsf{a}_1 + \dots + \mathsf{r}_k\mathsf{a}_k$ *for some* $\mathsf{r}_1, \dots, \mathsf{r}_k \in \mathsf{R}$*;*

(ii) if R *is a principal ideal ring then* S *possesses a greatest common divisor of the form* $\mathsf{d} = \mathsf{r}_1\mathsf{a}_1 + \dots + \mathsf{r}_k\mathsf{a}_k$ *for* $\mathsf{r}_1, \dots, \mathsf{r}_k \in \mathsf{R}$*;*

(iii) if R *is a unique factorisation domain then* S *possesses a greatest common divisor.*

Proof (i) Suppose that $(d) = (a_1, \dots, a_k)$. By Theorem 4.2.54 we have

$$(d) = \{rd \mid r \in \mathsf{R}\}.$$

Therefore, for each $j \in \{1, \dots, k\}$ there exists $s_j \in \mathsf{R}$ such that $a_j = s_j d$. Thus $d|a_j$ for each $j \in \{1, \dots, k\}$. By Exercise 4.2.16 we have

$$(a_1, \dots, a_k) = \{r_1a_1 + \dots + r_ka_k \mid r_1, \dots, r_k \in \mathsf{R}\}.$$

Therefore, if $r_1a_1 + \dots + r_ka_k \in (a_1, \dots, a_k$ we have

$$r_1a_1 + \dots + r_ka_k = (r_1s_1 + \dots + r_ks_k)d,$$

so d divides each element of $(a_1, \dots, a_k)$. If $d'|a_j$ for each $j \in \{1, \dots, k\}$ then $a_j = s'_jd'$ for some $s'_j \in \mathsf{R}$. Then, since $d \in (a_1, \dots, a_k)$, for some $r_j \in \mathsf{R}$, $j \in \{1, \dots, k\}$, we have

$$d = r_1a_1 + \dots + r_ka_k = (r_1s'_1 + \dots + r_ks'_k)d',$$

and so $d'|d$. Thus d is a greatest common divisor for S of the form $r_1a_1 + \dots + r_ka_k$.

Now suppose that $d = r_1a_1 + \dots + r_ka_k$ is a greatest common divisor for S for some $r_1, \dots, r_k \in \mathsf{R}$. Then $d \in (a_1, \dots, a_k)$ and so $(d) \subseteq (a_1, \dots, a_k)$. Now let $r'_1a_1 + \dots + r'_ka_k \in (a_1, \dots, a_k)$. Since $d|a_j$, $j \in \{1, \dots, k\}$, we have $a_j = s_jd$ for $s_j \in \mathsf{R}$. Then

$$r'_1a_1 + \dots + r'_ka_k = (r'_1s_1 + \dots + r'_ks_k)d,$$

and so $(a_1, \dots, a_k) \subseteq (d)$, giving this part of the result.

(ii) Since R is a principal ideal ring, there exists $d \in \mathsf{R}$ such that $(a_1, \dots, a_k) = (d)$, and the result then follows from part (i).

(iii) Let $(p_a)_{a\in A}$ be a selection of irreducibles. There will be a finite number of these, say $p_1, \dots, p_r$, such that

$$a_j = u_jp_1^{m_{j1}} \cdots p_r^{m_{jr}}$$

for $j \in \{1, \dots, k\}$ and $m_{j1}, \dots, m_{jr} \in \mathbb{Z}_{\geq 0}$, and for a unit $u_j \in \mathsf{R}$. For $l \in \{1, \dots, r\}$, denote

$$m_l = \min\{m_{1l}, \dots, m_{kl}\}.$$

We claim that $d = p_1^{m_1} \cdots p_r^{m_r}$ is a greatest common divisor for $\{a_1, \dots, a_k\}$. It is evident that $d|a_j$, $j \in \{1, \dots, k\}$. Suppose that $e|a_j$ for each $j \in \{1, \dots, k\}$. Then, by uniqueness of factorisation, $e = up_1^{n_1} \cdots p_r^{n_r}$ where $n_l \leq m_{jl}$ for each $l \in \{1, \dots, r\}$ and $j \in \{1, \dots, k\}$. Thus $n_l \leq m_l$, and so $e|d$. Thus d is a greatest common divisor. ■

Part (ii) of the preceding result has following corollary that is surprisingly useful, as we have seen already in Example 4.2.7–4, and as we shall see again in Corollary 4.4.36.

4.2.78 Corollary (Coprime elements in principal ideal rings) *If* R *is a principal ideal ring with identity then* $\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k} \in \mathsf{R}$ *are coprime if and only if there exists* $\mathrm{r}_1, \dots, \mathrm{r}_\mathrm{k} \in \mathsf{R}$ *such that* $\mathrm{r}_1\mathrm{a}_1 + \cdots + \mathrm{r}_\mathrm{k}\mathrm{a}_\mathrm{k} = 1_\mathsf{R}$.

The special case of the preceding result where a and b are coprime if and only if there exists $r, s \in \mathsf{R}$ such that $ra + sb = 1_\mathsf{R}$ is alternately called ***Euclid's Lemma*** since it appears (in slightly different form) in Euclid's *Elements*, or ***Bézout's identity*** after Bézout[2] who generalised Euclid's statement to principal ideal domains.

Closely related to the notion of greatest common divisor is the notion of least common multiple.

4.2.79 Definition (Least common multiple) Let R be a commutative ring and let $S \subseteq \mathsf{R}$. A ***least common multiple*** for S is an element $m \in \mathsf{R}$ such that

(i) $a|m$ for every $a \in S$ and

(ii) if $a|m'$ for every $a \in S$ then $m|m'$. •

As with the greatest common divisor, it is not generally the case that a subset of a ring will possess a least common multiple. Let us give some examples illustrating this.

4.2.80 Examples (Least common multiples)

1. As an example of a ring where two ring elements may not have a least common multiple, we work with the polynomial ring $\mathsf{R}[\xi]$ and the subring R_1 consisting of those polynomials of the form

$$a_0 + a_2\xi^2 + \cdots + a_k\xi^k, \qquad k \in \mathbb{Z}_{\geq 0} \setminus \{1\},\ a_0, a_2, \dots, a_k \in \mathsf{R}.$$

One can readily check that R_1 is a subring of $\mathsf{R}[\xi]$. Let us take $P_1 = \xi^2$ and $P_2 = \xi^3$. It is easy to see that the greatest common divisor of P_1 and P_2 is 1_R. We claim that P_1 and P_2 have no least common multiple. Suppose otherwise, and that P is a least common multiple for P_1 and P_2. Since $P_1|P_1P_2$ and $P_2|P_1P_2$, it

[2] Etienne Bézout (1730–1783) was a French mathematician who did work in the area of algebra.

follows that $P|P_1P_2$. We claim that $\frac{P_1P_2}{P}$ is a greatest common divisor for P_1 and P_2. Indeed, since $P_2|P$,

$$P_1 = \frac{P_1P_2}{P}\frac{P}{P_2} \quad\Longrightarrow\quad \frac{P_1P_2}{P}\bigg| P_1.$$

Similarly, $\frac{P_1P_2}{P}|P_2$. Suppose now that $Q|P_1$ and $Q|P_2$. Then $QP_2|P_1P_2$ and $QP_1|P_1P_2$. Note that QP is a least common multiple for QP_1 and QP_2 since P is a least common multiple for P_1 and P_2. Therefore, $QP|P_1P_2$ and so $Q|\frac{P_1P_2}{P}$, showing that $\frac{P_1P_2}{P}$ is indeed a greatest common divisor of P_1 and P_2. But this means that $\frac{P_1P_2}{P} = 1$, whence $P = \xi^5$. However,

$$\xi^6 = \xi^2\xi^4 = \xi^3\xi^3$$

is a common multiple of ξ^2 and ξ^3. But $\xi^5 \nmid \xi^6$ in R_1, and we arrive at a contradiction.

2. Consider the ring $\mathbb{R}$. We consider two cases for a subset $S \subseteq \mathbb{R}$.

   (a) Let $S \subseteq \mathbb{R}^*$ be any nonempty set of nonzero real numbers. We claim that every nonzero real number is a least common multiple for S. Indeed, if $m \in \mathbb{R}^*$ and $x \in S$ we can write $m = x(x^{-1}m)$, which shows that $x|m$ for every $s \in S$. Also, if $m' \in \mathbb{R}^*$ then we can write $m' = m(m^{-1}m')$. Since this particularly holds when m' has the property that $x|m'$ for every $x \in S$, it follows that any $m \in \mathbb{R}^*$ is a least common multiple for S.

   (b) If $0 \in S$ then we claim that 0 is the only least common multiple for S. Indeed, we can clearly write $0 = x0$ for any $x \in S$, which shows that $x|0$ for every $x \in S$. Also, if $x|m'$ for every $x \in S$, we in particular have $m' = a0$ for some $a \in \mathbb{R}$; thus $m' = 0$. Therefore, $0|m'$, and this shows that 0 is a least common multiple for S. We also showed (when we showed that $m' = 0$) that 0 is the only least common multiple. •

It is possible to give a characterisation of least common multiples for principal ideal domains. This mirrors the corresponding results, parts (ii) and (iii) of Proposition 4.2.77, for greatest common divisors.

4.2.81 Proposition (Existence of least common multiples) *For* R *a commutative unit ring and for* $\mathsf{S} = \{\mathsf{a}_1, \dots, \mathsf{a}_k\} \subseteq \mathsf{R}$*, the following statements hold:*

(i) if R *is a principal ideal domain then* S *possesses a least common multiple* m *which satisfies* $(\mathrm{m}) = \cap_{\mathrm{j}=1}^{\mathrm{k}}(\mathrm{a}_\mathrm{j})$*;*

(ii) if R *is a unique factorisation domain then* S *possesses a least common multiple.*

Proof (i) Since R is a principal ideal domain, and since $\cap_{j=1}^{k}(a_j)$ is an ideal by Theorem 4.2.52, there does indeed exist $m \in \mathsf{R}$ such that $(m) = \cap_{j=1}^{k}(a_j)$. Thus we need only show that m is a least common multiple for S. Since $m \in (m) \subseteq (a_j)$ for every $j \in \{1, \dots, k\}$, we can write $m = r_ja_j$ for some $r_j \in \mathsf{R}$. Thus $a_j|m$ for every $j \in \{1, \dots, k\}$.

Moreover, suppose that $a_j|m'$ for every $j \in \{1,\dots,k\}$. Then $m' = r'_j a_j$ for some $r_j \in \mathsf{R}$ for each $j \in \{1,\dots,k\}$. By Theorem 4.2.54 this means that $m' \in (a_j)$ for each $j \in \{1,\dots,k\}$. Thus $m' \in \cap_{j=1}^{k}(a_j)$, and so, again by Theorem 4.2.54, $m' = rm$ for some $r \in \mathsf{R}$. That is, $m|m'$, and so m is indeed a least common multiple.

(ii) Let $(p_a)_{a\in A}$ be a selection of irreducibles. There will be a finite number of these, say $p_1,\dots,p_r$, such that

$$a_j = u_j p_1^{m_{j1}} \cdots p_r^{m_{jr}}$$

for $j \in \{1,\dots,k\}$ and $m_{j1},\dots,m_{jr} \in \mathbb{Z}_{\geq 0}$, and for a unit $u_j \in \mathsf{R}$. For $l \in \{1,\dots,r\}$, denote

$$m_l = \max\{m_{1l},\dots,m_{kl}\}.$$

Similarly to the proof of Proposition 4.2.77(iii), one can show that $f = p_1^{m_1} \cdots p_r^{m_r}$ is a least common multiple for $\{a_1,\dots,a_k\}$. ■

Now let us turn to the matter of computing greatest common divisors (and, thus, least common multiples, at least in some cases). Note that part (ii) of Proposition 4.2.77 tells us that finite subsets $\{a_1,\dots,a_k\}$ of Euclidean rings possess greatest common divisors of the form $d = r_1a_1 + \cdots + r_ka_k$ for $r_1,\dots,r_k \in \mathsf{R}$. It turns out that there is also an algorithm for computing a greatest common divisor for a pair of ring elements in this case.

4.2.82 Theorem (Euclidean Algorithm) *Let* (R,δ) *be a Euclidean domain and let* $\mathrm{a},\mathrm{b} \in \mathsf{R}$ *with* $\mathrm{b} \neq 0_{\mathsf{R}}$. *Then there exists* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$, $\mathrm{q}_0,\mathrm{q}_1,\dots,\mathrm{q}_\mathrm{k} \in \mathsf{R}$, *and* $\mathrm{r}_0,\mathrm{r}_1,\dots,\mathrm{r}_\mathrm{k} \in \mathsf{R}\setminus\{0_{\mathsf{R}}\}$ *such that*

$$\begin{aligned}
\mathrm{a} &= \mathrm{q}_0\mathrm{r}_0 + \mathrm{r}_1, && \delta(\mathrm{r}_1) < \delta(\mathrm{r}_0),\\
\mathrm{r}_0 &= \mathrm{q}_1\mathrm{r}_1 + \mathrm{r}_2, && \delta(\mathrm{r}_2) < \delta(\mathrm{r}_1),\\
&\ \vdots \\
\mathrm{r}_{\mathrm{k}-2} &= \mathrm{q}_{\mathrm{k}-1}\mathrm{r}_{\mathrm{k}-1} + \mathrm{r}_\mathrm{k}, && \delta(\mathrm{r}_\mathrm{k}) < \delta(\mathrm{r}_{\mathrm{k}-1}),\\
\mathrm{r}_{\mathrm{k}-1} &= \mathrm{q}_\mathrm{k}\mathrm{r}_\mathrm{k}.
\end{aligned} \tag{4.5}$$

Moreover, r_k *so defined is a greatest common divisor for* $\{\mathrm{a},\mathrm{b}\}$.

Proof Take $r_0 = b$. A repeated application of the properties of a Euclidean domain ensures that there exists $q_j \in \mathsf{R}$, $j \in \mathbb{Z}_{\geq 0}$, and $r_j \in \mathsf{R}$, $j \in \mathbb{Z}_{>0}$, such that

$$\begin{aligned}
a &= q_0r_0 + r_1, && \delta(r_1) < \delta(r_0),\\
r_0 &= q_1r_1 + r_2, && \delta(r_2) < \delta(r_1),\\
&\ \vdots \\
r_{j-2} &= q_{j-1}r_{j-1} + r_j, && \delta(r_j) < \delta(r_{j-1}),\\
&\ \vdots
\end{aligned}$$

We need to prove that eventually $r_{k+1} = 0_{\mathsf{R}}$ for some $k \in \mathbb{Z}_{\geq 0}$. However, this follows from Proposition 4.2.41 since the sequence $(\delta(r_j))_{j\in\mathbb{Z}_{\geq 0}}$ is strictly decreasing.

Now we show that r_k is a greatest common divisor for $\{a, b\}$. We have $r_k | r_{k-1}$ by the last of equations (4.5). The second to last of these equations then gives $r_k | r_{k-2}$. Proceeding in this way we show that $r_k | r_j$, $j \in \{0, \dots, k-1\}$. In particular, $r_k | b$ and $r_k | a$ by the first of equations (4.5). Now suppose that $d | a$ and $d | b$. Then, by the first of equations (4.5), $d | (a - q_0 r_0)$ and so $d | r_1$. Using the second of equations (4.5) we similarly have $d | r_2$, and we may then proceed to show that $d | r_j$, $j \in \{0, 1, \dots, k\}$. Thus r_k is a greatest common divisor. ■

Let us illustrate the Euclidean Algorithm on an example.

4.2.83 Example (The Euclidean Algorithm) Let us consider the Euclidean domain $(\mathbb{Z}, \delta)$ with $\delta(j) = |j|$. Take $a = 762$ and $b = 90$. We then can readily compute

$$\begin{aligned} 762 &= 8 \cdot 90 + 42, \\ 90 &= 2 \cdot 42 + 6, \\ 42 &= 7 \cdot 6 \end{aligned}$$

We therefore conclude that 6 is a greatest common divisor for 762 and 90. •

The Euclidean Algorithm is also useful for computing an explicit form for Bézout's identity in Euclidean rings. The solutions to Bézout's identity also have a sometimes useful additional property.

4.2.84 Theorem (Bézout's identity using the Euclidean Algorithm) *If* (R, δ) *is a Euclidean domain and if* $\mathrm{a}, \mathrm{b} \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ *are coprime, let* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$*,* $\mathrm{q}_0, \mathrm{q}_1, \dots, \mathrm{q}_\mathrm{k} \in \mathsf{R}$*, and* $\mathrm{r}_0 = \mathrm{b}, \mathrm{r}_1, \dots, \mathrm{r}_{\mathrm{k}-1} \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ *be such that*

$$\begin{aligned} \mathrm{a} &= \mathrm{q}_0 \mathrm{r}_0 + \mathrm{r}_1, && \delta(\mathrm{r}_1) < \delta(\mathrm{r}_0), \\ \mathrm{r}_0 &= \mathrm{q}_1 \mathrm{r}_1 + \mathrm{r}_2, && \delta(\mathrm{r}_2) < \delta(\mathrm{r}_1), \\ &\ \vdots \\ \mathrm{r}_{\mathrm{k}-2} &= \mathrm{q}_{\mathrm{k}-1} \mathrm{r}_{\mathrm{k}-1} + \mathrm{u}, && \delta(\mathrm{u}) < \delta(\mathrm{r}_{\mathrm{k}-1}), \\ \mathrm{r}_{\mathrm{k}-1} &= \mathrm{q}_\mathrm{k} \mathrm{u}, \end{aligned}$$

where $\mathrm{u} \in \mathsf{R}$ *is a unit (this being the case since* a *and* b *are coprime). Then let* $\alpha_0 = 1_{\mathsf{R}}$ *and* $\beta_0 = -\mathrm{q}_{\mathrm{k}-1}$*, and recursively define* $\alpha_1, \dots, \alpha_{\mathrm{k}-1} \in \mathsf{R}$ *and* $\beta_1, \dots, \beta_{\mathrm{k}-1} \in \mathsf{R}$ *by*

$$\alpha_\mathrm{j} = \beta_{\mathrm{j}-1}, \quad \beta_\mathrm{j} = \alpha_{\mathrm{j}-1} - \mathrm{q}_{\mathrm{k}-1-\mathrm{j}} \beta_{\mathrm{j}-1}, \qquad \mathrm{j} \in \{1, \dots, \mathrm{k}-1\}.$$

If we take

$$\mathrm{r} = \begin{cases} 0_{\mathsf{R}}, & \delta(\mathrm{b}) = \delta(1_{\mathsf{R}}), \\ \mathrm{u}^{-1} \alpha_{\mathrm{k}-1}, & \delta(\mathrm{b}) > \delta(1_{\mathsf{R}}), \end{cases} \qquad \mathrm{s} = \begin{cases} \mathrm{b}^{-1}, & \delta(\mathrm{b}) = \delta(1_{\mathsf{R}}), \\ \mathrm{u}^{-1} \beta_{\mathrm{k}-1}, & \delta(\mathrm{b}) > \delta(1_{\mathsf{R}}), \end{cases}$$

then $\mathrm{ra} + \mathrm{sb} = 1_{\mathsf{R}}$*.*

Moreover, if $\mathsf{S} \subseteq \mathsf{R}$ *is a nontrivial, δ-closed, and δ-positive subsemiring, and if* a *and* b *additionally have the property that* $\mathrm{a}, \mathrm{b} \in \mathsf{S}$ *and that at least one of* a *and* b *is not a unit, then* *(i)* $q_0, q_1 \dots, q_k$ *and* $r_1, \dots, r_{k-1}$ *may be chosen to lie in* S *and,*

(ii) *if* $q_0, q_1 \dots, q_k$ *and* $r_1, \dots, r_{k-1}$ *are so chosen, then* r *and* s *as defined above additionally satisfy* $\delta(r) < \delta(b)$ *and* $\delta(s) < \delta(a)$.

Proof Let us first reduce to the case when $u = 1_R$. Multiply all equations in the Euclidean Algorithm for a and b by u^{-1}:

$$\begin{aligned}
u^{-1}a &= q_0u^{-1}r_0 + u^{-1}r_1, && \delta(u^{-1}r_1) < \delta(u^{-1}r_0),\\
u^{-1}r_0 &= q_1u^{-1}r_1 + u^{-1}r_2, && \delta(u^{-1}r_2) < \delta(u^{-1}r_1),\\
&\vdots\\
u^{-1}r_{k-2} &= q_{k-1}u^{-1}r_{k-1} + 1_R, && \delta(1_R) < \delta(u^{-1}r_{k-1}),\\
u^{-1}r_{k-1} &= q_k.
\end{aligned}$$

Note that the resulting equations hold if and only if the original equations hold, by virtue of R being an integral domain. The resulting equations are then the Euclidean Algorithm for $u^{-1}a$ and $u^{-1}b$, and at each step the remainders $r_0, r_1, \dots, r_{k-1}$ are multiplied by u^{-1}. The quotients $q_0, q_1, \dots, q_k$ remain the same, however. Thus the definitions of $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ and $\beta_0, \beta_1, \dots, \beta_{k-1}$ are unchanged from the Euclidean Algorithm for a and b. Applying the conclusions of the theorem to the modified Euclidean Algorithm then gives $r', s' \in R$ such that $r'(u^{-1}a) + s'(u^{-1}b) = 1_R$. Thus the conclusions of the first part of the theorem in the general case follow from those when $u = 1_R$ by taking $r = u^{-1}r'$ and $s = u^{-1}s'$. Also note by Proposition 4.2.41 that the relation $\delta(u^{-1}r_{j-1}) < \delta(u^{-1}r_j)$ is equivalent to the relation $\delta(r_{j-1}) < \delta(r_j)$, $j \in \{0, 1, \dots, k-1\}$. Therefore, the conclusions of the second part of the theorem in the general case also follow from those for the case when $u = 1_R$. Thus, in the remainder of the proof we suppose that $u = 1_R$.

Let us also eliminate the case where $\delta(b) = \delta(1_R)$. If this is the case then we have $a = qb + r$ with $\delta(r) = \delta(0_R)$, and so $r = 0_R$. Therefore, since b is a unit by Proposition 4.2.41 $q = ab^{-1}$. Now, taking $r = 0_R$ and $s = b^{-1}$, we have $ra + sb = 1_R$. Moreover, for the second part of the theorem, $\delta(r) < \delta(b)$ and $\delta(s) < \delta(a)$ since s is a unit and a is not, the latter by the hypotheses of the theorem. Thus the conclusions of the theorem hold when $\delta(b) = \delta(1_R)$. Thus, in the remainder of the proof we suppose that b is a nonzero nonunit.

We now prove the theorem by induction on k. If $k = 1$ then we have

$$\begin{aligned}
a &= q_0 \cdot r_0 + 1_R, && \delta(1_R) < \delta(r_0),\\
r_0 &= q_1.
\end{aligned}$$

Thus

$$1_R = 1_R \cdot a + (-q_0) \cdot b,$$

and the theorem holds with $r = \alpha_0 = 1_R$ and $s = \beta_0 = -q_0$. Now suppose the theorem true for $k \in \{1, \dots, m-1\}$ and consider the Euclidean Algorithm for a and $b = r_0$ of the form

$$\begin{aligned} a &= q_0 r_0 + r_1, && \delta(r_1) < \delta(r_0),\\ r_0 &= q_1 r_1 + r_2, && \delta(r_2) < \delta(r_1),\\ &\vdots \\ r_{m-2} &= q_{m-1} r_{m-1} + 1_{\mathsf{R}}, && \delta(1_{\mathsf{R}}) < \delta(r_{m-1}),\\ r_{m-1} &= q_m. \end{aligned}$$

By the induction hypothesis, the conclusions of the theorem hold for the last m equations. But the last m equations are the result of applying the Euclidean Algorithm in the case where "$a = r_0$" and "$b = r_1$." Thus, if we define $\alpha_0 = 1_{\mathsf{R}}$ and $\beta_0 = -q_{k-1}$, and recursively define $\alpha_1, \dots, \alpha_{m-2}$ and $\beta_1, \dots, \beta_{m-2}$ by

$$\alpha_j = \beta_{j-1}, \quad \beta_j = \alpha_{j-1} - q_{m-1-j}\beta_{j-1}, \qquad j \in \{1, \dots, m-2\},$$

and if we take $r' = \alpha_{m-2}$ and $s' = \beta_{m-2}$, then we have $r'r_0 + s'r_1 = 1_{\mathsf{R}}$. Since $r_0 = b$ we have

$$1_{\mathsf{R}} = \alpha_{m-2} r_0 + \beta_{m-2}(a - q_0 r_0) = (\alpha_{m-2} - q_0\beta_{m-2})b + \beta_{m-2}a,$$

and so the theorem holds with $r = \alpha_{m-1} = \beta_{m-2}$ and $s = \beta_{m-1} = \alpha_{m-2} - q_0\beta_{m-2}$, as desired.

Now we proceed to the second part of the theorem, supposing that $a, b \in \mathsf{S}$ for a δ-closed and δ-positive subsemiring $\mathsf{S} \subseteq \mathsf{R}$. Since $r_0 = b$, that q_0 and r_1 can be chosen to lie in S follows from the fact that S is δ-closed. This reasoning can then be applied to each line of the Euclidean Algorithm to ensure that all quotients and remainders can be chosen to lie in S. The following lemma records a useful property of these quotients and remainders.

1 Lemma *Using the notation of the theorem statement, suppose that* $\mathrm{a}, \mathrm{b} \in \mathsf{S}$ *and that* $\mathrm{q}_0, \mathrm{q}_1, \dots, \mathrm{q}_\mathrm{k}$ *and* $\mathrm{r}_1, \dots, \mathrm{r}_{\mathrm{k}-1}$ *are chosen to lie in* S*. Then, for* $\mathrm{j} \in \{0, 1, \dots, \mathrm{k}-1\}$*, either*

(i) $\alpha_\mathrm{j} \in \mathsf{S}$ *and* $-\beta_\mathrm{j} \in \mathsf{S}$ *or*

(ii) $-\alpha_\mathrm{j} \in \mathsf{S}$ *and* $\beta_\mathrm{j} \in \mathsf{S}$.

Proof The lemma is proved by induction on j. For $j = 0$ we have $\alpha_0 = 1_{\mathsf{R}} \in \mathsf{S}$ and $-\beta_0 = q_{k-1} \in \mathsf{S}$. Suppose the lemma true for $j \in \{0, 1, \dots, m\}$. We have two cases.

1. $\alpha_m \in \mathsf{S}$ and $-\beta_m \in \mathsf{S}$: We immediately have $-\alpha_{m+1} = -\beta_m \in \mathsf{S}$. Also, $\beta_{m+1} = \alpha_m - q_{k-2-m}\beta_m \in \mathsf{S}$ since $\alpha_m \in \mathsf{S}$ and $q_{k-m-2}(-\beta_m) \in \mathsf{S}$, using the semiring property of S.
2. $-\alpha_m \in \mathsf{S}$ and $\beta_m \in \mathsf{S}$: This case follows, *mutatis mutandis*, in the manner of the previous case. ▼

Now, the final thing we need to show is that r and s constructed as above from $a, b \in \mathsf{S}$ satisfy $\delta(r) < \delta(b)$ and $\delta(s) < \delta(a)$. We prove this by induction on k. For $k = 1$ we have $r = 1_{\mathsf{R}}$ and $s = -q_0$. Therefore,

$$\delta(r) = \delta(1_{\mathsf{R}}) < \delta(b)$$

since we are assuming that b is a nonzero nonunit. Also, since b is a nonzero nonunit,

$$\delta(s) = \delta(-q_0) < \delta(-q_0 b) = \delta(a - 1_{\mathsf{R}}) \leq \max\{\delta(a), \delta(1_{\mathsf{R}})\} \leq \delta(a),$$

using δ-positivity of S. So the final assertion of the theorem holds for $k = 1$. Now suppose that this assertion holds for $k \in \{1, \dots, m-1\}$ and consider the Euclidean Algorithm for a and b of the form

$$\begin{aligned}
a &= q_0 r_0 + r_1, & \delta(r_1) &< \delta(r_0),\\
r_0 &= q_1 r_1 + r_2, & \delta(r_2) &< \delta(r_1),\\
&\vdots\\
r_{m-2} &= q_{m-1} r_{m-1} + 1_{\mathsf{R}}, & \delta(1_{\mathsf{R}}) &< \delta(r_{m-1}),\\
r_{m-1} &= q_m.
\end{aligned}$$

Considering the last m equations, as in the first part of the proof we have the Euclidean Algorithm for "$a = r_0$" and "$b = r_1$." Therefore, considering $r', s' \in \mathsf{R}$ as constructed in the first part of the proof, we have $\delta(r') < \delta(r_1)$ and $\delta(s') < \delta(r_0)$. Again as in the first part of the proof, we take $r = s'$ and $s = r' - q_0 s'$ so that $ra + sb = 1_{\mathsf{R}}$. Then

$$\delta(r) = \delta(s') < \delta(r_0) = \delta(b).$$

It remains to show that $\delta(s) < \delta(a)$. First suppose that $\delta(a) < \delta(b)$. Then, by Proposition 4.2.44 we have $a = 0_{\mathsf{R}} \cdot b + a$ as the unique output of the Division Algorithm in S. Thus we must have $q_0 = 0_{\mathsf{R}}$ and $r_1 = a$. In this case,

$$\delta(s) = \delta(r') < \delta(r_1) = \delta(a),$$

giving the norm bound for s if $\delta(a) < \delta(b)$. Thus we consider the case when $\delta(b) \leq \delta(a)$. By the lemma we have either (1) $r' \in \mathsf{S}$ and $-q_0 s' \in \mathsf{S}$ or (2) $-r' \in \mathsf{S}$ and $q_0 s' \in \mathsf{S}$. Consider the case $r', -q_0 s' \in \mathsf{S}$. We then have

$$a = q_0 b + r_1, \quad s = -q_0 s' + r'$$

with $\delta(r_1) < \delta(b)$, $\delta(s') < \delta(b)$, and $\delta(r') < \delta(r_1)$. Since $a, q_0, b, r_1, s, -s', r' \in \mathsf{S}$ we use Theorem 4.2.50 to write these elements of S as uniquely defined polynomials in x, where

$$\delta(x) = \inf\{\delta(r) \mid r \in \mathsf{S},\ \delta(r) > \delta(1_{\mathsf{R}})\}.$$

Let us denote these polynomials by P_a, P_{q_0}, P_b, P_{r_1}, P_s, $P_{-s'}$, and $P_{r'}$. By Theorem 4.2.50 we have

$$\begin{aligned}
\delta(r_1) < \delta(b) \quad &\Longrightarrow \quad \deg(P_{r_1}) < \deg(P_b),\\
\delta(s') < \delta(b) \quad &\Longrightarrow \quad \deg(P_{s'}) < \deg(P_b),\\
\delta(r') < \delta(r_1) \quad &\Longrightarrow \quad \deg(P_{r'}) < \deg(P_{r_1}).
\end{aligned}$$

This immediately gives

$$\deg(P_a) = \deg(P_{q_0}) + \deg(P_b), \quad \deg(P_s) \leq \max\{\deg(P_{q_0}) + \deg(P_{s'}), \deg(P_{r'})\}.$$

If

$$\max\{\deg(P_{q_0}) + \deg(P_{s'}), \deg(P_{r'})\} = \deg(P_{q_0}) + \deg(P_{s'})$$

then

$$\deg(P_a) = \deg(P_{q_0}) + \deg(P_b) > \deg(P_{q_0}) + \deg(P_{s'}) \geq \deg(P_s)$$

if

$$\max\{\deg(P_{q_0}) + \deg(P_{s'}), \deg(P_{r'})\} = \deg(P_{r'})$$

then

$$\deg(P_a) = \deg(P_{q_0}) + \deg(P_b) > \deg(P_{r_1}) > \deg(P_{r'}) \geq \deg(P_s).$$

In either case we have $\deg(P_s) > \deg(P_a)$, and then we apply Theorem 4.2.50 again to give $\delta(s) < \delta(a)$ in the case when $r', -q_0 s' \in \mathsf{S}$. When $-r', q_0 s' \in \mathsf{S}$ then $-s \in \mathsf{S}$ and we write

$$a = q_0 b + r_1, \quad -s = q_0 s' + (-r').$$

The steps above may now be repeated to give $\delta(s) = \delta(-s) < \delta(a)$ in this case. ■

The preceding theorem is constructive, and the next example illustrates how the construction can be made. The example also serves to illustrate the proof.

4.2.85 Example (Bézout's identity) We take $\mathsf{R} = \mathbb{Z}$ and consider the elements $a = 770 = 2 \cdot 5 \cdot 7 \cdot 11$ and $b = 39 = 3 \cdot 13$. These integers are coprime since they have no common prime factors. We seek $r, s \in \mathbb{Z}$ such that $ra + sb = 1$. Let us first apply the Euclidean Algorithm:

$$\begin{aligned} 770 &= 19 \cdot 39 + 29, \\ 39 &= 1 \cdot 29 + 10, \\ 29 &= 2 \cdot 10 + 9, \\ 10 &= 1 \cdot 9 + 1, \\ 9 &= 9. \end{aligned}$$

The quotients are $q_0 = 19$, $q_1 = 1$, $q_2 = 2$, $q_3 = 1$ and $q_4 = 1$, and the remainders are $r_0 = 39$, $r_1 = 29$, $r_2 = 10$, $r_3 = 9$, and $r_4 = 1$. One can now immediately apply the recursive formula from Theorem 4.2.84 to find r and s. But let us proceed directly, by way of illustrating what is really going on with Theorem 4.2.84. The second to last of the equations from the Euclidean Algorithm gives us an expression for 1:

$$1 = 10 - 9.$$

The right-hand side of this expression is of the form $\alpha r_2 + \beta r_3$. As we go along, we wish to maintain this structure. Next, the third to last (or the third) equation in the Euclidean Algorithm gives an expression for 9 that we substitute into the preceding equation:

$$1 = 10 - 1 \cdot (29 - 2 \cdot 10) = 3 \cdot 10 - 1 \cdot 29.$$

Note that we express this as $\alpha r_1 - \beta r_2$. The second of the equations from the Euclidean Algorithm gives an expression for 10 that we substitute into the preceding equation:

$$1 = 3 \cdot (39 - 1 \cdot 29) - 1 \cdot 29 = 3 \cdot 39 - 4 \cdot 29.$$

The right-hand side of this expression has the form $\alpha \cdot r_0 + \beta r_1$. Noting that $r_0 = b$, we now only need to use the first equation from the Euclidean Algorithm to involve a:

$$1 = 3 \cdot 39 - 4 \cdot (770 - 19 \cdot 39) = -4 \cdot 770 + 79 \cdot 39.$$

Thus we have the desired result by taking $r = -4$ and $s = 79$. Note that for the ring $\mathbb{Z}$ we used δ as defined by $\delta(k) = |k|$. Therefore, $\delta(j - k) \leq \delta(j) + \delta(k)$ for all $j, k \in \mathbb{Z}$. And sure enough, consistent with the second part of Theorem 4.2.84, we have $\delta(r) < \delta(b)$ and $\delta(s) < \delta(a)$. •

4.2.12 Notes

Our Example 4.2.56 of a principal ideal domain that is not a Euclidean domain comes from the paper of Motzkin [1949]. The proof we give follows that in the paper of Campoli [1988].

The very particular Examples 4.2.56 and 4.2.63–4 seem like they come out of thin air. In fact, they come from the general field of algebraic number theory, and in particular from the study of certain fields called “quadratic number fields.” We refer the reader to [Theory 1987, Chapter 3] for more details, and for some general discussion that better motivates the computations in these examples.

Exercises

4.2.1 Prove Proposition 4.2.3.

4.2.2 Prove Proposition 4.2.5.

4.2.3 For a ring R, prove the following:

(a) $0_R \cdot r = r \cdot 0_R = 0_R$ for all $r \in R$;

(b) if R has unit element 1_R, then $(-1_R) \cdot r = -r$ for all $r \in R$;

(c) $(-r_1) \cdot r_2 = r_1 \cdot (-r_2) = -(r_1 \cdot r_2)$ for all $r_1, r_2 \in R$.

4.2.4 Show that in the ring $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$ it holds that “$1 = -1$.”

4.2.5 Prove Proposition 4.2.13.

4.2.6 Let R and S be rings. Show that, if $\phi\colon R \to S$ is an isomorphism, then ϕ^{-1} is a homomorphism, and so also an isomorphism.

4.2.7 Show that $\phi\colon R \to S$ is a monomorphism of rings if and only if $\ker(\phi) = \{0_R\}$.

4.2.8 Let $\phi\colon R \to S$ be an epimorphism of rings. Show that the map

$$r + \ker(\phi) \mapsto \phi(r)$$

is an isomorphism of the rings $R/\ker(\phi)$ and S.

4.2.9 Show that if $\phi\colon \mathbb{Z}_k \to \mathbb{Z}$ is a homomorphism for $k \in \mathbb{Z}_{>0}$, then $\operatorname{image}(\phi) = \{0\}$.

4.2.10 Let R be a unit ring with characteristic $k \in \mathbb{Z}_{>0}$. Show that $k = \inf\{k \in \mathbb{Z}_{>0} \mid k \cdot 1_{\mathsf{R}} = 0_{\mathsf{R}}\}$.

4.2.11 For a commutative unit ring R, show that the following statements are equivalent:

(i) R is an integral domain;

(ii) if $r, s \in \mathsf{R}$ have the property that $rs = 0_{\mathsf{R}}$, then either $r = 0_{\mathsf{R}}$ or $s = 0_{\mathsf{R}}$.

4.2.12 Let R be a commutative ring. Show that the product of nonzerodivisors $r_1, r_2 \in \mathsf{R}$ is a nonzerodivisor.

4.2.13 Show that if (R, δ) is a Euclidean domain and if $a, b \in \mathsf{R}$ have the property that $\delta(ab) < \delta(a)$, then $ab = 0_{\mathsf{R}}$, and consequently that $a = 0_{\mathsf{R}}$ or $b = 0_{\mathsf{R}}$.

4.2.14 Consider the Euclidean domain $(\mathbb{Z}, \delta)$ with $\delta(j) = |j|$. Find $j, k \in \mathbb{Z}_{>0}$ such that there exists $q_1, q_2, r_1, r_2 \in \mathbb{Z}$ such that

1. $j = q_1 k + r_1 = q_2 k + r_2$, and
2. $q_1 \neq q_2$ and $r_1 \neq r_2$.

4.2.15 Let (R, δ) be a Euclidean domain, let $\mathsf{S} \subseteq \mathsf{R}$ be a nontrivial δ-closed, and δ-positive subsemiring, and let $a, b \in \mathsf{S}$ with $a \neq 0_{\mathsf{R}}$ and with b a nonzero nonunit.

(a) Show that there exists $k \in \mathbb{Z}_{>0}$, $q_0, q_1, \dots, q_{k-1} \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$, and $r_0, r_1, \dots, r_{k-1} \in \mathsf{R}$ such that

$$\begin{aligned} a &= q_0 b + r_0, & \delta(r_0) &< \delta(b), \\ q_0 &= q_1 b + r_1, & \delta(r_1) &< \delta(b), \\ &\vdots \\ q_{k-2} &= q_{k-1} b + r_{k-1}, & \delta(r_{k-1}) &< \delta(b), \\ q_{k-1} &= 0_{\mathsf{R}} b + q_{k-1}, & \delta(q_{k-1}) &< \delta(b). \end{aligned}$$

Hint: *Refer to the proof of Theorem 4.2.48 to show that*

$$\delta(\mathrm{a}) > \delta(\mathrm{q}_0) > \delta(\mathrm{q}_1) > \cdots > \delta(\mathrm{q}_{\mathrm{k}-1}).$$

(b) Show that if we take $r_k = q_{k-1}$ then

$$a = r_0 + r_1 b + r_2 b^2 + \cdots + r_k b^k.$$

4.2.16 In a commutative ring R with unit with $S \subseteq \mathsf{R}$, show that

$$(S) = \{r_1 a_1 + \cdots + r_k a_k \mid k \in \mathbb{Z}_{>0},\ r_1, \dots, r_k \in \mathsf{R},\ a_1, \dots, a_k \in S\}.$$

4.2.17 Prove Proposition 4.2.60.

4.2.18 Let R be an integral domain.

(a) Show that p is prime if and only if up is prime for every unit u.

(b) Show that p is irreducible if and only if up is irreducible for every unit u.

4.2.19 Let R be a commutative unit ring and let $\mathscr{I}_{\mathsf{R}}$ be the set of irreducible elements of R. Show that the relation "$p_1 \sim p_2$ if $p_2 = up_1$ for some unit u" is an equivalence relation.

4.2.20 Show that there are an infinite number of positive primes in $\mathbb{Z}$.
Hint: *Suppose that* $\mathrm{p}_1, \dots, \mathrm{p}_\mathrm{k}$ *are the first* k *primes, and consider the prime factorisation of* $\mathrm{p}_1 \cdots \mathrm{p}_\mathrm{k} + 1$.

4.2.21 Let R be a commutative ring and let $S \subseteq \mathsf{R}$. Show that if d_1 and d_2 are greatest common divisors for S then $d_1 | d_2$ and $d_2 | d_1$. Show that if R is additionally an integral domain then d_1 and d_2 are associates.

4.2.22 Show that, if R is a principal ideal domain and if $S \subseteq \mathsf{R}$, then there exists a greatest common divisor for S.

4.2.23 Let R be a commutative ring and let $S \subseteq \mathsf{R}$. Show that if m_1 and m_2 are least common multiples for S then $m_1 | m_2$ and $m_2 | m_1$. Show that if R is additionally an integral domain then m_1 and m_2 are associates.

4.2.24 Show that, if R is a principal ideal domain and if $S \subseteq \mathsf{R}$ has the property that $\cap_{s \in S}(s) \neq \{0_{\mathsf{R}}\}$, then there exists a least common multiple for S.

Section 4.3

Fields

In this section we consider a special sort of ring, one whose nonzero elements are units. These special rings, called fields, are important to us because they form the backdrop for linear algebra, and as such are distinguished in the set of rings.

Do I need to read this section? Readers who are familiar with the basic arithmetic properties of real and numbers can probably omit reading this section. Certain of the ideas we discuss here will be important in our discussion of polynomials in Section 4.4, and so a reader wishing to learn about polynomials might benefit from first understanding fields in the degree of generality we present them in this section. •

4.3.1 Definitions and basic properties

The definition of a field proceeds easily once one has on hand the notion of a ring. However, in our definition we repeat the basic axiomatic structure so a reader will not have to refer back to Definition 4.2.1.

4.3.1 Definition A ***division ring*** is a unit ring in which every nonzero element is a unit, and a ***field*** is a commutative division ring. Thus a field is a set F with two binary operations, $(a_1, a_2) \mapsto a_1 + a_2$ and $(a_1, a_2) \mapsto a_1 \cdot a_2$, called ***addition*** and ***multiplication***, respectively, and which together satisfy the following rules:

(i) $(a_1 + a_2) + a_3 = a_1 + (a_2 + a_3)$, $a_1, a_2, a_3 \in \mathsf{F}$ (***associativity*** of addition);

(ii) $a_1 + a_2 = a_2 + a_1$, $a_1, a_2 \in \mathsf{F}$ (***commutativity*** of addition);

(iii) there exists $0_{\mathsf{F}} \in \mathsf{F}$ such that $a + 0_{\mathsf{F}} = a$, $a \in \mathsf{F}$ (***additive identity***);

(iv) for $a \in \mathsf{F}$, there exists $-a \in \mathsf{F}$ such that $a + (-a) = 0_{\mathsf{F}}$ (***additive inverse***);

(v) $(a_1 \cdot a_2) \cdot a_3 = a_1 \cdot (a_2 \cdot a_3)$, $a_1, a_2, a_3 \in \mathsf{F}$ (***associativity*** of multiplication);

(vi) $a_1 \cdot a_2 = a_2 \cdot a_1$, $a_1, a_2 \in \mathsf{F}$ (***commutativity*** of multiplication);

(vii) $a_1 \cdot (a_2 + a_3) = (a_1 \cdot a_2) + (a_1 \cdot a_3)$, $a_1, a_2, a_3 \in \mathsf{F}$ (***left distributivity***);

(viii) there exists $1_{\mathsf{F}} \in \mathsf{F}$ such that $1_{\mathsf{F}} \cdot a = a$, $a \in \mathsf{F}$ (***multiplicative identity***);

(ix) for $a \in \mathsf{F}$, there exists $a^{-1} \in \mathsf{F}$ such that $a^{-1} \cdot a = 1_{\mathsf{F}}$ (***multiplicative inverse***);

(x) $(a_1 + a_2) \cdot a_3 = (a_1 \cdot a_3) + (a_2 \cdot a_3)$, $a_1, a_2, a_3 \in \mathsf{F}$ (***right distributivity***). •

The following result gives some properties of fields that follow from the definitions or which follow from general properties of rings.

4.3.2 Proposition (Basic properties of fields) *Let* F *be a field and denote* $\mathsf{F}^* = \mathsf{F} \setminus \{0_{\mathsf{F}}\}$*. Then the following statements hold:*

(i) F^**, equipped with the binary operation of multiplication, is a group;*

(ii) F *is an integral domain;*

(iii) F *is a Euclidean domain;*

(iv) F *is a principal ideal domain;*

(v) F *is a unique factorisation domain.*

4.3.3 Remark (Fields as unique factorisation domains) It is worth commenting on the nature of fields as unique factorisation domains. The definition of a unique factorisation domain requires that one be able to factor nonzero nonunits as products of irreducibles. However, in fields there are neither any nonzero nonunits, nor any irreducibles. Therefore, fields are vacuous unique factorisation domains. •

Let us give some examples of fields.

4.3.4 Examples (Fields)

1. $\mathbb{Z}$ is not a field since the only units are -1 and 1.
2. $\mathbb{Q}$ is a field.
3. $\mathbb{R}$ is a field.
4. The ring $\mathbb{Z}_k$ is a field if and only if k is prime. This follows from our discussion in Example 4.2.7–4 of the units in $\mathbb{Z}_k$. However, let us repeat the argument here, using Bézout's Identity in a coherent manner. We rely on the fact that $\mathbb{Z}$ is a Euclidean domain (Theorem 4.2.45), and so a principal ideal domain (Theorem 4.2.55), and so a unique factorisation domain (Theorem 4.2.71).

   Suppose that k is prime and let $j \in \{1, \dots, k-1\}$. Then 1 is a greatest common divisor for $\{j, k\}$, and by Corollary 4.2.78 this means that there exists $l, m \in \mathbb{Z}$ such that $lj + mk = 1$. Therefore, $(j + k\mathbb{Z})(l + k\mathbb{Z}) = lj + k\mathbb{Z} = 1 + k\mathbb{Z}$, and so $j + k\mathbb{Z}$ is a unit.

   Now suppose that $\mathbb{Z}_k$ is a field and let $j \in \{1, \dots, k-1\}$. Then there exists $l \in \{1, \dots, k-1\}$ such that $(j + k\mathbb{Z})(l + k\mathbb{Z}) = 1 + k\mathbb{Z}$. Therefore, $jl + mk = 1$ for some $m \in \mathbb{Z}$, and by Corollary 4.2.78 we can conclude that j and k are relatively prime. Since this must hold for every $j \in \{1, \dots, k-1\}$, it follows from Proposition 4.2.70 that k is prime. •

4.3.2 Fraction fields

Corresponding to a commutative unit ring is a natural field given by "fractions" in R. The construction here strongly resembles the construction of the rational numbers from the integers, so readers may wish to review Section 2.1.1.

4.3.5 Definition (Fraction field) Let R be an integral domain and define an equivalence relation $\sim$ in $\mathsf{R} \times (\mathsf{R} \setminus \{0_{\mathsf{R}}\})$ by

$$(r,s) \sim (r',s') \qquad \iff \qquad rs' - r's = 0_{\mathsf{R}}$$

(the reader may verify in Exercise 4.3.1 that $\sim$ is indeed an equivalence relation). The set of equivalence classes under this equivalence relation is the ***fraction field*** of R, and is denoted by F_{R}. The equivalence class of (r,s) is denoted by $\frac{r}{s}$. •

Let us show that the name fraction *field* is justified.

4.3.6 Theorem (The fraction field is a field) *If* R *is an integral domain, then* F_{R} *is a field when equipped with the binary operations of addition and multiplication defined by*

$$\frac{\mathrm{r}_1}{\mathrm{s}_2} + \frac{\mathrm{r}_2}{\mathrm{s}_2} = \frac{\mathrm{r}_1\mathrm{s}_2 + \mathrm{r}_2\mathrm{s}_1}{\mathrm{s}_1\mathrm{s}_1}, \qquad \frac{\mathrm{r}_1}{\mathrm{s}_1} \cdot \frac{\mathrm{r}_1 \cdot \mathrm{r}_2}{\mathrm{s}_1 \cdot \mathrm{s}_2}.$$

Moreover, the map $\mathrm{r} \mapsto \frac{\mathrm{r}}{1_{\mathsf{R}}}$ *is a ring monomorphism from* R *to* F_{R}*.*

Proof If one defines the zero element in the field to be $\frac{0_{\mathsf{R}}}{1_{\mathsf{R}}}$, the unity element to be $\frac{1_{\mathsf{R}}}{1_{\mathsf{R}}}$, the additive inverse of $\frac{r}{s}$ to be $\frac{-r}{s}$, and the multiplicative inverse of $\frac{r}{s}$ to be $\frac{s}{r}$, then it is a matter of tediously checking the conditions of Definition 4.3.1 to see that F_{R} is a field. The final assertion is also easily checked. We leave the details of this to the reader as Exercise 4.3.2. ■

The only interesting example of a fraction field that we have encountered thus is the field $\mathbb{Q}$ which is obviously the fraction field of $\mathbb{Z}$. In Section 4.4.8 we will encounter the field of rational functions that is associated with a polynomial ring.

4.3.3 Subfields, field homomorphisms, and characteristic

All of the ideas in this section have been discussed in the more general setting of rings in Section 4.2. Therefore, we restrict ourselves to making the (obvious) definitions and pointing out the special features arising when one restricts attention to fields.

Since fields are also rings, the following definition is the obvious one.

4.3.7 Definition (Subfield) A nonempty subset K of a field F is a ***subfield*** if K is a subring of the ring F that (1) contains 1_{F} and (2) contains a^{-1} for every $a \in \mathsf{K} \setminus \{0_{\mathsf{F}}\}$. •

Of course, just as in Definition 4.2.12, a subset $\mathsf{K} \subseteq \mathsf{F}$ is a subfield if and only if (1) $a_1 + a_2 \in \mathsf{K}$ for all $a_1, a_2 \in \mathsf{K}$, (2) $a_1 \cdot a_2 \in \mathsf{K}$ for all $a_1, a_2 \in \mathsf{K}$, (3) $-a \in \mathsf{K}$ for all $a \in \mathsf{K}$, (4) $1_{\mathsf{F}} \in \mathsf{K}$, and (4) $a^{-1} \in \mathsf{K}$ for all nonzero $a \in \mathsf{K}$. Note that we do require that 1_{F} be an element of a subfield so as to ensure that subfields are actually fields (see Exercises 4.3.3 and 4.3.4).

Note that we have not made special mention of ideals which were so important to our characterisations of rings. The reason for this is that ideals for fields are simply not very interesting, as the following result suggests.

4.3.8 Proposition (Ideals of fields) *If* R *is a commutative unit ring with more than one element, then the following statements are equivalent:*

(i) R *is a field;*

(ii) $\{0_{\mathsf{R}}\}$ *is a maximal ideal of* R*;*

(iii) if I *is an ideal of* R*, then either* $\mathsf{I} = \{0_{\mathsf{R}}\}$ *or* $\mathsf{I} = \mathsf{R}$*.*

Proof (i) $\implies$ (ii) Suppose that I is an ideal of R for which $\{0_{\mathsf{R}}\} \subseteq \mathsf{I}$. If $\{0_{\mathsf{R}}\} \neq \mathsf{I}$ then let $a \in \mathsf{I} \setminus \{0_{\mathsf{R}}\}$. For any $r \in \mathsf{R}$ we then have $r = (ra^{-1})a$, meaning that $r \in \mathsf{I}$. Thus $\mathsf{I} = \mathsf{R}$, and so $\{0_{\mathsf{R}}\}$ is maximal.

(ii) $\implies$ (iii) This follows immediately by the definition of maximal ideal.

(iii) $\implies$ (i) Let $r \in \mathsf{R} \setminus \{0_{\mathsf{R}}\}$ and consider the ideal (r). Since $(r) \neq \{0_{\mathsf{R}}\}$ we must have $(r) = \mathsf{R}$. In particular, $1_{\mathsf{F}} = rs$ for some $s \in \mathsf{R}$, and so r is a unit. ■

The interesting relationship between fields and ideals, then, does not come from considering ideals of fields. However, there is an interesting connection of fields to ideals. This connection, besides being of interest to us in Section 4.6.5, gives some additional insight to the notion of maximal ideals. The result mirrors that for prime ideals given as Theorem 4.2.37.

4.3.9 Theorem (Quotients by maximal ideals are fields, and vice versa) *If* R *is a commutative unit ring with more than one element and if* $\mathsf{I} \subseteq \mathsf{R}$ *is an ideal, then the following two statements are equivalent:*

(i) I *is a maximal ideal;*

(ii) R/I *is a field.*

Proof Denote by $\pi_{\mathsf{I}} \colon \mathsf{R} \to \mathsf{R}/\mathsf{I}$ the canonical projection. Suppose that I is a maximal ideal and let $\mathsf{J} \subseteq \mathsf{R}/\mathsf{I}$ be an ideal. We claim that

$$\tilde{\mathsf{J}} = \{r \in \mathsf{R} \mid \pi_{\mathsf{I}}(r) \in \mathsf{J}\}$$

is an ideal in R. Indeed, let $r_1, r_2 \in \tilde{\mathsf{J}}$ and note that $\pi_{\mathsf{I}}(r_1 - r_2) = \pi_{\mathsf{I}}(r_1) - \pi_{\mathsf{I}}(r_2) \in \mathsf{J}$ since π_{I} is a ring homomorphism and since J is an ideal. Thus $r_1 - r_2 \in \tilde{\mathsf{J}}$. Now let $r \in \tilde{\mathsf{J}}$ and $s \in \mathsf{R}$ and note that $\pi_{\mathsf{I}}(sr) = \pi_{\mathsf{I}}(s)\pi_{\mathsf{I}}(r) \in \mathsf{J}$, again since π_{I} is a ring homomorphism and since J is an ideal. Thus $\tilde{\mathsf{J}}$ is an ideal. Clearly $\mathsf{I} \subseteq \tilde{\mathsf{J}}$ so that either $\tilde{\mathsf{J}} = \mathsf{I}$ or $\tilde{\mathsf{J}} = \mathsf{R}$. In the first case $\mathsf{J} = \{0_{\mathsf{R}} + \mathsf{I}\}$ and in the second case $\mathsf{J} = \mathsf{R}/\mathsf{I}$. Thus the only ideals of R/I are $\{0_{\mathsf{R}} + \mathsf{I}\}$ and R/I. That R/I is a field follows from Proposition 4.3.8.

Now suppose that R/I is a field and let J be an ideal of R for which $\mathsf{I} \subseteq \mathsf{J}$. We claim that $\pi_{\mathsf{I}}(\mathsf{J})$ is an ideal of R/I. Indeed, let $r_1 + \mathsf{I}, r_2 + \mathsf{I} \in \pi_{\mathsf{I}}(\mathsf{J})$. Then $r_1, r_2 \in \mathsf{J}$ and so $r_1 - r_2 \in \mathsf{J}$, giving $(r_1 - r_2) + \mathsf{I} \in \pi_{\mathsf{I}}(\mathsf{J})$. If $r + \mathsf{I} \in \pi_{\mathsf{I}}(\mathsf{J})$ and if $s + \mathsf{I} \in \mathsf{R}/\mathsf{I}$, then $r \in \mathsf{J}$ and so $sr \in \mathsf{J}$. Then $sr + \mathsf{I} \in \pi_{\mathsf{I}}(\mathsf{J})$, thus showing that $\pi_{\mathsf{I}}(\mathsf{J})$ is indeed an ideal. Since R/I is a field, by Proposition 4.3.8 we may conclude that either $\pi_{\mathsf{I}}(\mathsf{J}) = \{0_{\mathsf{R}} + \mathsf{I}\}$ or that $\pi_{\mathsf{I}}(\mathsf{J}) = \mathsf{R}/\mathsf{I}$. In the first case we have $\mathsf{J} \subseteq \mathsf{I}$ and hence $\mathsf{J} = \mathsf{I}$, and in the second case we have $\mathsf{J} = \mathsf{R}$. Thus I is maximal. ■

The definition of a homomorphism of fields follows from the corresponding definition for rings.

4.3.10 Definition (Field homomorphism, epimorphism, monomorphism, and isomorphism) For fields F and K, a map $\phi\colon \mathsf{F} \to \mathsf{K}$ is a ***field homomorphism*** (resp. ***epimorphism, monomorphism, isomorphism***) if it is a homomorphism (resp. epimorphism, monomorphism, isomorphism) of rings. If there exists an isomorphism from F to K, then F and K are ***isomorphic***. •

The definitions of kernel and image for field homomorphisms are then special cases of the corresponding definitions for rings, and the corresponding properties also follow, just as for rings.

For fields one adopts the notion of characteristic from rings. Thus a field has ***characteristic*** $\mathbf{k}$ if it has characteristic k as a ring. The next result gives the analogue of Proposition 4.2.30 for fields.

4.3.11 Proposition (Property of fields with given characteristic) *If F is a field then the following statements hold:*

(i) if F has characteristic zero then there exists a subfield K of F that is isomorphic to $\mathbb{Q}$;

(ii) if F has characteristic $\mathrm{k} \in \mathbb{Z}_{>0}$ then k is prime and there exists a subfield K of F that is isomorphic to $\mathbb{Z}_{\mathrm{k}}$.

Proof First suppose that F has characteristic zero. As in the proof of Proposition 4.2.30, let $\phi\colon \mathbb{Z} \to \mathsf{R}$ be the map $\phi(j) = j1_{\mathsf{F}}$, and recall that this map is a monomorphism, and so an isomorphism from $\mathbb{Z}$ to image(ϕ). For $j1_{\mathsf{F}} \in \text{image}(\phi) \setminus \{0_{\mathsf{F}}\}$, since F is a field there exists $(j1_{\mathsf{F}})^{-1} \in \mathsf{F}$ such that $(j1_{\mathsf{F}}) \cdot (j1_{\mathsf{F}})^{-1} = 1_{\mathsf{F}}$. We map then define a map $\bar{\phi}\colon \mathbb{Q} \to \mathsf{F}$ by $\bar{\phi}(\frac{j}{k}) = (j1_{\mathsf{F}})(k1_{\mathsf{F}})^{-1}$. First let us show that this map is well defined. Suppose that $\frac{j_1}{k_1} = \frac{j_2}{k_2}$, or equivalently that $j_1k_2 = j_2k_1$. Then, using Proposition 4.2.10,

$$(j_11_{\mathsf{F}})(k_21_{\mathsf{F}}) = 1_{\mathsf{F}}(j_1(k_21_{\mathsf{F}})) = (j_1k_2)1_{\mathsf{F}} = (j_2k_1)1_{\mathsf{F}} = 1_{\mathsf{F}}(j_2(k_11_{\mathsf{F}})) = (j_21_{\mathsf{F}})(k_11_{\mathsf{F}}).$$

Thus $(j_11_{\mathsf{F}})(k_11_{\mathsf{F}})^{-1} = (j_21_{\mathsf{F}})(k_21_{\mathsf{F}})^{-1}$, and so $\bar{\phi}(\frac{j_1}{k_1}) = \bar{\phi}(\frac{j_2}{k_2})$. Now let us show that $\bar{\phi}$ is a monomorphism. Suppose that $(j_11_{\mathsf{F}})(k_11_{\mathsf{F}})^{-1} = (j_21_{\mathsf{F}})(k_21_{\mathsf{F}})^{-1}$ so that, using Proposition 4.2.10, $(j_1k_2 - j_2k_1)1_{\mathsf{F}} = 0_{\mathsf{F}}$. Then it follows that $\frac{j_1}{k_1} = \frac{j_2}{k_2}$ since F has characteristic zero. Next we show that $\bar{\phi}$ is a homomorphism. We compute, after an application of Proposition 4.2.10,

$$\bar{\phi}(\tfrac{j_1}{k_1} + \tfrac{j_2}{k_2}) = ((j_1k_2 + j_2k_1)1_{\mathsf{F}})(k_1k_21_{\mathsf{F}})^{-1} = (j_1k_21_{\mathsf{F}})(k_1k_21_{\mathsf{F}})^{-1} + (j_2k_11_{\mathsf{F}})(k_1k_21_{\mathsf{F}})^{-1}.$$

Another application of Proposition 4.2.10 gives

$$(k_11_{\mathsf{F}})(k_21_{\mathsf{F}})\bar{\phi}(\tfrac{j_1}{k_1} + \tfrac{j_2}{k_2}) = (j_1k_21_{\mathsf{F}}) + (j_2k_11_{\mathsf{F}}),$$

which in turn gives

$$\bar{\phi}(\tfrac{j_1}{k_1} + \tfrac{j_2}{k_2}) = (k_11_{\mathsf{F}})^{-1}(k_21_{\mathsf{F}})((j_1k_21_{\mathsf{F}}) + (j_2k_11_{\mathsf{F}})) = (j_11_{\mathsf{F}})(k_11_{\mathsf{F}})^{-1} + (j_21_{\mathsf{F}})(k_21_{\mathsf{F}})^{-1},$$

or $\bar{\phi}(\frac{j_1}{k_1} + \frac{j_2}{k_2}) = \bar{\phi}(\frac{j_1}{k_1}) + \bar{\phi}(\frac{j_2}{k_2})$. We also have

$$\bar{\phi}(\tfrac{j_1}{k_1}\tfrac{j_2}{k_2}) = (j_1j_21_{\mathsf{F}})(k_1k_21_{\mathsf{F}})^{-1},$$

which gives, in turn,

$$(k_1 1_{\mathsf{F}})(k_2 1_{\mathsf{F}})\bar{\phi}(\tfrac{j_1}{k_1}\tfrac{j_2}{k_2}) = (j_1 1_{\mathsf{F}})(j_2 1_{\mathsf{F}})$$

and

$$\bar{\phi}(\tfrac{j_1}{k_1}\tfrac{j_2}{k_2}) = (j_1 1_{\mathsf{F}})(k_1 1_{\mathsf{F}})^{-1}(j_2 1_{\mathsf{F}})(k_2 1_{\mathsf{F}})^{-1},$$

or $\bar{\phi}(\frac{j_1}{k_1}\frac{j_2}{k_2}) = \bar{\phi}(\frac{j_1}{k_1})\bar{\phi}(\frac{j_2}{k_2})$. Thus image($\bar{\phi}$) is a subfield of F isomorphic to $\mathbb{Q}$ by the isomorphism $\bar{\phi}$.

For the second part of the result, suppose that $k = k_1k_2$ for $k_1, k_2 \in \{2, \dots, k-1\}$. Then, if F has characteristic k we have

$$0 = k1_{\mathsf{F}} = (k_1k_2)1_{\mathsf{F}} = (k_1 1_{\mathsf{F}})(k_2 1_{\mathsf{F}}).$$

Since F is an integral domain this means, by Exercise 4.2.11, that either $k_1 1_{\mathsf{F}} = 0$ or $k_2 1_{\mathsf{F}} = 0$. This contradicts the fact that F has characteristic k, and so it must not be possible to factor k as a product of positive integers in $\{2, \dots, k-1\}$. Thus k is prime. That F contains a subfield that is isomorphic to $\mathbb{Z}_k$ follows from Proposition 4.2.30. ■

We note that the construction in the proof of a subfield K isomorphic to $\mathbb{Q}$ or $\mathbb{Z}_k$ is explicit, and is by construction the smallest subfield of F. This subfield has a name.

4.3.12 Definition (Prime field) For a field F, the smallest subfield of F is the ***prime field*** of F and is denoted by F_0. •

Exercises

4.3.1 Show that the relation $\sim$ of Definition 4.3.5 is an equivalence relation.

4.3.2 Prove Theorem 4.3.6.

4.3.3 Give a subring of $\mathbb{R}$ that is not a subfield.

4.3.4 Show that, if K is a subfield of F, then K is a field using the binary operations of addition and multiplication of F, restricted to K.

4.3.5 Let F be a field with $\mathsf{K} \subseteq \mathsf{F}$. Show that K is a subfield if and only if

1. $1_{\mathsf{F}} \in \mathsf{K}$,
2. $a - b \in \mathsf{K}$ for each $a, b \in \mathsf{K}$, and
3. $ab^{-1} \in \mathsf{K}$ for each $a, b \in \mathsf{K}$ with $b \neq 0_{\mathsf{F}}$.

Section 4.4

Polynomials and rational functions

The reader no doubt has encountered polynomials before, at least as functions of a real variable. In this section we study polynomials, not as functions, but as algebraic objects. Polynomials, while interesting in their own right, intersect other areas of mathematics and its applications. For example, we shall see in Section 5.8 that polynomials play an important rôle in the understanding of linear maps on finite-dimensional vector spaces. We will also see in Section V-10.4 that an understanding of roots of polynomials is important in determining the stability of many varieties of linear systems.

Do I need to read this section? Readers familiar with polynomials, particularly with real coefficients, and the structure of their roots can probably forgo this section on a first read. However, the material will be important in Section 5.8, and in our construction of the complex numbers in Section 4.7. A good understanding of polynomials will really help in understanding why complex numbers are important, and where they "come from." •

4.4.1 Polynomials rings and their basic properties

We shall start by giving a quite formal definition of what we mean by a polynomial. Then we shall introduce the notation needed to make our definition more closely resemble what the reader may be used to thinking about when they think about polynomials.

4.4.1 Definition (Polynomials) Let R be a ring. A ***polynomial over* R** is a sequence $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ with the property that the set $\{j \in \mathbb{Z}_{\geq 0} \mid a_j \neq 0_{\mathsf{R}}\}$ is finite. If $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ are polynomials over R then their ***sum*** and ***product*** are the polynomials over R defined by

$$A + B = (a_j + b_j)_{j\in\mathbb{Z}_{\geq 0}}, \quad A \cdot B = \left(\sum_{j=0}^{k} a_j b_{k-j}\right)_{k\in\mathbb{Z}_{\geq 0}},$$

respectively. •

Before we get to introducing the natural notation for writing a polynomial, let us first give the essential algebraic structure of the set of polynomials over R.

4.4.2 Theorem (The set of polynomials over R is a ring) *If* R *is a ring then the set of polynomials over* R*, with the binary operations of addition and multiplication as in Definition 4.4.1, is a ring. Moreover,*

(i) if R *is commutative, then so too is the set of polynomials over* R*,*

(ii) *if* R *is a unit ring, then so too is the set of polynomials over* R*,*

(iii) *if* R *is has no nonzero zerodivisors, then so too does the set of polynomials over* R*, and*

(iv) *if* R *is an integral domain, then so too is the set of polynomials over* R*.*

Proof We should first make sure that the sum and product of two polynomials over R is again a polynomial over R. This is clearly true for the sum. To verify this for the product, let $A = (a_j)_{j\in\mathbb{Z}_{\ge0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\ge0}}$ be two polynomials over R and define

$$d_A = \sup\{j \in \mathbb{Z}_{>0} \mid a_j \neq 0_{\mathsf{R}}\}, \quad d_B = \sup\{j \in \mathbb{Z}_{>0} \mid b_j \neq 0_{\mathsf{R}}\}.$$

If $k > d_A + d_B$ we claim that

$$\sum_{j=0}^{k} a_j b_{k-j} = 0_{\mathsf{R}}.$$

Indeed, let $j \in \{0, 1, \dots, d_A\}$ so that $k - j > d_B$. Then we have $b_{k-j} = 0_{\mathsf{R}}$. Similarly, if $j \in \{d_A + 1, \dots, k\}$, then $a_j = 0_{\mathsf{R}}$. Therefore, $a_j b_{k-j} = 0_{\mathsf{R}}$ if $k > d_A + d_B$.

The commutativity and associativity properties of addition for rings clearly hold for addition as defined. Also, if we take the element $(0_{\mathsf{R}})_{j\in\mathbb{Z}_{\ge0}}$ to be the zero element, it has the necessary property. If $A = (a_j)_{j\in\mathbb{Z}_{\ge0}}$ then $-A \triangleq (-a_j)_{j\in\mathbb{Z}_{\ge0}}$ is readily seen to be the additive inverse for A. Let $A = (a_j)_{j\in\mathbb{Z}_{\ge0}}$, $B = (b_j)_{j\in\mathbb{Z}_{\ge0}}$, and $C = (c_j)_{j\in\mathbb{Z}_{\ge0}}$ be polynomials over R and compute

$$\begin{aligned}(A \cdot B) \cdot C &= \left(\sum_{j=0}^{k}\left(\sum_{l=0}^{j} a_l b_{j-l}\right) c_{k-j}\right)_{k\in\mathbb{Z}_{\ge0}} = \left(\sum_{\substack{j,l,m\\ j,l,m\ge0,\ j+l+m=k}} (a_j b_l) c_m\right)_{k\in\mathbb{Z}_{\ge0}}\\ &= \left(\sum_{\substack{j,l,m\\ j,l,m\ge0,\ j+l+m=k}} a_j (b_l c_m)\right)_{k\in\mathbb{Z}_{\ge0}} = \left(\sum_{j=0}^{k} a_j \left(\sum_{l=0}^{k-j} b_l c_{k-j-l}\right)\right)_{k\in\mathbb{Z}_{\ge0}}\\ &= A \cdot (B \cdot C).\end{aligned}$$

Thus multiplication is associative. Let us verify left distributivity:

$$A \cdot (B + C) = \left(\sum_{j=0}^{k} a_j (b_{j-k} + c_{j-k})\right)_{k\in\mathbb{Z}_{\ge0}} = \left(\sum_{j=0}^{k} a_j b_{k-j} + \sum_{j=0}^{k} a_j c_{k-j}\right)_{k\in\mathbb{Z}_{\ge0}} = A \cdot B + A \cdot C.$$

Right distributivity is similarly verified, and this shows that the set of polynomials over R is indeed a ring.

Next suppose that R is commutative. Then

$$A \cdot B = \left(\sum_{j=0}^{k} a_j b_{k-j}\right)_{k\in\mathbb{Z}_{\ge0}} = \left(\sum_{j=0}^{k} a_{k-j} b_j\right) = \left(\sum_{j=0}^{k} b_j a_{k-j}\right) = B \cdot A,$$

and so the ring of polynomials over R is also commutative.

If R has a unit 1_{R}, then the polynomial $(a_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$a_j = \begin{cases} 1_{\mathsf{R}}, & j = 0, \\ 0_{\mathsf{R}}, & j \neq 0 \end{cases}$$

is the multiplicative identity in the ring of polynomials over R.

Suppose that R has no nonzero zerodivisors and let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ be nonzero polynomials over R. We shall show that if neither A nor B are the zero polynomial, then both $A \cdot B$ and $B \cdot A$ are not the zero polynomial. This is equivalent to the assertion that there are no nonzero zerodivisors in the ring of polynomials over R. Since both A and B are nonzero, the nonnegative integers

$$d_A = \sup\{j \in \mathbb{Z}_{\geq 0} \mid a_j \neq 0_{\mathsf{R}}\}, \quad d_B = \sup\{j \in \mathbb{Z}_{\geq 0} \mid b_j \neq 0_{\mathsf{R}}\}$$

can be defined. Let $C = A \cdot B$ and $D = B \cdot A$, and write $C = (c_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $D = (d_j)_{j\in\mathbb{Z}_{\geq 0}}$. As can be seen from our computations above in showing that the product of two polynomials is again a polynomial, it follows that $c_j = 0_{\mathsf{R}}$ and $d_j = 0_{\mathsf{R}}$ for $j > d_A + d_B$. Moreover, one can directly compute that $c_{d_A+d_B} = a_{d_A} \cdot b_{d_B}$ and $d_{d_A+d_B} = b_{d_B} \cdot a_{d_A}$. Since R has no nonzerodivisors and since a_{d_A} and b_{d_B} are nonzero, it follows that $c_{d_A+d_B}$ and $d_{d_A+d_B}$ are also nonzero. Therefore, $A \cdot B$ and $B \cdot A$ are nonzero.

Finally, combining the previous three parts of the proof shows that if R is an integral domain, then so too is the ring of polynomials over R. ■

Before proceeding further it is convenient to define the degree of a polynomial. As we shall see after we introduce indeterminate notation, the degree is the "highest power" term in a polynomial.

4.4.3 Definition (Degree of a polynomial) Let R be a ring and let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ be a polynomial over R. Define

$$D(A) = \{j \in \mathbb{Z}_{\geq 0} \mid a_j \neq 0_{\mathsf{R}}\}.$$

The ***degree*** of A is

$$\deg(A) = \begin{cases} \sup D(A), & D(A) \neq \emptyset, \\ -\infty, & D(A) = \emptyset. \end{cases} \bullet$$

Note that we think of the map deg as taking values in $\{-\infty\} \cup \mathbb{Z}_{\geq 0} \subseteq \overline{\mathbb{R}}$, and we inherit the algebraic and order structure of $\overline{\mathbb{R}}$ as discussed in Section 2.2.5.

Now let us introduce some notation that will allow us to write polynomials in a manner that is more customary. To do so we introduce an object which plays the rôle of the "independent variable" if one thinks of polynomials as functions. However, since we do not want to think of polynomials as functions, we need to be a little cagey about this.

4.4.4 Definition (Indeterminate, $\mathsf{R}[\xi]$) Let R be a unit ring. The ***indeterminate*** in the ring of polynomials over R is the polynomial $\xi = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ defined by

$$\xi = \begin{cases} 1_{\mathsf{R}}, & j = 1, \\ 0, & j \neq 1. \end{cases}$$

The ring of polynomials over R with indeterminate ξ is denoted by $\mathsf{R}[\xi]$ (even if R is not a unit ring). •

The symbol "ξ" for the specific polynomial which we call the indeterminate is completely arbitrary. But it is this symbol which stands for the "independent variable." Indeed, many authors us the symbol "x" for the indeterminate. However, since we use x as the generic symbol for the independent variable for a function of a real variable, it seems more prudent to use something different. We shall sometimes use alternative symbols for the indeterminate, if it is convenient to do so.

In order to obtain the familiar representation of polynomials, the following result will also be helpful.

4.4.5 Proposition (R is a subring of $\mathsf{R}[\xi]$) *The map* $\iota_{\mathsf{R}}\colon \mathsf{R} \to \mathsf{R}[\xi]$ *defined by* $\iota_{\mathsf{R}}(\mathrm{r}) = (\mathrm{a_j}(\mathrm{r}))_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ *with*

$$\mathrm{a_j}(\mathrm{r}) = \begin{cases} \mathrm{r}, & \mathrm{j} = 0, \\ 0_{\mathsf{R}}, & \mathrm{j} \neq 0 \end{cases}$$

is a ring monomorphism.

Proof It is clear that ι_{R} is injective. It is a simple matter of using the definitions of addition and multiplication in $\mathsf{R}[\xi]$ to verify that $\iota_{\mathsf{R}}(r_1 + r_1) = \iota_{\mathsf{R}}(r_1) + \iota_{\mathsf{R}}(r_2)$ and $\iota_{\mathsf{R}}(r_1 \cdot r_2) = \iota_{\mathsf{R}}(r_1) \cdot \iota_{\mathsf{R}}(r_2)$. ■

Now let us see how we may combine the indeterminate and the fact that R is a subring of $\mathsf{R}[\xi]$ as the basis for a convenient representation of a general polynomial. Note that since the indeterminate ξ is just an element of the ring of polynomials, the expression ξ^k for $k \in \mathbb{Z}_{\geq 0}$ makes sense, just as defined prior to Proposition 4.2.10.

4.4.6 Proposition (Expressing polynomials using the indeterminate) *If* R *is a unit ring and if* $\mathrm{A} = (\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}} \in \mathsf{R}[\xi]$, *then*

$$\mathrm{A} = \iota_{\mathsf{R}}(\mathrm{a}_0) \cdot \xi^0 + \iota_{\mathsf{R}}(\mathrm{a}_1) \cdot \xi^1 + \cdots + \iota_{\mathsf{R}}(\mathrm{a}_{\deg(\mathrm{A})}) \cdot \xi^{\deg(\mathrm{A})}.$$

Proof For $k \in \mathbb{Z}_{\geq 0}$, a direct computation gives $\xi^k = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ where

$$b_j = \begin{cases} 1_{\mathsf{R}}, & j = k, \\ 0_{\mathsf{R}}, & j \neq k. \end{cases}$$

It is then easy to see that, for $k \in \{0, 1, \dots, \deg(A)\}$, $\iota_{\mathsf{R}}(a_k)\xi^k = (c_j)_{j\in\mathbb{Z}_{\geq 0}}$, where

$$c_j = \begin{cases} a_k, & j = k, \\ 0_{\mathsf{R}}, & j \neq k. \end{cases}$$

Since $a_j = 0_{\mathsf{R}}$ for $j > \deg(A)$, the result follows. ■

From now on, except when we feel the need to be pedantic (as we shall on occasion), we shall simply write "a" for "$\iota_{\mathsf{R}}(a)$," so that, if we also omit the "$\cdot$" for multiplication and recall that $\xi^0 = 1_{\mathsf{R}}$, an element A of $\mathsf{R}[\xi]$ is expressed as

$$A = a_{\deg(A)}\xi^{\deg(A)} + \cdots + a_1\xi + a_0$$

for $a_0, a_1, \dots, a_{\deg(A)} \in \mathsf{R}$. Note that this notation makes sense, even if R is not a unit ring, as it suppresses the need to use the fact that $\xi^0 = 1_{\mathsf{R}}$.

Let us introduce some terminology associated to polynomials.

4.4.7 Definition (Coefficients, constant and monic polynomial) Let R be a unit ring and let $A \in \mathsf{R}[\xi] \setminus \{0_{\mathsf{R}[\xi]}\}$, denoting

$$A = a_k\xi^k + \cdots + a_1\xi + a_0,$$

with $a_k = \deg(A)$.

(i) The ring elements $a_0, a_1, \dots, a_k$ are the ***coefficients*** of A.

(ii) The ***leading coefficient*** is a_k.

(iii) A is a ***constant polynomial*** if $k = 0$ (we also say that the zero polynomial is a constant polynomial).

(iv) A is a ***monic polynomial*** if $a_k = 1_{\mathsf{R}}$. •

4.4.2 Homomorphisms and polynomial rings

We have already seen that there is a natural monomorphism $\iota_{\mathsf{R}}\colon \mathsf{R} \to \mathsf{R}[\xi]$. We now consider various other relationships between homomorphisms and polynomials. Specifically, we are interested in determining ways in which ring homomorphisms can be used to define homomorphisms to and/or from polynomial rings.

The first result we state is one of a form that one might expect: namely that homomorphisms of rings induce homomorphisms of polynomial rings in a natural way.

4.4.8 Proposition (Homomorphisms of polynomials induced from homomorphisms of rings) *If R and S are rings and if $\phi\colon \mathsf{R} \to \mathsf{S}$ is a homomorphism, then the map $\phi_*\colon \mathsf{R}[\xi] \to \mathsf{S}[\eta]$ given by $\phi_*((\mathrm{a}_\mathrm{j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}) = (\phi(\mathrm{a}_\mathrm{j}))_{\mathrm{j}\in\mathbb{Z}_{\geq 0}}$ is a homomorphism of the polynomial rings. Moreover, ϕ_* is a monomorphism (resp. epimorphism) if ϕ is.*

Proof The first assertion is a straightforward application of the definitions of addition and multiplication in polynomial rings. The final assertion also follows directly from the definitions; see Exercise 4.4.2. ■

Another important homomorphism is the assignment of a R-valued function to a polynomial in $\mathsf{R}[\xi]$. To make sense of this, recall that R^{R} denotes the set of maps from R to itself, and that in Example 4.2.2–5 we showed that R^{R} was a ring.

4.4.9 Proposition (Polynomials as functions) *If* R *is a commutative ring, then the map* $\mathrm{Ev}_{\mathsf{R}}\colon \mathsf{R}[\xi] \to \mathsf{R}^{\mathsf{R}}$ *defined by*

$$\mathrm{Ev}_{\mathsf{R}}((\mathrm{a_j})_{\mathrm{j}\in\mathbb{Z}_{\geq 0}})(\mathrm{r}) = \mathrm{a}_0 + \sum_{\mathrm{j}=1}^{\infty} \mathrm{a_j r^j}$$

(noting that the sum is finite) is a homomorphism of rings, called the ***evaluation homomorphism****.*

Proof Let $A = (a_j)_{j\in\mathbb{Z}_{\geq 0}}$ and $B = (b_j)_{j\in\mathbb{Z}_{\geq 0}}$ be polynomials and compute

$$\begin{aligned}\mathrm{Ev}_{\mathsf{R}}(A+B)(r) &= (a_0+b_0) + \sum_{j=1}^{\infty}(a_j+b_j)r^j \\ &= \left(a_0 + \sum_{j=1}^{\infty} a_j r^j\right) + \left(b_0 + \sum_{j=1}^{\infty} b_j r^j\right) \\ &= \mathrm{Ev}_{\mathsf{R}}(A)(r) + \mathrm{Ev}_{\mathsf{R}}(B)(r),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Ev}_{\mathsf{R}}(A\cdot B)(r) &= a_0b_0 + \sum_{k=1}^{\infty}\sum_{j=0}^{k} a_j b_{k-j} r^k \\ &= a_0b_0 + \sum_{k=1}^{\infty}\sum_{j=1}^{k} (a_j r^j)(b_{k-j} r^{k-j}) \\ &= a_0b_0 + \left(\sum_{j=1}^{\infty} a_j r^j\right)\left(\sum_{k=1}^{\infty} b_k r^k\right) \\ &= \mathrm{Ev}_{\mathsf{R}}(A)(r)\mathrm{Ev}_{\mathsf{R}}(B)(r),\end{aligned}$$

Showing that Ev_{R} is a homomorphism. ■

Note that the ring must be commutative in order that the preceding result be true. To understand why, one should ascertain where commutativity is used in the proof. Note, however, that the map $\mathrm{Ev}_{\mathsf{R}}\colon \mathsf{R}[\xi] \to \mathsf{R}^{\mathsf{R}}$ can still be *defined*, even when R is not commutative, and indeed, we shall on occasion use this notation below. However, it is just not guaranteed to be a homomorphism in the noncommutative case.

The next example shows that, for certain rings, the evaluation homomorphism can have some unexpected behaviour.

4.4.10 Example (The evaluation homomorphism may not be injective) Let us consider the ring $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$, and consider the polynomial $A = (1+2\mathbb{Z})\xi^2 - (1+2\mathbb{Z})\xi$. We then have

$$\mathrm{Ev}_{\mathbb{Z}_2}(A)(k+2\mathbb{Z}) = (k^2+2\mathbb{Z}) - (k+2\mathbb{Z}).$$

Since there are only two elements in the ring $\mathbb{Z}_2$, $0+2\mathbb{Z}$ and $1+2\mathbb{Z}$, we can directly compute that $\mathrm{Ev}_{\mathbb{Z}_2}(A)(k+2\mathbb{Z}) = 0+2\mathbb{Z}$ for all $k+2\mathbb{Z} \in \mathbb{Z}_2$. That is to say, A evaluates to the zero function, even though A is itself nonzero. By Exercise 4.2.7 it follows that $\mathrm{Ev}_{\mathbb{Z}_2}$ is not injective. Thus a polynomial is *not* uniquely determined by the function corresponding to it.

The reader may explore a generalisation of this example in Exercise 4.4.4. •

4.4.3 Factorisation in polynomial rings

In this section we turn to the important topic of polynomial factorisation. The basis of this is polynomial long division which the reader probably learned in school. However, here we develop this in the abstract setting for polynomials over general rings. The idea, however, is exactly the same as the idea behind what one learns in school, so it is important to bear this in mind.

Before we begin with factorisation in earnest, it is important to say a few things about the properties of the degree of a polynomial. As we shall see, it is the degree that shall play the rôle of the map δ in Definition 4.2.38 of a Euclidean domain.

4.4.11 Proposition (Properties of degree) *If* R *is a ring and if* $\mathsf{A}, \mathsf{B} \in \mathsf{R}[\xi]$, *then the following statements hold:*

(i) $\deg(\mathsf{A}+\mathsf{B}) \leq \max\{\deg(\mathsf{A}), \deg(\mathsf{B})\}$;

(ii) $\deg(\mathsf{A}\cdot\mathsf{B}) \leq \deg(\mathsf{A}) + \deg(\mathsf{B})$;

(iii) *if the leading coefficient of either* A *or* B *is not a zerodivisor, then* $\deg(\mathsf{A}\cdot\mathsf{B}) = \deg(\mathsf{A}) + \deg(\mathsf{B})$.

Proof The first assertion is obvious from the definition of addition of polynomials, and the second follows as in the proof of the fact that the product of polynomials is a polynomial in Theorem 4.4.2. The last assertion follows since, as we saw when we proved in Theorem 4.4.2 that $\mathsf{R}[\xi]$ has no nonzero zerodivisors when R does not, the $(\deg(A)+\deg(B))$th term of $A\cdot B$ is the product of the leading coefficients of A and B. If these are not zerodivisors, then this product is nonzero, and so $\deg(A\cdot B) = \deg(A)+\deg(B)$. ■

The following examples illustrate that the inequalities in the preceding result cannot be replaced with equalities.

4.4.12 Examples (Properties of degree)

1. Consider the polynomials $A = \xi^3+\xi^2+\xi+1$ and $B = -\xi^3$ in $\mathbb{Z}[\xi]$. We then have $\deg(A) = \deg(B) = 3$, but $\deg(A+B) = 2$.
2. In $\mathbb{Z}_4$ consider the polynomials $A = (2+4\mathbb{Z})\xi^2 + (1+4\mathbb{Z})\xi + (2+4\mathbb{Z})$ and $B = (2+4\mathbb{Z})\xi^2$. We then compute
$$A\cdot B = (4+4\mathbb{Z})\xi^4 + (2+4\mathbb{Z})\xi^3 + (4+4\mathbb{Z})\xi^2 = (2+4\mathbb{Z})\xi^3.$$
Then $\deg(A\cdot B) = 3 < 4 = \deg(A)+\deg(B)$. •

The key idea in polynomial factorisation is now the following Division Algorithm that holds for certain types of polynomials.

4.4.13 Theorem (Division Algorithm for polynomials) *Let* R *be a unit ring and let* $\mathrm{A}, \mathrm{B} \in \mathsf{R}[\xi] \setminus \{0_{\mathsf{R}[\xi]}\}$ *be such that the leading coefficient of* B *is a unit in* R*. Then there exists unique* $\mathrm{Q}, \mathrm{R} \in \mathsf{R}[\xi]$ *such that* $\mathrm{A} = \mathrm{Q} \cdot \mathrm{B} + \mathrm{R}$ *and such that* $\deg(\mathrm{R}) < \deg(\mathrm{B})$*.*

Proof Throughout the proof we write

$$A = a_m\xi^m + \cdots + a_1\xi + a_0, \quad B = b_k\xi^k + \cdots + b_1\xi + b_0,$$

where a_m and b_k are nonzero. If $\deg(A) < \deg(B)$, then the theorem holds by taking $Q = 0_{\mathsf{R}[\xi]}$ and $R = A$. Thus we assume that $\deg(A) \geq \deg(B)$, and we proceed by induction on $\deg(A)$. If $\deg(A) = 0$ then $\deg(B) = 0$ and so $A = a_0$ and $B = b_0$, and by assumption $b_0 \in \mathsf{R}$ is a unit. The result holds taking $Q = a_0b_0^{-1}$ and $R = 0_{\mathsf{R}[\xi]}$. Suppose that the existence assertion holds when $\deg(A) = l$ for $l \in \{0, 1, \dots, m-1\}$, and suppose that $\deg(A) = m$. Also suppose that $\deg(B) = k \leq m$. Using the definition of polynomial multiplication we easily see that the polynomial $(a_mb_k^{-1}\xi^{m-k})B$ has degree m and leading coefficient a_m. Therefore the polynomial $A - (a_mb_k^{-1}\xi^{m-k})B$ has degree at most $m - 1$. By the induction hypotheses there exists polynomials Q' and R' with $\deg(R') < \deg(B)$ and such that

$$A - (a_mb_k^{-1}\xi^{m-k})B = Q' \cdot B + R'.$$

The existence part of the result now follows by taking $Q = Q' + (a_mb_k^{-1}\xi^{m-k})$ and $R = R'$.

To prove uniqueness, suppose that $A = Q_1 \cdot B + R_1 = Q_2 \cdot B + R_2$ where $\deg(R_1), \deg(R_2) < \deg(B)$. Then $(Q_1 - Q_2) \cdot B = R_2 - R_1$, and using Proposition 4.4.11 gives

$$\deg(Q_1 - Q_2) + \deg(B) = \deg(R_2 - R_1) \tag{4.6}$$

since the leading coefficient of B is a unit. But it also holds that $\deg(R_2 - R_1) \leq \max\{\deg(R_1), \deg(R_2)\} < \deg(B)$. Thus the only way that (4.6) can hold is if $\deg(Q_1 - Q_2) = -\infty$, and this implies also that $\deg(R_2 - R_1) = -\infty$. Therefore $Q_1 = Q_2$ and $R_1 = R_2$, as desired. ■

Note that the theorem does not say that $\mathsf{R}[\xi]$ is a Euclidean domain for any ring R since the hypotheses include restrictions on the leading coefficient of B. However, in some cases these restrictions on B always hold. For example, we have the following result.

4.4.14 Corollary (F[ξ] is a Euclidean domain) *Let* F *be a field and define* $\delta\colon \mathsf{F}[\xi] \to \mathbb{Z}_{\geq 0}$ *by*

$$\delta(\mathrm{A}) = \begin{cases} \deg(\mathrm{A}) + 1, & \mathrm{A} \neq 0_{\mathsf{R}[\xi]}, \\ 0, & \mathrm{A} = 0_{\mathsf{F}[\xi]}. \end{cases}$$

Then $(\mathsf{F}[\xi], \delta)$ *is a Euclidean domain, hence* $\mathsf{F}[\xi]$ *a principal ideal domain, and hence also a unique factorisation domain. Moreover,* $\mathsf{F}[\xi]$ *is a* δ*-closed and* δ*-positive subset of itself.*

Proof We have

$$\delta(A \cdot B) = \deg(A) + \deg(B) + 1 \geq \deg(A) + 1 = \delta(A)$$

if $A \cdot B \neq 0_{\mathsf{F}[\xi]}$. That $\mathsf{F}[\xi]$ is a Euclidean domain now follows immediately from Theorem 4.4.13 since, if B is not the zero polynomial, then its leading coefficient is a unit. That $\mathsf{F}[\xi]$ is a principal ideal domain and a unique factorisation domain follows from Theorems 4.2.55 and 4.2.71. For the final assertion it is clear that $\mathsf{F}[\xi]$ is δ-closed. We also note that, if both A and B are nonzero, then

$$\begin{aligned}\delta(A - B) = \deg(A - B) + 1 = \max\{\deg(A), \deg(B)\} + 1 \\ \leq \max\{\deg(A) + 1, \deg(B) + 1\} = \max\{\delta(A), \delta(B)\}.\end{aligned}$$

If either A or B is zero then we trivially have $\delta(A - B) \leq \max\{\delta(A), \delta(B)\}$. Thus $\mathsf{F}[\xi]$ is indeed a δ-positive subset of itself. ■

The fact that $\mathsf{F}[\xi]$ is a δ-closed and δ-positive subset of itself gives the following two corollaries of Proposition 4.2.44 and Theorem 4.2.48, respectively.

4.4.15 Corollary (Uniqueness of quotient and remainder in F[ξ]) *If* F *is a field and if* $\mathrm{A}, \mathrm{B} \in \mathsf{F}[\xi]$ *with* $\mathrm{B} \neq 0_{\mathsf{F}[\xi]}$*, then there exists unique* $\mathrm{Q}, \mathrm{R} \in \mathsf{F}[\xi]$ *such that* $\mathrm{A} = \mathrm{Q} \cdot \mathrm{B} + \mathrm{R}$ *and such that* $\deg(\mathrm{R}) < \deg(\mathrm{B})$*.*

4.4.16 Corollary (Base expansion in F[ξ]) *If* F *is a field and if* $\mathrm{A}, \mathrm{B} \in \mathsf{F}[\xi] \setminus \{0_{\mathsf{F}[\xi]}\}$ *with* $\deg(\mathrm{B}) \geq 1$*, then there exists unique* $\mathrm{k} \in \mathbb{Z}_{\geq 0}$ *and* $\mathrm{R}_0, \mathrm{R}_1, \dots, \mathrm{R}_\mathrm{k} \in \mathsf{F}[\xi]$ *such that*

(i) $\mathrm{R}_\mathrm{k} \neq 0_{\mathsf{F}[\xi]}$*,*

(ii) $\deg(\mathrm{R}_0), \deg(\mathrm{R}_1), \dots, \deg(\mathrm{R}_\mathrm{k}) < \deg(\mathrm{B})$*, and*

(iii) $\mathrm{A} = \mathrm{R}_0 + \mathrm{R}_1 \cdot \mathrm{B} + \mathrm{R}_2 \cdot \mathrm{B}^2 + \cdots + \mathrm{R}_\mathrm{k} \cdot \mathrm{B}^\mathrm{k}$*.*

It turns out that for polynomials rings over unique factorisation domains, even though they may not be polynomial rings over fields, are still unique factorisation domains, although they may no longer be principal ideal domains (see Example 4.2.74). Let us now develop a general result which indicates how this can arise. This will require a little buildup, starting with the following definition.

4.4.17 Definition (Primitive polynomial) Let R be a unique factorisation domain and let $A = \sum_{j=0}^{k} a_j \xi^j \in \mathsf{R}[\xi]$.

(i) A ***content*** of A is a greatest common divisor of $\{a_0, a_1, \dots, a_k\}$.

(ii) A is ***primitive*** if it has a content that is a unit in R. •

4.4.18 Example (Primitive polynomial) Consider the unique factorisation domain $\mathbb{Z}$ with its polynomial ring $\mathbb{Z}[\xi]$. If $A = 2\xi^2 + 4\xi - 8$ then 2 and −2 are contents of A since 2 and −2 are greatest common divisors of $\{2, 4, -8\}$. If $B = 3\xi + 5$ then B is primitive since 1 is a greatest common divisor of $\{3, 5\}$. •

We now record some results about polynomials over unique factorisation domains. The statement of the result relies on the fact that the polynomial ring over an integral domain R is naturally a subset of the polynomial ring over the fraction field F_R, a fact that follows from Theorem 4.3.6 and Proposition 4.4.8.

4.4.19 Proposition (Properties of polynomials over unique factorisation domains) *Let* R *be a unique factorisation domain with* F_{R} *its fraction field. Then the following statements hold for polynomials* $\mathrm{A}, \mathrm{B} \in \mathsf{R}[\xi] \subseteq \mathsf{F}_{\mathsf{R}}[\xi]$*.*

(i) $\mathrm{A} = \mathrm{c}_{\mathrm{A}}\mathrm{A}'$ *where* c_{A} *is a content of* A *and* $\mathrm{A}' \in \mathsf{R}[\xi]$ *is primitive;*

(ii) *if* c_{A} *and* c_{B} *are contents of* A *and* B*, respectively, then* $\mathrm{c}_{\mathrm{A}}\mathrm{c}_{\mathrm{B}}$ *is a content of* $\mathrm{A} \cdot \mathrm{B}$*;*

(iii) *if* A *and* B *are primitive, then* $\mathrm{A} \cdot \mathrm{B}$ *is primitive;*

(iv) *if* A *and* B *are primitive, then* $\mathrm{A}|\mathrm{B}$ *and* $\mathrm{B}|\mathrm{A}$ *in* $\mathsf{R}[\xi]$ *if and only if* $\mathrm{A}|\mathrm{B}$ *and* $\mathrm{B}|\mathrm{A}$ *in* $\mathsf{F}_{\mathsf{R}}[\xi]$*;*

(v) *if* A *is primitive and if* $\deg(\mathrm{A}) > 0$*, then* A *is irreducible in* $\mathsf{R}[\xi]$ *if and only if it is irreducible in* $\mathsf{F}_{\mathsf{R}}[\xi]$*.*

Proof (i) Write $A = \sum_{j=0}^{k} a_j \xi^j$ and write $a_j = c_A a'_j$ for $j \in \{0, 1, \dots, k\}$. Then the result follows by taking $A' = \sum_{j=0}^{k} a'_j \xi^j$.

(ii) By part (i), write $A = c_A A'$ and $B = c_B B'$ for A' and B' primitive. If c' is a content for $A' \cdot B'$, it is easy to see that $c_A c_B c'$ is a content for $A \cdot B = (c_A A') \cdot (c_B B')$. Thus it suffices to show that c' is a unit, i.e., that $A' \cdot B'$ is primitive. Suppose that $A' \cdot B'$ is not primitive and write $C = A' \cdot B' = (c_k = \sum_{j=0}^{k} a'_j b'_{k-j})_{k \in \mathbb{Z}_{\geq 0}}$, where $A' = (a'_j)_{j \in \mathbb{Z}_{\geq 0}}$ and $B' = (b'_j)_{j \in \mathbb{Z}_{\geq 0}}$. Suppose that $p \in \mathsf{R}$ is irreducible and that $p | c_j$ for all j. If $c_{A'}$ is a content for A' we have $p \nmid c_{A'}$ since $c_{A'}$ is a unit. Similarly, $p \nmid c_{B'}$ where $c_{B'}$ is a content for B'. Now define

$$\begin{aligned} n_{A'} &= \inf\{l \in \{0, 1, \dots, \deg(A)\} \mid p | a'_j,\ j \in \{0, 1, \dots, l\},\ p \nmid a'_l\}, \\ n_{B'} &= \inf\{l \in \{0, 1, \dots, \deg(B)\} \mid p | b'_j,\ j \in \{0, 1, \dots, l\},\ p \nmid b'_l\}. \end{aligned}$$

Note that $p | c_{n_{A'} + n_{B'}}$, and since

$$\begin{aligned} c_{n_{A'} + n_{B'}} = a'_0 b'_{n_{A'} + n_{B'}} &+ \dots + a'_{n_{A'}-1} b'_{n_{B'}+1} \\ &+ a'_{n_{A'}} b'_{n_{A'}} + a'_{n_{A'}+1} b'_{n_{B'}-1} + \dots + a'_{n_{A'}+n_{B'}} b'_0, \end{aligned}$$

$p | a'_{n_{A'}} b'_{n_{A'}}$, which implies that $p | a'_{n_{A'}}$ or $p | b'_{n_{A'}}$ since irreducibles are prime in unique factorisation domains (Proposition 4.2.70). This implies that either A' or B' is not primitive.

(iii) This follows directly from part (ii) since the product of units is again a unit.

(iv) Since $\mathsf{R} \subseteq \mathsf{F}_{\mathsf{R}}$, it is clear that if $A|B$ and $B|A$ in R, then $A|B$ and $B|A$ in F_{R}. Now suppose that $B|A$ in F_{R}. Then, by Proposition 4.2.60, $A = U \cdot B$ where $U \in \mathsf{F}_{\mathsf{R}}[\xi]$ is a unit. By Exercise 4.4.3 this means that $U = u$ for some $u \in \mathsf{F}_{\mathsf{R}}$, and ;et us write $u = \frac{a}{b}$ for $a, b \in \mathsf{R}$ with $b \neq 0_{\mathsf{R}}$. We thus have $bA = aB$. Since A and B are primitive, if c_A and c_B are contents for A and B, respectively, these must be units. Therefore, both b and bc_A are contents for bA and both a and ac_B are contents for aB. This means that $a = bv$ for a unit $v \in \mathsf{R}$ so that $bA = bvB$. Since $\mathsf{R}[\xi]$ is an integral domain by Theorem 4.4.2, this implies that $A = vB$ for a unit $v \in \mathsf{R}$ by Proposition 4.2.33. Now, by Proposition 4.2.60, $A|B$ and $B|A$ in $\mathsf{R}[\xi]$ since v is also a unit in $\mathsf{R}[\xi]$.

(v) Suppose that A is not irreducible in $\mathsf{F}_{\mathsf{R}}[\xi]$ and write $A = B \cdot C$ for $\mathsf{F}_{\mathsf{R}}[\xi]$ both nonunits. By Exercise 4.4.3 we must therefore have $\deg(B), \deg(C) \geq 1$. Write

$$B = \sum_{j=0}^{k} \frac{a_j}{b_j} \xi^j, \quad C = \sum_{j=0}^{l} \frac{c_j}{d_j} \xi^j$$

for $a_j, b_j \in \mathsf{R}$ with $b_j \neq 0_{\mathsf{R}}$ for $j \in \{0, 1, \dots, k\}$ and for $c_j, d_j \in \mathsf{R}$ with $d_j \neq 0_{\mathsf{R}}$ for $j \in \{0, 1, \dots, l\}$. Write $b = b_0 b_1 \cdots b_k$ and for $j \in \{0, 1, \dots, k\}$ define

$$\hat{b}_j = b_b b_1 \cdots b_{j-1} b_{j+1} \cdots b_k.$$

Define $B' = \sum_{j=0}^{k} a_j \hat{b}_j \xi^j \in \mathsf{R}[\xi]$ and write $B' = c_{B'} B''$ where $c_{B'}$ is a content of B' and where B'' is primitive, by part (i). A direct computation then shows that $B = \frac{1_{\mathsf{R}}}{b} B' = \frac{c_{B'}}{b} B''$. An entirely similar computation gives $C = \frac{c_{C'}}{d} C''$ where $c_{C'} \in \mathsf{R}$ and $C'' \in \mathsf{R}[\xi]$ is primitive. Therefore, since $A = B \cdot C$, we have $bdA = c_{B'} c_{C'} B'' \cdot C''$. Since A and $B'' \cdot C''$ are primitive, the latter by part (ii), it follows that both bd and $c_{B'} c_{C'}$ are contents for A. Thus $bd = u c_{B'} c_{C'}$ for a unit $u \in \mathsf{R}$. Thus $bdA = bduB'' \cdot C''$, or $A = uB'' \cdot C''$. Since $\deg(B'') = \deg(B) \geq 1$ and $\deg(C'') = \deg(C) \geq 1$, this implies that A is not irreducible in $\mathsf{R}[\xi]$ by Exercise 4.4.3.

Now suppose that A is irreducible in $\mathsf{F}_{\mathsf{R}}[\xi]$ and write $A = B \cdot C$ for $B, C \in \mathsf{R}[\xi] \subseteq \mathsf{F}_{\mathsf{R}}[\xi]$. Thus either B or C must be a unit in $\mathsf{F}_{\mathsf{R}}[\xi]$, and so by Exercise 4.4.3 we must have either $\deg(B) = 0$ or $\deg(C) = 0$. Suppose, without loss of generality, that $\deg(B) = 0$ so that $B = b_0 \in \mathsf{R} \setminus \{0\}$. Then, if c_C is a content for C, $b_0 c_C$ is a content for $A = B \cdot C$. Since A is primitive, $b_0 c_C$ must be a unit, and so, in particular, b_0 must be a unit in R by Exercise 4.1.2. Thus B is a unit in $\mathsf{R}[\xi]$, showing that A is irreducible in $\mathsf{R}[\xi]$. ■

Now, using the proposition, we can prove our main result concerning the factorisation properties of polynomials over unique factorisation domains.

4.4.20 Theorem (Polynomial rings over unique factorisation domains are unique factorisation domains) *If R is a unique factorisation domain, then $\mathsf{R}[\xi]$ is a unique factorisation domain.*

Proof Let $A \in \mathsf{R}[\xi]$ be a nonzero nonunit. If $\deg(A) = 0$ then A is an element of R under the natural inclusion of R in $\mathsf{R}[\xi]$ (see Proposition 4.4.5). In this case, A possesses a factorisation as a product of irreducibles since R is a unique factorisation domain. Now suppose that $\deg(A) \geq 1$, and by Proposition 4.4.19(i) write $A = c_A A'$ where $c_A \in \mathsf{R}$ is a content of A and where A' is primitive. If c_A is not a unit then write $c_A = c_{A,1} \cdots c_{A,l}$ where $c_{A,j} \in \mathsf{R}$, $j \in \{1, \dots, l\}$, are irreducible, this being possible since R is a unique factorisation domain. Note that the elements $c_{A,j}$, $j \in \{1, \dots, l\}$, are also irreducible thought of as elements of $\mathsf{R}[\xi]$ (why?). Now, since F_{R} is a unique factorisation domain by Corollary 4.4.14, write $A' = P'_1 \cdots P'_k$ where $P'_1, \dots, P'_k \in \mathsf{F}_{\mathsf{R}}[\xi]$ are irreducible. Now proceed as in the proof of Proposition 4.4.19(v) to show that, for $j \in \{1, \dots, k\}$, $P'_j = \frac{a_j}{b_j} P_j$ for $a_j, b_j \in \mathsf{R}$ with $b_j \neq 0_{\mathsf{R}}$ and with $P_j \in \mathsf{R}[\xi]$ primitive. Since $\frac{a_j}{b_j}$ is a unit in F_{R} and so in $\mathsf{F}_{\mathsf{R}}[\xi]$ by Exercise 4.4.3, by Exercise 4.2.18 it follows that P_j is irreducible in $\mathsf{F}_{\mathsf{R}}[\xi]$, and so in $\mathsf{R}[\xi]$ by Proposition 4.4.19(v). Writing $a = a_1 \cdots a_k$

and $b = b_1 \cdots b_k$, we have $A' = \frac{a}{b} P_1 \cdots P_k$, or $bA' = aP_1 \cdots P_k$. Since A' and $P_1 \cdots P_k$ are primitive (the latter by Proposition 4.4.19(iii)), it follows that $a = ub$ for u a unit in R. Therefore, if c_A is not a unit, we have

$$A = c_A A' = c_{A,1} \cdots c_{A,l}(uP_1)P_2 \cdots P_k,$$

where $c_{A,1}, \dots, c_{A,l} \in \mathsf{R} \subseteq \mathsf{R}[\xi]$ and $uP_1, P_2, \dots, P_k \in \mathsf{R}[\xi]$ are all irreducible in $\mathsf{R}[\xi]$ (noting, by Exercise 4.2.18 that uP_1 is irreducible). If c_A is a unit, then A is primitive already, and we can directly write

$$A = (uP_1)P_2 \cdots P_k,$$

where $(uP_1, P_2, \dots, P_k \in \mathsf{R}[\xi]$ are irreducible. This gives part (i) of Definition 4.2.66.

Now we verify part (ii) of Definition 4.2.66. We begin with a lemma. We already know from above that every element of $\mathsf{R}[\xi]$ possesses a factorisation as a product of irreducible. The lemma guarantees that the factorisation is of a certain form.

1 **Lemma** *If* R *is a unique factorisation domain and if* $\mathrm{A} \in \mathsf{R}[\xi]$ *is written as a product of irreducibles,* $\mathrm{A} = \mathrm{F}_1 \cdots \mathrm{F}_\mathrm{m}$*, then there exists irreducibles* $\mathrm{c}_1, \dots, \mathrm{c}_\mathrm{l} \in \mathsf{R}$ *and irreducibles* $\mathrm{P}_1, \dots, \mathrm{P}_\mathrm{k} \in \mathsf{R}[\xi]$ *such that* $\mathrm{l} + \mathrm{k} = \mathrm{m}$ *and such that* $\mathrm{F}_{\mathrm{j}_\mathrm{r}} = \mathrm{c}_\mathrm{r}$, $\mathrm{r} \in \{1, \dots, \mathrm{l}\}$*, and* $\mathrm{F}_{\mathrm{j}_{\mathrm{l}+\mathrm{s}}} = \mathrm{P}_\mathrm{s}$, $\mathrm{s} \in \{1, \dots, \mathrm{k}\}$*, where* $\{1, \dots, \mathrm{m}\} = \{\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{m}\}$.

Proof From the first part of the proof of the theorem we can write $A = c_{A,1} \cdots c_{A,l} P_1 \cdots P_k$ for irreducibles $c_{A,1}, \dots, c_{A,l} \in \mathsf{R}$ and irreducibles $P'_1, \dots, P'_k \in \mathsf{R}[\xi]$. We thus have

$$F_1 \cdots F_m = c_{A,1} \cdots c_{A,l} P'_1 \cdots P'_k.$$

Let $\{j_1, \dots, j_{l'}\}$ be the indices from $\{1, \dots, m\}$ such that $\deg(F_j) = 0$ if and only if $j \in \{1, \dots, j_{l'}\}$. Denote by $\{j_{l'+1}, \dots, j_m\}$ the remaining indices, so that $\deg(F_j) \geq 1$ if and only if $j \in \{j_{l'+1}, \dots, j_m\}$. Since the polynomials $F_{j_{l'+1}}, \dots, F_{j_m}$ are irreducible, they are primitive, so that $c_{A,1} \cdots c_{A,l}$ and $F_{j_1} \cdots F_{j_{l'}}$ are both contents for $P'_1 \cdots P'_k$. Thus there exists a unit $u \in \mathsf{R}$ such that $F_{j_1} \cdots F_{j_{l'}} = uc_{A,1} \cdots c_{A,l}$. By unique factorisation in R, $l' = l$ and there exists $\sigma \in \mathfrak{S}_l$ such that $F_{j_r} = u_{\sigma(r)} c_{A,\sigma(r)}$ for $r \in \{1, \dots, l\}$, and where $u_1, \dots, u_l$ are units in R. The result now follows by taking $c_r = u_{\sigma(r)} c_{A,\sigma(r)}$, $r \in \{1, \dots, l\}$ and $P_s = F_{j_{l+s}}$, $s \in \{1, \dots, k\}$. ▼

Now, using the lemma, let $c_1 \cdots c_l P_1 \cdots P_k$ and $c'_1 \cdots c'_{l'} P'_1 \cdots P'_{k'}$ be two factorisations of A by irreducibles, where $c_1, \dots, c_l, c'_1, \dots, c'_{l'} \in \mathsf{R}$ are irreducible and $P_1, \dots, P_k, P'_1, \dots, P'_{k'} \in \mathsf{R}[\xi]$ are irreducible. Since $P_1 \cdots P_k$ and $P'_1, \dots, P'_{k'}$ are primitive, $c_1 \cdots c_l$ and $c'_1 \cdots c'_{l'}$ are contents for A, and so there exists a unit $u \in \mathsf{R}$ such that $c_1 \cdots c_l = uc'_1 \cdots c'_{l'}$. Since R is a unique factorisation domain, $l = l'$ and there exists a permutation $\sigma \in \mathfrak{S}_l$ such that $c'_{\sigma(j)} = u_j c_j$ for $j \in \{1, \dots, l\}$, and for some set $u_1, \dots, u_l$ of units. Since $P_1 \cdots P_k$ and $P'_1 \cdots P'_{k'}$ have the same content, up to multiplication by a unit, it follows that $P'_1 \cdots P'_k = UP_1 \cdots P_k$ where $U \in \mathsf{R}[\xi]$ is a unit. Thus $U = v$ where $v \in \mathsf{R}$ is a unit by Exercise 4.4.3. Therefore, since $\mathsf{F}_{\mathsf{R}}[\xi]$ is a unique factorisation domain by Corollary 4.4.14, $k = k'$ and there exists a permutation $\sigma \in \mathfrak{S}_k$ such that $P'_{\sigma(j)} = v_j P_j$ for $j \in \{1, \dots, k\}$, and where v_j is a unit in F_{R}. Thus by Proposition 4.2.60, in $\mathsf{F}_{\mathsf{R}}[\xi]$, we have $P'_{\sigma(j)} | P_j$ and $P_j | P'_{\sigma(j)}$, $j \in \{1, \dots, k\}$. By Proposition 4.4.19(iv) we then have, in $\mathsf{R}[\xi]$, $P'_{\sigma(j)} | P_j$ and $P_j | P'_{\sigma(j)}$, $j \in \{1, \dots, k\}$. Therefore, by Proposition 4.2.60 again, there exists units $u_1, \dots, u_k \in \mathsf{R}$ such that $P'_{\sigma(j)} = u_j P_j$, $j \in \{1, \dots, k\}$. This then gives the uniqueness, up to units, of factorisation in $\mathsf{R}[\xi]$. ■

The following corollary is then of interest, e.g., in Example 4.2.74.

4.4.21 Corollary ($\mathbb{Z}[\xi]$ is a unique factorisation domain) *$\mathbb{Z}[\xi]$ is a unique factorisation domain.*

4.4.4 Roots, and prime and irreducible polynomials

The reader is probably familiar with the idea of a root of a polynomial, and in particular with certain facts about polynomials over $\mathbb{R}$, or perhaps over $\mathbb{C}$. We shall not discuss these two cases until Section 4.7.3. Here we merely say a few things about the relationships between polynomial roots and irreducible polynomials. As we shall see, the notion of a reducible polynomial is in general more complicated than one is led to believe by consideration of only polynomials over $\mathbb{R}$ and $\mathbb{C}$. While it is true that in these volumes we shall principally be interested in polynomials over $\mathbb{R}$, a little understanding of roots for general polynomials is useful in putting the special case into more context. Moreover, it also aids in really understanding why the complex numbers $\mathbb{C}$, constructed in Section 4.7.

Let us first define what is meant by a root.

4.4.22 Definition (Root of a polynomial) If R is a ring and if $A = \sum_{j=0}^{k} a_j \xi^j \in \mathsf{R}[\xi]$, a ***root*** of A is an element $r \in \mathsf{R}$ such that $\mathrm{Ev}_{\mathsf{R}}(A)(r) = \sum_{j=0}^{k} a_j r^j = 0_{\mathsf{R}}$. •

4.4.23 Remark (Roots and noncommutativity) If a ring is not commutative, one might also define a root as satisfying $\sum_{j=0}^{k} r^j a_j = 0_{\mathsf{R}}$. Sometimes the definition we give is referred to as a ***left root*** and the alternative definition as a ***right root***. We shall be interested in cases where this distinction is not important. •

Our immediate objective is to understand how roots of a polynomial relate to factorisations of the polynomial as products of polynomials. A first step towards this is the following result.

4.4.24 Proposition (Remainder Theorem) *If R is a unit ring, if $\mathrm{A} = \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{a_j} \xi^{\mathrm{j}} \in \mathsf{R}[\xi]$, and if $\mathrm{r} \in \mathsf{R}$, then there exists a unique polynomial $\mathrm{Q} \in \mathsf{R}[\xi]$ such that*

$$\mathrm{A} = \mathrm{Q} \cdot (\xi - \mathrm{r}) + \mathrm{Ev}_{\mathsf{R}}(\mathrm{A})(\mathrm{r}).$$

Proof If $A = 0_{\mathsf{R}[\xi]}$ then the result follows by taking $Q = 0_{\mathsf{R}[\xi]}$. If $A \neq 0_{\mathsf{R}[\xi]}$, then, by Theorem 4.4.13 there exists $Q, R \in \mathsf{R}[\xi]$ such that $A = Q \cdot (x - r) + R$, with $\deg(R) < 1$. Thus $R = r_0$ is a constant polynomial. Now write $Q = \sum_{j=0}^{k-1} q_j \xi^j$ and compute

$$\mathrm{Ev}_{\mathsf{R}}(A)(r) = -q_0 r + \sum_{j=0}^{k-1} (q_{j-1} - q_j r) + q_{k-1} r + r_0 = r_0,$$

as desired. The uniqueness follows from the uniqueness assertion of Theorem 4.4.13. ■

The remainder theorem then gives rise to the following extremely useful consequence of a polynomial having a root.

4.4.25 Proposition (Roots give rise to factorisation) *If* R *is a unit ring, the following statements concerning* $\mathrm{A} \in \mathsf{R}[\xi]$ *and* $\mathrm{r} \in \mathsf{R}$ *hold:*

(i) if r *is a root for* A *then* $(\xi - \mathrm{r})|\mathrm{A}$*;*

(ii) if R *is additionally commutative, then* r *is a root for* A *if* $(\xi - \mathrm{r})|\mathrm{A}$.

Proof (i) This follows direction from Proposition 4.4.24.

(ii) By Proposition 4.4.24 write $A = Q \cdot (\xi - r) + r_0$ where $r_0 = \mathrm{Ev}_{\mathsf{R}}(A)(r)$. Since $(\xi - r)|A$ write $A = B \cdot (\xi - r)$ for $B \in \mathsf{R}[\xi]$. Thus $Q \cdot (\xi - r) + r_0 = B \cdot (\xi - r)$. A direct computation shows that

$$\mathrm{Ev}_{\mathsf{R}}(Q \cdot (\xi - r))(r) = \mathrm{Ev}_{\mathsf{R}}(B \cdot (\xi - r)) = 0_{\mathsf{R}}$$

(cf. the proof of Proposition 4.4.24). Thus $r_0 = 0$ and so r is a root of A. ■

This factorisation then gives rise to the following characterisation of the set of roots of a polynomial, and is our first result of this type.

4.4.26 Proposition (Number of roots of a polynomial is bounded by its degree) *Let* R *be an integral domain and let* $\mathrm{A} \in \mathsf{R}[\xi]$ *have degree* k*. Then* A *has at most* k *distinct roots.*

Proof Denote by $r_1, \dots, r_l$ the distinct roots of A. We claim that

$$A = Q \cdot (\xi - r_l) \cdot \dots \cdot (\xi - r_1)$$

for some $Q \in \mathsf{R}[\xi]$. We prove this by induction on l. If $l = 1$ then the result holds by Proposition 4.4.25. Now suppose that the result holds for $l \in \{1, \dots, m\}$ and let $l = m+1$, supposing that $r_1, \dots, r_{m+1}$ are distinct roots for A. By the induction hypothesis we have

$$A = \tilde{Q} \cdot (\xi - r_m) \cdot \dots \cdot (\xi - r_1).$$

We must then have

$$\mathrm{Ev}_{\mathsf{R}}(\tilde{Q})(r_{m+1})(r_{m+1} - r_m) \cdots (r_{m+1} - r_1) = 0_{\mathsf{R}}.$$

Since $(r_{m+1})(r_{m+1} - r_m) \cdots (r_{m+1} - r_1) \neq 0_{\mathsf{R}}$ by virtue of R being an integral domain, it follows that $\mathrm{Ev}_{\mathsf{R}}(\tilde{Q})(r_{m+1}) = 0_{\mathsf{R}}$, again by virtue of R being an integral domain. Thus r_{m+1} is a root of $\tilde{Q}$, and so by Proposition 4.4.25 we have $\tilde{Q} = Q \cdot (\xi - r_{m+1})$, and so the result holds for $l = m + 1$. ■

Next we consider the case where a polynomial may not have all roots distinct. To make sense of what this even means requires the following definition.

4.4.27 Definition (Multiplicity of a root) Let R be an integral domain, let $A \in \mathsf{R}[\xi]$, and let $r \in \mathsf{R}$ be a root of A. The root r:

(i) has ***multiplicity*** $k \in \mathbb{Z}_{>0}$ if $(\xi - r)^k | A$ but $(\xi - r)^{k+1} \nmid A$;

(ii) is a ***simple root*** if it has multiplicity 1;

(iii) is a ***multiple root*** if it is not a simple root. •

It turns out that there is a simple check on when a polynomial has a multiple root. It involves introducing the notion of the derivative of a polynomial over a ring. To motivate this definition, note that the derivative of a polynomial function in calculus is again a polynomial function.

4.4.28 Definition (Formal derivative of a polynomial) If R is a unit ring and if $A = \sum_{j=0}^{k} a_j \xi^j \in \mathsf{R}[\xi]$, then the polynomial $A' = \sum_{j=1}^{k} j a_j \xi^{j-1}$ is the ***formal derivative*** of A. •

The formal derivative shares many of the properties of the usual derivative from calculus, but without requiring the notion of a limit for its definition. We refer the reader to Exercise 4.4.5 for a summary of some of these properties. Our main interest in the formal derivative is the following result, which the reader can check themselves for polynomial functions using the Chain Rule.

4.4.29 Proposition (Multiple roots are roots of the formal derivative) *Let* R *be a commutative unit ring and let* $\mathrm{A} \in \mathsf{R}[\xi]$. *Then the following statements hold:*

(i) if r *is a root of multiplicity greater than* 1 *then* r *is a root of* A′;

(ii) if additionally R *is an integral domain and if* r *is a root of* A *and* A′, *then* r *is a root of multiplicity greater than* 1 *of* A.

Proof (i) We have $A = (\xi - r)^k \cdot B$ for some $k \geq 2$ and for some $B \in \mathsf{F}[\xi]$. Then, referring to Exercise 4.4.5,

$$A' = k(\xi - r)^{k-1} \cdot B + (\xi - r)^k \cdot B' = (\xi - r)\Big(k(\xi - r)^{k-2} \cdot B + (\xi - r)^{k-1} \cdot B'\Big).$$

Thus $(\xi - r)|A'$ and so r is a root of A' by Proposition 4.4.25.

(ii) Since r is a root of A, we have $A = (\xi - r)^k \cdot B$ for some $k \geq 1$. Suppose that $k = 1$, Then, as in the preceding part of the proof, $A' = B + (\xi - r)B'$ and so, if r is a root of a', we must have that r is a root of B. But then $B = (\xi - r) \cdot C$ for some polynomial C, and the resulting contradiction allows us to conclude that $k > 1$. ■

Next we turn to the topic of irreducible and prime polynomials. As a segue from the topic of roots to this new topic, we have the following result.

4.4.30 Proposition (Roots and irreducibility) *If* R *is a commutative unit ring and if* $\mathrm{A} \in \mathsf{R}[\xi]$, *then the following statements hold:*

(i) if $\deg(\mathrm{A}) > 1$ *and if* A *has a root* $\mathrm{r} \in \mathsf{R}$, *then* A *is reducible;*

(ii) if R *is an integral domain, if* A *is primitive, and if* $\deg(\mathrm{A}) = 1$, *then* A *is irreducible.*

Proof (i) If A has a root $r \in \mathsf{R}$ then, by Proposition 4.4.25, we have $A = Q \cdot (\xi - r)$ for some $Q \in \mathsf{R}[\xi]$. If $\deg(A) > 1$ then, by Proposition 4.4.11, $1 < \deg(A) \leq \deg(Q) + 1$, meaning that $\deg(Q) \geq 1$, and so Q is not a unit in $\mathsf{R}[\xi]$ by Exercise 4.4.3. Thus A is reducible.

(ii) Write $A = B \cdot C$ for $B, C \in \mathsf{R}[\xi]$. Then $1 = \deg(A) = \deg(B) + \deg(C)$ by Proposition 4.4.11. We must then wither have $\deg(B) = 0$ and $\deg(C) = 1$, or $\deg(B) = 1$ and $\deg(C) = 0$. Thus either B or C, without loss of generality suppose B, must be a constant. Thus B must be a divisor of all coefficients of A, and since A is primitive, we conclude that B is a unit. Thus A is irreducible. ■

Let us give a few examples that illustrate the conclusions and necessarily omitted conclusions of the preceding result.

4.4.31 Examples (Roots and irreducibility)

1. The polynomial $A = \xi^3 - \xi^2 + \xi - 1$ is not reducible in $\mathbb{Z}[\xi]$ since it has 1 as a root. Note that $A = (\xi - 1)(\xi^2 + 1)$ is a factorisation of A into nonunits.
2. The polynomial $\xi^2 + 1$ is irreducible in $\mathbb{R}[\xi]$. Indeed, were A to be reducible then, since $\mathbb{R}$ is a field, we could write $A = (\xi - r_1)(\xi - r_2)$ for $r_1, r_2 \in \mathbb{R}$, meaning that A would have a root in $\mathbb{R}$. However, there is no real number whose square is equal to -1 since the square of a real number is always positive. (Of course, A is reducible in $\mathbb{C}[\xi]$, but this is the topic of Section 4.7.)
3. In $\mathbb{Z}[\xi]$ the polynomial $A = 2\xi + 2$ is reducible since $A = 2 \cdot (\xi + 1)$ and neither 2 nor $\xi + 1$ are units in $\mathbb{Z}[\xi]$. Note, however, that A is irreducible if thought of as a polynomial in $\mathbb{R}[\xi]$. •

Readers only used to polynomials over $\mathbb{R}$ or $\mathbb{C}$ will only have encountered polynomial rings whose irreducibles have degree 0, 1, or 2 (see Section 4.7.3). However, it is possible for irreducible polynomials to have arbitrarily large degree. In order to exhibit such a polynomial, we present the following general result.

4.4.32 Theorem (Eisenstein's Criterion) *Let* R *be a unique factorisation domain and let* $\mathrm{A} = \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{a_j}\xi^{\mathrm{j}}$ *be a polynomial in* $\mathsf{R}[\xi]$ *of degree* $\mathrm{k} \geq 1$. *If there exists an irreducible* $\mathrm{p} \in \mathsf{R}$ *such that*

(i) $\mathrm{p}|\mathrm{a_j}$ *for* $\mathrm{j} \in \{0, 1, \dots, \mathrm{k} - 1\}$,

(ii) $\mathrm{p} \nmid \mathrm{a_k}$, *and*

(iii) $\mathrm{p}^2 \nmid \mathrm{a_0}$,

then A *is irreducible in* $\mathsf{F}_{\mathsf{R}}[\xi]$. *Moreover, if* A *is primitive, then* A *is irreducible in* $\mathsf{R}[\xi]$.

Proof By Proposition 4.4.19(i), write $A = c_A A'$ with c_A a content for A and A' primitive in $\mathsf{R}[\xi]$. Since c_A is nonzero, it is a unit in F_{R}, and so the result will follow if A' is irreducible in $\mathsf{F}_{\mathsf{R}}[\xi]$ by Exercise 4.2.18. By Proposition 4.4.19(v) it then suffices to show that A' is irreducible in $\mathsf{R}[\xi]$. So suppose that $A' = B \cdot C$ and write

$$A' = \sum_{j=0}^{k} a'_j \xi^j, \quad B = \sum_{j=0}^{r} b_j \xi^j, \quad C = \sum_{j=0}^{s} c_j \xi^j$$

with $r, s \geq 1$ and with $b_r, c_s \neq 0_{\mathsf{R}}$. Since $p \nmid a_k$ we have $p \nmid c_A$. Therefore, for $j \in \{0, 1, \dots, k-1\}$, $p|a'_j$ if and only if $p|a_j$. In particular, $p|a'_0 = b_0 c_0$. Since p is a prime in R by Proposition 4.2.70, it follows that either $p|b_0$ or $p|c_0$. Suppose without loss of generality that $|b_0$. Since $p^2 \nmid a_0$ and since $p \nmid c_A$, $p^2 \nmid a'_0$, from which we conclude that $p \nmid c_0$. Define

$$n_B = \inf\{l \in \mathbb{Z}_{\geq 0} \mid p|b_j, j \in \{0, 1, \dots, l\}, p \nmid b_l\}.$$

Note that n_B is well defined since $p \nmid b_j$ for some j, since otherwise p would divides every coefficient of $B \cdot C$, contradicting the fact that A' is primitive. Now note that

$$a'_{n_B} = b_0 c_{n_B} + b_1 c_{n_B - 1} + \cdots + b_{n_B - 1} c_1 + b_{n_B} c_0.$$

Since $p|a_{n_B}$, $p|a'_{n_B}$, from which we conclude, using the definition of n_B, that $p|b_{n_B} c_0$, which means that either $p|b_{n_B}$ or $p|c_0$, which is a contradiction. Thus we cannot write $A' = B \cdot C$ where $\deg(B), \deg(C) \geq 1$, and so A' is irreducible in $\mathsf{R}[\xi]$.

The final assertion is one half of Proposition 4.4.19(v). ■

Now let us give an interesting (and important, although for reasons that we will not touch upon) example of an irreducible polynomial.

4.4.33 Example (Irreducibility of certain cyclotomic polynomials) For $k \in \mathbb{Z}_{>0}$ the ***cyclotomic polynomial*** of degree k is the polynomial in $\mathbb{Z}[\xi]$ given by

$$\Phi_k = \xi^k + \xi^{k-1} + \cdots + \xi + 1.$$

We claim that, if $p \in \mathbb{Z}_{>0}$ is prime, then Φ_{p-1} is irreducible. The most enlightening proof of this fact involves Galois theory. We give a less enlightening, but more direct proof. We nonetheless do rely on something we have yet to develop, namely rational functions which we introduce in Section 4.4.8.

We use a sequence of lemmata to prove the irreducibility of Φ_{p-1}. The first lemma expresses the cyclotomic polynomial as a rational function.

1 Lemma $\Phi_{\mathrm{k}} = \dfrac{\xi^{\mathrm{k}+1} - 1}{\xi - 1}$.

Proof For $k = 1$ we have

$$\frac{\xi^2 - 1}{\xi - 1} = \frac{(\xi + 1)(\xi - 1)}{\xi - 1} = \xi + 1 = \Phi_1.$$

Now suppose the result true for $k \in \{1, \dots, l-1\}$ and compute

$$\Phi_l = \sum_{j=0}^{l} \xi^j = \xi^l + \frac{\xi^l - 1}{\xi - 1} = \frac{\xi^l(\xi - 1) + \xi^l - 1}{\xi - 1} = \frac{\xi^{l+1} - \xi^l + \xi^l - 1}{\xi - 1} = \frac{\xi^{l+1} - 1}{\xi - 1}. \qquad ▼$$

2 Lemma *If* $\mathrm{k} \in \mathbb{Z}_{>0}$*, then* $\frac{(\xi+1)^{\mathrm{k}}-1}{\xi}$ *is a polynomial in* $\mathbb{Z}[\xi]$*. If* k *is additionally prime, then this polynomial is irreducible.*

Proof We use the Binomial Theorem, Proposition 4.2.11, to compute

$$(\xi+1)^k - 1 = -1 + \sum_{j=0}^{k} \frac{k!}{j!(k-j)!}\xi^j = \xi \underbrace{\sum_{j=0}^{k-1} \frac{k!}{(j+1)!(k-j-1)!}\xi^j}_{A}.$$

This shows that $\frac{(\xi+1)^k-1}{\xi}$ is a polynomial for $k \in \mathbb{Z}_{>0}$; it is the polynomial A. Now suppose that $k = p$ is prime and let us write $A = \sum_{j=0}^{p-1} a_j\xi^j$. We then have

1. $p|a_j$ for $j \in \{0,1,\dots,p-2\}$ since $a_j = \frac{p!}{(j+1)!(p-j-1)!}$ and since p is prime so that $(j+1) \nmid p$ and $(p-j-1) \nmid p$,
2. $p \nmid a_{p-1}$ since $a_{p-1} = 1$, and
3. $p^2 \nmid a_0$ since $a_0 = \frac{p!}{(p-1)!}$ and since p is prime so that $(p-1) \nmid p$.

The irreducibility of $\frac{(\xi+1)^p-1}{\xi}$ now follows from Eisenstein's Criterion. ▼

3 Lemma *If* $\Phi_{\mathrm{k}-1}$ *is reducible then* $\frac{(\xi+1)^{\mathrm{k}}-1}{\xi}$ *is reducible.*

Proof From Lemma 1 we have

$$\Phi_{k-1} = \frac{\xi^k-1}{\xi-1} = \frac{(\eta+1)^k-1}{\eta}$$

where $\eta = \xi - 1$. Suppose that Φ_{k-1} is reducible and write $\Phi_{k-1} = A \cdot B$ where neither A nor B is a unit. If either $\deg(A) = 0$ or $\deg(B) = 0$ then it follows that A or B are contents for Φ_k. Since Φ_k is primitive, this means that either A or B must be a unit. Therefore we may suppose that $\deg(A), \deg(B) \geq 1$. Write

$$A = \sum_{j=0}^{r} a_j\xi^j, \quad B = \sum_{j=0}^{s} b_j\xi^j.$$

We then have

$$\frac{(\eta+1)^k-1}{\eta} = \frac{\xi^k-1}{\xi-1} = \left(\sum_{j=0}^{r} a_j\xi^j\right)\left(\sum_{j=0}^{s} b_j\xi^j\right) = \left(\sum_{j=0}^{r} a_j(\eta+1)^j\right)\left(\sum_{j=0}^{s} b_j(\eta+1)^j\right).$$

Now note that $\sum_{j=0}^{r} a_j(\eta+1)^j$ and $\sum_{j=0}^{s} b_j(\eta+1)^j$ are polynomials in η of degree $r \geq 1$ and $s \geq 1$, respectively. Therefore $\frac{(\eta+1)^k-1}{\eta} = A' \cdot B'$ for polynomials $A', B' \in \mathbb{Z}[\eta]$ of positive degree. Thus $\frac{(\eta+1)^k-1}{\eta}$ is reducible. ▼

Combining Lemmas 2 and 3 immediately allows us to conclude that Φ_{p-1} is irreducible in $\mathbb{Z}[\xi]$. In particular, since there are infinitely many positive prime numbers by Exercise 4.2.20, this shows that there are irreducible polynomials in $\mathbb{Z}[\xi]$ of arbitrarily large degree. Note that, since Φ_{p-1} is primitive, it follows from Proposition 4.4.19(v) that Φ_{p-1} is also irreducible in $\mathbb{Q}[\xi]$. •

As a final matter, we give the relationship between irreducibles and primes in polynomial rings. First of all, since a ring is always a subring of its polynomial ring by Proposition 4.4.5, any relationships between irreducibles and primes in the polynomial ring must be inherited from the ring itself. Thus the following result is about the best one can expect in terms of relating irreducible and prime polynomials, given our understanding of the relationship between irreducibles and primes in general rings.

4.4.34 Proposition (Relationship between irreducible and prime polynomials) *If* R *is an integral domain, then the following statements hold:*

(i) if $A \in \mathsf{R}[\xi]$ *is prime then it is irreducible;*

(ii) if R *is additionally a unique factorisation domain, then* $A \in \mathsf{R}[\xi]$ *is prime if it is irreducible.*

Proof (i) Since $\mathsf{R}[\xi]$ is an integral domain if R is an integral domain (Theorem 4.4.2), this part of the result follows from Proposition 4.2.65.

(ii) Since $\mathsf{R}[\xi]$ is a unique factorisation domain if R is a unique factorisation domain (Theorem 4.4.20), this part of the result follows from Proposition 4.2.70. ■

4.4.5 Greatest common divisors of polynomials

In this section we essentially restate, for organisational purposes, some of the relevant conclusions of Section 4.2.11 in the setting of polynomials rings. We also illustrate an application of the Euclidean Algorithm for polynomials, since it will be of interest to be able to compute greatest common divisors for polynomials.

First let us state a result that allows us to conclude the existence and character of greatest common divisors in polynomial rings in many cases.

4.4.35 Proposition (Existence and form of greatest common divisors for polynomials) *For* R *be a commutative unit ring and for* $\mathsf{S} = \{A_1, \dots, A_k\} \subseteq \mathsf{R}[\xi]$*, the following statements hold:*

(i) the following statements for $D \in \mathsf{R}[\xi]$ *are equivalent:*

(a) $(D) = (A_1, \dots, A_k)$*;*

(b) D *is a greatest common divisor for* S *of the form* $D = R_1 \cdot A_1 + \cdots + R_k A_k$ *for some* $R_1, \dots, R_k \in \mathsf{R}[\xi]$*;*

(ii) if R *is a field then* S *possesses a greatest common divisor of the form* $D = R_1 \cdot A_1 + \cdots + R_k A_k$ *for some* $R_1, \dots, R_k \in \mathsf{R}[\xi]$*;*

(iii) if R *is a unique factorisation domain then* S *possesses a greatest common divisor.*

Proof (i) This follows from Proposition 4.2.77 since $\mathsf{R}[\xi]$ is a commutative unit ring if R is (Theorem 4.4.2).

(ii) If R is a field then $\mathsf{R}[\xi]$ is a Euclidean domain, and hence a principal ideal domain, by Corollary 4.4.14. This part of the result now follows from Proposition 4.2.77.

(iii) If R is a unique factorisation domain then so too is $\mathsf{R}[\xi]$ (Theorem 4.4.20). This part of the result then follows from Proposition 4.2.77. ■

For fields, this gives the following useful corollary.

4.4.36 Corollary (Bézout's identity for polynomials) *If* F *is a field and if* $\mathrm{A}_1, \dots, \mathrm{A}_k \in \mathsf{F}[\xi]$ *are coprime polynomials, then there exists* $\mathrm{R}_1, \dots, \mathrm{R}_k \in \mathsf{F}[\xi]$ *such that* $\mathrm{R}_1 \cdot \mathrm{A}_1 + \dots + \mathrm{R}_k \cdot \mathrm{A}_k = 1_{\mathsf{F}}$.

Let us first observe that for polynomials over many rings one can, as can be done in the ring $\mathbb{Z}$, select a distinguished member from a collection of greatest common divisors.

4.4.37 Proposition (Selecting from greatest common divisors of polynomials) *If* F *is a field and if* $\mathrm{S} \subseteq \mathsf{F}[\xi]$*, then there exists a unique* $\mathrm{D} \in \mathsf{F}[\xi]$ *with the properties*

(i) D *is monic and*

(ii) D *is a greatest common divisor for* S.

In this case we say that D *is* the ***greatest common divisor*** *for* S.

Proof By Exercise 4.2.22 we know that S possesses a greatest common divisor since $\mathsf{F}[\xi]$ is a principal ideal domain by Corollary 4.4.14. By Exercises 4.2.21 and 4.4.3 we know that, if $D' \in \mathsf{F}[\xi]$ is a greatest common divisor for S, then the set of greatest common divisors has the form

$$\{uD' \mid u \text{ is a unit in } \mathsf{R}\}.$$

In particular, if we take $u = d_k^{-1}$ where d_k is the leading coefficient of D', then we see that the polynomial $D = uD'$ is monic, and also a greatest common divisor. The uniqueness of D is established as follows. If D is a monic greatest common divisor for S then $D = uD'$ for some unit u. If d_k is the leading coefficient of D' we immediately see that $ud_k = 1_{\mathsf{F}}$, so giving $u = d_k^{-1}$. ■

Of course, a similar statement holds for least common multiples, and we state this here, noting that the proof is rather like that for greatest common divisors, making use of Exercises 4.2.24 and 4.2.23.

4.4.38 Proposition (Selecting from least common multiple of polynomials) *If* F *is a field and if* $\mathrm{S} \subseteq \mathsf{F}[\xi]$*, then there exists a unique* $\mathrm{D} \in \mathsf{F}[\xi]$ *with the properties*

(i) D *is monic and*

(ii) D *is a least common multiple for* S.

In this case we say that D *is* the ***least common multiple*** *for* S.

Now let us illustrate the Euclidean Algorithm for polynomials over a field.

4.4.39 Example (The Euclidean Algorithm for polynomials) Consider the field $\mathbb{R}$ and the polynomials

$$A = 10\xi^6 + 55\xi^5 + 105\xi^4 + 81\xi^3 + 19\xi^2 + 2, \quad B = 2\xi^5 + 11\xi^4 + 21\xi^3 + 16\xi^2 + 3\xi - 1.$$

Doing the tedious polynomial long division gives

$$\begin{aligned} 10\xi^6 + 55\xi^5 + 105\xi^4 + 81\xi^3 + 19\xi^2 + 2 &= (5\xi)(2\xi^5 + 11\xi^4 + 21\xi^3 + 16\xi^2 + 3\xi - 1) \\ &\quad + (\xi^3 + 4\xi^2 + 5\xi + 2), \\ 2\xi^5 + 11\xi^4 + 21\xi^3 + 16\xi^2 + 3\xi - 1 &= (2\xi^2 + 3\xi - 1)(\xi^3 + 4\xi^2 + 5\xi + 2) \\ &\quad + (\xi^2 + 2\xi + 1), \\ \xi^3 + 4\xi^2 + 5\xi + 2 &= (\xi + 2)(\xi^2 + 2\xi + 1), \end{aligned}$$

from which we conclude that the greatest common divisor of A and B is $\xi^2+2\xi+1$. •

It is also true that one can, following Theorem 4.2.84, use the Euclidean Algorithm to find polynomials that satisfy the Bézout identity corresponding to a pair of coprime polynomials. Let us record the result, and then illustrate it with an example.

4.4.40 Proposition (Bézout's identity for polynomials using the Euclidean Algorithm) *Let* F *be a field and let* $A, B \in \mathsf{F}[\xi]$ *be coprime polynomials. Then there exists* $R, S \in \mathsf{F}[\xi]$ *such that*

(i) $R \cdot A + S \cdot B = 1_{\mathsf{F}[\xi]}$ *and*

(ii) $\deg(R) < \deg(B)$ *and* $\deg(S) < \deg(A)$.

Proof Recall from the proof of Corollary 4.4.14 that for the Euclidean domain $\mathsf{F}[\xi]$ we define $\delta\colon \mathsf{F}[\xi] \to \mathbb{Z}_{\geq 0}$ by

$$\delta(A) = \begin{cases} \deg(A) + 1, & A \neq 0_{\mathsf{R}[\xi]}, \\ 0, & A = 0_{\mathsf{F}[\xi]}. \end{cases}$$

Therefore,

$$\begin{aligned} \delta(A - B) = \delta(A + (-B)) = \deg(A + (-B)) + 1 \leq \max\{\deg(A), \deg(B)\} + 1 \\ < \deg(A) + \deg(B) + 2 < \delta(A) + \delta(B). \end{aligned}$$

The result now follows immediately from Theorem 4.2.84. ■

Now let us illustrate how to apply the Euclidean Algorithm for polynomials. The computations are a little tedious, albeit straightforward. Moreover, they can be implemented systematically in a symbolic manipulation program, enabling quick computation, at least for low-degree polynomials.

4.4.41 Example (Solving the Bézout identity for polynomials) We consider the field $\mathbb{Q}$ (or $\mathbb{R}$, but the computations are the same) with polynomials $A = x^4 + 6x^3 + 12x^2 + 11x + 6 = (x^2 + x + 1)(x + 2)(x + 3)$ and $B = x^2 + 12x + 35 = (x + 5)(x + 7)$. Since these polynomials have no common prime factors, they are coprime. Let us write out the Euclidean Algorithm:

$$\begin{aligned} x^4 + 6x^3 + 12x^2 + 11x + 6 &= (x^2 - 6x + 49)(x^2 + 12x + 35) + (-367x - 1709), \\ (x^2 + 12x + 35) &= (-\tfrac{1}{367}x - \tfrac{2695}{134689})(-367x - 1709) + \tfrac{108360}{134689} \\ (-367x - 1709) &= -\tfrac{49430863}{108360}x - \tfrac{230183501}{108360}. \end{aligned}$$

Note that this tells us that $u = \frac{108360}{134689}$ is a greatest common divisor for A and B. Since any unit multiplied by a greatest common divisor is also a greatest common divisor (Exercise 4.2.21), it follows that 1 is a greatest common divisor, and so A and B are coprime. One can now apply Theorem 4.2.84 directly. One defines $\alpha_0 = 1$ and $\beta_0 = \frac{1}{367}x + \frac{2695}{134689}$ and then

$$\begin{aligned} \alpha_1 &= \beta_0 = \tfrac{1}{367}x + \tfrac{2695}{134689}, \\ \beta_1 &= \alpha_0 - (x^2 - 6x + 49)\beta_0 = -\tfrac{1}{367}x^3 - \tfrac{493}{134689}x^2 - \tfrac{1813}{134689}x - \tfrac{266744}{134689}. \end{aligned}$$

Taking $R = u\alpha_1$ and $S = u\beta_1$ gives $R \cdot A + S \cdot B = 1$, as desired. Furthermore, note that $\deg(R) = 1 < 2 = \deg(B)$ and $\deg(S) = 3 < 4 = \deg(A)$, as predicted by Theorem 4.2.84. •

4.4.6 Quotients of polynomial rings by ideals

As we shall see in Section 4.6, to construct a field which contains the roots of a given polynomial, one uses quotients of polynomial rings by prime ideals. In this section we consider quotients of polynomial rings by general ideals. Let us remind the reader of some facts about ideals of polynomial rings. From Corollary 4.4.14 we know that $\mathsf{F}[\xi]$ is a Euclidean domain, and hence a principal ideal domain. Therefore, every ideal in $\mathsf{F}[\xi]$ is generated by some polynomial A. Moreover, if $A \in \mathsf{F}[\xi]$ then we can write $A = aA'$ where $s \in \mathsf{F}^*$ and where A' is a monic polynomial (simply take a to be the leading coefficient of A). Since $A|A'$ and $A'|A$ by Proposition 4.2.60, $(A) = (A')$ by Proposition 4.2.61. Therefore, in considering principal ideals (A) in $\mathsf{F}[\xi]$, we may without loss of generality suppose that A is monic. If A is monic and has degree 0, then A is a unit in $\mathsf{F}[\xi]$. Then, by Proposition 4.2.61, $(A) = \mathsf{F}[\xi]$. This case will not be of interest to us, since we will be considering the quotient ring $\mathsf{F}[\xi]/(A)$, and if $(A) = \mathsf{F}[\xi]$, then this quotient ring is the zero ring. Thus it is most interesting to consider the case when $\deg(A) \geq 1$.

Our first result gives a convenient representation for elements in the quotient ring of the polynomial ring and the ideal generated by a general monic polynomial. The result also explicitly describes the ring structure of the quotient.

4.4.42 Proposition (Quotients of polynomial rings by principal ideals) *Let* R *be a commutative unit ring and let* A *be a monic polynomial of degree* $\mathrm{k} \geq 1$*, and write* $\mathrm{A} = \xi^{\mathrm{k}} - \sum_{\mathrm{j}=0}^{\mathrm{k}-1} \mathrm{a_j}\xi^{\mathrm{j}}$*. Recursively define* $\alpha_0, \alpha_1, \dots, \alpha_{\mathrm{k}-1} \in \mathsf{F}$ *by* $\alpha_0 = 1_{\mathsf{R}}$ *and*

$$\alpha_{\mathrm{m}} = \sum_{\mathrm{j}=1}^{\mathrm{m}} \mathrm{a}_{\mathrm{k}-\mathrm{j}}\alpha_{\mathrm{m}-\mathrm{j}}.$$

Then, given $\mathrm{B} \in \mathsf{R}[\xi]$ *there exists unique* $\mathrm{b}_0, \mathrm{b}_1, \dots, \mathrm{b}_{\mathrm{k}-1} \in \mathsf{R}$ *such that*

$$\mathrm{B} + (\mathrm{A}) = \mathrm{b}_0 + \mathrm{b}_1\xi + \cdots + \mathrm{b}_{\mathrm{k}-1}\xi^{\mathrm{k}-1} + (\mathrm{A}).$$

Moreover, if

$$\begin{aligned}\mathrm{B} + (\mathrm{A}) &= \mathrm{b}_0 + \mathrm{b}_1\xi + \cdots + \mathrm{b}_{\mathrm{k}-1}\xi^{\mathrm{k}-1} + (\mathrm{A}),\\ \mathrm{C} + (\mathrm{A}) &= \mathrm{c}_0 + \mathrm{c}_1\xi + \cdots + \mathrm{c}_{\mathrm{k}-1}\xi^{\mathrm{k}-1} + (\mathrm{A}) \in \mathsf{R}[\xi]/(\mathrm{A}),\end{aligned}$$

then

$$\begin{aligned}(\mathrm{B} + (\mathrm{A})) + (\mathrm{C} + (\mathrm{A})) &= (\mathrm{b}_0 + \mathrm{c}_0) + (\mathrm{b}_1 + \mathrm{c}_1)\xi + \cdots + (\mathrm{b}_{\mathrm{k}-1} + \mathrm{c}_{\mathrm{k}-1})\xi^{\mathrm{k}-1} + (\mathrm{A}),\\ (\mathrm{B} + (\mathrm{A})) \cdot (\mathrm{C} + (\mathrm{A})) &= \sum_{\mathrm{l}=0}^{\mathrm{k}-1}\sum_{\mathrm{j}=0}^{\mathrm{l}} \mathrm{b_j c_{l-j}}\xi^{\mathrm{l}}\\ &\quad + \sum_{\mathrm{l}=0}^{\mathrm{k}-2}\sum_{\mathrm{j}=0}^{\mathrm{l}} \mathrm{b}_{\mathrm{k}-\mathrm{j}}\mathrm{c}_{\mathrm{k}-(\mathrm{l}-\mathrm{j})} \sum_{\mathrm{r}=0}^{\mathrm{k}-\mathrm{j}-1} \alpha_{\mathrm{r}} \sum_{\mathrm{s}=0}^{\mathrm{j}+1} \mathrm{a_s}\xi^{\mathrm{s}+(\mathrm{k}-\mathrm{j}-1)-\mathrm{r}} + (\mathrm{A}).\end{aligned}$$

Proof If $B = 0_{\mathsf{R}[\xi]}$ then take $b_0 = b_1 = \cdots = b_{k-1} = 0_{\mathsf{R}}$. These are clearly the only elements of R for which

$$(A) = b_0 + b_1\xi + \cdots + b_{k-1}\xi^{k-1} + (A).$$

Now suppose that $\deg(B) \geq 0$. Since A is monic, its leading coefficient is not a zerodivisor (it is a unit), and so we can apply Theorem 4.4.13 to conclude that there exists $Q, R \in \mathsf{R}[\xi]$ such that $B = Q \cdot A + R$, with $\deg(R) < k$. Thus we can write $R = \sum_{j=0}^{k-1} b_j\xi^j$. Since $Q \cdot A \in (A)$ we then have $B + (A) = R + (A)$¡ and this gives the existence part of the result. To prove uniqueness, suppose that

$$b_0 + b_1\xi + \cdots + b_{k-1}\xi^{k-1} + (A) = b'_0 + b'_1\xi + \cdots + b'_{k-1}\xi^{k-1} + (A).$$

Then

$$b_0 + b_1\xi + \cdots + b_{k-1}\xi^{k-1} = b'_0 + b'_1\xi + \cdots + b'_{k-1}\xi^{k-1} + C \cdot A$$

for some $C \in \mathsf{R}[\xi]$ by Theorem 4.2.54. But by Proposition 4.4.11 this implies that $\deg(C \cdot A) = \deg(C) + \deg(A) < k$. Since $\deg(A) = k$ we must therefore have $\deg(C) = -\infty$ or $C = 0_{\mathsf{R}[\xi]}$. Thus $b_j = b'_j$ for $j \in \{0, 1, \dots, k-1\}$.

Now let us prove that ring addition and multiplication in $\mathsf{R}[\xi]$ have the stated form. It is clear by definition of addition in quotient rings that addition has the form given. For multiplication we use the following lemma.

1 Lemma *For* $\mathrm{m} \in \{0, 1, \dots, \mathrm{k}-1\}$,

$$\xi^{\mathrm{k+m}} + (\mathrm{A}) = \sum_{\mathrm{l=0}}^{\mathrm{m}} \alpha_{\mathrm{l}} \sum_{\mathrm{j=0}}^{\mathrm{k-1-m+l}} \mathrm{a_j} \xi^{\mathrm{j+m-l}} + (\mathrm{A}).$$

Proof We have,

$$\xi^k + (A) = \sum_{j=0}^{k-1} a_j \xi^j + (A),$$

and one can check that this agrees with the lemma for $m = 0$. Now suppose that the expression in the lemma for $\xi^{k+m} + (A)$ holds for $m \in \{0, 1, \dots, r\}$ and compute

$$\begin{aligned}
\xi^{k+r+1} + (A) = \xi\xi^{k+r} + (A) &= \sum_{l=0}^{r} \alpha_l \sum_{j=0}^{k-1-r+l} a_j \xi^{j+r-l+1} + (A) \\
&= \sum_{l=0}^{r} \alpha_l \sum_{j=0}^{k-2-r+l} a_j \xi^{j+r-l+1} + \sum_{l=0}^{r} \alpha_l a_{k+l-(r+1)} \xi^k + (A) \\
&= \sum_{l=0}^{r} \alpha_l \sum_{j=0}^{k-2-r+l} a_j \xi^{j+r-l+1} + \sum_{l=0}^{r} \alpha_l a_{k+l-(r+1)} \sum_{s=0}^{k-1} a_s \xi^s + (A) \\
&= \sum_{l=0}^{r} \alpha_l \sum_{j=0}^{k-2-r+l} a_j \xi^{j+r-l+1} + \sum_{l=1}^{r+1} a_{k-j} \alpha_{r+1-j} \sum_{s=0}^{k-1} a_s \xi^s + (A) \\
&= \sum_{l=0}^{r+1} \alpha_l \sum_{j=0}^{k-1-(r+1)+l} a_j \xi^{j+(r+1)-l} + (A),
\end{aligned}$$

which may be checked to be the conclusion of the lemma for $m = r + 1$. ▼

We then compute

$$\begin{aligned}
(B + (A)) \cdot (C + (A)) &= \left(\sum_{j=0}^{k-1} b_j \xi^j \right) \left(\sum_{j=0}^{k-1} c_j \xi^j \right) + (A) \\
&= \sum_{l=0}^{k-1} \sum_{j=0}^{l} b_j c_{l-j} \xi^l + \sum_{l=0}^{k-2} \sum_{j=0}^{l} b_{k-j} c_{k-(l-j)} \xi^{2(k-1)-j} + (A) \\
&= \sum_{l=0}^{k-1} \sum_{j=0}^{l} b_j c_{l-j} \xi^l + \sum_{l=0}^{k-2} \sum_{j=0}^{l} b_{k-j} c_{k-(l-j)} \xi^{k+(k-j-1)} + (A) \\
&= \sum_{l=0}^{k-1} \sum_{j=0}^{l} b_j c_{l-j} \xi^l \\
&\quad + \sum_{l=0}^{k-2} \sum_{j=0}^{l} b_{k-j} c_{k-(l-j)} \sum_{r=0}^{k-j-1} \alpha_r \sum_{s=0}^{j+1} a_s \xi^{s+(k-j-1)-r} + (A).
\end{aligned}$$

This is the expression stated in the proposition. ■

Note that multiplication in the quotient ring $\mathsf{R}[\xi]/(A)$ can be very complicated. Indeed, it is far from clear that it even has the properties needed for multiplication in a ring, and the only reason we know it does is that this follows from the development of the general theory. In any event, these quotient rings are at least a good way to generate rings with a very complicated ring structure. However, we also have other reasons for being interested in these rings.

Let us illustrate the proposition on an example. A more useful, and in fact simpler, example will come up in Section 4.7.2.

4.4.43 Example (Quotient of a polynomial ring by a principal ideal) We consider the ring $\mathbb{Z}$ with its polynomial ring $\mathbb{Z}[\xi]$, and we recall from Example 4.4.33 the cyclotomic polynomial $\Phi_2 = \xi^2 + \xi + 1$. By Proposition 4.4.42 we know that every element of the quotient ring $\mathbb{Z}[\xi]/(\Phi_2)$ is uniquely expressed as $b_0 + b_1\xi + (\Phi_2)$ for $b_0, b_1 \in \mathbb{Z}$. If

$$B + (\Phi_2) = b_0 + b_1\xi + (\Phi_2), C + (\Phi_2) = c_0 + c_1 + (\Phi_2) \in \mathbb{Z}[\xi]/(\Phi_2),$$

then we have

$$(B + (\Phi_2)) + (C + (\Phi_2)) = (b_0 + c_0) + (b_1 + c_1)\xi + (\Phi_2).$$

To express the product of $B + (\Phi_2)$ and $C + (\Phi_2)$ we note that, in the terminology of Proposition 4.4.42, $a_0 = a_1 = -1$. Rather than apply the cumbersome product formula of Proposition 4.4.42 (which does have the virtue of being implementable in a symbolic computation program), let us compute the product directly. We first compute

$$\xi^2 + (\Phi_2) = -\xi - 1 + (\Phi_2).$$

Then we have

$$\begin{aligned}(B + (\Phi_2)) \cdot (C + (\Phi_2)) &= (b_0 + b_1\xi)(c_0 + c_1\xi) + (\Phi_2)\\ &= b_0c_0 + (b_0c_1 + b_1c_0)\xi + b_1c_1\xi^2 + (\Phi_2)\\ &= (b_0c_0 - b_1c_1) + (b_0c_1 + b_1c_0 - b_1c_2)\xi + (\Phi_2).\end{aligned}$$

One could, if one wished, check directly that this sum and product satisfy the conditions for a ring. However, the general theory gives us this conclusion for free. •

Now let us state some facts about the character of the quotient ring $\mathsf{F}[\xi]/(A)$.

4.4.44 Theorem (Properties of the quotient ring $\mathsf{F}[\xi]/(\mathrm{A})$) *If F is a field and if $\mathrm{A} \in \mathsf{F}[\xi]$ is a monic polynomial of degree at least 1, then the following statements hold:*

(i) the ring $\mathsf{F}/(\mathrm{A})$ contains a subring isomorphic to F;

(ii) if A is irreducible then $\mathsf{F}[\xi]/(\mathrm{A})$ is a field;

(iii) if A *is reducible then* $\mathsf{F}[\xi]/(\mathrm{A})$ *is not an integral domain.*

Proof For the first assertion, let $\pi_{(A)}\colon \mathsf{F}[\xi] \to \mathsf{F}[\xi]/(A)$ be the canonical projection mapping $B \in \mathsf{F}[\xi]$ to $B + (A)$. We claim that the restriction of $\pi_{(A)}$ to $\mathsf{F} \subseteq \mathsf{F}[\xi]$ is injective. Indeed, suppose that $\pi_{(A)}(a) = 0_{\mathsf{F}[\xi]/(A)}$ for some $a \in \mathsf{F}$. Then $a \in (A)$, whence, by Theorem 4.2.54, $a = B \cdot A$ for some $B \in \mathsf{F}[\xi]$. Since $\deg(a) = \deg(B) + \deg(A)$ and since $\deg(A) \geq 1$, it holds that $\deg(a) = \deg(B) = -\infty$, and so $a = 0_{\mathsf{F}}$. Thus $\pi_{(A)}|\mathsf{F}$ is injective. Therefore, $\mathrm{image}(\pi_{(A)})$ is a subring of $\mathsf{F}[\xi]/(A)$ isomorphic to F.

Note that, by Theorem 4.2.64, Theorems 4.2.37 and 4.3.9, and Proposition 4.4.34, the following statements are equivalent:

1. A is prime;
2. A is irreducible;
3. (A) is prime;
4. (A) is maximal;
5. $\mathsf{F}[\xi]/(A)$ is a field;
6. $\mathsf{F}[\xi]/(A)$ is an integral domain.

From this, the last two statements of the theorem follow immediately. ■

Before we proceed, let us consider Example 4.4.43 in the context of the theorem.

4.4.45 Example (Quotient of a polynomial ring by a principal ideal (cont'd)) We now consider the field $\mathbb{Q}$ with its polynomial ring $\mathbb{Q}[\xi]$. Since $\mathbb{Z}$ is a subring of $\mathbb{Q}$, $\mathbb{Z}[\xi]$ is a subring of $\mathbb{Q}[\xi]$ (Proposition 4.4.8), and so the polynomial $\Phi_2 = \xi^2 + \xi + 1$ can be thought of as a polynomial, not only in $\mathbb{Z}[\xi]$, but in $\mathbb{Q}[\xi]$. As we showed in Example 4.4.33, Φ_2 is irreducible in $\mathbb{Z}[\xi]$. By Proposition 4.4.19(v), since Φ_2 is primitive in $\mathbb{Z}[\xi]$, Φ_2 is also irreducible in $\mathbb{Q}[\xi]$. Therefore, by Theorem 4.4.44, $\mathbb{Q}[\xi]/(\Phi_2)$ is a field. As a result, the ring operations from Example 4.4.43 actually define a field if one allows b_0, b_1, c_0, and c_1 to be rational numbers rather than integers. Again, it is not so easy to verify directly the fact that the product admits a multiplicative inverse, but this follows directly from Theorem 4.4.44. •

4.4.7 Polynomials in multiple indeterminates

In this section we give a very quick definition of polynomials in more than one indeterminate. The case of the single indeterminate worked out in detail will hopefully serve as adequate preparation for the general case.

Let $X = (\xi_i)_{i\in I}$ be an arbitrary family of indeterminates, i.e., some set indexed by a set I. Denote by $(\mathbb{Z}^I_{\geq 0})_0$ the set of maps $\phi\colon I \to \mathbb{Z}_{\geq 0}$ such that the set $\{\xi_i \mid \phi(\xi_i) \neq 0\}$ is finite. Now denote by $\mathsf{F}[X]$ the set of maps $\Psi\colon (\mathbb{Z}^I_{\geq 0})_0 \to \mathsf{F}$ such that the set $\{\phi \mid \Psi(\phi) \neq 0_{\mathsf{F}}\}$ is finite. If $X = (\xi_1, \dots, \xi_k)$ then we shall write $\mathsf{F}[X] = \mathsf{F}[\xi_1, \dots, \xi_k]$. We define the structure of a commutative ring on $\mathsf{F}[X]$ by

$$(\Psi_1 + \Psi_2)(\phi) = \Psi_1(\phi) + \Psi_2(\phi), \quad (\Psi_1\Psi_2)(\phi) = \Psi_1(\phi)\Psi_2(\phi).$$

One should think of $\mathsf{F}[X]$ as being the polynomial ring with indeterminates X and coefficients in F. The monomials are just elements of $(\mathbb{Z}_{\geq 0}^{I})_0$, and so an element ϕ can be written as

$$\xi_{i_1}^{j_1} \cdots \xi_{i_k}^{i_k},$$

where ϕ is defined by

$$\phi(\xi_i) = \begin{cases} j_l, & i = i_l,\ l \in \{1, \dots, k\}, \\ 0, & \text{otherwise.} \end{cases}$$

An element of $\mathsf{F}[X]$ is then a finite linear combination of these monomials with coefficients in F. In order to express such a linear combination in a simple way, it is convenient to use so-called "multi-index notation." We let $X = (\xi_i)_{i\in I}$ be a set of indeterminates indexed by a set I. An element of $(\mathbb{Z}_{\geq 0}^{I})_0$ defines a family $(j_i)_{i\in I}$ of elements of $\mathbb{Z}_{\geq 0}$, only finitely many of which are nonzero. Such a family we call a **I-*multi-index***. If $J = (j_i)_{i\in I}$ is an I-multi-index, let $j_{i_1}, \dots, j_{i_{k(J)}}$ be the nonzero terms in the family. Then we denote by

$$\xi^J = \xi_{i_1}^{j_{i_1}} \cdots \xi_{i_{k(J)}}^{j_{i_{k(J)}}}$$

the corresponding monomial. Then an element $A \in \mathsf{F}[X]$ has the form

$$A = \sum_{J \text{ an } I\text{-multi-index}} C_A(J)\xi^J,$$

for a function $C_A \colon (\mathbb{Z}_{\geq 0}^{I})_0 \to \mathsf{F}$ which is zero except for finitely many values of the argument. This is the ***multi-index form*** for A.

We leave it to the reader to check that $\mathsf{F}[X]$ is indeed a commutative ring. This is relatively easy using multi-index notation.

Given $A \in \mathsf{F}[X]$ with $X = (\xi_i)_{i\in I}$ we can define an evaluation homomorphism Ev_{F} which assigns to a polynomial $A \in \mathsf{F}[X]$ an F-valued function on the vector space F_0^I (see Example 4.5.43) as follows. Let $A \in \mathsf{F}[X]$ and write

$$A = \sum_{J \text{ an } I\text{-multi-index}} C_A(J)\xi^J.$$

If $v \in \mathsf{F}_0^I$ then we think of v as a map from I to F which is zero except for finitely many values of its argument. Thus an element $v \in \mathsf{F}_0^I$ defines a family $(v_i)_{i\in I}$ of elements of F, only finitely many of which are nonzero. If J is an I-multi-index with nonzero elements $j_{i_1}, \dots, j_{i_{k(J)}}$, then we write

$$v^J = v_{i_1}^{j_{i_1}} \cdots v_{i_{k(J)}}^{j_{i_{k(J)}}}.$$

We then define

$$\mathrm{Ev}_{\mathsf{F}}(A)(v) = \sum_{J \text{ an } I\text{-multi-index}} C_A(J)v^J.$$

It is straightforward to check that this is a homomorphism of rings.

4.4.8 Rational functions

Rational functions are simply fractions of polynomials. As with polynomials, we regard rational functions (despite their name) as being algebraic objects rather than functions of an independent variable.

Let us give the formal definition, recalling from Theorem 4.3.6 that the fraction field of an integral domain is a field, and from Theorem 4.4.2 that the ring of polynomials over an integral domain is an integral domain.

4.4.46 Definition (Rational function) If R is an integral domain, then the ***field of rational functions*** over R is the fraction field of the integral domain $\mathsf{R}[\xi]$. The field of rational functions over R is denoted by $\mathsf{R}(\xi)$. •

Let us say a few words about the manner in which rational functions are represented. As per Definition 4.3.5, we denote by $\frac{N}{D}$ the equivalence class associated with $(N, D) \in \mathsf{R}\times\mathsf{R}\setminus\{0_{\mathsf{R}}\}$. The following result gives a means of choosing a standard representative from each equivalence class for a rational function over a field.

4.4.47 Proposition (Coprime fractional representative) *If F is a field and if $\mathrm{R} \in \mathsf{F}(\xi)$ is a rational function over F, then there exists unique polynomials $\mathrm{N}, \mathrm{D} \in \mathsf{F}[\xi]$ such that*

(i) D *is monic,*

(ii) N *and* D *are coprime, and*

(iii) $\mathrm{R} = \frac{\mathrm{N}}{\mathrm{D}}$.

*We call $\frac{\mathrm{N}}{\mathrm{D}}$ the **coprime fractional representative** of* R.

Proof Suppose that $R = \frac{N'}{D'}$ (here and throughout the proof, N' does *not* mean the formal derivative of N). Let Q be the greatest common divisor for N' and D' so that $N' = QN''$ and $D' = QD''$. Then, since $N'D'' = N''D'$, $\frac{N'}{D'} = \frac{N''}{D''}$. Moreover, N'' and D'' are coprime. Write

$$N'' = \sum_{j=0}^{k} c_j \xi^j, \quad D'' = \sum_{j=0}^{l} p_j \xi^j,$$

with $p_l \neq 0_{\mathsf{F}}$. Note that $\frac{p_l^{-1}N''}{p_l^{-1}D''} = \frac{N''}{D''}$ since $N''(p_l^{-1}D'') = (p_l^{-1}N'')D''$. Therefore, $R = \frac{p_l^{-1}N''}{p_l^{-1}D''}$. Since $p_l^{-1}D''$ is monic, the existence part of the result follows.

Now suppose that $\frac{N}{D}$ and $\frac{N'}{D'}$ are two coprime fractional representatives. Then $ND' = N'D$. Since $\mathsf{F}[\xi]$ is a unique factorisation domain let us write

$$N = P_1 \cdots P_k, \quad N' = P'_1 \cdots P'_{k'}, \quad D = Q_1 \cdots Q_l, \quad D' = Q'_1 \cdots Q'_{l'}$$

for irreducible polynomials P_r, $r \in \{1, \dots, k\}$, P'_r, $r \in \{1, \dots, k'\}$, Q_s, $s \in \{1, \dots, l\}$, and Q'_s, $s \in \{1, \dots, l'\}$, all of degree at least 1. Since N and D are coprime and N' and D' are coprime, P_r and Q_s are coprime for all $r \in \{1, \dots, k\}$ and $s \in \{1, \dots, l\}$, and P'_r and Q'_s are coprime for all $r \in \{1, \dots, k'\}$ and $s \in \{1, \dots, l'\}$. In particular, for $r \in \{1, \dots, k\}$ and

$s \in \{1,\dots,l\}$, it cannot hold that $P_r = uQ_s$ for some unit $u \in \mathsf{F}$, and, for $r \in \{1,\dots,k'\}$ and $s \in \{1,\dots,l'\}$, it cannot hold that $P'_r = uQ'_s$ for some unit $u \in \mathsf{F}$. Now note that

$$ND' = N'D = P_1 \cdots P_k Q'_1 \cdots Q'_{l'} = P'_1 \cdots P'_{k'} Q_1 \cdots Q_l.$$

The rightmost two expressions must be factorisations of $ND' = N'D$, and so $k+l' = k'+l$ and the terms differ only by multiplication by units in $\mathsf{F}[\xi]$, i.e., by units in F by Exercise 4.4.3. By our previous observations about the coprimeness of P_r and Q_s, $r \in \{1,\dots,k\}$, $s \in \{1,\dots,l\}$, and P'_r and Q'_s, $r \in \{1,\dots,k'\}$, $s \in \{1,\dots,l'\}$, it must then be the case that $k = k'$, $l = l'$, and there exists units $u_1,\dots,u_l \in \mathsf{F}$ and $\sigma \in \mathfrak{S}_l$ such that $Q'_{\sigma(s)} = u_s Q_s$ for $s \in \{1,\dots,l\}$. In particular,

$$Q'_{\sigma(1)} \cdots Q'_{\sigma(l)} = u_1 \cdots u_l Q_1 \cdots Q_l,$$

or $D' = u_1 \cdots u_l D$. Since both D and D' are monic it follows that $u_1 \cdots u_l = 1_{\mathsf{F}}$, and so $D' = D$. Then $ND = N'D$, whence $N' = N$ by Proposition 4.2.33 since $\mathsf{F}[\xi]$ is an integral domain. ■

The following simple definitions are of some importance in linear system theory, cf. Sections V-?? and V-??.

4.4.48 Definition (Relative degree, proper, strictly proper, biproper) If F is a field and if $R \in \mathsf{F}[\xi]$ has coprime fractional representative $\frac{N}{D}$, then the ***relative degree*** of R is $\deg(D) - \deg(N)$. The rational function R is

(i) ***proper*** if its relative degree is nonnegative,

(ii) ***strictly proper*** if its relative degree is positive, and

(iii) ***biproper*** if its relative degree is zero. •

In many applications in system theory one deals only with rational functions that are proper, and often strictly proper.

Other straightforward and useful language related to the coprime fractional representative is the following.

4.4.49 Definition (Poles and zeros) Let F be a field, let $R \in \mathsf{F}(\xi)$, and let $\frac{N}{D}$ be the coprime fractional representative for R. An element $a \in \mathsf{F}$ is a ***zero*** of R if it is a root of N and is a ***pole*** of R if it is a root of D. The zeros of R are denoted by $\mathrm{Z}(R)$ and the poles of R are denoted by $\mathrm{P}(R)$. The ***multiplicity*** of a zero (resp. pole) is its multiplicity as a root of N (resp. D). •

As is the case with polynomials, it is wise to not think of rational functions as functions, per se, but as algebraic objects. However, as with polynomials, it is possible to assign to a rational function a function in the usual sense. This is done as follows, cf. Proposition 4.4.9.

4.4.50 Definition (Rational functions as functions) Let F be a field and let $R \in \mathsf{F}(\xi)$ have coprime fractional representative $R = \frac{N}{D}$. Define $\mathrm{Ev}_{\mathsf{F}}(R)\colon \mathsf{F} \setminus \mathrm{P}(R) \to \mathsf{F}$ by

$$\mathrm{Ev}_{\mathsf{F}}(R)(a) = \frac{\mathrm{Ev}_{\mathsf{F}}(N)(a)}{\mathrm{Ev}_{\mathsf{F}}(D)(a)}. \qquad \bullet$$

The main result that we prove in this section is the so-called partial fraction decomposition for rational functions.

4.4.51 Theorem (Partial fraction decomposition) *Let* F *be a field, let* $\mathrm{R} \in (\xi)$*, and suppose that* $\mathrm{R} = \frac{\mathrm{N}}{\mathrm{D}}$ *is a coprime fractional representative.*

There exists

(i) m *irreducible monic polynomials* $\mathrm{D}_1, \dots, \mathrm{D}_\mathrm{m} \in \mathsf{F}[\xi]$*,*

(ii) $\mathrm{k}_1, \dots, \mathrm{k}_\mathrm{m} \in \mathbb{Z}_{>0}$*,*

(iii) $\mathrm{k}_1 + \dots + \mathrm{k}_\mathrm{m}$ *polynomials* $\mathrm{N}_{1,1}(\mathrm{x}), \dots, \mathrm{N}_{1,\mathrm{k}_1}, \dots, \mathrm{N}_{\mathrm{m},1}, \dots, \mathrm{N}_{\mathrm{m},\mathrm{k}_\mathrm{m}} \in \mathsf{F}[\xi]$*, and*

(iv) a polynomial $\mathrm{Q} \in \mathsf{F}[\xi]$ *of degree* $\deg(\mathrm{N}) - \deg(\mathrm{D})$ *(take* $\mathrm{Q} = 0$ *if* $\deg(\mathrm{N}) - \deg(\mathrm{D}) < 0$*),*

with the properties

(v) $\mathrm{D}_1, \dots, \mathrm{D}_\mathrm{m}$ *are coprime,*

(vi) $\deg(\mathrm{N}_{\mathrm{r},\mathrm{s}}) < \deg(\mathrm{D}_\mathrm{r})$ *for* $\mathrm{r} \in \{1, \dots, \mathrm{m}\}$ *and* $\mathrm{s} \in \{1, \dots, \mathrm{k}_\mathrm{r}\}$*, and*

(vii) $\displaystyle \mathrm{R} = \sum_{\mathrm{r}=1}^{\mathrm{m}} \sum_{\mathrm{s}=1}^{\mathrm{k}_\mathrm{r}} \frac{\mathrm{N}_{\mathrm{r},\mathrm{s}}}{\mathrm{D}_\mathrm{r}^\mathrm{s}} + \mathrm{Q}.$

Furthermore, the objects described in (i)–(iv) are the unique such objects with the properties (v)–(vii).

Proof Let us first write $R = \frac{\tilde{N}}{D} + \tilde{Q}$ for some polynomials $\tilde{N}$ and Q. If $\deg(N) < \deg(D)$ then we take $\tilde{N} = N$ and $\tilde{Q} = 0_{\mathsf{F}[\xi]}$. If $\deg(N) \geq \deg(D)$ use the Division Algorithm to write $N = \tilde{Q} \cdot D + \tilde{N}$ where $\deg(\tilde{N}) < \deg(D)$. Then $R = \frac{\tilde{N}}{D} + \tilde{Q}$, as desired.

Since D is monic and since $\mathsf{F}[\xi]$ is a unique factorisation domain, we can write

$$D = D_1' \cdots D_k'$$

for irreducible polynomials $D_1', \dots, D_k' \in \mathsf{F}$. For $j \in \{1, \dots, k\}$ we can write $D_j' = u_j D_j''$ where $u_j \in \mathsf{F}$ is a unit and where D_j'' is monic. Since D is monic, $u_1 \cdots u_k = 1_{\mathsf{F}}$. Therefore, D is a product of monic irreducible polynomials. Note that two polynomials D_{j_1}'' and D_{j_2}'', $j_1, j_2 \in \{1, \dots, k\}$, are coprime if and only if they are distinct by virtue of the fact that they are irreducible. Thus we can collect the distinct polynomials from the set $\{D_1'', \dots, D_k''\}$, and we denote these by $D_1, \dots, D_m$. We can then write

$$D = D_1^{k_1} \cdots D_m^{k_m},$$

noting that $D_1, \dots, D_m$ are monic, irreducible, and coprime. Thus we can write

$$\frac{\tilde{N}}{D} = \frac{\tilde{N}}{D_1^{k_1} \cdots D_m^{k_m}}.$$

Since $D_1^{k_1}$ and $D_2^{k_2} \cdots D_m^{k_m}$ are coprime, by Corollary 4.4.36 we have

$$D_1^{k_1} A_1 + (D_2^{k_2} \cdots D_m^{k_m}) B_1 = 1_{\mathsf{F}} \tag{4.7}$$

for some $A_1, B_1 \in \mathsf{F}[\xi]$. Then

$$\frac{\tilde{N}}{D} = \frac{\tilde{N}_1}{D_1^{k_1}} + \frac{\tilde{N} A_1}{D_2^{k_2} \cdots D_m^{k_m}},$$

where $\tilde{N}_1 = \tilde{N} B_1$. Note that $\tilde{N}$ and D_1 are coprime and that B_1 and D_1 are coprime, the latter by virtue of (4.7). Therefore, $\tilde{N}_1$ and D_1 are coprime. Now, since $D_2^{k_2}$ and $D_3^{k_3} \cdots D_m^{k_m}$ are coprime we can write

$$D_2^{k_2} A_2 + (D_3^{k_3} \cdots D_m^{k_m}) B_2 = 1_{\mathsf{F}} \tag{4.8}$$

for some $A_2, B_2 \in \mathsf{F}[\xi]$. Thus we have

$$\frac{\tilde{N} A_1}{D_2^{k_2} \cdots D_m^{k_m}} = \frac{\tilde{N}_2}{D_2^{k_2}} + \frac{\tilde{N} A_1 A_2}{D_3^{k_3} \cdots D_m^{k_m}},$$

where $\tilde{N}_2 = \tilde{N} A_1 B_2$. Now we have $\tilde{N}$ and D_2 coprime, A_1 and D_2 coprime (by (4.7)), and B_2 and D_2 coprime (by (4.8)). Thus $\tilde{N}_2$ and D_2 are coprime. Thus

$$\frac{\tilde{N}}{D} = \frac{\tilde{N}_1}{D_1^{k_1}} + \frac{\tilde{N}_2}{D_2^{k_2}} + \frac{\tilde{N} A_1 A_2}{D_3^{k_3} \cdots D_m^{k_m}}.$$

This can be continued m times (we leave the formal but straightforward induction argument to the reader) to give

$$\frac{\tilde{N}}{D} = \sum_{r=1}^{m} \frac{\tilde{N}_r}{D_r^{k_r}}$$

for polynomials $\tilde{N}_1, \dots, \tilde{N}_m$ such that $\tilde{N}_r$ and D_r are coprime for each $r \in \{1, \dots, m\}$. At this point we are not guaranteed that $\deg(\tilde{N}_r) < \deg(D_r^{k_r})$ for $r \in \{1, \dots, m\}$ (in fact, this will generally not be the case). However, if $\deg(\tilde{N}_r) \geq \deg(D_r^{k_r})$, then use the Division Algorithm: $\tilde{N}_r = \tilde{Q}_r D_r^{k_r} + N_r$ where $\deg(N_r) < \deg(D_r^{k_r})$. Then we have

$$R = \sum_{r=1}^{m} \frac{N_r}{D_r^{k_r}} + Q, \tag{4.9}$$

where $Q = \tilde{Q} + \tilde{Q}_1 + \cdots + \tilde{Q}_m$. We claim that, for $r \in \{1, \dots, m\}$, N_r and D_r are coprime. Suppose not. Then, since D_r is irreducible, $N_r = D_r A_r$ for some $A_r \in \mathsf{F}[\xi]$. In this case we have $\tilde{N}_r = D_r(\tilde{Q}_r D_r^{k_r - 1} + A_r)$, and so $D_r | \tilde{N}_r$, which we know is not true. Thus we indeed know that

1. $D_1, \dots, D_m$ are monic, irreducible, and coprime,
2. $\deg(N_r) < \deg(D_r^{k_r})$ for $r \in \{1, \dots, m\}$, and
3. N_r and D_r are coprime for $r \in \{1, \dots, m\}$.

Let us stop at this point to address the matter of uniqueness of the representation (4.9). In the statement of the lemma we refer to $m \in \mathbb{Z}_{>0}$ and $Q, D_1, \dots, D_m, N_1, \dots, N_m \in \mathsf{F}[\xi]$ as we have constructed them.

1 Lemma *Let* $\mathrm{R} \in \mathsf{F}(\xi)$ *be expressed as*

$$\mathrm{R} = \sum_{\mathrm{r}=1}^{\mathrm{m}'} \frac{\mathrm{N}'_{\mathrm{r}}}{(\mathrm{D}'_{\mathrm{r}})^{\mathrm{k}'_{\mathrm{r}}}} + \mathrm{Q}',$$

where

(i) $\mathrm{k}'_{\mathrm{r}} \in \mathbb{Z}_{>0}$, $\mathrm{r} \in \{1, \dots, \mathrm{m}'\}$,

(ii) $\mathrm{D}'_1, \dots, \mathrm{D}'_{\mathrm{m}'} \in \mathsf{F}[\xi]$ *are monic, irreducible, and coprime,*

(iii) $\mathrm{Q}', \mathrm{N}'_1, \dots, \mathrm{N}'_{\mathrm{m}'} \in \mathsf{F}[\xi]$,

(iv) $\deg(\mathrm{N}'_{\mathrm{r}}) < \deg((\mathrm{D}'_{\mathrm{r}})^{\mathrm{k}'_{\mathrm{r}}})$ *for* $\mathrm{r} \in \{1, \dots, \mathrm{m}'\}$,

(v) N'_{r} *and* D'_{r} *are coprime for* $\mathrm{r} \in \{1, \dots, \mathrm{m}\}$, *and*

(vi) $\mathrm{k}'_1, \dots, \mathrm{k}'_{\mathrm{m}'} \in \mathbb{Z}_{>0}$.

Then $\mathrm{Q}' = \mathrm{Q}$, $\mathrm{m}' = \mathrm{m}$ *and there exists* $\sigma \in \mathfrak{S}_{\mathrm{m}}$ *such that* $\mathrm{D}'_{\mathrm{r}} = \mathrm{D}_{\sigma(\mathrm{r})}$, $\mathrm{N}'_{\mathrm{r}} = \mathrm{N}_{\sigma(\mathrm{r})}$, *and* $\mathrm{k}'_{\mathrm{r}} = \mathrm{k}_{\sigma(\mathrm{r})}$ *for* $\mathrm{r} \in \{1, \dots, \mathrm{m}\}$.

Proof Define distinct polynomials $\tilde{D}_1, \dots, \tilde{D}_{\tilde{m}}$ so that

$$\{\tilde{D}_1, \dots, \tilde{D}_{\tilde{m}}\} = \{D_1, \dots, D_m, D'_1, \dots, D'_{m'}\}.$$

We can then write

$$\sum_{r=1}^{m} \frac{N_r}{D_r^{k_r}} + Q = \sum_{j=1}^{\tilde{m}} \frac{\tilde{N}_r}{\tilde{D}_r^{\tilde{k}_r}} + Q$$

and

$$\sum_{r=1}^{m'} \frac{N'_r}{(D'_r)^{k_r}} + Q' = \sum_{j=1}^{\tilde{m}} \frac{\tilde{N}'_r}{(\tilde{D}'_r)^{\tilde{k}'_r}} + Q'$$

for polynomials $\tilde{N}_r, \tilde{N}'_r$, and for $\tilde{k}_r, \tilde{k}'_r \in \mathbb{Z}_{\geq 0}$, $r \in \{1, \dots, \tilde{m}\}$. We adopt the convention in writing these formulae that if $\tilde{k}_r$ or $\tilde{k}'_r$ is zero, then this means that $\tilde{N}_r$ or $\tilde{N}'_r$, respectively, is zero. Thus we reduce the problem to showing that if

$$\sum_{j=1}^{\tilde{m}} \frac{\tilde{N}_r}{\tilde{D}_r^{\tilde{k}_r}} + Q = \sum_{j=1}^{\tilde{m}} \frac{\tilde{N}'_r}{(\tilde{D}'_r)^{\tilde{k}'_r}} + Q', \tag{4.10}$$

then $\tilde{k}'_r = \tilde{k}_r$, $\tilde{N}'_r = \tilde{N}_r$, $r \in \{1, \dots, \tilde{m}\}$, and $Q' = Q$.

Let $j \in \{1, \dots, \tilde{m}\}$. If $\tilde{k}_j = \tilde{k}'_j$, then the term corresponding to $\tilde{D}_j$ does not appear in either of the expressions in (4.10), so there is nothing more to do in this case. So suppose that one of $\tilde{k}_j$ or $\tilde{k}'_j$ is nonzero; without loss of generality make it $\tilde{k}_j$. Also without loss of generality suppose that $\tilde{k}_j \geq \tilde{k}'_j$. Now rearrange (4.10):

$$\frac{\tilde{N}_j}{\tilde{D}_j^{\tilde{k}_j}} - \frac{\tilde{N}'_j}{\tilde{D}_j^{\tilde{k}'_j}} = \sum_{r \in \{1, \dots, \tilde{m}\} \setminus \{j\}} \left(\frac{\tilde{N}'_r}{\tilde{D}_r^{\tilde{k}'_r}} - \frac{\tilde{N}_r}{\tilde{D}_r^{\tilde{k}_r}} \right) + (Q' - Q). \tag{4.11}$$

Let $M \in \mathsf{F}[\xi]$ be a least common multiple for the polynomials $\tilde{D}_r^{\tilde{k}_r}, \tilde{D}_r^{\tilde{k}'_r}, r \in \{1, \dots, \tilde{m}\} \setminus \{j\}$. Since the polynomials $\tilde{D}_1, \dots, \tilde{D}_{\tilde{m}}$ are all irreducible, they are coprime, and so $\tilde{D}_j \nmid M$. Now multiply (4.11) by $M D_j^{\tilde{k}_j}$:

$$M(\tilde{N}_j - \tilde{D}_j^{\tilde{k}_j - \tilde{k}'_j} \tilde{N}'_j) = A\tilde{D}_j^{\tilde{k}_j}$$

for some polynomial A. If $\tilde{k}'_j < \tilde{k}_j$ then we have

$$M\tilde{N}_j = \tilde{D}_j(A\tilde{D}_j^{\tilde{k}_j - 1} + \tilde{D}_j^{\tilde{k}_j - \tilde{k}'_j - 1} \tilde{N}'_j),$$

implying that $\tilde{D}_j | \tilde{N}_j$, which contradicts our assumption that $\tilde{N}_j$ and $\tilde{D}_j$ are coprime. It must therefore hold that $\tilde{k}'_j = \tilde{k}_j$. Upon showing this we then conclude that $D_j^{\tilde{k}_j} | (\tilde{N}_j - \tilde{N}'_j)$. Since $\deg(\tilde{N}_j - \tilde{N}'_j) < \deg(D_j^{\tilde{k}_j})$, this can only hold when $\tilde{N}_j - \tilde{N}'_j = 0_{\mathsf{F}[\xi]}$. It is now obvious that $Q' = Q$, and the lemma is proved. ▼

Fix $r \in \{1, \dots, m\}$. Now, by Corollary 4.4.16, write

$$N_r = A_{r,0} + A_{r,1} D_r + A_{r,2} D_r^2 + \dots + A_{r,m_r} D_r^{m_r}$$

for some unique $m_r \in \mathbb{Z}_{\geq 0}$ and unique $A_{r,j}$, $j \in \{0, 1, \dots, m_r\}$, and where $\deg(A_{r,j}) < \deg(D_r)$ for $j \in \{0, 1, \dots, m_r\}$. Since $\deg(N_r) < \deg(D_r^{k_r})$, by Proposition 4.4.11 we have $m_r < k_r$. If $m_r < k_r - 1$ then define $A_{r,j} = 0_{\mathsf{F}[\xi]}$ for $j \in \{m_r + 1, \dots, k_r - 1\}$. Then we have

$$\frac{N_r}{D_r^{k_r}} = \frac{A_{r,0}}{D_r^{k_r}} + \frac{A_{r,1}}{D_r^{k_r - 1}} + \dots + \frac{A_{r,k_r - 1}}{D_r}.$$

Defining $N_{r,s} = A_{r,k_r - s}$ then gives the unique representation of $\frac{N_r}{D_r^{k_r}}$ as

$$\frac{N_r}{D_r^{k_r}} = \sum_{s=1}^{k_r} \frac{N_{r,s}}{D_r^s}.$$

This then gives the theorem. ■

The matter of computing the partial fraction expansion is, in general, a little tedious. In cases where the field F is algebraically closed, or more generally, where the denominator in the coprime fractional representative splits in F, there are procedures that enable straightforward computation of the partial fraction decomposition (see Exercise 4.4.7). Otherwise, one is essentially left with the construction in the proof of Theorem 4.4.51, which is constructive, provided that one can factor the denominator. Let us illustrate the computation of a partial fraction expansion with an example.

4.4.52 Example (Partial fraction decomposition) Consider the rational function

$$R = \frac{\xi^5 + 3\xi^2 + 2}{\xi^4 + 3\xi^3 + 4\xi^2 + 3\xi + 1}$$

as an element of $\mathbb{Q}(\xi)$. First let us write this as a sum of a strictly proper rational function and a polynomial. We do this by seeking polynomials $Q, P \in \mathbb{Q}(\xi)$ such that

$$\xi^5 + 3\xi^2 + 2 = Q(\xi^4 + 3\xi^3 + 4\xi^2 + 3\xi + 1) + P.$$

This is the Division Algorithm for polynomials, and we leave the details to the reader:

$$Q = \xi - 3, \quad P = 5\xi^3 + 12\xi^2 + 8\xi + 5.$$

Thus we have

$$R = \frac{5\xi^3 + 12\xi^2 + 8\xi + 5}{\xi^4 + 3\xi^3 + 4\xi^2 + 3\xi + 1} + \xi - 3.$$

The next step is to write the denominator as a product of irreducible polynomials. This is always the problematic step; for general fields (even for $\mathbb{Q}$) it is difficult to even determine whether a polynomial is irreducible, never mind write it as a product of irreducibles. However, in this case, one might be able to, by hook or by crook or by computer, notice that

$$\xi^4 + 3\xi^3 + 4\xi^2 + 3\xi + 1 = (\xi + 1)^2(\xi^2 + \xi + 1),$$

and that the polynomials $\xi + 1$ and $\xi^2 + \xi + 1$ are irreducible, the first obviously, and the second by Example 4.4.33. Next, noting that the polynomials $(\xi + 1)^2$ and $\xi^2 + \xi + 1$ are coprime, we seek polynomials A and B such that

$$A(\xi + 1)^2 + B(\xi^2 + \xi + 1) = 1.$$

To do this, we use the Euclidean Algorithm as suggested by Theorem 4.2.84. Doing the (in this case simple) polynomial long division gives the following as the Euclidean Algorithm:

$$\begin{aligned}
(\xi + 1)^2 &= 1(\xi^2 + \xi + 1) + \xi, \\
\xi^2 + \xi + 1 &= (\xi + 1)\xi + 1, \\
\xi &= 1\xi.
\end{aligned}$$

Now, to compute A and B, we follow the prescription of Theorem 4.2.84. We define $\alpha_0 = 1$ and $\beta_0 = -(\xi + 1)$, and then $\alpha_1 = -(\xi + 1)$ and $\beta_1 = 1 + 1(\xi + 1) = \xi + 2$. Thus we take $A = -(\xi + 1)$ and $B = \xi + 2$. That is,

$$1 = -(\xi + 1)(\xi + 1)^2 + (\xi + 2)(\xi^2 + \xi + 1),$$

and so

$$\frac{5\xi^3+12\xi^2+8\xi+5}{(\xi+1)^2(\xi^2+\xi+1)} = -\frac{(\xi+1)(5\xi^3+12\xi^2+8\xi+5)}{\xi^2+\xi+1} + \frac{(\xi+2)(5\xi^3+12\xi^2+8\xi+5)}{(\xi+1)^2}. \tag{4.12}$$

To put this expression into the desired form, for each of the summands on the right hand side we expand the numerator as a sum of powers of the denominator, as per Corollary 4.4.16. We do this by successively applying the Division Algorithm. Thus we write

$$(\xi+1)(5\xi^3+12\xi^2+8\xi+5) = (5\xi^2+12\xi+3)(\xi^2+\xi+1)+(-2\xi+2),$$

which gives

$$-\frac{(\xi+1)(5\xi^3+12\xi^2+8\xi+5)}{\xi^2+\xi+1} = \frac{2\xi-2}{\xi^2+\xi+1} - 5\xi^2-12\xi-3. \tag{4.13}$$

We also compute

$$(\xi+2)(5\xi^3+12\xi^2+8\xi+5) = (5\xi^3+17\xi^2+15\xi+6)(\xi+1)+4$$

and

$$5\xi^3+17\xi^2+15\xi+6 = (5\xi^2+12\xi+3)(\xi+1)+3.$$

One could continue this process again on $5\xi^2+12\xi+3$. However, as we shall see, this is unnecessary. In any case, we have

$$\begin{aligned}(\xi+2)(5\xi^3+12\xi^2+8\xi+5) &= ((5\xi^2+12\xi+3)(\xi+1)+3)(\xi+1)+4\\ &= (5\xi^2+12\xi+3)(\xi+1)^2+3(\xi+1)+4.\end{aligned}$$

Therefore,

$$\frac{(\xi+2)(5\xi^3+12\xi^2+8\xi+5)}{(\xi+1)^2} = \frac{3}{\xi+1} + \frac{4}{(\xi+1)^2} + 5\xi^2+12\xi+3. \tag{4.14}$$

Substituting (4.13) and (4.14) into (4.12) gives

$$\frac{5\xi^3+12\xi^2+8\xi+5}{(\xi+1)^2(\xi^2+\xi+1)} = \frac{2\xi-2}{\xi^2+\xi+1} + \frac{3}{\xi+1} + \frac{4}{(\xi+1)^2},$$

which in turn gives

$$R = \frac{2\xi-2}{\xi^2+\xi+1} + \frac{3}{\xi+1} + \frac{4}{(\xi+1)^2} + \xi - 3$$

as the partial fraction decomposition for R.

The method we have used here is a little cumbersome, but it is entirely systematic (once one writes the denominator in the coprime fractional representative as a product of irreducibles), and also illustrates the proof of Theorem 4.4.51. In practice, when writing the partial fraction expansion for rational functions over the field of real numbers (which is the most commonly encountered situation in these volumes), one can benefit from the use of Exercises 4.4.7 and 4.4.8. •

Exercises

4.4.1 Let R be a commutative unit ring. Show that if $\mathsf{R}[\xi]$ is an integral domain then R is an integral domain.

4.4.2 Prove Proposition 4.4.8.

4.4.3 Let R be an integral domain.

(a) Show that $A \in \mathsf{R}[\xi]$ is a unit if and only if $A = a_0$ is a constant polynomial where a_0 is a unit in R.

(b) Now suppose that R is additionally a field. Show that $A \in \mathsf{F}[\xi]$ is a unit if and only if it is a nonzero constant polynomial.

4.4.4 If R is a commutative ring with a finite number of elements, show that Ev_{R} is not injective.

4.4.5 Let R be an integral domain, let $a \in \mathsf{R}$, and let $A, B \in \mathsf{R}[\xi]$. Prove the following properties of the formal derivative:

(a) $(aA)' = a(A')$;

(b) $(A + B)' = A' + B'$;

(c) $(A \cdot B)' = A \cdot B' + A' \cdot B$;

(d) $(A^k)' = kA^{k-1} \cdot A'$, $k \in \mathbb{Z}_{\geq 0}$.

4.4.6 Determine the partial fraction expansion of $\frac{\xi^3+1}{(\xi+1)^4}$.

The next exercise illustrates what is referred to as the ***Heaviside coverup***[3] method for computing the partial fraction expansion in some cases. This method only works when the denominator of the coprime fractional representative splits in the field.

4.4.7 Let F be a field, let $R \in \mathsf{F}(\xi)$, and let $\frac{N}{D}$ be the coprime fractional representative for R. We suppose that D splits in F so that we can write $D = D_1^{k_1} \cdots D_m^{k_m}$ for monic coprime degree 1 polynomials $D_r = (\xi - a_r)$, $r \in \{1, \dots, m\}$, and for $k_1, \dots, k_m \in \mathbb{Z}_{>0}$. It then follows that the partial fraction decomposition of R is of the form

$$R = \sum_{r=1}^{m} \sum_{s=1}^{k_r} \frac{n_{r,s}}{D_r^s} + Q$$

for $Q \in \mathsf{F}[\xi]$ and for $n_{r,s} \in \mathsf{F}$, $r \in \{1, \dots, m\}$, $s \in \{1, \dots, k_r\}$.

(a) Show that, if $k_r = 1$, then

$$n_{r,1} = \frac{\mathrm{Ev}_{\mathsf{F}}(ND_r)(a_r)}{\mathrm{Ev}_{\mathsf{F}}(D_1^{k_1} \cdots D_{r-1}^{k_{r-1}} D_{r+1}^{k_{r+1}} \cdots D_m^{k_m})(a_r)}$$

[3] After Oliver Heaviside (1850–1925), an Englishman who, although having terminated his formal education before its completion, made contributions to the understanding of electromagnetism. He is remembered for his “operational calculus” which he invented to solve differential equations. This technique, while lacking in rigour, did allow the use of algebraic manipulations in solving differential equations.

for each $r \in \{1, \dots, m\}$.

(b) Explain how to extend the previous computation to the computation of $n_{r,1}, \dots, n_{r,k_r}$ when $k_r > 1$.
Hint: *Write*

$$\frac{\mathrm{N}}{\mathrm{D}_1^{\mathrm{k}_1} \cdots \mathrm{D}_{\mathrm{r}-1}^{\mathrm{k}_{\mathrm{r}-1}} \mathrm{D}_{\mathrm{r}}^{\mathrm{k}_{\mathrm{r}}} \mathrm{D}_{\mathrm{r}+1}^{\mathrm{k}_{\mathrm{r}+1}} \cdots \mathrm{D}_{\mathrm{m}}^{\mathrm{k}_{\mathrm{m}}}} = \frac{1}{\mathrm{D}_{\mathrm{r}}^{\mathrm{k}_{\mathrm{r}}-1}} \left(\frac{\mathrm{N}}{\mathrm{D}_1^{\mathrm{k}_1} \cdots \mathrm{D}_{\mathrm{r}-1}^{\mathrm{k}_{\mathrm{r}-1}} \mathrm{D}_{\mathrm{r}} \mathrm{D}_{\mathrm{r}+1}^{\mathrm{k}_{\mathrm{r}+1}} \cdots \mathrm{D}_{\mathrm{m}}^{\mathrm{k}_{\mathrm{m}}}} \right)$$

and apply part (a) to the expression in the parentheses on the right hand side.

(c) Use the previous two parts of the exercise to compute the partial fraction expansion of $\frac{1}{(\xi+1)^2(\xi+2)}$ as an element of (say) $\mathbb{R}(\xi)$.

The preceding exercise always applies to rational functions defined over the field $\mathbb{C}$ of complex numbers since, as we shall see in Theorem 4.7.6, this field is algebraically closed. For rational functions defined over the field $\mathbb{R}$ of real numbers, although the method of Exercise 4.4.7 may not apply directly (when the denominator polynomial does not split over $\mathbb{R}$), it is possible to think of an element of $\mathbb{R}(\xi)$ as being an element of $\mathbb{C}(\xi)$, and thus apply the methods of Exercise 4.4.7. The next exercise shows how one can then take the resulting partial fraction expansion in $\mathbb{C}(\xi)$ and convert it to a partial fraction expansion in $\mathbb{R}(\xi)$. Various ideas from Section 4.7.3 will be useful here. In particular, the reader will wish to recall that if $r \in \mathbb{C}$ is a root of $A \in \mathbb{R}[\xi] \subseteq \mathbb{C}[\xi]$, then $\bar{r}$, the complex conjugate of r, is also a root of A.

4.4.8 Note that since $\mathbb{C}$ is algebraically closed, if $R \in \mathbb{C}(\xi)$ then the partial fraction expansion of R has the form

$$R = \sum_{r=1}^{m} \sum_{s=1}^{k_r} \frac{n_{r,s}}{(s - a_r)^s}$$

for $n_{r,s} \in \mathbb{C}$, $r \in \{1, \dots, m\}$, $s \in \{1, \dots, k_r\}$, and for $a_r \in \mathbb{C}$, $r \in \{1, \dots, m\}$. Moreover, the coefficients $n_{r,s}$, $r \in \{1, \dots, m\}$, $s \in \{1, \dots, k_r\}$, can be computed as in Exercise 4.4.7. In this exercise we additionally suppose that $R \in \mathbb{R}(\xi) \subseteq \mathbb{C}(\xi)$.

(a) Argue that the roots $a_1, \dots, a_m$ of the factors in the denominators can be ordered such that $a_{j+l} = \bar{a}_j$, $j \in \{1, \dots, l\}$, and such that $a_{2l+1}, \dots, a_m \in \mathbb{R}$. Also argue that, when the roots are so ordered, $k_{j+l} = k_j$, $j \in \{1, \dots, l\}$.

For the rest of the exercise, we suppose the roots to have been ordered as in part (a). The coefficients $n_{r,s}$, $r \in \{2l + 1, \dots, m\}$, $s \in \{1, \dots, k_r\}$, can be computed as in Exercise 4.4.7, so we concern ourselves with the complex roots.

(b) Show that $n_{j+l,s} = \bar{n}_{j,s}$, $j \in \{1, \dots, l\}$, $s \in \{1, \dots, k_j\}$.

Now fix $j \in \{1, \dots, l\}$ and $s \in \{1, \dots, k_j\}$, and for notational simplicity define $a = a_j$ and $n = n_{j,s}$. Corresponding to a root a, $j \in \{1, \dots, l\}$, of the denominator will be a term in the partial fraction expansion of the form

$$\frac{n}{(\xi - a)^s} + \frac{\bar{n}}{(\xi - \bar{a})^s}. \tag{4.15}$$

(c) Show how to convert the expression (4.15) into a term in the partial fraction expansion of R as an element of $\mathbb{R}(\xi)$. Thus the denominators should be irreducible polynomials *over* $\mathbb{R}$.

Section 4.5

Vector spaces

One of the more important structures that we will use at a fairly high degree of generality is that of a vector space. As with almost everything we have encountered in this chapter, a vector space is a set equipped with certain operations. In the case of vector spaces, one of these operations melds the vector space together with another algebraic structure, in this case a field. A typical first encounter with vector spaces deals primarily with the so-called finite-dimensional case. In this case, a great deal, indeed, pretty much everything, can be said about the structure of these vector spaces. However, in these volumes we shall also encounter so-called infinite-dimensional vector spaces. A study of the structure of these gets rather more detailed than the finite-dimensional case. In this section we deal only with algebraic matters. Important additional structure in the form of a topology is the topic of Chapter III-6.

Do I need to read this section? If you are not already familiar with the idea of an abstract vector space, then you need to read this section. If you are, then it can be bypassed, and perhaps referred to as needed. Parts of this section are also good ones for readers looking for simple proofs that illustrate certain techniques for proving things. These ceases to become true when we discuss bases, since we take an abstract approach motivated by the fact that many of the vector spaces we deal with in these volumes are infinite-dimensional. •

4.5.1 Definitions and basic properties

Throughout this section we let F be a general field, unless otherwise stated. The fields of most interest to us will be $\mathbb{R}$ (see Section 2.1) and $\mathbb{C}$ (see Section 4.7). However, most constructions done with vector spaces are done just as conveniently for general fields as for specific ones.

4.5.1 Definition (Vector space) Let F be a field. A ***vector space*** over F, or an **F-*vector space***, is a nonempty set V with two operations: (1) ***vector addition***, denoted by $\mathsf{V} \times \mathsf{V} \ni (v_1, v_2) \mapsto v_1 + v_2 \in \mathsf{V}$, and (2) ***scalar multiplication***, denoted by $\mathsf{F} \times \mathsf{V}(a, v) \mapsto av \in \mathsf{V}$. Vector addition and scalar multiplication must satisfy the following rules:

(i) $v_1 + v_2 = v_2 + v_1$, $v_1, v_2 \in \mathsf{V}$ (***commutativity***);

(ii) $v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$, $v_1, v_2, v_3 \in \mathsf{V}$ (***associativity***);

(iii) there exists an vector $0_{\mathsf{V}} \in \mathsf{V}$ with the property that $v + 0_{\mathsf{V}} = v$ for every $v \in \mathsf{V}$ (***zero vector***);

(iv) for every $v \in \mathsf{V}$ there exists a vector $-v \in \mathsf{V}$ such that $v + (-v) = 0_{\mathsf{V}}$ (***negative vector***);

(v) $a(bv) = (ab)v$, $a, b \in \mathsf{F}$, $v \in \mathsf{V}$ (***associativity***);

(vi) $1_{\mathsf{F}}v = v$, $v \in \mathsf{V}$;

(vii) $a(v_1 + v_2) = av_1 + av_2$, $a \in \mathsf{F}$, $v_1, v_2 \in \mathsf{V}$ (***distributivity***);

(viii) $(a_1 + a_2)v = a_1v + a_2v$, $a_1, a_2 \in \mathsf{F}$, $v \in \mathsf{V}$ (***distributivity*** again).

A ***vector*** in a vector space V is an element of V. •

We have already encountered some examples of vector spaces. Let us indicate what some of these are, as well as introduce some important new examples of vector spaces. The verifications that the stated sets are vector spaces is routine, and we leave this to the reader in the exercises.

4.5.2 Examples (Vector spaces)

1. Consider a set $0_{\mathsf{V}} = \{v\}$ with one element. There are no choices for the F-vector space structure in this case. We must have $v + v = v$, $av = v$ for every $a \in \mathsf{F}$, $-v = v$, and $0_{\mathsf{V}} = v$. One can then verify that $\{v\}$ is then indeed an F-vector space. This vector space is called the ***trivial vector space***, and is sometimes denoted by $\{0\}$, reflecting the fact that the only vector in the vector space is the zero vector.
2. Let F^n denote the n-fold Cartesian product of F with itself. Let us denote a typical element of F^n by $(v_1, \dots, v_n)$. We define vector addition in F^n by

$$(u_1, \dots, u_n) + (v_1, \dots, v_n) = (u_1 + v_1, \dots, u_n + v_n)$$

and we define scalar multiplication in F^n by

$$a(v_1, \dots, v_n) = (av_1, \dots, av_n).$$

The vector spaces $\mathbb{R}^n$ and $\mathbb{C}^n$, over $\mathbb{R}$ and $\mathbb{C}$, respectively, will be of particular importance to us. The reader who has no previous knowledge of vector spaces would be well served by spending some time understanding the geometry of vector addition and scalar multiplication in, say, $\mathbb{R}^2$.

3. Let us denote by F^∞ the set of sequences in F. Thus an element of F^∞ is a sequence $(a_j)_{j\in\mathbb{Z}_{>0}}$ with $a_j \in \mathsf{F}$, $j \in \mathbb{Z}_{>0}$. We define vector addition and scalar multiplication by

$$(a_j)_{j\in\mathbb{Z}_{>0}} + (b_j)_{j\in\mathbb{Z}_{>0}} = (a_j + b_j)_{j\in\mathbb{Z}_{>0}}, \quad a(a_j)_{j\in\mathbb{Z}_{>0}} = (aa_j)_{j\in\mathbb{Z}_{>0}},$$

respectively. This can be verified to make F^∞ into an F-vector space. It is tempting to think of things like $\mathsf{F}^\infty = \lim_{n\to\infty} \mathsf{F}^n$, but one must exercise care, since the limit needs definition. This is the realm of Chapter III-6.

4. Let us denote by F_0^∞ the subset of F^∞ consisting of sequences for which all but a finite number of terms is zero. Vector addition and scalar multiplication are defined for F_0^∞ are defined just as for F^∞. It is just as straightforward to verify that these operations make F_0^∞ an F-vector space.

5. If K is a field extension of F (see Definition 4.6.1) and if V is a K-vector space, then V is also an F-vector space with the operation of vector addition being exactly that of V as a K-vector space, and with scalar multiplication simply being the restriction of scalar multiplication by K to F.
6. The set $\mathsf{F}[\xi]$ of polynomials over F is an F-vector space. Vector addition is addition in the usual sense of polynomials, and scalar multiplication is multiplication of polynomials, using the fact that F is a subring of $\mathsf{F}[\xi]$ consisting of the constant polynomials.
7. Denote by $\mathsf{F}_k[\xi]$ the polynomials over F of degree at most k. Using the same definitions of vector addition and scalar multiplication as were used for the F-vector space $\mathsf{F}[\xi]$ in the preceding example, $\mathsf{F}_k[\xi]$ is an F-vector space.
8. Let S be a set and let V be an F-vector space. As in Definition 1.3.1, let V^S be the set of maps from S to V. Let us define vector addition and scalar multiplication in V^S by
$$(f+g)(x) = f(x) + g(x), \quad (af)(x) = a(f(x))$$
for $f, g \in \mathsf{V}^S$ and $a \in \mathsf{F}$. One may directly verify that these operations indeed satisfy the conditions to make V^S into an F-vector space.
9. Let $I \subseteq \mathbb{R}$ be an interval and let $\mathsf{C}^0(I;\mathbb{R})$ denote the set of continuous $\mathbb{R}$-valued functions on I. Following the preceding example, define vector addition and scalar multiplication in $\mathsf{C}^0(I;\mathbb{R})$ by
$$(f+g)(x) = f(x) + g(x), \quad (af)(x) = a(f(x)), \qquad f, g \in \mathsf{C}^0(I;\mathbb{R}),\ a \in \mathbb{R},$$
respectively. With these operations, one can verify that $\mathsf{C}^0(I;\mathbb{R})$ is a $\mathbb{R}$-vector space. •

Let us now prove some elementary facts about vector spaces.

4.5.3 Proposition (Properties of vector spaces) *Let F be a field and let V be an F-vector space. The following statements hold:*

(i) there exists exactly one vector $0_{\mathsf{V}} \in \mathsf{V}$ such that $\mathrm{v} + 0_{\mathsf{V}} = \mathrm{v}$ for all $\mathrm{v} \in \mathsf{V}$;

(ii) for each $\mathrm{v} \in \mathsf{V}$ there exists exactly one vector $-\mathrm{v} \in \mathsf{V}$ such that $\mathrm{v} + (-\mathrm{v}) = 0_{\mathsf{V}}$;

(iii) $\mathrm{a}0_{\mathsf{V}} = 0_{\mathsf{V}}$ for all $\mathrm{a} \in \mathsf{F}$;

(iv) $0_{\mathsf{F}}\mathrm{v} = 0_{\mathsf{V}}$ for each $\mathrm{v} \in \mathsf{V}$;

(v) $\mathrm{a}(-\mathrm{v}) = (-\mathrm{a})\mathrm{v} = -(\mathrm{a}\mathrm{v})$ for all $\mathrm{a} \in \mathsf{F}$ and $\mathrm{v} \in \mathsf{V}$;

(vi) if $\mathrm{a}\mathrm{v} = 0_{\mathsf{V}}$, then either $\mathrm{a} = 0_{\mathsf{F}}$ or $\mathrm{v} = 0_{\mathsf{V}}$.

Proof Parts (i) and (ii) follow in the same manner as part (i) of Proposition 4.1.6.

(iii) For some $v \in \mathsf{V}$ we compute
$$av = a(v + 0_{\mathsf{V}}) = av + a0_{\mathsf{V}}.$$

Therefore,
$$av + (-(av)) = av + (-(av)) + a0_{\mathsf{V}} \quad \implies \quad 0_{\mathsf{V}} = 0_{\mathsf{V}} + a0_{\mathsf{V}} = a0_{\mathsf{V}},$$

which gives the result.

(iv) For some $a \in \mathsf{F}$ we compute

$$av = (a + 0_{\mathsf{F}})v = av + 0_{\mathsf{F}}v.$$

Therefore,

$$av + (-(av)) = av + (-(av)) + 0_{\mathsf{F}}v \quad\Longrightarrow\quad 0_{\mathsf{V}} = 0_{\mathsf{V}} + 0_{\mathsf{F}}v = 0_{\mathsf{F}}v,$$

giving the result.

(v) We have

$$0_{\mathsf{V}} = a0_{\mathsf{V}} = a(v + (-v)) = av + a(-v).$$

Therefore, $a(-v) = -(av)$. Similarly,

$$0_{\mathsf{V}} = 0_{\mathsf{F}}v = (a - a)v = av + (-a)v.$$

Therefore $(-a)v = -(av)$.

(vi) Suppose that $av = 0_{\mathsf{V}}$. If $a = 0_{\mathsf{F}}$ then there is nothing to prove. If $a \neq 0_{\mathsf{F}}$ then we have

$$0_{\mathsf{V}} = a^{-1}0_{\mathsf{V}} = a^{-1}(av) = (a^{-1}a)v = 1_{\mathsf{F}}v = v,$$

which gives the result. ■

In this section it will be convenient to have on hand the notion of a homomorphism of vector spaces. This is a topic about which we will have much to say in Chapter 5, but here we simply give the definition.

4.5.4 Definition (Linear map) Let F be a field and let U and V be F-vector spaces. An **F-*homomorphism*** of U and V, or equivalently an **F-*linear map*** between U and V, is a map $\mathsf{L}\colon \mathsf{U} \to \mathsf{V}$ having the properties that

(i) $\mathsf{L}(u_1 + u_2) = \mathsf{L}(u_1) + \mathsf{L}(u_2)$ for every $u_1, u_2 \in \mathsf{U}$ and

(ii) $\mathsf{L}(au) = a\mathsf{L}(u)$ for every $a \in \mathsf{F}$ and $u \in \mathsf{U}$.

An F-homomorphism L is an **F-*monomorphism*** (resp. **F-*epimorphism***, **F-*isomorphism***) if L is injective (resp. surjective, bijective). If there exists an isomorphism between F-vector spaces U and V, then U and V are **F-*isomorphic***. An F-homomorphism from V to itself is called an **F-*endomorphism*** of V. The set of F-homomorphisms from U to V is denoted by $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$, and the set of F-endomorphisms of V is denoted by $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$. •

We shall frequently simply call an “F-homomorphism” or an “F-linear map “ a “homomorphism” or a “linear map” when F is understood. We postpone to Section 5.4 an exposition of the properties of linear maps, as well as a collection of illustrative examples. In this section we shall principally encounter a few examples of isomorphisms.

4.5.2 Subspaces

As with most algebraic objects, with vector spaces it is interesting to talk about subsets that respect the structure.

4.5.5 Definition Let F be a field. A nonempty subset U of an F-vector space V is a ***vector subspace***, or simply a ***subspace***, if $u_1 + u_2 \in \mathsf{U}$ for all $u_1, u_2 \in \mathsf{U}$ and if $au \in \mathsf{U}$ for all $a \in \mathbb{F}$ and all $u \in \mathsf{U}$. •

As we saw with subgroups and subrings, subspaces are themselves vector spaces.

4.5.6 Proposition (A vector subspace is a vector space) *Let F be a field. A nonempty subset $\mathsf{U} \subseteq \mathsf{V}$ of an F-vector space V is a subspace if and only if U is a vector space using the operations of vector addition and scalar multiplication in V, restricted to U.*

Proof This is Exercise 4.5.11. ■

Let us give some examples of subspaces. We leave the straightforward verifications of our claims as exercises.

4.5.7 Examples (Subspaces)

1. For each $n \in \mathbb{Z}_{>0}$, F^n can be regarded as a subspace of F_0^∞ by tacking on zeros to the n-tuple in F^n to get a sequence indexed by $\mathbb{Z}_{>0}$.
2. The subset F_0^∞ of F^∞ is a subspace.
3. For each $k \in \mathbb{Z}_{\geq 0}$, $\mathsf{F}_k[\xi]$ is a subspace of $\mathsf{F}[\xi]$. However, the set of polynomials of degree k is *not* a subspace of $\mathsf{F}[\xi]$. Why?
4. In Exercise 4.5.10 the reader can verify that, for $r \in \mathbb{Z}_{>0}$, the set $\mathsf{C}^r(I; \mathbb{R})$ of r-times continuously differentiable $\mathbb{R}$-valued functions defined on an interval I is a $\mathbb{R}$-vector space. In fact, it is a subspace of $\mathsf{C}^0(I; \mathbb{R})$. •

Analogously with homomorphisms of groups and rings, there are two natural subspaces associated with a homomorphism of vector spaces.

4.5.8 Definition (Kernel and image of linear map) Let F be a vector space, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$.

(i) The ***image*** of L is $\mathrm{image}(\mathsf{L}) = \{\mathsf{L}(u) \mid u \in \mathsf{U}\}$.

(ii) The ***kernel*** of L is $\ker(\mathsf{L}) = \{u \in \mathsf{U} \mid \mathsf{L}(u) = 0_{\mathsf{V}}\}$. •

It is straightforward to verify that the image and kernel are subspaces.

4.5.9 Proposition (Kernel and image are subspaces) *Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$. Then $\mathrm{image}(\mathsf{L})$ and $\ker(\mathsf{L})$ are subspaces of V and U, respectively.*

Proof This is Exercise 4.5.16. ■

An important sort of subspace arises from taking sums of vectors with arbitrary coefficients in the field over which the vector space is defined. To make this more formal, we have the following definition.

4.5.10 Definition (Linear combination) Let F be a field and let V be an F-vector space. If $S \subseteq \mathsf{V}$ is nonempty, a ***linear combination*** from S is an element of V of the form

$$c_1 v_1 + \cdots + c_k v_k,$$

where $c_1, \dots, c_k \in \mathsf{F}$ and $v_1, \dots, v_k \in S$. We call $c_1, \dots, c_k$ the ***coefficients*** in the linear combination. •

The important feature of the set of linear combinations from a subset of a vector space is that they form a subspace.

4.5.11 Proposition (The set of linear combinations is a subspace) *If* F *is a field, if* V *is an* F*-vector space, and if* $\mathsf{S} \subseteq \mathsf{V}$ *is nonempty, then the set of linear combinations from* S *is a subspace of* V*. Moreover, this subspace is the smallest subspace of* V *containing* S*.*

Proof Let

$$B = b_1 u_1 + \cdots + b_l v_l, \quad C = c_1 v_1 + \cdots + c_k v_k$$

be linear combinations from S and let $a \in \mathsf{F}$. Then

$$B + C = b_1 u_1 + \cdots + b_l u_l + c_1 v_1 + \cdots + c_k v_k$$

is immediately a linear combination from S with vectors $u_1, \dots, u_l, v_1, \dots, v_k$ and coefficients $b_1, \dots, b_l, c_1, \dots, c_k$. Also

$$aC = (ac_1) v_1 + \cdots + (ac_k) v_k$$

is a linear combination from S with vectors $v_1, \dots, v_k$ and coefficients $ac_1, \dots, ac_k$. Thus $B + C$ and aC are linear combinations from S.

Now let U be a subspace of V containing S. If $c_1 v_1 + \cdots + c_k v_k$ is a linear combination from S then, since $S \subseteq \mathsf{U}$ and since U is a subspace, $c_1 v_1 + \cdots + c_k v_k \in \mathsf{U}$. Therefore, U contains the set of linear combinations from S, and hence follows the second assertion of the proposition. ■

Based on the preceding result we have the following definition. Note that the definition is "geometric," whereas the proposition gives a more concrete version in that the explicit form of elements of the subspace are given.

4.5.12 Definition (Subspace generated by a set) If F is a field, if V is an F-vector space, and if $S \subseteq \mathsf{V}$ is nonempty, then the ***subspace generated by*** **S** is the smallest subspace of V containing S. This subspace is denoted by $\operatorname{span}_{\mathsf{F}}(S)$. •

We close this section with a definition of a "shifted subspace" which will come up in our discussion in Sections 5.1.8 and 5.4.8.

4.5.13 Definition (Affine subspace) Let F be a field and let V be an F-vector space. A subset $\mathsf{A} \subseteq \mathsf{V}$ is an ***affine subspace*** if there exists $v_0 \in \mathsf{V}$ and a subspace U of V such that

$$\mathsf{A} = \{v_0 + u \mid u \in \mathsf{U}\}.$$

The subspace U is the ***linear part*** of A. •

Intuitively, an affine subspace is a subspace U shifted by the vector v_0. Let us give some simple examples of affine subspaces.

4.5.14 Examples (Affine subspaces)

1. Every subspace is also an affine subspace “shifted” by the zero vector.
2. If U is a subspace of a vector space V and if $u_0 \in \mathsf{U}$, then the affine subspace

$$\{u_0 + u \mid u \in \mathsf{U}\}$$

is simply the subspace U. That is to say, if we shift a subspace by an element of itself, the affine subspace is simply a subspace.

3. Let $\mathsf{V} = \mathbb{R}^2$. The vertical line

$$\{(1, 0) + (0, y) \mid y \in \mathbb{R}\}$$

through the point $(1, 0)$ is an affine subspace. •

4.5.3 Linear independence

The notion of linear independence lies at the heart of understanding much of the theory of vector spaces, and the associated topic of linear algebra which we treat in detail in Chapter 5. The precise definition we give for linear independence is one that can be difficult to understand on a first encounter. However, it is important to understand that this definition has, in actuality, been carefully crafted to be maximally useful; the definition in its precise form is used again and again in proofs in this section and in Chapter 5.

4.5.15 Definition (Linearly independent) Let F be a field and let V be an F-vector space.

(i) A finite family $(v_1, \dots, v_k)$ of vectors in V is ***linearly independent*** if the equality

$$c_1 v_1 + \cdots + c_k v_k = 0_{\mathsf{V}}, \qquad c_1, \dots, c_k \in \mathsf{F},$$

is satisfied only if $c_1 = \cdots = c_k = 0_{\mathsf{F}}$.

(ii) A finite set $S = \{x_j \mid j \in \{1, \dots, k\}\}$ is linearly independent if the finite family corresponding to the set is linearly independent.

(iii) An nonempty family $(v_a)_{a \in A}$ of vectors in V is ***linearly independent*** if every finite subfamily of $(v_a)_{a \in A}$ is linearly independent.

(iv) A nonempty subset $S \subseteq \mathsf{V}$ is ***linearly independent*** if every nonempty finite subset of S is linearly independent.

(v) A nonempty family $(v_a)_{a \in A}$ if vectors in V is ***linearly dependent*** if it is not linearly independent.

(vi) A nonempty subset $S \subseteq \mathsf{V}$ is ***linearly dependent*** if it is not linearly independent. •

The definition we give is not quite the usual one since we define linear independence and linear dependence for both sets of vectors and families of vectors. Corresponding to any set $S \subseteq \mathsf{V}$ of vectors is a family of vectors in a natural way: $(v)_{v \in S}$. Thus one can, in actuality, get away with only defining linear independence and linear dependence for families of vectors. However, since most references will consider sets of vectors, we give both flavours of the definition. Let us see with a simple example that only dealing with sets of vectors may not suffice.

4.5.16 Example (Sets of vectors versus families of vectors) Let F be a field and let $\mathsf{V} = \mathsf{F}^2$. Define $v_1 = (1_{\mathsf{F}}, 0_{\mathsf{F}})$ and $v_2 = (1_{\mathsf{F}}, 0_{\mathsf{F}})$. Then the family (v_1, v_2) is linearly dependent since $1_{\mathsf{F}}v_1 - 1_{\mathsf{F}}v_2 = 0_{\mathsf{V}}$. However, since $\{v_1, v_2\} = \{(1_{\mathsf{F}}, 0_{\mathsf{F}}\}$, this set is, in fact, linearly independent. •

As can easily be gleaned from this example, the distinction between linearly independent sets and linearly independent families only arises when the family contains the same vector in two places. We shall frequently talk about sets rather than families, accepting that in doing so we disallow the possibility of considering that two vectors in the set might be the same.

There is a potential inconsistency with the above definition of a general linearly independent set. Specifically, if $S = (v_1, \dots, v_k)$ is a finite family of vectors, then Definition 4.5.15 proposes two definitions of linear independence, one from part (i) and one from part (iv). To resolve this we prove the following result.

4.5.17 Proposition (Subsets of finite linearly independent sets are linearly independent) *Let F be a field, let V be an F-vector space, and let $(\mathrm{v}_1, \dots, \mathrm{v}_\mathrm{k})$ be linearly independent according to part (i) of Definition 4.5.15. Then any nonempty subfamily of $(\mathrm{v}_1, \dots, \mathrm{v}_\mathrm{k})$ is linearly independent.*

Proof Let $(v_{j_1}, \dots, v_{j_l})$ be a nonempty subfamily of $(v_1, \dots, v_k)$ and suppose that

$$c_1 v_{j_1} + \dots + c_l v_{j_l} = 0_{\mathsf{V}}.$$

Let $\{j_{l+1}, \dots, j_k\}$ be a distinct set of indices for which $\{1, \dots, k\} = \{j_1, \dots, j_l, j_{l+1}, j_k\}$. Then

$$c_1 v_{j_1} + \dots + c_l v_{j_l} + 0_{\mathsf{F}} v_{j_{l+1}} + \dots + 0_{\mathsf{F}} v_{j_k} = 0_{\mathsf{V}}.$$

Since the set $(v_1, \dots, v_k)$ is linearly independent, it follows that $c_1 = \dots = c_l = 0_{\mathsf{F}}$, giving the result. ■

Let us give some examples of linearly independent and linearly dependent sets to illustrate the ideas.

4.5.18 Examples (Linear independence)

1. In the F-vector space F^n consider the n vectors $\boldsymbol{e}_1, \dots, \boldsymbol{e}_n$ defined by

$$\boldsymbol{e}_j = (0, \dots, 0, \underbrace{1_{\mathsf{F}}}_{j\text{th position}}, 0, \dots, 0).$$

We claim that these vectors are linearly independent. Indeed, suppose that

$$c_1 \boldsymbol{e}_1 + \dots + c_n \boldsymbol{e}_n = 0_{\mathsf{F}^n}$$

for $c_1, \dots, c_n \in \mathsf{F}$. Using the definition of vector addition and scalar multiplication in F^n this means that

$$(c_1, \dots, c_n) = (0, \dots, 0),$$

which immediately gives $c_1 = \dots = c_n = 0_{\mathsf{F}}$. This gives linear independence, as desired.

2. In the F-vector space F_0^∞ define vectors $\boldsymbol{e}_j$, $j \in \mathbb{Z}_{>0}$, by asking that $\boldsymbol{e}_j$ be the sequence consisting of zeros except for the jth term in the sequence, which is 1_{F}. We claim that the family $(\boldsymbol{e}_j)_{j\in\mathbb{Z}_{>0}}$ is linearly independent. Indeed. let $\boldsymbol{e}_{j_1}, \dots, \boldsymbol{e}_{j_k}$ be a finite subset of $(\boldsymbol{e}_j)_{j\in\mathbb{Z}_{>0}}$. Then suppose that

$$c_1\boldsymbol{e}_{j_1} + \dots + c_k\boldsymbol{e}_{j_k} = 0_{\mathsf{F}_0^\infty}$$

for $c_1, \dots, c_k \in \mathsf{F}$. Using the definition of vector addition and scalar multiplication in F_0^∞, the linear combination $c_1\boldsymbol{e}_{j_1} + \dots + c_k\boldsymbol{e}_{j_k}$ is equal to the sequence $(a_l)_{j\in\mathbb{Z}_{>0}}$ in F given by

$$a_l = \begin{cases} c_r, & l = j_r \text{ for some } r \in \{1, \dots, k\}, \\ 0_{\mathsf{F}}, & \text{otherwise.} \end{cases}$$

Clearly this sequence is equal to zero if and only if $c_1 = \dots = c_k = 0_{\mathsf{F}}$, thus showing that $(\boldsymbol{e}_j)_{j\in\mathbb{Z}_{>0}}$ is linearly independent.

3. Since F_0^∞ is a subspace of F^∞, it follows easily that the family $(\boldsymbol{e}_j)_{j\in\mathbb{Z}_{>0}}$ is linearly independent in F^∞.

4. In the F-vector space $\mathsf{F}_k[\xi]$ of polynomials of degree at most k the family $(1, \xi, \dots, \xi^k)$ is linearly independent. Indeed, suppose that

$$c_0 + c_1\xi + \dots + c_k\xi^k = 0_{\mathsf{F}[\xi]} \tag{4.16}$$

for $c_0, c_1, \dots, c_k \in \mathsf{F}$. One should now recall the definition of $\mathsf{F}[\xi]$ as sequences in F for which a finite number of elements in the sequence are nonzero. The elements in the sequence, recall, are simply the coefficients of the polynomial. Therefore, a polynomial is the zero polynomial if and only if all of its coefficients are zero. In particular, (4.16) holds if and only if $c_0 = c_1 = \dots = c_k = 0_{\mathsf{F}}$.

5. In the vector space $\mathsf{F}[\xi]$ we claim that the set $(\xi^j)_{j\in\mathbb{Z}_{\geq 0}}$ is linearly independent. To see why this is so, choose a finite subfamily $(\xi^{j_1}, \dots, \xi^{j_k})$ from the family $(\xi^j)_{j\in\mathbb{Z}_{\geq 0}}$ and suppose that

$$c_1\xi^{j_1} + \dots + c_k\xi^{j_k} = 0_{\mathsf{F}[\xi]} \tag{4.17}$$

for some $c_1, \dots, c_k \in \mathsf{F}$. As we argued in the previous example, a polynomial is zero if and only if all of its coefficients is zero. Therefore, (4.17) holds if and only if $c_1 = \dots = c_k = 0_{\mathsf{F}}$, thus showing linear independence of the family $(\xi^j)_{j\in\mathbb{Z}_{\geq 0}}$.

6. In the $\mathbb{R}$-vector space $\mathsf{C}^0([0,\pi];\mathbb{R})$ define vectors (i.e., functions) $\cos_j\colon I \to \mathbb{R}$, $j \in \mathbb{Z}_{\geq 0}$, and $\sin_j\colon I \to \mathbb{R}$, $j \in \mathbb{Z}_{>0}$, by

$$\cos_j = \cos(jx), \quad \sin_j(x) = \sin(jx).$$

We claim that the family $(\cos_j)_{j\in\mathbb{Z}_{\geq 0}} \cup (\sin_j)_{j\in\mathbb{Z}_{>0}}$ is linearly independent. To see this, suppose that a finite linear combination of these vectors vanishes:

$$a_1 \cos_{j_1} + \cdots + a_l \cos_{j_l} + b_1 \sin_{k_1} + \cdots + b_m \sin_{k_m} = 0_{\mathsf{C}^0([0,2\pi];\mathbb{R})}, \tag{4.18}$$

for $a_1,\dots,a_l,b_1,\dots,b_m \in \mathbb{R}$. Now multiply (4.18) by the function $\cos_{j_r}$ for some $r \in \{1,\dots,l\}$ and integrate both sides of the equation over the interval $[0,2\pi]$:

$$\begin{aligned} a_1 \int_0^{2\pi} \cos_{j_1}(x)\cos_{j_r}(x)\,\mathrm{d}x + \cdots + a_l \int_0^{2\pi} \cos_{j_l}(x)\cos_{j_r}(x)\,\mathrm{d}x \\ + b_1 \int_0^{2\pi} \sin_{k_1}(x)\cos_{j_r}(x)\,\mathrm{d}x + \cdots + b_m \int_0^{2\pi} \sin_{k_m}(x)\cos_{j_r}(x)\,\mathrm{d}x = 0. \end{aligned} \tag{4.19}$$

Now we recall the following trigonometric identities

$$\cos(a)\cos(b) = \tfrac{1}{2}(\cos(a-b) + \cos(a+b)), \quad \cos(a)\sin(b) = \tfrac{1}{2}(\sin(a+b) - \sin(a-b)),$$
$$\sin(a)\sin(b) = \tfrac{1}{2}(\cos(a-b) - \cos(a+b)),$$
$$\cos^2(a) = \tfrac{1}{2}(1+\cos(2a)), \quad \sin^2(a) = \tfrac{1}{2}(1-\cos(2a),$$

for $a,b \in \mathbb{R}$. The above identities are easily proved using Euler's formula $\mathrm{e}^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x)$ and properties of the exponential function. We recommend that the reader learn these derivations and then overwrite that portion of their memory used for storing trigonometric identities with something useful like, say, sports statistics or lines from their favourite movies. The above trigonometric identities can now be used, along with the derivative (and hence integral, by the Fundamental Theorem of Calculus) rules for trigonometric functions to derive the following identities for $j, k \in \mathbb{Z}_{>0}$:

$$\int_0^{2\pi} \cos(jx)\cos(kx)\,\mathrm{d}x = \begin{cases} 0, & j \neq k, \\ \pi, & j = k, \end{cases}$$

$$\int_0^{2\pi} \cos(jx)\sin(kx)\,\mathrm{d}x = 0,$$

$$\int_0^{2\pi} \sin(jx)\sin(kx)\,\mathrm{d}x = \begin{cases} 0, & j \neq k, \\ \pi, & j = k, \end{cases}$$

$$\int_0^{2\pi} \cos(0x)\cos(0x)\,\mathrm{d}x = 2\pi,$$

$$\int_0^{2\pi} \cos(0x)\cos(kx)\,\mathrm{d}x = 0,$$

$$\int_0^{2\pi} \cos(0x)\sin(kx)\,\mathrm{d}x = 0.$$

Applying these identities to (4.19) gives $\pi a_r = 0$ if $j_r \neq 0$ and gives $2\pi a_r = 0$ if $j_r = 0$. In either case we deduce that $a_r = 0$, $r \in \{1, \dots, l\}$. In like manner, multiplying (4.18) by $\sin_{k_s}$, $s \in \{1, \dots, m\}$, and integrating over the interval $[0, 2\pi]$ gives $b_s = 0$, $s \in \{1, \dots, m\}$. This shows that the coefficients in the linear combination (4.18) are zero, and, therefore, that the set $(\cos_j)_{j \in \mathbb{Z}_{\geq 0}} \cup (\sin_j)_{j \in \mathbb{Z}_{>0}}$ is indeed linearly independent. •

The reader will hopefully have noticed strong similarities between Examples 1 and 4 and between Examples 2 and 5. This is not an accident, but is due to the fact that the vector spaces F^{k+1} and $\mathsf{F}_k[\xi]$ are isomorphic and that the vector spaces F_0^∞ and $\mathsf{F}[\xi]$ are isomorphic. The reader is asked to explicitly write isomorphisms of these vector spaces in Exercise 4.5.21.

Let us now prove some facts about linearly independent and linearly dependent sets.

4.5.19 Proposition (Properties of linearly (in)dependent sets) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{S} \subseteq \mathsf{V}$ *be nonempty. Then the following statements hold:*

(i) if $\mathsf{S} = \{\mathrm{v}\}$ *for some* $\mathrm{v} \in \mathsf{V}$*, then* S *is linearly independent if and only if* $\mathrm{v} \neq 0_{\mathsf{V}}$*;*

(ii) if $0_{\mathsf{V}} \in \mathsf{S}$ *then* S *is linearly dependent;*

(iii) if S *is linearly independent and if* $\mathsf{T} \subseteq \mathsf{S}$ *is nonempty, then* T *is linearly independent;*

(iv) if S *is linearly dependent and if* $\mathsf{T} \subseteq \mathsf{V}$*, then* $\mathsf{S} \cup \mathsf{T}$ *is linearly dependent;*

(v) if S *is linearly independent, if* $\{\mathrm{v}_1, \dots, \mathrm{v}_k\} \subseteq \mathsf{S}$*, and if*

$$\mathrm{a}_1\mathrm{v}_1 + \cdots + \mathrm{a}_k\mathrm{v}_k = \mathrm{b}_1\mathrm{v}_1 + \cdots + \mathrm{b}_k\mathrm{v}_k$$

for $\mathrm{a}_1, \dots, \mathrm{a}_k, \mathrm{b}_1, \dots, \mathrm{b}_k \in \mathsf{F}$*, then* $\mathrm{a}_j = \mathrm{b}_j$*,* $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$*;*

(vi) *if* S *is linearly independent and if* $\mathrm{v} \notin \mathrm{span}_{\mathsf{F}}(\mathrm{S})$, *then* $\mathrm{S} \cup \{\mathrm{v}\}$ *is linearly independent.*

Proof (i) Note that $c0_{\mathsf{V}} = 0_{\mathsf{V}}$ if and only if $c = 0_{\mathsf{F}}$ by Proposition 4.5.3(vi). This is exactly equivalent to what we are trying to prove.

(ii) If $0_{\mathsf{V}} \in S$ then the finite subset $\{0_{\mathsf{V}}\}$ is linearly dependent by part (i).

(iii) Let $\{v_1, \dots, v_k\} \subseteq T \subseteq S$ and suppose that

$$c_1 v_1 + \dots c_k v_k = 0_{\mathsf{V}}$$

for $c_1, \dots, c_k \in \mathsf{F}$. Since $\{v_1, \dots, v_k\} \subseteq S$ and since S is linearly independent, it follows that $c_1 = \dots = c_k = 0_{\mathsf{F}}$.

(iv) Since S is linearly dependent there exists vectors $\{v_1, \dots, v_k\} \subseteq S$ and $c_1, \dots, c_k \in \mathsf{F}$ not all zero such that

$$c_1 v_1 + \dots + c_k v_k = 0_{\mathsf{V}}.$$

Since $\{v_1, \dots, v_k\} \subseteq S \cup T$, it follows that $S \cup T$ is linearly dependent.

(v) If

$$a_1 v_1 + \dots + a_k v_k = b_1 v_1 + \dots + a_k v_k,$$

then

$$(a_1 - b_1) v_1 + \dots + (a_k - b_k) v_k = 0_{\mathsf{V}}.$$

Since the set $\{v_1, \dots, v_k\}$ is linearly independent, it follows that $a_j - b_j = 0_{\mathsf{F}}$ for $j \in \{1, \dots, k\}$, which gives the result.

(vi) Let $\{v_1, \dots, v_k\} \subseteq S \cup \{v\}$. If $\{v_1, \dots, v_k\} \subseteq S$ then the set is immediately linearly independent. If $\{v_1, \dots, v_k\} \not\subseteq S$, then we may without loss of generality suppose that $v_k = v$. Suppose that

$$c_1 v_1 + \dots + c_{k-1} v_{k-1} + c_k v_k = 0_{\mathsf{V}}.$$

First suppose that $c_k \neq 0_{\mathsf{F}}$. Then

$$v_k = -c_k^{-1} c_1 v_1 + \dots + c_k^{-1} c_{k-1} v_{k-1},$$

which contradicts the fact that $v_k \notin \mathrm{span}_{\mathsf{F}}(S)$. Thus we must have $c_k = 0_{\mathsf{F}}$. However, since S is linearly independent, it immediately follows that $c_1 = \dots = c_{k-1} = 0_{\mathsf{F}}$. Thus $S \cup \{v\}$ is linearly independent. ■

4.5.4 Basis and dimension

The notion of the dimension of a vector space, which is derived from the concept of a basis, is an important one. Of particular importance is the dichotomy between vector spaces whose dimension is finite and those whose dimension is infinite. Essentially, finite-dimensional vector spaces, particularly those defined over $\mathbb{R}$, behave in a manner which often correspond somehow to our intuition. In infinite dimensions, however, our intuition can often lead us astray. And in these volumes we will be often interested in infinite-dimensional vector spaces. This infinite-dimensional case is complicated, and any sort of understanding will require understanding much of Chapter III-6.

For now, we get the ball rolling by introducing the idea of a basis.

4.5.20 Definition (Basis for a vector space) Let F be a field and let V be a vector space over F. A ***basis*** for V is a subset $\mathscr{B}$ of V with the properties that

(i) $\mathscr{B}$ is linearly independent and

(ii) $\operatorname{span}_{\mathsf{F}}(\mathscr{B}) = \mathsf{V}$. •

4.5.21 Remark (Hamel[4] basis) Readers who have had a first course in linear algebra should be sure to note that we do not require a basis to be a finite set. Nonetheless, the definition we give is probably exactly the same as the one encountered in a typical first course. What is different is that we have defined the notion of linear independence and the notion associated with the symbol "$\operatorname{span}_{\mathsf{F}}(\cdot)$" in a general way. Sometimes the word "basis" is reserved for finite sets of vectors, with the notion we give being called a ***Hamel basis***. •

Let us first prove that every vector space possesses a basis in the sense that we have defined the notion.

4.5.22 Theorem (Every vector space possesses a basis) *If* F *is a field and if* V *is an* F*-vector space, then there exists a basis for* V.

Proof Let $\mathscr{C}$ be the collection of subsets of V that are linearly independent. Such collections exist since, for example, $\{v\} \in \mathscr{C}$ if $v \in \mathsf{V}$ is nonzero. Place a partial order $\preceq$ on $\mathscr{C}$ by asking that $S_1 \preceq S_2$ if $S_1 \subseteq S_2$. Let $\mathscr{S} \subseteq \mathscr{C}$ be a totally ordered subset. Note that $\cup_{S\in\mathscr{S}} S$ is an element of $\mathscr{C}$. Indeed, let $\{v_1,\dots,v_k\} \subseteq \cup_{S\in\mathscr{S}} S$. Then $v_j \in S_j$ for some $S_j \in \mathscr{S}$. Let $j_0 \in \{1,\dots,k\}$ be chosen such that S_{j_0} is the largest of the sets $S_1,\dots,S_k$ according to the partial order $\preceq$, this being possible since $\mathscr{S}$ is totally ordered. Then $\{v_1,\dots,v_k\} \subseteq S_{j_0}$ and so $\{v_1,\dots,v_k\}$ is linearly independent since S_{j_0} is linearly independent. It is also evident that $\cup_{S\in\mathscr{S}} S$ is an upper bound for $\mathscr{S}$. Thus every totally ordered subset of $\mathscr{C}$ possesses an upper bound, and so by Zorn's Lemma possesses a maximal element. Let $\mathscr{B}$ be such a maximal element. By construction $\mathscr{B}$ is linearly independent. Let $v \in \mathsf{V}$ and suppose that $v \notin \operatorname{span}_{\mathsf{F}}(\mathscr{B})$. Then by Proposition 4.5.19(vi), $\mathscr{B} \cup \{v\}$ is linearly independent and $\mathscr{B} \subseteq \mathscr{B}\{v\}$. This contradicts the fact that $\mathscr{B}$ is maximal, and so it must hold that if $v \in \mathsf{V}$, then $v \in \operatorname{span}_{\mathsf{F}}(\mathscr{B})$. That is to say, $\operatorname{span}_{\mathsf{F}}(\mathscr{B}) = \mathsf{V}$. ■

One of the important properties of a basis is the following result.

4.5.23 Proposition (Unique representation of vectors in bases) *If* F *is a field, if* V *is an* F*-vector space, and if* $\mathscr{B}$ *is a basis for* V*, then, for* $\mathrm{v} \in \mathsf{V}$ *there exists a unique finite subset* $\{\mathrm{v}_1,\dots,\mathrm{v}_k\} \subseteq \mathscr{B}$ *and unique nonzero coefficients* $\mathrm{c}_1,\dots,\mathrm{c}_k \in \mathsf{F}$ *such that*

$$\mathrm{v} = \mathrm{c}_1\mathrm{v}_1 + \cdots + \mathrm{c}_k\mathrm{v}_k.$$

Proof Let $v \in \mathsf{V}$. Since $\operatorname{span}_{\mathsf{F}}(\mathscr{B}) = \mathsf{V}$, there exists $\{u_1,\dots,u_l\} \subseteq \mathscr{B}$ and $a_1,\dots,a_l \in \mathsf{F}$ such that

$$v = a_1u_1 + \cdots + a_lu_l. \tag{4.20}$$

[4] Georg Karl Wilhelm Hamel (1877–1954) was a German mathematician whose contributions to mathematics were in the areas of function theory, mechanics, and the foundations of mathematics

Moreover, given the vectors $\{u_1, \dots, u_l\}$, the coefficients $a_1, \dots, a_l$ in (4.20) are unique. Let $\{v_1, \dots, v_k\} \subseteq \{u_1, \dots, u_l\}$ be these vectors for which the corresponding coefficient in (4.20) is nonzero. Denote by $c_1, \dots, c_k$ the coefficients in (4.20) corresponding to the vectors $\{v_1, \dots, v_k\}$. This gives the existence part of the result.

Suppose that $\{v'_1, \dots, v'_{k'}\} \subseteq \mathscr{B}$ and $c'_1, \dots, c'_{k'} \in \mathsf{F}^*$ satisfy

$$v = c'_1 v'_1 + \dots + c'_{k'} v'_{k'}.$$

Now take $\{w_1, \dots, w_m\}$ to be a set of vectors such that $\{w_1, \dots, w_m\} = \{v_1, \dots, v_k\} \cup \{v'_1, \dots, v'_{k'}\}$. Note that

$$\{v_1, \dots, v_k\}, \{v'_1, \dots, v'_{k'}\} \subseteq \{w_1, \dots, w_m\}.$$

Since $\{w_1, \dots, w_m\} \subseteq \mathscr{B}$ it is linearly independent. Therefore, by Proposition 4.5.19(v), there exists unique coefficients $b_1, \dots, b_m \in \mathsf{F}$ such that

$$v = b_1 w_1 + \dots + b_m w_m.$$

But we also have

$$v = c_1 v_1 + \dots + c_k v_k = c'_1 v'_1 + \dots + c'_{k'} v'_{k'}.$$

Therefore, it must hold that $\{v_1, \dots, v_k\} = \{v'_1, \dots, v'_{k'}\} = \{w_1, \dots, w_m\}$, and from this the result follows. ■

One of the more useful characterisations of bases is the following result.

4.5.24 Theorem (Linear maps are uniquely determined by their values on a basis) *Let F be a field, let V be an F-vector space, and let $\mathscr{B} \subseteq \mathsf{V}$ be a basis. Then, for any F-vector space W and any map $\phi\colon \mathscr{B} \to \mathsf{W}$ there exists a unique linear map $\mathsf{L}_\phi \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}; \mathsf{W})$ such that the diagram*

$$\begin{tikzcd}
\mathscr{B} \arrow[r, "\phi"] \arrow[d] & \mathsf{W} \\
\mathsf{V} \arrow[ur, "\mathsf{L}_\phi"'] &
\end{tikzcd}$$

commutes, where the vertical arrow is the inclusion.

Proof Denote $\mathscr{B} = \{e_i\}_{i \in I}$. If $v \in \mathsf{V}$ we have $v = \sum_{i \in I} v_i e_i$ for $v_i \in \mathsf{F}$, $i \in I$, all but finitely many of which are zero. Then define

$$\mathsf{L}_\phi(v) = \sum_{i \in I} v_i \phi(e_i).$$

This map is linear since

$$\mathsf{L}_\phi(u + v) = \sum_{i \in I} (u_i + v_i)\phi(e_i) = \sum_{i \in I} u_i \phi(e_i) + \sum_{i \in I} v_i \phi(e_i) = \mathsf{L}_\phi(u) + \mathsf{L}_\phi(v)$$

and

$$\mathsf{L}_\phi(av) = \sum_{i \in I} a v_i \phi(e_i) = a \sum_{i \in I} v_i \phi(e_i) = a\mathsf{L}_\phi(v),$$

where all manipulations make sense by virtue of the sums being finite. This gives the existence part of the theorem.

Suppose that $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{W})$ is another linear map for which the diagram in the theorem statement commutes. This implies that $\mathsf{L}(e_i) = \mathsf{L}_\phi(e_i)$ for $i \in I$. Now, if $v = \sum_{i\in I} v_i e_i$ is a finite linear combination of basis elements, then

$$\mathsf{L}\left(\sum_{i\in I} v_i e_i\right) = \sum_{i\in I} v_i \mathsf{L}(e_i) = \sum_{i\in I} v_i \mathsf{L}_\phi(e_i) = \mathsf{L}_\phi\left(\sum_{i\in I} v_i e_i\right),$$

giving $\mathsf{L} = \mathsf{L}_\phi$. ■

The theorem is very useful, and indeed often used, since it tells us that to define a linear map one need only define it on each vector of a basis.

As we shall shortly see, the notion of the dimension of a vector space relies completely on a certain property of any two bases for a vector space, namely that they have the same cardinality.

4.5.25 Theorem (Different bases have the same size) *If F is a field, if V is an F-vector space, and if $\mathscr{B}_1$ and $\mathscr{B}_2$ are two bases for V, then $\mathrm{card}(\mathscr{B}_1) = \mathrm{card}(\mathscr{B}_2)$.*

Proof The proof is broken into two parts, the first for the case when one of $\mathscr{B}_1$ and $\mathscr{B}_2$ is finite, and the second the case when both $\mathscr{B}_1$ and $\mathscr{B}_2$ are infinite.

Let us first prove the following lemma.

1 Lemma *If $\{\mathrm{v}_1,\dots,\mathrm{v}_\mathrm{n}\}$ is a basis for V then any set of $\mathrm{n}+1$ vectors in V is linearly dependent.*

Proof We prove the lemma by induction on n. In the case when $n = 1$ we have $\mathsf{V} = \mathrm{span}_{\mathsf{F}}(v_1)$. Let $u_1, u_2 \in \mathsf{V}$ so that $u_1 = a_1 v_1$ and $u_2 = a_2 v_1$ for some $a_1, a_2 \in \mathsf{F}$. If either u_1 or u_2 is zero then the set $\{u_1, u_2\}$ is immediately linearly dependent by Proposition 4.5.19(ii). Thus we can assume that a_1 and a_2 are both nonzero. In this case we have

$$a_2 u_1 - a_1 u_2 = a_2(a_1 v_1) - a_1(a_2 v_1) = 0_{\mathsf{V}},$$

so that $\{u_1, u_2\}$ is not linearly independent. Now suppose that the lemma holds for $n \in \{1,\dots,k\}$ and let $\{v_1,\dots,v_{k+1}\}$ be a basis for V. Consider a set $\{u_1,\dots,u_{k+2}\}$ and write

$$u_s = \sum_{r=1}^{k+1} a_{rs} v_r, \qquad s \in \{1,\dots,k+2\}.$$

First suppose that $a_{1s} = 0_{\mathsf{F}}$ for all $s \in \{1,\dots,k+2\}$. It then holds that $\{u_1,\dots,u_{k+2}\} \subseteq \mathrm{span}_{\mathsf{F}}(v_2,\dots,v_{k+1})$. By the induction hypothesis, since $\mathrm{span}_{\mathsf{F}}(v_2,\dots,v_{k+1})$ has basis $\{v_2,\dots,v_{k+1}\}$, it follows that $\{u_1,\dots,u_{k+1}\}$ is linearly dependent, and so $\{u_1,\dots,u_{k+2}\}$ is also linearly dependent by Proposition 4.5.19(iv). Thus we suppose that not all of the coefficients a_{1s}, $s \in \{1,\dots,k+2\}$ is zero. For convenience, and without loss of generality, suppose that $a_{11} \neq 0_{\mathsf{F}}$. Then

$$a_{11}^{-1} u_1 = v_1 + a_{11}^{-1} a_{21} v_2 + \dots + a_{11}^{-1} a_{k+1,1} v_{k+1}.$$

We then have

$$u_s - a_{11}^{-1}a_{1s}u_1 = \sum_{r=2}^{k+1}(a_{rs} + a_{1s}a_{11}^{-1}a_{r1})v_r, \qquad s \in \{2, \dots, k+2\}.$$

meaning that $u_s - a_{11}^{-1}a_{1s}u_1 \in \operatorname{span}_{\mathsf{F}}(v_2, \dots, v_{k+1})$ for $s \in \{2, \dots, k+2\}$. By the induction hypothesis it follows that the set $\{u_2 - a_{11}^{-1}a_{12}u_1, \dots, u_{k+2} - a_{11}^{-1}a_{1,k+2}u_1\}$ is linearly dependent. We claim that this implies that $\{u_1, u_2, \dots, u_{k+2}\}$ is linearly dependent. Indeed, let $c_2, \dots, c_{k+2} \in \mathsf{F}$ be not all zero and such that

$$c_2(u_2 - a_{11}^{-1}a_{12}u_1) + \dots + c_{k+2}(u_{k+2} - a_{11}^{-1}a_{1,k+2}u_1) = 0_{\mathsf{V}}.$$

Then

$$(-c_2a_{11}^{-1}a_{12} - \dots - c_{k+2}a_{11}^{-1}a_{1,k+2})u_1 + c_2u_2 + \dots + c_{k+2}u_{k+2} = 0_{\mathsf{V}}.$$

Since not all of the coefficients $c_2, \dots, c_{k+2}$ are zero, it follows that $\{u_1, u_2, \dots, u_{k+2}\}$ is linearly dependent. This completes the proof. ▼

Now consider the case when either $\mathscr{B}_1$ or $\mathscr{B}_2$ is finite. Thus, without loss of generality suppose that $\mathscr{B}_1 = \{v_1, \dots, v_n\}$. It follows that $\mathscr{B}_2$ can have at most n elements. Thus $\mathscr{B}_2 = \{u_1, \dots, u_m\}$ for $m \leq n$. But, since $\mathscr{B}_2$ is a basis, it also holds that $\mathscr{B}_1$ must have at most m elements. Thus $n \leq m$, and so $m = n$ and thus $\operatorname{card}(\mathscr{B}_1) = \operatorname{card}(\mathscr{B}_2)$.

Now let us turn to the general case when either or both of $\mathscr{B}_1$ and $\mathscr{B}_2$ are infinite. For $u \in \mathscr{B}_1$ let $\mathscr{B}_2(u)$ be the unique finite subset $\{v_1, \dots, v_k\}$ of $\mathscr{B}_2$ such that

$$u = c_1v_1 + \dots + c_kv_k$$

for some $c_1, \dots, c_k \in \mathsf{F}^*$. We now prove a lemma.

2 **Lemma** *If* $\mathrm{v} \in \mathscr{B}_2$ *then there exists* $\mathrm{u} \in \mathscr{B}_1$ *such that* $\mathrm{v} \in \mathscr{B}_2(\mathrm{u})$*.*

Proof Suppose otherwise. Thus suppose that there exists $v \in \mathscr{B}_2$ such that, for every $u \in \mathscr{B}_1$, $v \notin \mathscr{B}_2(u)$. We claim that $\mathscr{B}_1 \cup \{v\}$ is then linearly independent. Indeed, let $\{v_1, \dots, v_k\} \subseteq \mathscr{B}_1 \cup \{v\}$. If $\{v_1, \dots, v_k\} \subseteq \mathscr{B}_1$ then we immediately have that $\{v_1, \dots, v_k\}$ is linearly independent. So suppose that $\{v_1, \dots, v_k\} \not\subseteq \mathscr{B}_1$, and suppose without loss of generality that $v_k = v$. Let $c_1, \dots, c_k \in \mathsf{F}$ satisfy

$$c_1v_1 + \dots + c_kv_k = 0_{\mathsf{V}}.$$

If $c_k \neq 0_{\mathsf{F}}$ then

$$v = -c_k^{-1}c_1v_1 + \dots - c_k^{-1}c_{k-1}v_{k-1},$$

implying that $v \in \operatorname{span}_{\mathsf{F}}(v_1, \dots, v_{k-1})$. We can thus write v as a linear combination of vectors from the finite subsets $\mathscr{B}_2(v_j)$, $j \in \{1, \dots, k-1\}$. Let $\{w_1, \dots, w_m\}$ be a set of distinct vectors with the property that

$$\{w_1, \dots, w_m\} = \cup_{j=1}^{k-1}\mathscr{B}_2(v_j).$$

Thus $\mathscr{B}_2(v_j) \subseteq \{w_1, \dots, w_m\}$ for $j \in \{1, \dots, k-1\}$. It then follows that $v \in \operatorname{span}_{\mathsf{F}}(w_1, \dots, w_m)$. However, since $v \notin \{w_1, \dots, w_m\}$ by our assumption that $v \notin \mathscr{B}_2(u)$

for every $u \in \mathscr{B}_1$, it follows that $\{v, w_1, \dots, w_m\}$ is linearly independent, which is a contradiction. Therefore, $c_k = 0_{\mathsf{F}}$.

On the other hand, if $c_k = 0_{\mathsf{F}}$ then it immediately follows that $c_1 = \dots = c_{k-1} = 0_{\mathsf{F}}$ since $\{v_1, \dots, v_{k-1}\} \subseteq \mathscr{B}_1$ and since $\mathscr{B}_1$ is linearly independent. Therefore, $\mathscr{B}_1 \cup \{v\}$ is indeed linearly independent. In particular, $v \notin \operatorname{span}_{\mathsf{F}}(\mathscr{B}_1)$, contradicting the fact that $\mathscr{B}_1$ is a basis. ▼

From the lemma we know that $\mathscr{B}_2 = \cup_{u \in \mathscr{B}_1} \mathscr{B}_2(u)$. By the definition of multiplication of cardinal numbers, and using the fact that $\operatorname{card}(\mathbb{Z}_{>0})$ exceeds every finite cardinal number, we have

$$\operatorname{card}(\mathscr{B}_2) \leq \operatorname{card}(\mathscr{B}_1) \operatorname{card}(\mathbb{Z}_{>0}).$$

By Corollary 1.7.18 it follows that $\operatorname{card}(\mathscr{B}_2) \leq \operatorname{card}(\mathscr{B}_1)$. By interchanging the rôles of $\mathscr{B}_1$ and $\mathscr{B}_2$ we can also show that $\operatorname{card}(\mathscr{B}_1) \leq \operatorname{card}(\mathscr{B}_2)$. By the Cantor–Schröder–Bernstein Theorem, $\operatorname{card}(\mathscr{B}_1) = \operatorname{card}(\mathscr{B}_2)$. ■

Let us give some other useful constructions concerning bases. The proofs we give are valid for arbitrary bases. We invite the reader to give proofs in the case of finite bases in Exercise 4.5.18.

4.5.26 Theorem (Bases and linear independence) *Let* F *be a field and let* V *be an* F*-vector space. For a subset* $\mathsf{S} \subseteq \mathsf{V}$*, the following statements hold:*

(i) if S *is linearly independent, then there exists a basis* $\mathscr{B}$ *for* V *such that* $\mathsf{S} \subseteq \mathscr{B}$*;*

(ii) if $\operatorname{span}_{\mathsf{F}}(\mathsf{S}) = \mathsf{V}$*, then there exists a basis* $\mathscr{B}$ *for* V *such that* $\mathscr{B} \subseteq \mathsf{S}$*.*

Proof (i) Let $\mathscr{C}(S)$ be the collection of linearly independent subsets of V which contain S. Since $S \in \mathscr{C}(S)$, $\mathscr{C}(S) \neq \emptyset$. The set $\mathscr{C}(S)$ can be partially ordered by inclusion. Thus $S_1 \preceq S_2$ if $S_1 \subseteq S_2$. Just as in the proof of Theorem 4.5.22, every totally ordered subset of $\mathscr{C}(S)$ has an upper bound, and so $\mathscr{C}(S)$ possesses a maximal element $\mathscr{B}$ by Zorn's Lemma. This set may then be shown to be a basis just as in the proof of Theorem 4.5.22.

(ii) Let $\mathscr{D}(S)$ be the collection of linearly independent subsets of S, and partially order $\mathscr{D}(S)$ by inclusion, just as we partially ordered $\mathscr{C}(S)$ in part (i). JUst as in the proof of Theorem 4.5.22, every totally ordered subset of $\mathscr{D}(S)$ has an upper bound, and so $\mathscr{D}(S)$ possesses a maximal element $\mathscr{B}$. We claim that every element of S is a linear combination of elements of $\mathscr{B}$. Indeed, if this were not the case, then there exists $v \in S$ such that $v \notin \operatorname{span}_{\mathsf{F}}(\mathscr{B})$. Then $\mathscr{B} \cup \{v\}$ is linear independent by Proposition 4.5.19(vi), and is also contained in S. This contradicts the maximality of $\mathscr{B}$, and so we indeed have $S \subseteq \operatorname{span}_{\mathsf{F}}(\mathscr{B})$. Therefore,

$$\operatorname{span}_{\mathsf{F}}(\mathscr{B}) = \operatorname{span}_{\mathsf{F}}(S) = \mathsf{V},$$

giving the theorem. ■

Now it makes sense to talk about the dimension of a vector space.

4.5.27 Definition (Dimension, finite-dimensional, infinite-dimensional) Let F be a field, let V be an F-vector space, and let $\mathscr{B}$ be a basis for V. The ***dimension*** of the vector space V, denoted by $\dim_{\mathsf{F}}(\mathsf{V})$, is the cardinal number $\operatorname{card}(\mathscr{B})$. If $\mathscr{B}$ is finite then V is ***finite-dimensional***, and otherwise V is ***infinite-dimensional***. We will slightly abuse notation and write $\dim_{\mathsf{F}}(\mathsf{V}) = \infty$ whenever V is infinite-dimensional. •

Let us give some examples of vector spaces of various dimensions.

4.5.28 Examples (Basis and dimension)

1. The trivial vector space $\mathsf{V} = \{0_{\mathsf{V}}\}$ consisting of the zero vector has $\varnothing$ as a basis.
2. The F-vector space F^n has as a basis the set $\mathscr{B} = \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ defined in Example 4.5.18–1. In that example, $\mathscr{B}$ was shown to be linearly independent. Also, since
$$(v_1, \dots, v_n) = v_1\boldsymbol{e}_1 + \cdots + v_n\boldsymbol{e}_n,$$
it follows that $\operatorname{span}_{\mathsf{F}}(\mathscr{B}) = \mathsf{F}^n$. Thus $\dim_{\mathsf{F}}(\mathsf{F}^n) = n$. The basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ is called the ***standard basis***.
3. The subspace F_0^∞ of F^∞ has a basis which is easily described. Indeed, it is easy to verify that $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}}$ is a basis for F_0^∞. We adopt the notation from the finite-dimensional case and call this the ***standard basis***.
4. We next consider the F-vector space F^∞. Since $\mathsf{F}_0^\infty \subseteq \mathsf{F}^\infty$, and since the standard basis $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}}$ is linearly independent in F^∞, we know by Theorem 4.5.26 that we can extend the standard basis for F_0^∞ to a basis for F^∞. This extension is nontrivial since, for example, the sequence $\{1_{\mathsf{F}}\}_{j\in\mathbb{Z}_{>0}}$ in F cannot be written as a finite linear combination of standard basis vectors. Thus the set $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}} \cup \{\{1_{\mathsf{F}}\}_{j\in\mathbb{Z}_{>0}}\}$ is linearly independent. This linearly set shares with the standard basis the property of being countable. It turns out, in fact, that any basis for F^∞ has the cardinality of $\mathbb{R}$, and so the process of tacking on linearly independent vectors to the standard basis for F_0^∞ will take a long time to produce a basis for F^∞. We will not understand this properly until Section 5.7, where we will see that F^∞ is the algebraic dual of F_0^∞, and so thereby derive by general means the dimension of F^∞. For the moment we merely say that F^∞ is a much larger vector space than is F_0^∞.
5. In $\mathsf{F}_k[\xi]$, it is easy to verify that $\{1, \xi, \dots, \xi^k\}$ is a basis. Indeed, we have already shown that the set is linearly independent. It follows from the definition of $\mathsf{F}_k[\xi]$ that the set also generates $\mathsf{F}_k[\xi]$.
6. The set $\{\xi^j\}_{j\in\mathbb{Z}_{\geq 0}}$ forms a basis for $\mathsf{F}[\xi]$. Again, we have shown linear independence, and that this set generates $\mathsf{F}[\xi]$ follows by definition. •

4.5.29 Remark (Nonuniqueness of bases) Generally, it will not be the case that a vector spaces possesses a "natural" basis, although one might argue that the bases of Example 4.5.28 are fairly natural. But, even in cases where one might have a basis that is somehow distinguished, it is useful to keep in mind that other bases are possible, and that one should be careful not to rely overly on the comfort offered by a specific basis representation. In particular, if one is in the business of proving theorems using bases, one should make sure that what is being proved is independent of basis, if this is in fact what is intended. At this point in our presentation we do not have enough machinery at hand to explore this idea fully. Also, in Section 5.4.5 we shall discuss the matter of changing bases. •

Finally, let us prove the more or less obvious fact that dimension is preserved by isomorphism.

4.5.30 Proposition (Dimension characterises a vector space) *If* F *is a field and if* V_1 *and* V_2 *are* F*-vector spaces, then the following statements are equivalent:*

(i) V_1 *and* V_2 *are isomorphic;*

(ii) $\dim_{\mathsf{F}}(\mathsf{V}_1) = \dim_{\mathsf{F}}(\mathsf{V}_2)$.

Proof (i) $\Longrightarrow$ (ii) Let $\mathsf{L}\colon \mathsf{V}_1 \to \mathsf{V}_2$ be an isomorphism and let $\mathscr{B}_1$ be a basis for V_1. We claim that $\mathscr{B}_2 = \mathsf{L}(\mathscr{B}_1)$ is a basis for V_2. Let us first show that $\mathscr{B}_2$ is linearly independent. Let $v_1 = \mathsf{L}(u_1), \dots, v_k = \mathsf{L}(u_k) \in \mathscr{B}_2$ be distinct and suppose that

$$c_1 v_1 + \dots = c_k v_k = 0_{\mathsf{V}_2}$$

for $c_1, \dots, c_k \in \mathsf{F}$. Since L is linear we have

$$\mathsf{L}(c_1 u_1 + \dots + c_k u_k) = 0_{\mathsf{V}_2}.$$

Since L is injective, by Exercise 4.5.23 we have

$$c_1 u_1 + \dots + c_k u_k = 0_{\mathsf{V}_1},$$

showing that $c_1 = \dots = c_k = 0_{\mathsf{F}}$. Thus $\mathscr{B}_2$ is linearly independent. Moreover, for $v \in \mathsf{V}_2$ let $u = \mathsf{L}^{-1}(v)$ and then let $u_1, \dots, u_k \in \mathscr{B}_1$ and $c_1, \dots, c_k \in \mathsf{F}$ satisfy $u = c_1 u_1 + \dots + c_k u_k$. Then

$$\mathsf{L}(u) = c_1 \mathsf{L}(u_1) + \dots + c_k \mathsf{L}(u_k)$$

since L is linear. Therefore $v \in \operatorname{span}_{\mathsf{F}}(\mathscr{B}_2)$, and so $\mathscr{B}_2$ is indeed a basis. Since $\mathsf{L}|\mathscr{B}_1$ is a bijection onto $\mathscr{B}_2$ we have $\operatorname{card}(\mathscr{B}_2) = \operatorname{card}(\mathscr{B}_1)$, and this is the desired result.

(ii) $\Longrightarrow$ (i) Suppose that $\mathscr{B}_1$ and $\mathscr{B}_2$ are bases for V_1 and V_2, respectively, with the same cardinality. Thus there exists a bijection $\phi\colon \mathscr{B}_1 \to \mathscr{B}_2$. Now, by Theorem 4.5.24, define $\mathsf{L} \in \operatorname{Hom}_{\mathsf{F}}(\mathsf{V}_1; \mathsf{V}_2)$ by asking that $\mathsf{L}|\mathscr{B}_1 = \phi$. We claim that L is an isomorphism. To verify injectivity, suppose that $\mathsf{L}(u) = 0_{\mathsf{V}_2}$ for $u \in \mathsf{V}_1$. Write

$$u = c_1 u_1 + \dots + c_k u_k$$

for $c_1, \dots, c_k \in \mathsf{F}$ and $u_1, \dots, u_k \in \mathscr{B}_1$. Then

$$0_{\mathsf{V}_2} = c_1 \mathsf{L}(u_1) + \dots + c_k \mathsf{L}(u_k),$$

giving $c_j = 0_{\mathsf{F}}$, $j \in \{1, \dots, k\}$, since $\mathsf{L}(u_1), \dots, \mathsf{L}(u_k)$ are distinct elements of $\mathscr{B}_2$, and so linearly independent. Thus L is injective by Exercise 4.5.23. For surjectivity, let $v \in \mathsf{V}_2$ and write

$$v = c_1 v_1 + \cdots + c_k v_k$$

for $c_1, \dots, c_k \in \mathsf{F}$ and $v_1, \dots, v_k \in \mathscr{B}_2$. Then, if we define

$$u = c_1 \phi^{-1}(v_1) + \cdots + c_k \phi^{-1}(v_k) \in \mathsf{V}_2$$

we readily verify that $\mathsf{L}(u) = v$. ■

4.5.5 Intersections, sums, and products

In this section we investigate means of manipulating multiple subspaces and vector spaces. We begin by defining some constructions associated to subspaces of a vector space.

4.5.31 Definition (Sum and intersection) Let F be a field, let V be an F-vector space, and let $(\mathsf{U}_j)_{j \in J}$ be a family of subspaces of V indexed by a set J.

(i) The ***sum*** of $(\mathsf{U}_j)_{j \in J}$ is the subspace generated by $\cup_{j \in J} \mathsf{U}_j$, and is denoted by $\sum_{j \in J} \mathsf{U}_j$.

(ii) The ***intersection*** of $(\mathsf{U}_j)_{j \in J}$ is the set $\cap_{j \in J} \mathsf{U}_j$ (i.e., the set theoretic intersection). •

4.5.32 Notation (Finite sums of subspaces) If $\mathsf{U}_1, \dots, \mathsf{U}_k$ are a finite number of subspaces of an F-vector space V, then we will sometimes write

$$\sum_{j=1}^{k} \mathsf{U}_j = \mathsf{U}_1 + \cdots + \mathsf{U}_k. \qquad \bullet$$

4.5.33 Notation (Sum of subsets) We will also find it occasionally useful to be able to talk about sums of subsets that are not subspaces. Thus, if $(A_i)_{i \in I}$ is a family of subsets of an F-vector space V we denote by

$$\sum_{i \in I} A_i = \{v_{i_1} + \cdots + v_{i_k} \mid i_1, \dots, i_k \in I \text{ distinct}, \ v_{i_j} \in A_{i_j}, \ j \in \{1, \dots, k\}, \ k \in \mathbb{Z}_{>0}\}.$$

Thus $\sum_{i \in I} A_i$ consists of finite sums of vectors from the subsets A_i, $i \in I$. Following our notation above, if $I = \{1, \dots, k\}$ then we write

$$\sum_{i \in I} A_i = A_1 + \cdots + A_k. \qquad \bullet$$

The sum and intersection are the subspace analogues of the set theoretic union and intersection, with the analogue being exact in the case of intersection. Note that the union of subspaces need not be a subspace (see Exercise 4.5.17). It is true that the intersection of subspaces is a subspace.

4.5.34 Proposition (Intersections of subspaces are subspaces) *If* F *is a field, if* V *is an* F*-vector space, and if* $(\mathsf{U}_\mathrm{j})_{\mathrm{j}\in\mathrm{J}}$ *is a family of subspaces, then* $\cap_{\mathrm{j}\in\mathrm{J}}\mathsf{U}_\mathrm{j}$ *is a subspace.*

Proof If $v \in \cap_{j\in J}\mathsf{U}_a$ and if $a \in \mathsf{F}$ then $av \in \mathsf{U}_j$ for each $j \in J$. Thus $av \in \cap_{j\in J}\mathsf{U}_j$. If $v_1, v_2 \in \cap_{j\in J}\mathsf{U}_j$ then $v_1 + v_2 \in \mathsf{U}_j$ for each $j \in J$. Thus $v_1 + v_2 \in \cap_{j\in J}\mathsf{U}_j$. ■

Note that, by definition, if $(\mathsf{U}_j)_{j\in J}$ is a family of subspaces of an F-vector space V, and if $v \in \sum_{j\in J}\mathsf{U}_j$, then there exists a finite set $j_1, \dots, j_k \in J$ of indices and vectors $u_{j_l} \in \mathsf{U}_{j_l}$, $l \in \{1, \dots, k\}$, such that $v = u_{j_1} + \cdots + u_{j_k}$. In taking sums of subspaces, there is an important special instance when this decomposition is unique.

4.5.35 Definition (Internal direct sum of subspaces) Let F be a field, let V be an F-vector space, and let $(\mathsf{U}_j)_{j\in J}$ be a collection of subspaces of V. The vector space V is the ***internal direct sum*** of the subspaces $(\mathsf{U}_j)_{j\in J}$, and we write $\mathsf{V} = \bigoplus_{j\in J}\mathsf{U}_j$, if, for any $v \in \mathsf{V} \setminus \{0_\mathsf{V}\}$, there exists unique indices $\{j_1, \dots, j_k\} \subseteq J$ and unique nonzero vectors $u_{j_l} \in \mathsf{U}_{j_l}$, $l \in \{1, \dots, k\}$, such that $v = u_{j_1} + \cdots + u_{j_k}$. Each of the subspaces U_j, $j \in J$, is a ***summand*** in the internal direct sum. •

The following property of internal direct sums is useful.

4.5.36 Proposition (Representation of the zero vector in an internal direct sum of subspaces) *Let* F *be a field, let* V *be an* F*-vector space, and suppose that* V *is the internal direct sum of the subspaces* $(\mathsf{U}_\mathrm{j})_{\mathrm{j}\in\mathrm{J}}$*. If* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{k} \in \mathrm{J}$ *are distinct and if* $\mathrm{u}_{\mathrm{j}_\mathrm{l}} \in \mathsf{U}_{\mathrm{j}_\mathrm{l}}$, $\mathrm{l} \in \{1, \dots, \mathrm{k}\}$, *satisfy*

$$\mathrm{u}_{\mathrm{j}_1} + \cdots + \mathrm{u}_{\mathrm{j}_\mathrm{k}} = 0_\mathsf{V},$$

then $\mathrm{u}_{\mathrm{j}_\mathrm{l}} = 0_\mathsf{V}$, $\mathrm{l} \in \{1, \dots, \mathrm{k}\}$.

Proof Suppose that not all of the vectors $u_{j_1}, \dots, u_{j_k}$ are zero. Without loss of generality, then, suppose that $u_{j_1} \neq 0_\mathsf{V}$. Then

$$u_{j_1}, \text{ and } u_{j_1} + u_{j_1} + u_{j_2} + \cdots + u_{j_m} + u_{j_{m+1}}$$

are both representations of u_{j_1} as finite sums of vectors from the subspaces $(\mathsf{U}_j)_{j\in J}$. By the definition of internal direct sum it follows that $u_{j_1} = 2u_{j_1}$ and $u_{j_2} = \cdots = u_{j_k} = 0_\mathsf{V}$. Thus $u_{j_1} = 0_\mathsf{V}$, which is a contradiction. ■

The following alternative characterisation of the internal direct sum is sometimes useful.

4.5.37 Proposition (Characterisation of internal direct sum for vector spaces) *Let* F *be a field, let* V *be an* F*-vector space, and let* $(\mathsf{U}_\mathrm{j})_{\mathrm{j}\in\mathrm{J}}$ *be a collection of subspaces of* V*. Then* $\mathsf{V} = \bigoplus_{\mathrm{j}\in\mathrm{J}}\mathsf{U}_\mathrm{j}$ *if and only if*

(i) $\mathsf{V} = \sum_{\mathrm{j}\in\mathrm{J}}\mathsf{U}_\mathrm{j}$ *and,*

(ii) for any $\mathrm{j}_0 \in \mathrm{J}$*, we have* $\mathsf{U}_{\mathrm{j}_0} \cap \left(\sum_{\mathrm{j}\in\mathrm{J}\setminus\{\mathrm{j}_0\}}\mathsf{U}_\mathrm{j}\right) = \{0_\mathsf{V}\}$.

Proof Suppose that $\mathsf{V} = \bigoplus_{j\in J}\mathsf{U}_j$. By definition we have $\mathsf{V} = \sum_{j\in J}\mathsf{U}_j$. Let $j_0 \in J$ and suppose that $v \in \mathsf{U}_{j_0} \cap \left(\sum_{j\in J\setminus\{j_0\}}\mathsf{U}_j\right)$. Define $\mathsf{V}_{j_0} = \sum_{j\in J\setminus\{j_0\}}\mathsf{U}_j$ and note that $\mathsf{V}_{j_0} =$

$\bigoplus_{j\in J\setminus\{j_0\}} \mathsf{U}_j$. If $v \neq 0_{\mathsf{V}}$ then there exists unique indices $j_1, \dots, j_k \in J \setminus \{j_0\}$ and unique nonzero vectors $u_{j_l} \in \mathsf{U}_{j_l}$, $l \in \{1, \dots, k\}$, such that $v = u_{j_1} + \dots + u_{j_l}$. However, since we also have $v = v$, this contradicts the fact that there exists a unique collection $j'_1, \dots, j'_{k'} \in J$ of indices and unique nonzero vectors $u_{j'_l} \in \mathsf{U}_{j'_l}$, $l' \in \{1, \dots, k'\}$, such that $v = u_{j'_1} + \dots + u_{j'_k}$. Thus we must have $v = 0_{\mathsf{V}}$.

Now suppose that (i) and (ii) hold. Let $v \in \mathsf{V} \setminus \{0_{\mathsf{V}}\}$. It is then clear from (i) that there exists indices $j_1, \dots, j_k \in J$ and nonzero vectors $u_{j_l} \in \mathsf{U}_{j_l}$, $l \in \{1, \dots, k\}$, such that $v = u_{j_1} + \dots + u_{j_k}$. Suppose that $j'_1, \dots, j'_{k'}$ and $u'_{j'_1}, \dots, u'_{j'_{k'}}$ is another collection of indices and nonzero vectors such that $v = u'_{j'_1} + \dots + u'_{j'_{k'}}$. Then

$$0_{\mathsf{V}} = u_{j_1} + \dots + u_{j_k} - (u'_{j'_1} + \dots + u'_{j'_{k'}}).$$

By Proposition 4.5.36 it follows that if $l \in \{1, \dots, k\}$ and $l' \in \{1, \dots, k'\}$ satisfy $j_l = j'_{l'}$, then $u_{j_l} = u'_{j'_{l'}}$. If for $l \in \{1, \dots, k\}$ there exists no $l' \in \{1, \dots, k'\}$ such that $j_l = j'_{l'}$, then we must have $u_{j_l} = 0_{\mathsf{V}}$. Also, if for $l' \in \{1, \dots, k'\}$ there exists no $l \in \{1, \dots, k\}$ such that $j'_{l'} = j_l$, then we must have $u'_{j'_{l'}} = 0_{\mathsf{V}}$. From this we conclude that $\mathsf{V} = \bigoplus_{j\in J} \mathsf{U}_j$. ■

The notion of internal direct sum has the following important relationship with the notion of a basis.

4.5.38 Theorem (Bases and internal direct sums for vector spaces) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathscr{B}$ *be a basis for* V*, and define a family* $(\mathsf{U}_{\mathrm{u}})_{\mathrm{u}\in\mathscr{B}}$ *of subspaces by* $\mathsf{U}_{\mathrm{u}} = \mathrm{span}_{\mathsf{F}}(\mathrm{u})$*. Then* $\mathsf{V} = \bigoplus_{\mathrm{u}\in\mathscr{B}} \mathsf{U}_{\mathrm{u}}$.

Proof Let $v \in \mathsf{V}$. Since $\mathsf{V} = \mathrm{span}_{\mathsf{F}}(\mathscr{B})$, there exists $v_1, \dots, v_k \in \mathscr{B}$ and unique $c_1, \dots, c_k \in \mathsf{F}^*$ such that $v = c_1v_1 + \dots + c_kv_k$. Therefore, $u_j = c_jv_j \in \mathsf{U}_j$ for $j \in \{1, \dots, k\}$. Thus $u_1, \dots, u_k$ are the unique nonzero elements of the subspaces $(\mathsf{U}_u)_{u\in\mathscr{B}}$ such that $v = u_1 + \dots + u_k$. ■

Up to this point we have considered only operations on subspaces of a given vector space. Next we consider ways of combining vector spaces that are not necessarily subspaces of a certain vector space. The reader will at this point wish to recall the notion of a general Cartesian product as given in Section 1.6.2. Much of what will be needed in these volumes relies only on finite Cartesian products, so readers not wishing to wrap their minds around the infinite case can happily consider the following constructions only for finite collections of vector spaces.

4.5.39 Definition (Direct product and direct sum of vector spaces) Let F be a field and let $(\mathsf{V}_j)_{j\in J}$ be a family of F-vector spaces.

(i) The ***direct product*** of the family $(\mathsf{V}_j)_{j\in J}$ is the F-vector space $\prod_{j\in J} \mathsf{V}_j$ with vector addition and scalar multiplication defined by

$$(f_1 + f_2)(j) = f_1(j) + f_2(j), \quad (af)(j) = a(f(j))$$

for $f, f_1, f_2 \in \prod_{j\in J} \mathsf{V}_j$ and for $a \in \mathsf{F}$.

(ii) The ***direct sum*** of the family $(\mathsf{V}_j)_{j\in J}$ is the subspace $\bigoplus_{j\in J} \mathsf{V}_j$ of $\prod_{j\in J} \mathsf{V}_j$ consisting of those elements $f \colon J \to \cup_{j\in J}\mathsf{V}_j$ for which the set $\{j \in J \mid f(j) \neq 0_{\mathsf{V}_j}\}$ is finite. Each of the vector spaces V_j, $j \in J$, is a ***summand*** in the direct sum. •

4.5.40 Notation (Finite direct products and sums) In the case when the index set J is finite, say $J = \{1, \dots, k\}$, we clearly have $\prod_{j=1}^{k} \mathsf{V}_j = \bigoplus_{j=1}^{k} \mathsf{V}_j$. We on occasion adopt the convention of writing $\mathsf{V}_1 \oplus \dots \oplus \mathsf{V}_k$ for the resulting vector space in this case. This version of the direct sum (or equivalently direct product) is the one that we will most frequently encounter. •

Let us connect the notion of a direct sum with the notion of an internal direct sum as encountered in Definition 4.5.35. This also helps to rectify the potential inconsistency of multiple uses of the symbol $\bigoplus$. The reader will want to be sure they understand infinite Cartesian products in reading this result.

4.5.41 Proposition (Internal direct sum and direct sum of vector spaces) *Let* F *be a field, let* V *be an* F*-vector space, and let* $(\mathsf{U}_j)_{j\in J}$ *be a family of subspaces of* V *such that* V *is the internal direct sum of these subspaces. Let* $i_{\mathsf{U}_j} \colon \mathsf{U}_j \to \mathsf{V}$ *be the inclusion. Then the map from the direct sum* $\bigoplus_{j\in J} \mathsf{U}_j$ *to* V *defined by*

$$f \mapsto \sum_{j\in J} i_{\mathsf{U}_j} f(j)$$

(noting that the sum is finite) is an isomorphism.

Proof Let us denote the map in the statement of the proposition by L. For $f, f_1, f_2 \in \bigoplus_{j\in J} \mathsf{U}_j$ and for $a \in \mathsf{F}$ we have

$$\mathsf{L}(f_1 + f_2) = \sum_{j\in J} (f_1 + f_2)(j) = \sum_{j\in J} (f_1(j) + f_2(j)) = \sum_{j\in J} f_1(j) + \sum_{j\in J} f_2(j) = \mathsf{L}(f_1) + \mathsf{L}(f_2)$$

and

$$\mathsf{L}(af) = \sum_{j\in J} (af)(j) = \sum_{j\in J} a(f(j)) = a \sum_{j\in J} f(j) = a\mathsf{L}(f),$$

using the fact that all sums are finite. This proves linearity of $a\mathsf{L}$.

Next suppose that $\mathsf{L}(f) = 0_{\mathsf{V}}$. By Proposition 4.5.36 it follows that $f(j) = 0_{\mathsf{V}}$ for each $j \in J$. This gives injectivity of L by Exercise 4.5.23. If $v \in \mathsf{V}$, we can write $v = u_{j_1} + \dots + u_{j_k}$ for $j_1, \dots, j_k \in J$ and for $u_{j_l} \in \mathsf{U}_{j_l}$, $l \in \{1, \dots, k\}$. If we define $f \in \bigoplus_{j\in J} \mathsf{U}_j$ by $f(j_l) = u_{j_l}$, $l \in \{1, \dots, k\}$ and $f(j) = 0_{\mathsf{V}}$ for $j \notin \{j_1, \dots, j_k\}$, then $\mathsf{L}(f) = v$, showing that L is surjective. ■

4.5.42 Notation ("Internal direct sum" versus "direct sum") In the setup of the proposition, the direct sum $\bigoplus_{j\in J} \mathsf{U}_j$ is sometimes called the ***external direct sum*** of the subspaces $(\mathsf{U}_j)_{j\in J}$. The proposition says that the external direct sum is isomorphic to the internal direct sum. We shall often simply say "direct sum" rather than explicitly indicating the nature of the sum. •

Let us give an important example of a direct sum.

4.5.43 Example (The direct sum of copies of F) Let J be an arbitrary index set and let $\bigoplus_{j\in J}\mathsf{F}$ be the direct sum of "J copies" of the field F. In the case when $J=\{1,\dots,n\}$ we have $\bigoplus_{j\in J}\mathsf{F}=\mathsf{F}^n$ and in the case when $J=\mathbb{Z}_{>0}$ we have $\bigoplus_{j\in J}\mathsf{F}=\mathsf{F}_0^\infty$. Thus this example generalises two examples we have already encountered. For $j\in J$ define $\boldsymbol{e}_j\colon J\to\mathsf{F}$ by

$$\boldsymbol{e}_j(j')=\begin{cases}1_{\mathsf{F}}, & j'=j,\\ 0_{\mathsf{F}}, & j'\neq j.\end{cases}$$

(Recall the definition of the Cartesian product to remind yourself that $\boldsymbol{e}_j\in\bigoplus_{j\in J}\mathsf{F}$.) We claim that $\{\boldsymbol{e}_j\}_{j\in J}$ is a basis for $\bigoplus_{j\in J}\mathsf{F}$. First let us show that the set is linearly independent. Let $j_1,\dots,j_k\in J$ be distinct and suppose that, for every $j'\in J$,

$$c_1\boldsymbol{e}_{j_1}(j')+\cdots+c_k\boldsymbol{e}_{j_k}(j')=0_{\mathsf{F}}$$

for some $c_1,\dots,c_k\in\mathsf{F}$. Then, taking $j'=j_l$ for $l\in\{1,\dots,k\}$ we obtain $c_l=0_{\mathsf{F}}$. This gives linear independence. It is clear by definition of the direct sum that

$$\operatorname{span}_{\mathsf{F}}(\{\boldsymbol{e}_j\}_{j\in J})=\bigoplus_{j\in J}\mathsf{F}.$$

We call $\{\boldsymbol{e}_j\}_{j\in J}$ the ***standard basis*** for $\bigoplus_{j\in J}\mathsf{F}$. •

4.5.44 Notation (Alternative notation for direct sums and direct products of copies of F) There will be times when it is convenient to use notation that is less transparent, but more compact, than the notation $\prod_{j\in J}\mathsf{F}$ and $\bigoplus_{j\in J}\mathsf{F}$. The notation we adopt, motivated by Examples 4.5.2–3 and 4 is

$$\prod_{j\in J}\mathsf{F}=\mathsf{F}^J,\quad \bigoplus_{j\in J}\mathsf{F}=\mathsf{F}_0^J.$$

For the direct product, this notation is in fact perfect, since, as sets, $\prod_{j\in J}\mathsf{F}$ and F^J are identical. •

The importance of the direct sum is now determined by the following theorem.

4.5.45 Theorem (Vector spaces are isomorphic to direct sums of one-dimensional subspaces) *Let F be a field, let V be an F-vector space, and let $\mathscr{B}\subseteq\mathsf{V}$ be a basis. Let $\{\mathbf{e}_{\mathrm{u}}\}_{\mathrm{u}\in\mathscr{B}}$ be the standard basis for $\bigoplus_{\mathrm{u}\in\mathscr{B}}\mathsf{F}$ and define a map $\iota_{\mathscr{B}}\colon\{\mathbf{e}_{\mathrm{u}}\}_{\mathrm{u}\in\mathscr{B}}\to\mathscr{B}$ by $\iota_{\mathscr{B}}(\mathbf{e}_{\mathrm{u}})=\mathrm{u}$. Then there exists a unique F-isomorphism $\iota_{\mathsf{V}}\colon\bigoplus_{\mathrm{u}\in\mathscr{B}}\mathsf{F}\to\mathsf{V}$ such that the following diagram commutes:*

$$\begin{tikzcd}
\{\mathbf{e}_{\mathrm{u}}\}_{\mathrm{u}\in\mathscr{B}} \arrow[r, "\iota_{\mathscr{B}}"] \arrow[d] & \mathscr{B} \arrow[d] \\
\bigoplus_{\mathrm{u}\in\mathscr{B}}\mathsf{F} \arrow[r, "\iota_{\mathsf{V}}"'] & \mathsf{V}
\end{tikzcd}$$

where the vertical arrows represent the inclusion maps.

Proof First we define the map ι_{V}. Denote a typical element of $\bigoplus_{u\in\mathscr{B}}\mathsf{F}$ by

$$c_1\boldsymbol{e}_{u_1}+\cdots+c_k\boldsymbol{e}_{u_k}$$

for $c_1,\dots,c_k\in\mathsf{F}$ and distinct $u_1,\dots,u_k\in\mathscr{B}$. We define

$$\iota_{\mathsf{V}}(c_1\boldsymbol{e}_{u_1}+\cdots+c_k\boldsymbol{e}_{u_k})=c_1u_1+\cdots+c_ku_k.$$

It is then a simple matter to check that ι_{V} is a linear map. We also claim that it is an isomorphism. To see that it is injective suppose that

$$\iota_{\mathsf{V}}(c_1\boldsymbol{e}_{u_1}+\cdots+c_k\boldsymbol{e}_{u_k})=0_{\mathsf{V}}.$$

Then, by Proposition 4.5.36 and by the definition of ι_{V}, we have $c_1=\cdots=c_k=0_{\mathsf{F}}$. Thus the only vector mapping to zero is the zero vector, and this gives injectivity by Exercise 4.5.23. The proof of surjectivity is similarly straightforward. If $v\in\mathsf{V}$ then we can write $v=c_1u_1+\cdots+c_ku_k$ for some $c_1,\dots,c_k\in\mathsf{F}$ and $u_1,\dots,u_k\in\mathscr{B}$. Then the vector $c_1\boldsymbol{e}_{u_1}+\cdots+c_k\boldsymbol{e}_{u_k}\in\bigoplus_{u\in\mathscr{B}}\mathsf{F}$ maps to v under ι_{V}. The commutativity of the diagram in the theorem is checked directly. ■

4.5.46 Remark (Direct sums versus direct products) Note that the theorem immediately tells us that, when considering vector spaces, one can without loss of generality suppose that the vector space is a direct sum of copies of the field F. Thus direct sums are, actually, the most general form of vector space. Thinking along these lines, it becomes natural to wonder what is the value of considering direct products. First of all, Theorem 4.5.45 tells us that the direct product can be written as a direct sum, although not using the standard basis, cf. Example 4.5.28–4. The importance of the direct product will not become apparent until Section 5.7 when we discuss algebraic duals. •

Theorem 4.5.45 has the following corollary which tells us the relationship between the dimension of a vector space and its cardinality.

4.5.47 Corollary (The cardinality of a vector space) *If F is a field and if V is an F-vector space then*

(i) $\mathrm{card}(\mathsf{V})=\mathrm{card}(\mathsf{F})^{\dim_{\mathsf{F}}(\mathsf{V})}$ *if both* $\dim_{\mathsf{F}}(\mathsf{V})$ *and* $\mathrm{card}(\mathsf{F})$ *are finite and*

(ii) $\mathrm{card}(\mathsf{V})=\max\{\mathrm{card}(\mathsf{F}),\dim_{\mathsf{F}}(\mathsf{V})\}$ *if either* $\dim_{\mathsf{F}}(\mathsf{V})$ *or* $\mathrm{card}(\mathsf{F})$ *is infinite.*

Proof By Theorem 4.5.45, and since the dimension and cardinality of isomorphic vector spaces obviously agree (the former by Proposition 4.5.30), we can without loss of generality take the case when $\mathsf{V}=\bigoplus_{j\in J}\mathsf{F}$. We let $\{\boldsymbol{e}_j\}_{j\in J}$ be the standard basis. If J is finite then $\mathrm{card}(\mathsf{V})=\mathrm{card}(\mathsf{F})^{\mathrm{card}(J)}$ by definition of cardinal multiplication. If $\mathrm{card}(\mathsf{F})$ is finite then the result follows immediately. If $\mathrm{card}(\mathsf{F})$ is infinite then

$$\mathrm{card}(\mathsf{F})^{\mathrm{card}(J)}=\mathrm{card}(\mathsf{F})=\max\{\mathrm{card}(\mathsf{F}),\mathrm{card}(J)\}$$

by Theorem 1.7.17. This gives the result when $\dim_{\mathsf{F}}(\mathsf{V})$ is finite.

For the case when $\mathrm{card}(J)$ is infinite, we use the following lemma.

1 **Lemma** *If* F *is a field and if* V *is an infinite-dimensional* F*-vector space, then* $\mathrm{card}(\mathsf{V}) = \mathrm{card}(\mathsf{F}) \cdot \dim_{\mathsf{F}}(\mathsf{V})$*.*

Proof As in the proof of the theorem, we suppose that $\mathsf{V} = \bigoplus_{j \in J} \mathsf{F}$. We use the fact that every vector in V is a finite linear combination of standard basis vectors. Thus

$$\mathsf{V} = \{0_{\mathsf{V}}\} \cup \Big(\cup_{k \in \mathbb{Z}_{>0}} \{c_1 \boldsymbol{e}_{j_1} + \cdots + c_k \boldsymbol{e}_{j_k} \mid c_1, \dots, c_k \in \mathsf{F}^*,\ j_1, \dots, j_k \in J \text{ distinct}\}\Big). \tag{4.21}$$

Note that

$$\mathrm{card}(\{c_1 \boldsymbol{e}_{j_1} + \cdots + c_k \boldsymbol{e}_{j_k} \mid c_1, \dots, c_k \in \mathsf{F}^*,\ j_1, \dots, j_k \in J \text{ distinct}\}) \\ = ((\mathrm{card}(\mathsf{F}) - 1)\,\mathrm{card}(J))^k.$$

Thus, noting that the union in (4.21) is disjoint,

$$\mathrm{card}(\mathsf{V}) = \sum_{k=0}^{\infty} ((\mathrm{card}(\mathsf{F}) - 1)\,\mathrm{card}(J))^k.$$

By Theorem 1.7.17 we have

$$\mathrm{card}(\mathsf{V}) = \mathrm{card}(J) \sum_{k=0}^{\infty} (\mathrm{card}(\mathsf{F}) - 1).$$

If $\mathrm{card}(\mathsf{F})$ is finite then $\mathrm{card}(\mathsf{F}) \geq 2$ (since F contains a unit and a zero), and so, in this case, $\sum_{k=0}^{\infty}(\mathrm{card}(\mathsf{F}) - 1) = \mathrm{card}(\mathbb{Z}_{>0})$. If $\mathrm{card}(\mathsf{F})$ is infinite then $\sum_{k=0}^{\infty}(\mathrm{card}(\mathsf{F}) - 1) = \mathrm{card}(\mathsf{F})$ by Theorem 1.7.17. In either case we have $\mathrm{card}(\mathsf{V}) = \mathrm{card}(\mathsf{F}) \cdot \mathrm{card}(J)$. ▼

We now have two cases.

1. *J is infinite and F is finite:* In this case we have

$$\mathrm{card}(J) \cdot \mathrm{card}(\mathsf{F}) \leq \mathrm{card}(J) \cdot \mathrm{card}(J) = \mathrm{card}(J)$$

by Theorem 1.7.17, and we clearly have $\mathrm{card}(J) \cdot \mathrm{card}(\mathsf{F}) \geq \mathrm{card}(J)$. Thus $\mathrm{card}(J) \cdot \mathrm{card}(\mathsf{F}) = \mathrm{card}(J)$.

2. *J and F are both infinite:* In this case, by Theorem 1.7.17, we have

$$\mathrm{card}(J) \cdot \mathrm{card}(\mathsf{F}) = \max\{\mathrm{card}(J), \mathrm{card}(\mathsf{F})\},$$

and the result follows. ■

We also have the following corollary to Theorem 4.5.45, along with Proposition 4.5.30, which gives an essential classification of vector spaces.

4.5.48 Corollary (Characterisation of isomorphic vector spaces) *If* F *is a field,* F*-vector spaces* V_1 *and* V_2 *are* F*-isomorphic if and only if* $\dim_{\mathsf{F}}(\mathsf{V}_1) = \dim_{\mathsf{F}}(\mathsf{V}_2)$*.*

Let us make Theorem 4.5.45 concrete in a simple case, just to bring things down to earth for a moment. The reader should try to draw the parallels between the relatively simple example and the more abstract proof of Theorem 4.5.45.

4.5.49 Example (Direct sum representations of finite-dimensional vector spaces) Let V be an n-dimensional vector space. By Theorem 4.5.45 we know that V is isomorphic to F^n. Moreover, the theorem explicitly indicates how an isomorphism is assigned by a basis. Thus let $\{e_1, \dots, e_n\}$ be a basis for V and let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ be the standard basis for F^n. Then we define the map

$$\iota_{\mathscr{B}} \colon \{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\} \to \{e_1, \dots, e_n\}$$

by $\iota_{\mathscr{B}}(\boldsymbol{e}_j) = e_j$, $j \in \{1, \dots, n\}$. The associated isomorphism $\iota_{\mathsf{V}} \colon \mathsf{F}^n \to \mathsf{V}$ is then given by

$$\iota_{\mathsf{V}}(v_1, \dots, v_n) = v_1 e_1 + \cdots + v_n e_n.$$

The idea is simply that linear combinations of the standard basis are mapped to linear combinations of the basis for V with the coefficients preserved. •

Let us conclude our discussions in this section by understanding the relationship between direct sums and dimension. Note that, given Proposition 4.5.41, the result applies to both internal direct sums and direct sums, although it is only stated for internal direct sums.

4.5.50 Proposition (Dimension and direct sum) *Let F be a field, let V be an F-vector space, let $(\mathsf{U}_\mathsf{j})_{\mathsf{j}\in\mathsf{J}}$ be a family of F-vector spaces such that $\mathsf{V} = \bigoplus_{\mathsf{j}\in\mathsf{J}} \mathsf{U}_\mathsf{j}$, and let $(\mathscr{B}_\mathsf{j})_{\mathsf{j}\in\mathsf{J}}$ be such that $\mathscr{B}_\mathsf{j}$ is a basis for U_j. Then $\cup_{\mathsf{j}\in\mathsf{J}}\mathscr{B}_\mathsf{j}$ is a basis for V. In particular,*

$$\dim_\mathsf{F}(\mathsf{V}) = \dim_\mathsf{F}(\mathsf{U}_1) + \cdots + \dim_\mathsf{F}(\mathsf{U}_\mathsf{k}).$$

Proof Let $v \in \mathsf{V}$. Then there exists unique $j_1, \dots, j_k \in J$ and nonzero $u_{j_l} \in \mathsf{U}_{j_l}$, $j \in \{1, \dots, k\}$, such that $v = u_{j_1} + \cdots + u_{j_k}$. For each $l \in \{1, \dots, k\}$ there exists unique $c_1^l, \dots, c_k^l \in \mathsf{F}^*$ and unique $u_1^l, \dots, u_{k_l}^l \in \mathscr{B}_{j_l}$ such that

$$u_{j_l} = c_1^l u_1^l + \cdots + c_{k_l}^l u_{k_l}^l.$$

Then we have

$$v = \sum_{l=1}^{k} \sum_{r=1}^{k_l} c_r^l u_r^l$$

as a representation of v as a finite linear combination of elements of $\cup_{j\in J}\mathscr{B}_j$ with nonzero coefficients. Moreover, this is the unique such representation since, at each step in the construction, the representations were unique. ■

4.5.6 Complements and quotients

We next consider another means of construction vector spaces from subspaces. We first address the question of when, given a subspace, there exists another subspace which gives a direct sum representation of V.

4.5.51 Definition (Complement of a subspace) If F is a field, if V is an F-vector space, and if U is a subspace of V, a ***complement*** of U in V is a subspace W of V such that $\mathsf{V} = \mathsf{U} \oplus \mathsf{W}$. •

Complements of subspaces always exist.

4.5.52 Theorem (Subspaces possess complements) *If F is a field, if V is an F-vector space, and if U is a subspace of V, then there exists a complement of U.*

Proof Let $\mathscr{B}'$ be a basis for U. By Theorem 4.5.26 there exists a basis $\mathscr{B}$ for V such that $\mathscr{B}' \subseteq \mathscr{B}$. Let $\mathscr{B}'' = \mathscr{B} \setminus \mathscr{B}'$ and define $\mathsf{W} = \operatorname{span}_{\mathsf{F}}(\mathscr{B}'')$. We claim that W is a complement of U in V. First let $v \in \mathsf{V}$. Then, since $\mathscr{B}$ is a basis for V, there exists $c'_1, \dots, c'_{k'}, c''_1, \dots, c''_{k''} \in \mathsf{F}$, $u'_1, \dots, u'_{k'} \in \mathscr{B}'$, and $u''_1, \dots, u''_{k''} \in \mathscr{B}''$ such that

$$v = \underbrace{c'_1 u'_1 + \dots + c'_{k'} u'_{k'}}_{\in \mathsf{U}} + \underbrace{c''_1 u''_1 + \dots + c''_{k''} u''_{k''}}_{\in \mathsf{W}} .$$

Thus $\mathsf{V} = \mathsf{U} + \mathsf{W}$. Next let $v \in \mathsf{U} \cap \mathsf{W}$. If $v \neq 0_{algV}$ then there exists unique $u'_1, \dots, u'_{k'} \in \mathscr{B}'$ and $u''_1, \dots, u''_{k''} \in \mathscr{B}''$ and unique $c'_1, \dots, c'_{k'}, c''_1, \dots, c''_{k''} \in \mathsf{F}$ such that

$$v = c'_1 u'_1 + \dots + c'_{k'} u'_{k'} = c''_1 u''_1 + \dots + c''_{k''} u''_{k''}.$$

This, however, contradicts the uniqueness of the representation of v as a finite linear combination of elements of $\mathscr{B}$ with nonzero coefficients. Thus $v = 0_{\mathsf{V}}$. Therefore, $\mathsf{V} = \mathsf{U} \oplus \mathsf{W}$ by Proposition 4.5.37. ■

For the same reason that a vector space possesses multiple bases, it is also the case that a strict subspace i.e., one not equal to the entire vector space, will generally possess multiple complements. Thus, while complements exist, there is not normally a natural such choice, except in the presence of additional structure (the most common such structure being an inner product, something not discussed until Chapter III-4). However, there is a unique way in which one can associate a new vector space to a subspace in such a way that this new vector space has some properties of a complement.

4.5.53 Definition (Quotient by a subspace) Let F be a field, let V be an F-vector space, and let U be a subspace of V. The ***quotient*** of V by U is the set of equivalence classes in V under the equivalence relation

$$v_1 \sim v_2 \qquad \Longleftrightarrow \qquad v_1 - v_2 \in \mathsf{U}.$$

We denote by V/U the quotient of V by U, and we denote by $\pi_{\mathsf{V}/\mathsf{U}} \colon \mathsf{V} \to \mathsf{V}/\mathsf{U}$ the map, called the ***canonical projection***, assigning to $v \in \mathsf{V}$ its equivalence class. •

Thinking of V as an Abelian group with product defined by vector addition, the quotient V/U is simply the set of cosets of the subgroup U; see Definition 4.1.16. We shall adapt the notation for groups to denote a typical element in V/U by

$$v + \mathsf{U} = \{v + u \mid u \in \mathsf{U}\}.$$

Since V is Abelian, by Proposition 4.1.20 it follows that V/U possesses a natural Abelian group structure. It also possesses a natural vector space structure, as the following result indicates.

4.5.54 Proposition (The quotient by a subspace is a vector space) *Let F be a field, let V be an F-vector space, and let U be a subspace of V. The operations of vector addition and scalar multiplication in V/U defined by*

$$(\mathrm{v}_1 + \mathsf{U}) + (\mathrm{v}_2 + \mathsf{U}) = (\mathrm{v}_1 + \mathrm{v}_2) + \mathsf{U}, \quad \mathrm{a}(\mathrm{v} + \mathsf{U}) = (\mathrm{av}) + \mathsf{U}, \qquad \mathrm{v}, \mathrm{v}_1, \mathrm{v}_2 \in \mathsf{V},\ \mathrm{a} \in \mathsf{F},$$

respectively, satisfy the axioms for an F-vector space.

Proof We define the zero vector in V/U by $0_{\mathsf{V}/\mathsf{U}} = 0_{\mathsf{V}} + \mathsf{U}$ and we define the negative of a vector $v + \mathsf{U}$ by $(-v) + \mathsf{U}$. It is then a straightforward matter to check the axioms of Definition 4.5.1, a matter which we leave to the interested reader. ■

The following "universal" property of quotients is useful.

4.5.55 Proposition (A "universal" property of quotient spaces) *Let F be a field, let V be an F-vector space, and let U be a subspace of V. If W is another F-vector space and if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}; \mathsf{W})$ has the property that $\ker(\mathsf{L}) \subseteq \mathsf{U}$, then there exists $\overline{\mathsf{L}} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}/\mathsf{U}; \mathsf{W})$ such that the diagram*

$$\begin{tikzcd}
\mathsf{V} \arrow[r, "\mathsf{L}"] \arrow[d, "\pi_{\mathsf{V}/\mathsf{U}}"'] & \mathsf{W} \\
\mathsf{V}/\mathsf{U} \arrow[ur, "\overline{\mathsf{L}}"'] &
\end{tikzcd}$$

commutes. Moreover, if $\overline{\mathsf{L}}' \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}/\mathsf{U}; \mathsf{W})$ is such that the preceding diagram commutes, then $\overline{\mathsf{L}}' = \overline{\mathsf{L}}$.

Proof We define $\overline{\mathsf{L}}(v + \mathsf{U}) = \mathsf{L}(v)$. This map is well-defined since, if $v' + \mathsf{U} = v + \mathsf{U}$ then $v' = v + u$ for $u \in \mathsf{U}$, whence

$$\overline{\mathsf{L}}(v' + \mathsf{U}) = \mathsf{L}(v') = \mathsf{L}(v + u) = \mathsf{L}(v) = \overline{\mathsf{L}}(v + \mathsf{U}).$$

One verifies directly that

$$\overline{\mathsf{L}}((v_1 + \mathsf{U}) + (v_2 + \mathsf{U})) = \overline{\mathsf{L}}(v_1 + \mathsf{U}) + \overline{\mathsf{L}}(v_2 + \mathsf{U}), \qquad \overline{\mathsf{L}}(a(v + \mathsf{U})) = a\overline{\mathsf{L}}(v + \mathsf{U}),$$

giving linearity of $\overline{\mathsf{L}}$. For the final assertion of the proposition, the commuting of the diagram exactly says that $\overline{\mathsf{L}}'(v + \mathsf{U}) = \mathsf{L}(v)$, as desired. ■

Next we consider the relationship between complements and quotient spaces.

4.5.56 Theorem (Relationship between complements and quotients) *Let F be a field, let V be an F-vector space, and let U be a subspace of V with a complement W. Then the map $\iota_{\mathsf{U},\mathsf{W}} \colon \mathsf{W} \to \mathsf{V}/\mathsf{U}$ defined by*

$$\iota_{\mathsf{U},\mathsf{W}}(\mathrm{w}) = \mathrm{w} + \mathsf{U}$$

is an isomorphism. In particular, $\dim_{\mathsf{F}}(\mathsf{W}) = \dim_{\mathsf{F}}(\mathsf{V}/\mathsf{U})$ for any complement W of U in V.

Proof The map $\iota_{\mathsf{U},\mathsf{W}}$ is readily checked to be linear, and we leave this verification to the reader. Suppose that $w + \mathsf{U} = 0_{\mathsf{V}} + \mathsf{U}$ for $w \in \mathsf{W}$. This implies that $w \in \mathsf{U}$, which gives $w = 0_{\mathsf{V}}$ by Proposition 4.5.37; thus $\iota_{\mathsf{U},\mathsf{W}}$ is injective by Exercise 4.5.23. Now let $v + \mathsf{U} \in \mathsf{V}/\mathsf{U}$. Since $\mathsf{V} = \mathsf{U} \oplus \mathsf{W}$ we can write $v = u + w$ for $u \in \mathsf{U}$ and $w \in \mathsf{W}$. Since $v - w \in \mathsf{U}$ we have $v + \mathsf{U} = w + \mathsf{U}$. Thus $\iota_{\mathsf{U},\mathsf{W}}$ is also surjective.

The final assertion follows from Propositions 4.5.30 and 4.5.50. ■

The preceding result gives the dimension of the quotient, and the next result reinforces this by giving an explicit basis for the quotient.

4.5.57 Proposition (Basis for quotient) *Let* F *be a field, let* V *be an* F*-vector space, and let* U *be a subspace of* V*. If* $\mathscr{B}$ *is a basis for* V *with the property that there exists a subset* $\mathscr{B}' \subseteq \mathscr{B}$ *with the property that* $\mathscr{B}'$ *is a basis for* U*, then*

$$\{\mathrm{v} + \mathsf{U} \mid \mathrm{v} \in \mathscr{B} \setminus \mathscr{B}'\}$$

is a basis for V/U.

Proof Let $\mathscr{B}''$ be such that $\mathscr{B} = \mathscr{B}' \cup \mathscr{B}''$ and $\mathscr{B}' \cup \mathscr{B}'' = \emptyset$. If $v \in \mathsf{V}$ then we can write

$$v = c_1 u_1 + \cdots + c_k u_k + d_1 v_1 + \cdots + d_l v_l$$

for $c_1, \dots, c_k, d_1, \dots, d_l \in \mathsf{F}$, for $u_1, \dots, u_k \in \mathscr{B}'$, and for $v_1, \dots, v_l \in \mathscr{B}''$. Then

$$\begin{aligned} v + \mathsf{U} &= (c_1 u_1 + \cdots + c_k u_k + d_1 v_1 + \cdots + d_l v_l) + \mathsf{U} \\ &= (d_1 v_1 + \cdots + d_l v_l) + \mathsf{U} = (d_1 v_1 + \mathsf{U}) + \cdots + (d_l v_l + \mathsf{U}), \end{aligned}$$

showing that $\{v + \mathsf{U} \mid v \in \mathscr{B}''\}$ generates V/U. To show linear independence, suppose that

$$(d_1 v_1 + \mathsf{U}) + \cdots + (d_l v_l + \mathsf{U}) = 0_{\mathsf{V}} + \mathsf{U}$$

for $v_1, \dots, v_l \in \mathscr{B}''$ and $d_1, \dots, d_l \in \mathsf{F}$. Then $d_1 v_1 + \cdots + d_l v_l \in \mathsf{U}$, and so $d_1 v_1 + \cdots + d_l v_l = 0_{\mathsf{V}}$ by Proposition 4.5.37. Since $\mathscr{B}''$ is linearly independent by Proposition 4.5.19(iii), it follows that $d_1 = \cdots = d_l = 0_{\mathsf{F}}$, and so $\{v + \mathsf{U} \mid v \in \mathscr{B}''\}$ is linearly independent. ■

The preceding theorem motivates the following definition.

4.5.58 Definition (Codimension of a subspace) Let F be a field, let V be an F-vector space, and let U be a subspace of V. The ***codimension*** of U, denoted by $\mathrm{codim}_{\mathsf{F}}(\mathsf{U})$, is $\dim_{\mathsf{F}}(\mathsf{V}/\mathsf{U})$. •

Combining Proposition 4.5.50 and Theorem 4.5.56 immediately gives the following result.

4.5.59 Corollary (Dimension and codimension of a subspace) *If* F *is a field, if* V *is an* F*-vector space, and if* U *is a subspace of* V*, then* $\dim_{\mathsf{F}}(\mathsf{V}) = \dim_{\mathsf{F}}(\mathsf{U}) + \mathrm{codim}_{\mathsf{F}}(\mathsf{U})$.

4.5.7 Complexification of $\mathbb{R}$-vector spaces

It will often be useful to regard a vector space defined over $\mathbb{R}$ as being defined over $\mathbb{C}$. This is fairly straightforward to do.

4.5.60 Definition (Complexification of a $\mathbb{R}$-vector space) If V is a $\mathbb{R}$-vector space, the ***complexification*** of V is the $\mathbb{C}$-vector space $\mathsf{V}_{\mathbb{C}}$ defined by

(i) $\mathsf{V}_{\mathbb{C}} = \mathsf{V} \times \mathsf{V}$,

and with the operations of vector addition and scalar multiplication defined by

(ii) $(u_1, u_2) + (v_1, v_2) = (u_1 + v_1, u_2 + v_2)$, $u_1, u_2, v_1, v_2 \in \mathsf{V}$, and

(iii) $(a + ib)(u, v) = (au - bv, av + bu)$ for $a, b \in \mathbb{R}$ and $u, v \in \mathsf{V}$. •

We recall from Example 4.5.2–5 that any $\mathbb{C}$-vector space is also a $\mathbb{R}$-vector space by simply restricting scalar multiplication to $\mathbb{R}$. It will be convenient to regard V as a subspace of the $\mathbb{R}$-vector space $\mathsf{V}_{\mathbb{C}}$. There are many ways one might do this. For example, we can identify V with the either of the two subspaces

$$\{(u, v) \in \mathsf{V}_{\mathbb{C}} \mid v = 0_{\mathsf{V}}\}, \quad \{(u, v) \in \mathsf{V}_{\mathbb{C}} \mid u = 0_{\mathsf{V}}\},$$

and there are many other possible choices. However, the subspace on the left is the most natural one for reasons that will be clear shortly. We thus define the monomorphism $\iota_{\mathsf{V}} \colon \mathsf{V} \to \mathsf{V}_{\mathbb{C}}$ of $\mathbb{R}$-vector spaces by $\iota(v) = (v, 0_{\mathsf{V}})$, and we note that $\operatorname{image}(\iota_{\mathsf{V}})$ is a subspace of $\mathsf{V}_{\mathbb{C}}$ that is isomorphic to V.

The following result records that $\mathsf{V}_{\mathbb{C}}$ has the desired properties.

4.5.61 Proposition (Properties of complexification) *If V is a $\mathbb{R}$-vector space then the complexification $\mathsf{V}_{\mathbb{C}}$ has the following properties:*

(i) $\mathsf{V}_{\mathbb{C}}$ is a $\mathbb{C}$-vector space and $\dim_{\mathbb{C}}(\mathsf{V}_{\mathbb{C}}) = \dim_{\mathbb{R}}(\mathsf{V})$;

(ii) $\mathsf{V}_{\mathbb{C}}$ is a $\mathbb{R}$-vector space and $\dim_{\mathbb{R}}(\mathsf{V}_{\mathbb{C}}) = 2\dim_{\mathbb{R}}(\mathsf{V})$;

(iii) every element of $\mathsf{V}_{\mathbb{C}}$ can be uniquely expressed as $\iota_{\mathsf{V}}(\mathrm{u}) + \mathrm{i}\,\iota_{\mathsf{V}}(\mathrm{v})$ for some $\mathrm{u}, \mathrm{v} \in \mathsf{V}$.

Proof (i) The verification of the axioms for $\mathsf{V}_{\mathbb{C}}$ to be a $\mathbb{C}$-vector space is straightforward and relatively unilluminating, so we leave the reader to fill in the details. Let us verify that $\dim_{\mathbb{C}}(\mathsf{V}) = \dim_{\mathbb{R}}(\mathsf{V})$. Let $\mathscr{B}$ be a basis for V and define

$$\mathscr{B}_{\mathbb{C}} = \{(u, 0_{\mathsf{V}}) \mid u \in \mathscr{B}\}.$$

We claim that $\mathscr{B}_{\mathbb{C}}$ is a basis for $\mathsf{V}_{\mathbb{C}}$ as a $\mathbb{C}$-vector space. To show linear independence of $\mathscr{B}_{\mathbb{C}}$, suppose that

$$(a_1 + ib_1)(u_1, 0_{\mathsf{V}}) + \cdots + (a_k + ib_k)(u_k, 0_{\mathsf{V}}) = (0_{\mathsf{V}}, 0_{\mathsf{V}})$$

for $a_1, \dots, a_k, b_1, \dots, b_k \in \mathbb{R}$. Using the definition of scalar multiplication this implies that

$$(a_1 u_1, b_1 u_1) + \cdots + (a_k u_k, b_k u_k) = (0_{\mathsf{V}}, 0_{\mathsf{V}}).$$

Linear independence of $\mathscr{B}$ then implies that $a_j = b_j = 0$ for $j \in \{1, \dots, k\}$, so giving linear independence of $\mathscr{B}_{\mathbb{C}}$. Now let $(u, v) \in \mathsf{V}_{\mathbb{C}}$. There then exists $u_1, \dots, u_k \in \mathscr{B}$ and $a_1, \dots, a_k, b_1, \dots, b_k \in \mathbb{R}$ such that

$$u = a_1 u_1 + \cdots + a_k u_k, \quad v = b_1 u_1 + \cdots + b_k u_k.$$

We then have

$$(u,v) = (a_1u_1 + \cdots + a_ku_k, b_1u_1 + \cdots + b_ku_k) = (a_1u_1, b_1u_1) + \cdots + (a_ku_k, b_ku_k).$$

Using the rules for scalar multiplication in $\mathsf{V}_{\mathbb{C}}$ this gives

$$(u,v) = (a_1 + \mathrm{i}b_1)(u_1, 0_{\mathsf{V}}) + \cdots + (a_k + \mathrm{i}b_k)(u_k, 0_{\mathsf{V}}).$$

Thus $\mathscr{B}_{\mathbb{C}}$ spans $\mathsf{V}_{\mathbb{C}}$, and so is a basis for $\mathsf{V}_{\mathbb{C}}$.

(ii) That $\mathsf{V}_{\mathbb{C}}$ is a $\mathbb{R}$-vector space follows from Example 4.5.2–5. Note that scalar multiplication in the $\mathbb{R}$-vector space $\mathsf{V}_{\mathbb{C}}$, i.e., restriction of $\mathbb{C}$ scalar multiplication to $\mathbb{R}$, is defined by $a(u,v) = (au, av)$. Thus $\mathsf{V}_{\mathbb{C}}$ as a $\mathbb{R}$-vector space is none other than $\mathsf{V} \oplus \mathsf{V}$. That $\dim_{\mathbb{R}}(\mathsf{V}_{\mathbb{C}}) = 2\dim_{\mathbb{R}}(\mathsf{V})$ then follows from Proposition 4.5.50.

(iii) Using the definition of $\mathbb{C}$ scalar multiplication we have

$$\mathrm{i}\iota_{\mathsf{V}}(v) = \mathrm{i}(v, 0_{\mathsf{V}}) = (0_{\mathsf{V}}, v).$$

Thus we clearly have

$$(u,v) = \iota_{\mathsf{V}}(u) + \mathrm{i}\,\iota_{\mathsf{V}}(v),$$

giving the existence of the stated representation. Now, if

$$\iota_{\mathsf{V}}(u_1) + \mathrm{i}\,\iota_{\mathsf{V}}(v_1) = \iota_{\mathsf{V}}(u_2) + \mathrm{i}\,\iota_{\mathsf{V}}(v_2),$$

then $(u_1, v_1) = (u_2, v_2)$, and so $u_1 = u_2$ and $v_1 = v_2$, giving uniqueness of the representation. ■

The final assertion in the proposition says that we can think of $(u,v) \in \mathsf{V}_{\mathbb{C}}$ as $(u, 0_{\mathsf{V}}) + \mathrm{i}(v, 0_{\mathsf{V}})$. With this as motivation, we shall use the notation $(u,v) = u + \mathrm{i}v$ when it is convenient. This then leads to the following definitions which adapt those for complex numbers to the complexification of a $\mathbb{R}$-vector space.

4.5.62 Definition (Real part, imaginary part, complex conjugation) Let V be a $\mathbb{R}$-vector space with $\mathsf{V}_{\mathbb{C}}$ its complexification.

(i) The ***real part*** of $(u,v) \in \mathsf{V}_{\mathbb{C}}$ is $\mathrm{Re}(u,v) = u$.

(ii) The ***imaginary part*** of $(u,v) \in \mathsf{V}_{\mathbb{C}}$ is $\mathrm{Im}(u,v) = v$.

(iii) The representation $u + \mathrm{i}v$ of $(u,v) \in \mathsf{V}_{\mathbb{C}}$ is the ***canonical representation***.

(iv) ***Complex conjugation*** is the map $\sigma_{\mathsf{V}} \colon \mathsf{V}_{\mathbb{C}} \to \mathsf{V}_{\mathbb{C}}$ defined by $\sigma_{\mathsf{V}}(u,v) = (u, -v)$. •

Using the canonical representation of elements in the complexification, $\mathbb{C}$-scalar multiplication in $\mathsf{V}_{\mathbb{C}}$ can be thought of as applying the usual rules for $\mathbb{C}$ multiplication to the expression $(a + \mathrm{i}b)(u + \mathrm{i}v)$:

$$(a + \mathrm{i}b)(u + \mathrm{i}v) = (au - bv) + \mathrm{i}(bu + av).$$

This is a helpful mnemonic for remembering the scalar multiplication rule for $\mathsf{V}_{\mathbb{C}}$.

It is easy to show that $\sigma_{\mathsf{V}} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V}\mathbb{C})$, but that $\sigma_{\mathsf{V}} \notin \mathrm{End}_{\mathbb{C}}(\mathsf{V}_{\mathbb{C}})$ (see Exercise 4.5.25). Moreover, complex conjugation has the following easily verified properties.

The following example should be thought of, at least in the finite-dimensional case, as the typical one.

4.5.63 Example ($\mathbb{R}^n_{\mathbb{C}} = \mathbb{C}^n$) We take the $\mathbb{R}$-vector space $\mathbb{R}^n$ and consider its complexification $\mathbb{R}^n_{\mathbb{C}}$. The main point to be made here is the following lemma.

1 Lemma *The map* $(x_1, \dots, x_n) + i(y_1, \dots, y_n) \mapsto (x_1 + iy_1, \dots, x_n + iy_n)$ *is a* $\mathbb{C}$*-isomorphism of* $\mathbb{R}^n_{\mathbb{C}}$ *with* $\mathbb{C}^n$*.*

Proof This follows by the definition of vector addition and $\mathbb{C}$-scalar multiplication in $\mathbb{R}^n_{\mathbb{C}}$. ▼

Let us look at some of the constructions associated with complexification in order to better understand them. First note that $\mathbb{R}^n_{\mathbb{C}}$ has the structure of both a $\mathbb{R}$- and $\mathbb{C}$-vector space. One can check that a basis for $\mathbb{R}^n_{\mathbb{C}}$ as a $\mathbb{R}$-vector space is given by the set

$$\{e_1 + i\mathbf{0}, \dots, e_n + i\mathbf{0}, \mathbf{0} + ie_1, \dots, \mathbf{0} + ie_n\},$$

and a basis for $\mathbb{R}^n_{\mathbb{C}}$ as a $\mathbb{C}$-vector space is given by the set

$$\{e_1 + i\mathbf{0}, \dots, e_n + i\mathbf{0}\},$$

where $\{e_1, \dots, e_n\}$ is the standard basis for $\mathbb{R}^n$. It is also clear that

$$\mathrm{Re}(x + iy) = x, \quad \mathrm{Im}(x + iy) = y, \quad \sigma_{\mathbb{R}^n}(x + iy) = x - iy.$$

The idea in this example is, essentially, that one can regard the complexification of $\mathbb{R}^n$ as the vector space obtained by "replacing" the real entries in a vector with complex entries. •

4.5.8 Extending the scalars for a vector space

In Section 4.5.7 we saw how one can naturally regard a $\mathbb{R}$-vector space as a $\mathbb{C}$-vector space. In this section we generalise this idea to general field extensions, as it will be useful in studying endomorphisms of finite-dimensional vector spaces in Section 5.8. This development relies on the tensor product which itself is a part of multilinear algebra. Thus a reader will need to make a diversion ahead to Section 5.6 in order to understand the material in this section.

While we have not yet discussed field extensions (we do so formally and in detail in Section 4.6), the notion is a simple one. A field K that contains a field F as a subfield is an ***extension*** of F. As we will show in Proposition 4.6.2, and is easily seen in any case, K is an F-vector space. We shall make essential use of this fact in this section. Indeed, the key idea in complexification comes from understanding the $\mathbb{R}$-vector space structure of $\mathbb{C}$. Here we generalise this idea.

We may now define the extension of an F-vector space to an extension K of F. This definition will seem odd at first glance, relying as it does on the tensor product. It is only after we explore it a little that it will (hopefully) seem "correct."

4.5.64 Definition (Extension of scalars for a vector space) Let F be a field, let K be an extension of F, and let V be an F-vector space. The ***extension*** of V to K is

$$\mathsf{V}_{\mathsf{K}} = \mathsf{K} \otimes \mathsf{V}. \qquad \bullet$$

At this point, we certainly understand all the symbols in the definition. However, it is not so clear what V_{K} really is. To begin to understand it, let us first show that it has the structure of a vector space over K; it is this structure that is of most interest to us.

4.5.65 Proposition (V_{K} is an K-vector space) *Let K be an extension of a field F and let V be an F-vector space. Using vector addition and scalar multiplication defined by vector addition in $\mathsf{K} \otimes \mathsf{V}$ (as an F-vector space) and $\mathrm{b}(\mathrm{a} \otimes \mathrm{v}) = (\mathrm{ab}) \otimes \mathrm{v}$, $\mathrm{a}, \mathrm{b} \in \mathsf{K}$, $\mathrm{v} \in \mathsf{V}$, respectively, $\mathsf{K} \otimes \mathsf{V}$ is a vector space over K.*

Proof First let us show that the definition of scalar multiplication in K is well-defined. We note that for $b \in \mathsf{K}$ the map $\phi_b \colon \mathsf{K} \times \mathsf{V} \to \mathsf{K} \otimes \mathsf{V}$ defined by $\phi_b(a, v) = (ba) \otimes v$ is bilinear. Thus there exists a unique linear map $\mathsf{L}_{\phi_b} \colon \mathsf{K} \otimes \mathsf{V} \to \mathsf{K} \otimes \mathsf{V}$ satisfying $\mathsf{L}_{\phi_b}(a \otimes v) = (ba) \otimes v$. Now, if

$$a_1 \otimes v_1 + \cdots + a_k \otimes v_k$$

is an arbitrary element of $\mathsf{K} \otimes \mathsf{V}$, it follows that

$$\mathsf{L}_{\phi_b}(a_1 \otimes v_1 + \cdots + a_k \otimes v_k) = (ba_1) \otimes v_1 + \cdots + (ba_k) \otimes v_k$$

since L_{ϕ_b} is linear. Thus scalar multiplication is well-defined on all of $\mathsf{K} \otimes \mathsf{V}$. To show that vector addition and scalar multiplication satisfy the usual axioms for a vector space is now straightforward, and we leave the details of this to the reader. ■

Let us show that this complicated notion of scalar extension agrees with complexification.

4.5.66 Example ($\mathsf{V}_{\mathbb{C}} = \mathbb{C} \otimes \mathsf{V}$) We let V be a $\mathbb{R}$-vector space with complexification $\mathsf{V}_{\mathbb{C}}$. Let us show that "$\mathsf{V}_{\mathbb{C}} = \mathsf{V}_{\mathbb{C}}$;" i.e., that complexification as in Section 4.5.7 agrees with extension of scalars as in Definition 4.5.64. To see this we define an isomorphism $\iota_{\mathbb{C}}$ from $\mathsf{V}_{\mathbb{C}}$ (the complexification as in Section 4.5.7) to $\mathbb{C} \otimes \mathsf{V}$ by

$$\iota_{\mathbb{C}}(u, v) = 1 \otimes u + \mathrm{i} \otimes v.$$

Let us show that this is an isomorphism of $\mathbb{C}$-vector spaces. First we note that

$$\begin{aligned} \iota_{\mathbb{C}}((u_1, v_1) + (u_2, v_2)) &= \iota_{\mathbb{C}}(u_1 + u_2, v_1 + v_2) = 1 \otimes (u_1 + u_2) + \mathrm{i}(v_1 + v_2) \\ &= (1 \otimes u_1 + \mathrm{i}v_1) + (1 \otimes u_2 + \mathrm{i} \otimes v_2) = \iota_{\mathbb{C}}(u_1, v_1) + \iota_{\mathbb{C}}(u_2, v_2) \end{aligned}$$

and

$$\begin{aligned} \iota_{\mathbb{C}}((a + \mathrm{i}b)(u, v)) &= \iota_{\mathbb{C}}(au - bv, av + bu) = 1 \otimes (au - bv) + \mathrm{i}(av + bu) \\ &= 1 \otimes (au) + 1 \otimes (-bv) + \mathrm{i} \otimes (av) + \mathrm{i} \otimes (bu) \\ &= a \otimes u + (-b) \otimes v + (\mathrm{i}a) \otimes v + (\mathrm{i}b) \otimes u \\ &= a(1 \otimes u + \mathrm{i} \otimes v) + \mathrm{i}b(1 \otimes u + \mathrm{i} \otimes v) \\ &= (a + \mathrm{i}b)(1 \otimes u + \mathrm{i} \otimes v) = (a + \mathrm{i}b)\iota_{\mathbb{C}}(u, v), \end{aligned}$$

so showing that $\iota_{\mathbb{C}}$ is a $\mathbb{C}$-linear. To show that $\iota_{\mathbb{C}}$ is injective, suppose that $\iota_{\mathbb{C}}(u, v) = 0_{\mathbb{C}\otimes\mathsf{V}}$. Thus

$$1 \otimes u + \mathrm{i} \otimes v = 1 \otimes 0_{\mathsf{V}} + \mathrm{i} \otimes 0_{\mathsf{V}},$$

and so $u = v = 0_{\mathsf{V}}$. Thus $\iota_{\mathbb{C}}$ is injective by Exercise 4.5.23. To show that $\iota_{\mathbb{C}}$ is surjective, it suffices (why?) to show that $(a + ib) \otimes v \in \mathrm{image}(\iota_{\mathbb{C}})$ for each $a, b \in \mathbb{R}$ and $v \in \mathsf{V}$. This follows since

$$\iota_{\mathbb{C}}(av, bv) = 1 \otimes (av) + \mathrm{i} \otimes (bv) = a \otimes v + (\mathrm{i}b) \otimes v = (a + \mathrm{i}b) \otimes v.$$

Note that $1 \otimes u + \mathrm{i} \otimes v$ is the corresponding decomposition of $(u, v) \in \mathsf{V}_{\mathbb{C}}$ into its real and imaginary parts. If one keeps this in mind, and uses the usual rules for manipulating tensor products, it is easy to see why $\mathbb{C} \otimes \mathsf{V}$ is, indeed, the complexification of V. •

4.5.9 Notes

Exercises

4.5.1 Verify the vector space axioms for Example 4.5.2–1.

4.5.2 Verify the vector space axioms for Example 4.5.2–2.

4.5.3 Verify the vector space axioms for Example 4.5.2–3.

4.5.4 Verify the vector space axioms for Example 4.5.2–4.

4.5.5 Verify the vector space axioms for Example 4.5.2–5.

4.5.6 Verify the vector space axioms for Example 4.5.2–6.

4.5.7 Verify the vector space axioms for Example 4.5.2–7.

4.5.8 Verify the vector space axioms for Example 4.5.2–8.

4.5.9 Verify the vector space axioms for Example 4.5.2–9.

4.5.10 Let $I \subseteq \mathbb{R}$, let $r \in \mathbb{Z}_{>0}$, and denote by $\mathsf{C}^r(I; \mathbb{R})$ the set of $\mathbb{R}$-valued functions on I that are r-times continuously differentiable. Define vector addition and scalar multiplication in such a way that $\mathsf{C}^r(I; \mathbb{R})$ is a $\mathbb{R}$-vector space.

4.5.11 Prove Proposition 4.5.6.

4.5.12 Verify the claim of Example 4.5.7–1.

4.5.13 Verify the claim of Example 4.5.7–2.

4.5.14 Verify the claim of Example 4.5.7–3.

4.5.15 Verify the claim of Example 4.5.7–4.

4.5.16 Prove Proposition 4.5.9.

4.5.17 Do the following.

(a) Give an example of a vector space V and two subspaces U_1 and U_2 of V such that $\mathsf{U}_1 \cup \mathsf{U}_2$ is not a subspace.

(b) If V is an F-vector space and if $\mathsf{U}_1, \dots, \mathsf{U}_k$ are subspaces of V, show that $\cup_{j=1}^k \mathsf{U}_j$ is a subspace if and only if there exists $j_0 \in \{1, \dots, k\}$ such that $\mathsf{U}_j \subseteq \mathsf{U}_{j_0}$ for $j \in \{1, \dots, k\}$.

(c) If V is an F-vector space and if $(\mathsf{U}_j)_{j \in J}$ is an arbitrary family of subspaces, give conditions, analogous to those of part (b), that ensure that $\cup_{j \in J} \mathsf{U}_j$ is a subspace.

4.5.18 Prove Theorem 4.5.26 in the case when $\dim_{\mathsf{F}}(\mathsf{V}) < \infty$.

4.5.19 Let F be a field, let $\mathsf{V}$and W be F-vector spaces, let $\mathscr{B} \subseteq \mathsf{V}$ be a basis, let $\phi \colon \mathscr{B} \to \mathsf{W}$ be a map, and let $\mathsf{L}_\phi \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}; \mathsf{W})$ be the unique linear map determined as in Theorem 4.5.24.

(a) Show that L_ϕ is injective if and only if the family $(\phi(v))_{v \in \mathscr{B}}$ is linearly independent.

(b) Show that L_ϕ is surjective if and only if $\mathrm{span}_{\mathsf{F}}(\phi(\mathscr{B})) = \mathsf{W}$.

4.5.20 Let F be a field and let V be an F-vector space. If U is a subspace of V and if $v_1, v_2 \in \mathsf{V}$, show that the affine subspaces

$$\{v_1 + u \mid u \in \mathsf{U}\}, \quad \{v_2 + u \mid u \in \mathsf{U}\}$$

agree if and only if $v_1 - v_2 \in \mathsf{U}$.

4.5.21 Construct explicit isomorphisms between the following pairs of F-vector spaces:

(a) F^{k+1} and $\mathsf{F}_k[\xi]$;

(b) F_0^∞ and $\mathsf{F}[\xi]$.

4.5.22 Construct an explicit $\mathbb{R}$-isomorphism between $\mathbb{R}^\infty$ and the set $\mathbb{R}[[\xi]]$ of $\mathbb{R}$-formal power series.

4.5.23 Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$. Show that L is injective if and only if $\ker(\mathsf{L}) = \{0_{\mathsf{U}}\}$.

4.5.24 Let F be a field and let V be an F-vector space with U a strict subspace of V.

(a) Show that, if $\dim_{\mathsf{F}}(\mathsf{V}) < \infty$, then $\dim_{\mathsf{F}}(\mathsf{U}) < \dim_{\mathsf{F}}(\mathsf{V})$.

(b) Give examples of F, V, and U as above such that $\dim_{\mathsf{F}}(\mathsf{U}) = \dim_{\mathsf{F}}(\mathsf{V})$.

4.5.25 Let V be a $\mathbb{R}$-vector space with $\mathsf{V}_{\mathbb{C}}$ its complexification. Show that the complex conjugation σ_{V} is a $\mathbb{R}$-linear map of $\mathsf{V}_{\mathbb{C}}$, but not a $\mathbb{C}$-linear map.

Section 4.6

Field extensions

In this section we study in a little detail the notion of a field extension. This is a big industry for algebraists going under the name of "Galois Theory." We shall not have too much occasion to use the elements of this theory, so we give a very bare bones presentation, referring to the references in Section 4.6.10 for more details.

The principle point behind studying field extensions is that in a given field there are polynomials that do not have roots in the field; equivalently, there are irreducible polynomials of degree greater than one. For example, the polynomial $\xi^2 - 2$ is irreducible in $\mathbb{Q}[\xi]$. In order to better understand polynomials in a given field, it may then be useful to consider a larger field in which a given polynomial, or maybe all polynomials, have roots.

Do I need to read this section? The results in this section lend some context to our construction in Section 4.7 of the complex numbers. Section 4.6.5 is particularly useful in this respect, although it is also interesting to know just what an algebraic closure is in order to understand the phrase, "The complex numbers are the algebraic closure of the real numbers." Many of the results in this section will be used only in Sections 5.8.12 and 5.8.14. In particular, this is the only place we will have occasion to use notions of normal, separable, and Galois field extensions. •

4.6.1 Definitions and basic constructions

The study of extensions is a little different from what one very often does in algebra, which is to study *sub*objects; an extension is a *super*object. The reason for this is that with field extensions, the emphasis is on constructing extensions that contain roots of polynomials, since we know that it is possible that a polynomial may not have roots in the field in which one is working.

First the definition.

4.6.1 Definition (Field extension) An ***extension*** of a field F is a field K of which F is a subfield. •

The following algebraic property of extensions is where one begins to look for much of their interesting structure.

4.6.2 Proposition (Algebraic properties of extensions) *If K is an extension of a field F, then K is an F-algebra with the operations of addition, scalar multiplication, and product defined by addition in the field K, restriction of scalar multiplication in K to F in the first argument, and product in K, respectively.*

Proof This is an elementary verification of the axioms; see Exercise 4.6.2. ■

Note that if K is an extension of a field F and if $A \in \mathsf{F}[\xi]$ is a polynomial with coefficients in F, then we regard A as a polynomial with coefficients in K by virtue of the fact that $\mathsf{F} \subseteq \mathsf{K}$ (cf. Example 4.5.2–5).

Since an extension of a field is a vector space over the field, we have the following definition.

4.6.3 Definition (Degree of extension, finite extension) If K is an extension of a field F, the ***degree*** of K is $[\mathsf{K} : \mathsf{F}] = \dim_{\mathsf{F}}(\mathsf{K})$. An extension is ***finite*** if $[\mathsf{K} : \mathsf{F}]$ is finite. •

Since one can have nested extensions, it is possible to talk about the degree of these.

4.6.4 Proposition (The degree of nested extensions) *Let F be a field, let K be an extension of F, and let L be an extension of K. Then $[\mathsf{L} : \mathsf{F}] = [\mathsf{L} : \mathsf{K}][\mathsf{K} : \mathsf{F}]$.*

Proof Let $\{a_i\}_{i\in I}$ be a basis for K over F and let $\{b_j\}_{j\in J}$ be a basis for L over K. We claim that $\{a_ib_j\}_{(i,j)\in I\times J}$ is a basis for L over F, noting that this will give the result. Let $\beta \in \mathsf{L}$ and write

$$\beta = \sum_{j\in J} \beta_j b_j$$

with only finitely many of the $\beta_j \in \mathsf{K}$, $j \in J$, being nonzero. For each $j \in J$ write

$$\beta_j = \sum_{i\in I} \alpha_{ij} a_i$$

where only finitely many of the $\alpha_{ij} \in \mathsf{F}$, $i \in I$, being nonzero. Then we have

$$\beta = \sum_{i\in I}\sum_{j\in J} \alpha_{ij} a_a b_j,$$

showing that the set $\{a_ib_j\}_{(i,j)\in I\times J}$ generates L over F. Now suppose that

$$\sum_{i\in I}\sum_{j\in J} \alpha_{ij} a_i b_j = 0_{\mathsf{L}}$$

for some $\alpha_{ij} \in \mathsf{F}$, $(i, j) \in I \times J$, only finitely many of which are nonzero. Then, since $\{b_j\}_{j\in J}$ is linearly independent,

$$\sum_{i\in I} \alpha_{ij} a_i = 0_{\mathsf{K}}, \qquad j \in J.$$

Since $\{a_i\}_{i\in I}$ is a basis, $\alpha_{ij} = 0_{\mathsf{F}}$, $(i, j) \in I\times J$, giving $\{a_ib_j\}_{(i,j)\in I\times J}$ as linearly independent. ■

A common sort of extension comes from "extending" a field in the smallest possible manner to include a certain set of new numbers. The following definition makes this precise.

4.6.5 Definition (Adjoining elements to a field) Let F be a field with K an extension of F. For a subset $S \subseteq \mathsf{K}$ let us denote by $\mathsf{F}[S]$ (resp. $\mathsf{F}(S)$) the smallest subring (resp. subfield) of K containing $\mathsf{F} \cup S$. The field $\mathsf{F}(S)$ is the field obtained by ***adjoining*** the elements from S. We will also say that $\mathsf{F}(S)$ is the extension of F ***generated*** by S. •

If $S = \{a_1, \dots, a_k\}$ then we shall denote $\mathsf{F}[S] = \mathsf{F}[a_1, \dots, a_k]$ and $\mathsf{F}(S) = \mathsf{F}(a_1, \dots, a_k)$. It is possible to describe $\mathsf{F}[S]$ and $\mathsf{F}(S)$ fairly explicitly. In order to do so, we need to recall some notation concerning evaluating polynomials with arbitrary sets of indeterminates. We refer to the end of Section 4.4.7.

4.6.6 Theorem (Characterisation of F[S] and F(S)) *Let* F *be a field with extension* K *and let* $\mathsf{S} \subseteq \mathsf{K}$*. Then the following statements hold:*

(i) $\mathsf{F}[\mathsf{S}]$ *consists of the elements of* K *of the form*

$$\mathrm{Ev}_{\mathsf{K}}(\mathrm{A})(\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k}), \qquad \mathrm{k} \in \mathbb{Z}_{\geq 0},\ \mathrm{A} \in \mathsf{F}[\xi_1, \dots, \xi_\mathrm{k}],\ \mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k} \in \mathrm{S};$$

(ii) $\mathsf{F}(\mathsf{S})$ *consists of the elements of* K *of the form*

$$\frac{\mathrm{Ev}_{\mathsf{K}}(\mathrm{A})(\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k})}{\mathrm{Ev}_{\mathsf{K}}(\mathrm{B})(\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k})}, \qquad \mathrm{k} \in \mathbb{Z}_{\geq 0},\ \mathrm{A}, \mathrm{B} \in \mathsf{F}[\xi_1, \dots, \xi_\mathrm{k}],$$
$$\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k} \in \mathrm{S},\ \mathrm{Ev}_{\mathsf{K}}(\mathrm{B})(\mathrm{a}_1, \dots, \mathrm{a}_\mathrm{k}) \neq 0_{\mathsf{K}}.$$

Proof (i) Let A_S be the set of elements of K of the form

$$\mathrm{Ev}_{\mathsf{K}}(A)(a_1, \dots, a_k), \qquad k \in \mathbb{Z}_{\geq 0},\ A \in \mathsf{F}[\xi_1, \dots, \xi_k],\ a_1, \dots, a_k \in S.$$

Then $S \subseteq A_S$ and $\mathsf{F} \subseteq A_S$. Moreover, one can directly check that A_S is a subring of K, essentially because it is the image of the polynomial ring with indeterminates S under the evaluation homomorphism. Thus $\mathsf{F}[S] \subseteq A_S$ since $\mathsf{F}[S]$ is the smallest subring of K containing S and F. Now note that since $S \subseteq \mathsf{F}[S]$ and since $\mathsf{F}[S]$ is a subring,

$$a_1^{j_1} \cdots a_k^{j_k} \in \mathsf{F}[S], \qquad k \in \mathbb{Z}_{>0},\ a_1, \dots, a_k \in S,\ j_1, \dots, j_k \in \mathbb{Z}_{\geq 0}. \tag{4.22}$$

Since $\mathsf{F} \subseteq \mathsf{F}[S]$ it then follows that all finite F-linear combinations of elements of the form (4.22) are in $\mathsf{F}[S]$. But this exactly means that $A_S \subseteq \mathsf{F}[S]$.

(ii) Let B_S be the set of elements of K of the form

$$\frac{\mathrm{Ev}_{\mathsf{K}}(A)(a_1, \dots, a_k)}{\mathrm{Ev}_{\mathsf{K}}(B)(a_1, \dots, a_k)}, \qquad k \in \mathbb{Z}_{\geq 0},\ A, B \in \mathsf{F}[\xi_1, \dots, \xi_k],$$
$$a_1, \dots, a_k \in S,\ \mathrm{Ev}_{\mathsf{K}}(B)(a_1, \dots, a_k) \neq 0_{\mathsf{K}}.$$

Note that $S \subseteq B_S$ and $\mathsf{F} \subseteq B_S$. We claim that B_S is a subfield of K. Consider two elements $\alpha, \beta \in B_S$ given by

$$\alpha = \frac{\mathrm{Ev}_{\mathsf{K}}(A)(a_1, \dots, a_k)}{\mathrm{Ev}_{\mathsf{K}}(B)(a_1, \dots, a_k)}, \quad \beta = \frac{\mathrm{Ev}_{\mathsf{K}}(C)(b_1, \dots, b_m)}{\mathrm{Ev}_{\mathsf{K}}(D)(b_1, \dots, b_m)}.$$

Then

$$\alpha + \beta = \frac{\mathrm{Ev}_{\mathsf{K}}(AD + BC)(a_1, \dots, a_k, b_1, \dots, b_m)}{\mathrm{Ev}_{\mathsf{K}}(BD)(a_1, \dots, a_k, b_1, \dots, b_m)},$$
$$\alpha\beta = \frac{\mathrm{Ev}_{\mathsf{K}}(AC)(a_1, \dots, a_k, b_1, \dots, b_m)}{\mathrm{Ev}_{\mathsf{K}}(BD)(a_1, \dots, a_k, b_1, \dots, b_m)},$$

giving $\alpha + \beta, \alpha\beta \in B_S$. Thus B_S is a subfield and so $\mathsf{F}(S) \subseteq B_S$. Since $\mathsf{F}[S] \subseteq \mathsf{F}(S)$ and since $\mathsf{F}(S)$ is a subfield, all elements of K of the form $\alpha^{-1}\beta$ are in $\mathsf{F}(S)$ for $\alpha, \beta \in \mathsf{F}[S]$ with $\beta \neq 0_{\mathsf{K}}$. In particular, $B_S \subseteq \mathsf{F}(S)$. ■

Said in plain English, elements of $\mathsf{F}[S]$ are polynomial functions of variables in S and elements of $\mathsf{F}(S)$ are rational functions of variables in S.

Let us give some examples of what happens when one adjoins numbers to a field. In these examples it is convenient for us to suppose that the reader is familiar with complex numbers, although we have not yet formally introduced them.

4.6.7 Examples (Adjoining elements to a field)

1. We consider the extension $\mathbb{R}$ of $\mathbb{Q}$. Recall from Example 2.1.15 that $\sqrt{2}$ is irrational. Then $\mathbb{Q}(\sqrt{2})$ is an extension of $\mathbb{Q}$. By Theorem 4.6.6 we know that elements of $\mathbb{Q}(\sqrt{2})$ are of the form

$$\frac{\mathrm{Ev}_{\mathbb{R}}(A)(\sqrt{2})}{\mathrm{Ev}_{\mathbb{R}}(B)(\sqrt{2})}, \qquad A, B \in \mathbb{Q}[\xi].$$

Note that, for any polynomials $A, B \in \mathbb{Q}[\xi]$, we have

$$\mathrm{Ev}_{\mathbb{R}}(A)(\sqrt{2}) = \alpha_1\sqrt{2} + \alpha_0, \quad \mathrm{Ev}_{\mathbb{R}}(B)(\sqrt{2}) = \beta_1\sqrt{2} + \beta_0, \qquad \alpha_0, \alpha_1, \beta_0, \beta_1 \in \mathbb{Q}.$$

(why?). Therefore, we have

$$\frac{\mathrm{Ev}_{\mathbb{R}}(A)(\sqrt{2})}{\mathrm{Ev}_{\mathbb{R}}(B)(\sqrt{2})} = \frac{\alpha_1\sqrt{2} + \alpha_0}{\beta_1\sqrt{2} + \beta_0}\frac{\beta_1\sqrt{2} - \beta_0}{\beta_1\sqrt{2} - \beta_0} = \gamma_1\sqrt{2} + \gamma_0$$

where $\gamma_0, \gamma_1 \in \mathbb{Q}$ are given by

$$\gamma_1 = \frac{\alpha_0\beta_1 - \alpha_1\beta_0}{2\beta_1^2 - \beta_0^2}, \quad \gamma_0 = \frac{2\alpha_1\beta_1 - \alpha_0\beta_0}{2\beta_1^2 - \beta_0^2}.$$

The upshot is that

$$\mathbb{Q}(\sqrt{2}) = \{\alpha_1\sqrt{2} + \alpha_0 \mid \alpha_0, \alpha_1 \in \mathbb{Q}\}. \tag{4.23}$$

In this example we have worked this out in detail. However, we shall see subsequently that there is a general picture that gives this specific conclusion.

2. We consider the extension $\mathbb{R}$ of $\mathbb{Q}$, and we adjoin $\sqrt[4]{2}$ to $\mathbb{R}$ to get the field $\mathbb{Q}(\sqrt[4]{2})$. By Theorem 4.6.6 we have $\mathbb{Q}(\sqrt[4]{2})$ as the elements of $\mathbb{R}$ of the form

$$\frac{\mathrm{Ev}_{\mathbb{R}}(A)(\sqrt[4]{2})}{\mathrm{Ev}_{\mathbb{R}}(B)(\sqrt[4]{2})}, \qquad A, B \in \mathbb{Q}[\xi].$$

We could proceed directly, as above, to show that

$$\mathbb{Q}(\sqrt[4]{2}) = \{\alpha_3 2^{3/4} + \alpha_2 2^{1/2} + \alpha_1 2^{1/4} + \alpha_0 \mid \alpha_0, \alpha_1, \alpha_2, \alpha_3 \in \mathbb{Q}\}.$$

However, we shall see below (Example 4.6.27) that there is a general idea from which this follows easily. •

4.6.2 Automorphisms of field extensions

This means that an extension K of F has two natural structures: (1) the structure of a field and (2) the structure of an F-vector space. One is thus interested in mappings of field extensions that preserve both of these structures. The following definition encodes this.

4.6.8 Definition (F-homomorphism, F-automorphism) Let K and K' be extensions of a field F. An ***F-homomorphism*** is a homomorphism $\phi\colon \mathsf{K} \to \mathsf{K}'$ of fields that is also a homomorphism of F-vector spaces. An ***F-automorphism*** of K is an invertible F-homomorphism of K with itself. The set of F-automorphisms is the ***Galois group*** of K over F, and is denoted by $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$. •

The following result gives a useful alternative characterisation of F-homomorphisms.

4.6.9 Proposition (Characterisation of F-homomorphisms) *Let F be a field and let K and K' be extensions of F. A homomorphism $\phi\colon \mathsf{K} \to \mathsf{K}'$ of fields is an F-homomorphism if and only if $\phi|\mathsf{F} = \mathrm{id}_{\mathsf{F}}$.*

Proof This is Exercise 4.6.4. ■

One of the most easily understood examples of a field extension is the extension of the real numbers to the complex numbers. While we will not formally discuss the complex numbers until Section 4.7, let us give these as an example to illustrate the Galois group, along with some more elementary examples.

4.6.10 Examples (F-automorphism)

1. Consider the extension $\mathbb{Q}(\sqrt{2})$ of $\mathbb{Q}$. As we saw in Example 4.6.7–1 we have

$$\mathbb{Q}(\sqrt{2}) = \{\alpha_1\sqrt{2} + \alpha_0 \mid \alpha_0, \alpha_1 \in \mathbb{Q}\}.$$

If we think of $\mathbb{Q}(\sqrt{2})$ as a two-dimensional $\mathbb{Q}$-vector space with basis $\{1, \sqrt{2}\}$, then a $\mathbb{Q}$-automorphism must have the form

$$\phi\colon \alpha_1\sqrt{2} + \alpha_0 \mapsto (a_{11}\alpha_1 + a_{12}\alpha_0)\sqrt{2} + (a_{21}\alpha_1 + a_{22}\alpha_0)$$

for $a_{11}, a_{12}, a_{21}, a_{22} \in \mathbb{Q}$. If ϕ is to restrict to the identity map on $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2})$ then we must have

$$a_{12}\alpha_0\sqrt{2} + a_{22}\alpha_0 = \alpha_0$$

for every $\alpha_0 \in \mathbb{Q}$. Thus $a_{12} = 0$ and $a_{22} = 1$. If ϕ is to be a homomorphism of the field $\mathbb{Q}(\sqrt{2})$ then

$$2 = \phi(2) = \phi(\sqrt{2}\sqrt{2}) = \phi(\sqrt{2})\phi(\sqrt{2}) = (a_{11}\sqrt{2} + a_{21})^2 = 2a_{11}^2 + a_{21}^2 + 2\sqrt{2}a_{11}a_{21}.$$

Thus we must have $a_{11} = 0$ or $a_{21} = 0$. If $a_{11} = 0$ then this gives $a_{21}^2 = 2$ which cannot be solved for $a_{21} \in \mathbb{Q}$. Thus we must have $a_{21} = 0$ and so $a_{11}^2 = 1$.

In summary, there are two $\mathbb{Q}$-automorphisms of $\mathbb{Q}(\sqrt{2})$:

$$\alpha_1\sqrt{2} + \alpha_0 \mapsto \alpha_1\sqrt{2} + \alpha_0, \quad \alpha_1\sqrt{2} + \alpha_0 \mapsto -\alpha_1\sqrt{2} + \alpha_0.$$

In this example we have done this "by hand." As we shall see, however, there are often simpler ways to determine the group of automorphisms of a field extension.

2. Now let us consider the extension $\mathbb{Q}(\sqrt[4]{2})$ of $\mathbb{Q}$. As we asserted in Example 4.6.7–2,

$$\mathbb{Q}(\sqrt[4]{2}) = \{\alpha_3 2^{3/4} + \alpha_2 2^{1/2} + \alpha_1 2^{1/4} + \alpha_0 \mid \alpha_0, \alpha_1, \alpha_2, \alpha_3 \in \mathbb{Q}\}.$$

For this example, let us not work out explicitly what the automorphisms look like. Instead we merely write them down; there are two of them:

$$\begin{aligned} \alpha_3 2^{3/4} + \alpha_2 2^{1/2} + \alpha_1\sqrt[4]{2} + \alpha_0 &\mapsto \alpha_3 2^{3/4} + \alpha_2 2^{1/2} + \alpha_1\sqrt[4]{2} + \alpha_0, \\ \alpha_3 2^{3/4} + \alpha_2 2^{1/2} + \alpha_1\sqrt[4]{2} + \alpha_0 &\mapsto \alpha_3 2^{3/4} + \alpha_2 2^{1/2} - \alpha_1\sqrt[4]{2} + \alpha_0. \end{aligned}$$

We shall see how this arises in Section 4.6.5.

3. The extension $\mathbb{C}$ of the real numbers is two-dimensional, having $\{1, \mathrm{i}\}$ as a basis. Let us use this basis to represent a $\mathbb{R}$-automorphism ϕ of $\mathbb{C}$ by

$$\phi(\alpha_1\mathrm{i} + \alpha_0) = (a_{11}\alpha_1 + a_{12}\alpha_1)\mathrm{i} + (a_{21}\alpha_1 + a_{22}\alpha_0).$$

The condition that $\phi|\mathbb{R} = \mathrm{id}_{\mathbb{R}}$ gives $a_{11} = 1$ and $a_{21} = 0$. If ϕ is a homomorphism of the field $\mathbb{C}$ we must have

$$-1 = \phi(-1) = \phi(\mathrm{ii}) = \phi(\mathrm{i})\phi(\mathrm{i}) = (a_{11}\mathrm{i} + a_{21})^2 = -a_{11}^2 + a_{21}^2 + 2a_{11}a_{21}\mathrm{i}.$$

Thus $a_{11}a_{21} = 0$. If $a_{11} = 0$ then we must have $a_{21}^2 = -1$ which cannot be solved for $a_{21} \in \mathbb{R}$. Thus $a_{21} = 0$ which gives $a_{11}^2 = 1$. Thus we have two $\mathbb{R}$-automorphisms of $\mathbb{C}$:

$$\alpha_1\mathrm{i} + \alpha_0 \mapsto \alpha_1\mathrm{i} + \alpha_0, \quad \alpha_1\mathrm{i} + \alpha_0 \mapsto -\alpha_1\mathrm{i} + \alpha_0. \qquad \bullet$$

Thus we see that elements of the Galois group can be viewed as generalisations of complex conjugation. This will be explored in more detail in Section 4.6.6.

As Proposition 4.6.9 shows, an F-automorphism of K fixes points in F. Thus the subfield F of K is fixed by every F-automorphism. One can then ask whether other subfields of F are fixed by F-automorphisms. More generally, one has the following notion.

4.6.11 Definition (Fixed field) Let F be a field and let G be a subgroup of the group of field isomorphisms of F with itself. The ***fixed field*** associated with G is

$$\mathsf{F}^{\mathsf{G}} = \{a \in \mathsf{F} \mid \phi(a) = a \text{ for every } \phi \in \mathsf{G}\}. \qquad \bullet$$

One can then interpret Proposition 4.6.9 as saying that $\mathsf{F} \subseteq \mathsf{K}^{\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})}$. It is easy to show that the fixed field is indeed a field.

4.6.3 Algebraic field extensions

In the preceding section we saw that there was an important interaction between polynomials and field extensions. Namely, we saw that there is always a field extension that contains the roots of a prescribed polynomial. The converse may not be true. That is, it is not always the case that a field extension is obtained by adding the roots of polynomials.

The following definition captures this idea.

4.6.12 Definition (Algebraic extension, transcendental extension) Let F be a field with K an extension of F.

(i) An element $a \in \mathsf{K}$ is ***algebraic*** over F if there exists a nonzero polynomial $A \in \mathsf{F}[\xi]$ which has a as a root in K.

(ii) An element $a \in \mathsf{K}$ is ***transcendental*** over F if it is not algebraic over F.

(iii) The extension K is ***algebraic*** over F if every element of K is algebraic.

(iv) The extension K is ***transcendental*** if it is not algebraic. •

Let us give some examples of algebraic and transcendental elements of a field.

4.6.13 Examples (Algebraic and transcendental elements)

1. The element $\sqrt{2} \in \mathbb{R}$ is algebraic over $\mathbb{Q}$, it being a root in $\mathbb{R}$ of the polynomial $\xi^2 - 2 \in \mathbb{Q}[\xi]$. In like manner $\sqrt[4]{2} \in \mathbb{R}$ is algebraic over $\mathbb{Q}$ since it is a root of the polynomial $\xi^4 - 2$.
2. One can show that $\mathrm{e}, \pi \in \mathbb{R}$ are transcendental over $\mathbb{Q}$. This is nontrivial. We refer to Section 2.4.1 for more discussion of this.
3. The element $\mathrm{i} \in \mathbb{C}$ (we have not yet talked about the complex numbers, but will do so in Section 4.7) is algebraic over $\mathbb{R}$, it being the root of the polynomial $\xi^2 - 1 \in \mathbb{R}[\xi]$. There are no transcendental elements in $\mathbb{C}$ over $\mathbb{R}$. This is a consequence of the fact that $\mathbb{C}$ is algebraically closed (see Definition 4.6.22), something we will prove in Theorem 4.7.6. •

The case when the degree of an extension is finite is of particular interest.

4.6.14 Proposition (Finite-dimensional field extensions are algebraic) *Let* F *be a field and let* K *be an extension of* F*. If* $[\mathsf{K} : \mathsf{F}]$ *is finite then* K *is algebraic.*

Proof Suppose that K is an n-dimensional F-vector space. If $a \in \mathsf{K}$ this means that the set $\{1_{\mathsf{K}}, a, \dots, a^n\}$ must be linearly dependent. Therefore, there exists $c_0, c_1, \dots, c_n \in \mathsf{F}$ such that

$$c_n a^n + \dots + c_1 a + c_0 = 0_{\mathsf{K}}.$$

Therefore, if we define $A = c_n \xi^n + \dots + c_1 \xi + c_0$, a is a root of A in K. ■

Associated with an element $a \in \mathsf{K}$ that is algebraic over a field F is an important polynomial. Let us denote by

$$\mathsf{I}_a = \{A \in \mathsf{F}[\xi] \mid \operatorname{Ev}_{\mathsf{K}}(A)(a) = 0_{\mathsf{K}}\}$$

the set of polynomials in $\mathsf{F}[\xi]$ which have a as a root in K. This set is not simply $\{0_{\mathsf{F}[\xi]}\}$ since a is algebraic. It is straightforward to verify that I_a is an ideal (Exercise 4.6.5). Therefore, since $\mathsf{F}[\xi]$ is a principal ideal domain, we have $\mathsf{I}_a = (M_a)$ for some nonzero, nonconstant polynomial $M_a \in \mathsf{F}[\xi]$. Moreover, if we additionally require that M_a be monic, this uniquely defines M_a (why?).

4.6.15 Definition (Minimal polynomial of an algebraic element) Let F be a field, let K be an extension of F, and let $a \in \mathsf{K}$ be algebraic over F. The ***minimal polynomial*** of a is the unique monic polynomial with the property that $\mathsf{I}_a = (M_a)$. •

We shall see another instance of the notion of the minimal polynomial in Section 5.8.4.

The minimal polynomial has the following important property.

4.6.16 Proposition (The minimal polynomial is irreducible) *Let* F *be a field, let* K *be an extension of* F*, and let* $\mathrm{a} \in \mathsf{K}$ *be algebraic. Then the minimal polynomial* $\mathrm{M_a}$ *is irreducible.*

Proof Suppose that M_a is not irreducible so that $M_a = AB$ for $A, B \in \mathsf{F}[\xi]$ monic and of degree at least one. Then

$$\operatorname{Ev}_{\mathsf{K}}(AB)(a) = \operatorname{Ev}_{\mathsf{K}}(A)(a)\operatorname{Ev}_{\mathsf{K}}(B)(a) = 0_{\mathsf{K}}.$$

Thus either $\operatorname{Ev}_{\mathsf{K}}(A)(a) = 0_{\mathsf{K}}$ or $\operatorname{Ev}_{\mathsf{K}}(B)(a) = 0_{\mathsf{K}}$; without loss of generality suppose the former. Then $M_a | A$. But this is absurd since A is nonzero and $\deg(A) < \deg(M_a)$, giving a contradiction. ■

4.6.17 Proposition (Characterisation of F(a) for algebraic a) *Let* F *be a field, let* K *be an extension of* F*, and let* $\mathrm{a} \in \mathsf{K}$ *be algebraic. Then the following statements hold:*

(i) $\mathsf{F}[\mathrm{a}] = \mathsf{F}(\mathrm{a})$*;*

(ii) there exists an F*-isomorphism from* $\mathsf{F}(\mathrm{a})$ *to* $\mathsf{F}[\xi]/(\mathrm{M_a})$*;*

(iii) $\{1_{\mathsf{K}}, \mathrm{a}, \dots, \mathrm{a}^{\mathrm{k}-1}\}$ *is a basis for the* F*-vector space* $\mathsf{F}(\mathrm{a})$*, where* $\mathrm{k} = \deg(\mathrm{M_a})$*.*

Proof (i) Define a ring epimorphism $\phi_a \colon \mathsf{F}[\xi] \to \mathsf{F}[a]$ (this is a ring epimorphism by Theorem 4.6.6) by $\phi_a(A) = \mathrm{Ev}_{\mathsf{K}}(A)(a)$, and note that $\mathsf{F}[a]$ is isomorphic to $\mathsf{F}[\xi]/\ker(\phi_a)$ (see Exercise 4.2.8). Since $\ker(\phi_a)$ is an ideal by Proposition 4.2.26 and since $\mathsf{F}[\xi]$ is a principal ideal domain, we must have $\ker(\phi_a) = (\tilde{M}_a)$ for some unique monic polynomial $\tilde{M}_a \in \mathsf{F}[\xi]$. Indeed, it is clear by definition of the minimal polynomial that, in fact, $\tilde{M}_a = M_a$. We claim that $\mathsf{F}[a]$ is an integral domain. Indeed, suppose that

$$\mathrm{Ev}_{\mathsf{K}}(A)(a) \cdot \mathrm{Ev}_{\mathsf{K}}(B)(a) = \mathrm{Ev}_{\mathsf{K}}(AB)(a) = 0_{\mathsf{K}}.$$

Then $AB \in (M_a)$ and so $M_a | AB$. Since M_a is irreducible, either $M_a | A$ or $M_a | B$. Thus either $\mathrm{Ev}_{\mathsf{K}}(A)(a) = 0_{\mathsf{K}}$ or $\mathrm{Ev}_{\mathsf{K}}(B)(a) = 0_{\mathsf{K}}$, giving $\mathsf{F}[a]$ as an integral domain. We know that $\mathsf{F}[\xi]/\ker(\phi_a)$, by virtue of being isomorphic to $\mathsf{F}[a]$, is an integral domain. Therefore, by Theorem 4.2.37, $\ker(\phi_a)$ is a prime ideal and so a maximal ideal by Proposition 4.2.21. This means that $\mathsf{F}[a]$ is a subfield of K by Theorem 4.3.9. Thus $\mathsf{F}[a]$ is a subfield of K containing F and a. Therefore, $\mathsf{F}(a) \subseteq \mathsf{F}[a]$ since $\mathsf{F}(a)$ is the smallest subfield containing F and a. It is clear from Theorem 4.6.6 that $\mathsf{F}[a] \subseteq \mathsf{F}(a)$, giving $\mathsf{F}[a] = \mathsf{F}(a)$ as desired.

(ii) This was proved en route to proving part (i).

(iii) First let us show that $\{1_{\mathsf{K}}, a, \dots, a^{k-1}\}$ is linearly independent. Suppose that

$$c_{k-1}a^{k-1} + \cdots + c_1 a + c_0 = 0_{\mathsf{K}}.$$

This means that a is a root of the polynomial

$$P = c_{k-1}\xi^{k-1} + \cdots + c_1\xi + c_0.$$

Thus $P \in \mathsf{I}_a$ and so $M_a | P$ since $\mathsf{I}_a = (M_a)$. Since $\deg(P) < \deg(M_a)$ we must have $P = 0_{\mathsf{F}[\xi]}$.

To show that $\{1_{\mathsf{K}}, a, \dots, a^{k-1}\}$ spans $\mathsf{F}(a)$ we use part (i). From this and Theorem 4.6.6, if $b \in \mathsf{F}(a)$ then we have $b = \mathrm{Ev}_{\mathsf{K}}(B)(a)$ for some $B \in \mathsf{F}[\xi]$. We can use the Euclidean Algorithm to write $B = QM_a + R$ for $Q, R \in \mathsf{F}[\xi]$ with $\deg(R) < \deg(M_a)$. Therefore, $\mathrm{Ev}_{\mathsf{K}}(B)(a) = \mathrm{Ev}_{\mathsf{K}}(R)(a)$, meaning that

$$b = b_{k-1}a^{k-1} + \cdots + b_1 a + b_0$$

for some $b_0, b_1, \dots, b_{k-1} \in \mathsf{F}$, as desired. ■

The main point to take away from the previous theorem is that $\mathsf{F}(a)$ is a finite-dimensional F-vector space when a is algebraic. Therefore, combining this with Proposition 4.6.14 gives the following result.

4.6.18 Corollary (F(a) is algebraic) *If* F *is a field, if* K *is an extension of* F*, and if* $\mathrm{a} \in \mathsf{K}$ *is algebraic over* F*, then* $\mathsf{F}(\mathrm{a})$ *is an algebraic extension of* K*.*

Now suppose that we have a finite subset $S = \{a_1, \dots, a_m\} \subseteq \mathsf{K}$ of a field extension K of F. We denote $\mathsf{F}(S) = \mathsf{F}(a_1, \dots, a_m)$. It is straightforward to see that

$$\mathsf{F}(a_1, \dots, a_m) = \mathsf{F}(a_1, \dots, a_{m-1})(a_m).$$

Therefore, much can be deduced about $\mathsf{F}(a_1, \dots, a_m)$ by considering only the case when $m = 1$. In particular, we have the following result.

4.6.19 Proposition (Characterisation of $\mathsf{F}(\mathsf{a}_1, \dots, \mathsf{a}_\mathsf{m})$) *Let* F *be a field, let* K *be an extension of* F*, and let* $\mathsf{a}_1, \dots, \mathsf{a}_\mathsf{m} \in \mathsf{K}$ *be algebraic. Then* $\mathsf{F}(\mathsf{a}_1, \dots, \mathsf{a}_\mathsf{m})$ *is a finite-dimensional* F*-vector space, and so is an algebraic extension of* F*.*

Proof We prove this by induction on m. The case $m = 1$ is covered by Proposition 4.6.17. So suppose that $\mathsf{F}(a_1, \dots, a_{m-1})$ is a finite-dimensional F-vector space with basis $\{b_1, \dots, b_r\}$. Now let $b \in \mathsf{F}(a_1, \dots, a_m) = \mathsf{F}(a_1, \dots, a_{m-1})(a_m)$ and, by Proposition 4.6.17, write

$$b = c_{k-1}a_m^{k-1} + \dots + c_1 a_m + c_0$$

for $c_0, c_1, \dots, c_{k-1} \in \mathsf{F}(a_1, \dots, a_{m-1})$. Then we can write

$$b = \sum_{j=0}^{k-1} c_j a_m^j = \sum_{j=0}^{k-1} \sum_{l_j=1}^{m} c_{j,l_j} b_{l_j} a_m^j,$$

for $c_{j,l_j} \in \mathsf{F}$, $j \in \{0, 1, \dots, k-1\}$, $l_j \in \{1, \dots, m\}$. This shows that every element of $\mathsf{F}(a_1, \dots, a_m)$ is a linear combination of the finite set of elements $b_{l_j} a_m^j$, $j \in \{0, 1, \dots, k-1\}$, $l_j \in \{1, \dots, m\}$ of $\mathsf{F}(a_1, \dots, a_m)$.

The final assertion follows from Proposition 4.6.14. ■

A key result that we shall require concerning algebraic extensions is the following.

4.6.20 Proposition (The set of algebraic elements is a field) *Let* F *be a field and let* K *be an extension of* F*. Then the set* $\mathsf{A}(\mathsf{F}, \mathsf{K})$ *of algebraic elements of* K *over* F *is a subfield of* K*.*

Proof Let $a, b \in \mathsf{A}(\mathsf{F}, \mathsf{K})$ and note that $\mathsf{F}(a, b)$ is an algebraic extension of F containing a and b. Thus $\mathsf{F}(a, b) \subseteq \mathsf{A}(\mathsf{F}, \mathsf{K})$. Thus $a - b, ab^{-1} \in \mathsf{F}(a, b) \subseteq \mathsf{K}$ by Exercise 4.3.5. Thus K is a subfield, again by Exercise 4.3.5. ■

It is useful to see how adjoining an element to a field behaves under isomorphisms of the field.

4.6.21 Proposition (Adjoining elements and isomorphisms) *Let* F_1 *and* F_2 *be fields with* $\phi \colon \mathsf{F}_1 \to \mathsf{F}_2$ *an isomorphism, and let* K_1 *and* K_2 *be extensions of* F_1 *and* F_2*, respectively. For* $\mathrm{A} \in \mathsf{F}_1[\xi]$ *given by*

$$\mathrm{A} = \mathrm{a}_\mathrm{k}\xi^\mathrm{k} + \dots + \mathrm{a}_1\xi + \mathrm{a}_0,$$

define $\phi_*\mathrm{A} \in \mathsf{F}_2[\xi]$ *by*

$$\phi_*\mathrm{A} = \phi(\mathrm{a}_\mathrm{k})\xi^\mathrm{k} + \dots + \phi(\mathrm{a}_1)\xi + \phi(\mathrm{a}_0).$$

If A *is irreducible, if* $\mathrm{r}_1 \in \mathsf{K}_1$ *is a root of* A*, and if* $\mathrm{r}_2 \in \mathsf{K}_2$ *is a root of* $\phi_*\mathrm{A}$*, then there exists a unique isomorphism* $\psi \colon \mathsf{F}_1(\mathrm{r}_1) \to \mathsf{F}_2(\mathrm{r}_2)$ *such that* $\psi|\mathsf{F}_1 = \phi$ *and such that* $\psi(\mathrm{r}_1) = \mathrm{r}_2$*.*

Proof Let M_{r_1} and M_{r_2} be the minimal polynomials of r_1 and r_2, respectively. We claim that $A = aM_{r_1}$ and $\phi_*A = \phi(a)M_{r_2}$ for $a \in \mathsf{F}_1^*$. Since r_1 is a root of A in K_1, $M_{r_1}|A$. Since A is irreducible this means that $A = aM_{r_1}$ for a nonzero. By similar arguments, since ϕ_*A is irreducible, $\phi_*A = a'M_{r_2}$. Since a' is then the leading coefficient of ϕ_*A we must have $a' = \phi(a)$, as desired. Note that $M_{r_2} = \phi_*M_{r_1}$ as a consequence of these arguments.

By Theorem 4.6.6 and Proposition 4.6.17 we write elements of $\mathsf{F}_1(r_1)$ and $\mathsf{F}_2(r_2)$ as $\mathrm{Ev}_{\mathsf{K}_1}(B_1)(r_1)$ and $\mathrm{Ev}_{\mathsf{K}_2}(B_2)(r_2)$, respectively, for $B_a \in \mathsf{F}_a[\xi]$, $a \in \{1, 2\}$. Then define ψ by

$$\psi(\mathrm{Ev}_{\mathsf{K}_1}(B)(r_1)) = \mathrm{Ev}_{\mathsf{K}_2}(\phi_* B)(r_2).$$

First we should show that ψ is well-defined. Suppose that

$$\mathrm{Ev}_{\mathsf{K}_1}(B)(r_1) = \mathrm{Ev}_{\mathsf{K}_1}(B')(r_1)$$

for $B, B' \in \mathsf{F}_1[\xi]$. Then $B - B' \in \ker(\phi_{r_1})$ where $\phi_{r_1} \colon \mathsf{F}_1[\xi] \to \mathsf{F}_1(r_1)$ is defined by $\phi_{r_1}(B) = \mathrm{Ev}_{\mathsf{K}_1}(B)(r_1)$. Thus, by Proposition 4.6.17, $M_{r_1} | (B - B')$. Since $M_{r_2} = \phi_* M_{r_1}$ this implies that $M_{r_2} | (\phi_* B - \phi_* B')$. Thus

$$\mathrm{Ev}_{\mathsf{K}_2}(\phi_* B)(r_2) = \mathrm{Ev}_{\mathsf{K}_2}(\phi_* B')(r_2),$$

showing that ψ is well-defined. To see that ψ is a bijective note that $\psi(r_1^j) = r_2^j$, meaning that ψ maps the ordered basis $\{1_{\mathsf{F}_1}, r_1, \dots, r_1^{k-1}\}$ for $\mathsf{F}_1(r_1)$ bijectively onto the ordered basis $\{1_{\mathsf{F}_2}, r_2, \dots, r_2^{k-1}\}$ for $\mathsf{F}_2(r_2)$. It is easy to verify directly that ψ is a field homomorphism.

To obtain the uniqueness of ψ we note that, if $\psi' \colon \mathsf{F}_1(r_1) \to \mathsf{F}_2(r_2)$ is a field isomorphism which maps r_1 to r_2 and restricts to ϕ on F_1, it follows immediately that

$$\psi'(\mathrm{Ev}_{\mathsf{K}_1}(B)(r_1)) = \mathrm{Ev}_{\mathsf{K}_2}(\phi_* B)(r_2),$$

as desired. ■

4.6.4 Algebraic closure

In this section we investigate a specific algebraic field extension, one in which all polynomials are guaranteed to have roots. The following definition describes what we are after.

4.6.22 Definition (Algebraically closed, algebraic closure) Let K be a field. If every polynomial in K splits over K, then K is ***algebraically closed***. If F is an algebraically closed algebraic extension of K, then F is an ***algebraic closure*** of K. •

It turns out that every field possesses an algebraic closure, and that, moreover, all algebraic closures are isomorphic.

4.6.23 Theorem (Existence and uniqueness of the algebraic closure) *If F is a field then there exists a field $\bar{\mathsf{F}}$ that is an algebraic closure of F. Moreover, if $\bar{\mathsf{F}}_1$ and $\bar{\mathsf{F}}_2$ are algebraic closures of F, then there exists an F-isomorphism from F_1 to F_2.*

Proof The proof relies on the notion of the polynomial ring with an arbitrary collection of indeterminates; see Section 4.4.7.

Let

$$X = \{\xi_A \mid A \in \mathsf{F}[\xi] \text{ monic of degree at least } 1\}$$

be a set of indeterminates indexed by the nonconstant monic polynomials. If $A \in X$ then we have the single indeterminate $\xi_A \in X$. Since A is a polynomial in a single indeterminate, we can think of the indeterminate not as ξ, but as ξ_A. In this way we can think of A as an element in $\mathsf{F}[\xi_A] \subseteq \mathsf{F}[X]$. We denote this element by $A(\xi_A)$. Let I be the ideal generated by $\{A(\xi_A) \mid \xi_A \in X\}$. We claim that $\mathsf{I} \subset \mathsf{F}[X]$. Suppose otherwise so that $1_{\mathsf{F}[X]} \in \mathsf{I}$. Thus, by Theorem 4.2.54, there exists $A_1, \dots, A_k \in \mathsf{F}[\xi]$ and $\Psi_1, \dots, \Psi_k \in \mathsf{F}[X]$ such that

$$\Psi_1 A_1(\xi_{A_1}) + \cdots + \Psi_k A_k(\xi_{A_k}) = 1_{\mathsf{F}[X]}.$$

Let us write the polynomials $\Psi_1, \dots, \Psi_k$ as sums of monomials:

$$\Psi_l = \sum c^l_{j_1 \cdots j_{ms_l}} \xi^{j_1}_{B^l_1} \cdots \xi^{j_{ms_l}}_{B^l_{ms_l}}, \qquad l \in \{1, \dots, k\}.$$

Now, by Theorem 4.6.26, let K be an extension of F such that the polynomials $A_1, \dots, A_k$ split in $\mathsf{K}[\xi]$. Let $r_j \in \mathsf{K}$ be a root of A_j, $j \in \{1, \dots, k\}$. Now consider the evaluation of the indeterminates by letting ξ_A take the value r_j if $A = A_j$ for some $j \in \{1, \dots, k\}$, and by letting ξ_A take the value 0_{K} otherwise. Then denote by $F \colon \mathsf{K} \to \mathsf{K}$ the function defined by evaluating

$$\Psi_1 A_1(\xi_{A_1}) + \cdots + \Psi_k A_k(\xi_{A_k})$$

using these values for the indeterminates. Since $A_j(\xi_{A_j}) = 0_{\mathsf{K}}$ using these values for the indeterminates, we have $0_{\mathsf{K}} = 1_{\mathsf{K}}$. Thus our assumption that $1_{\mathsf{F}[X]} \in \mathsf{I}$ is invalid.

By Theorem 4.2.19 it follows that there is a maximal ideal J in $\mathsf{F}[X]$ containing I. Let us take $\mathsf{K}(\mathsf{F}) = \mathsf{F}[X]/\mathsf{J}$ which is a field by Theorem 4.3.9. We claim that the restriction to F of the projection π from $\mathsf{F}[X]$ to $\mathsf{K}(\mathsf{F})$ is injective. Indeed, suppose that $\pi(a) = 0_{\mathsf{K}(\mathsf{F})}$ for some $a \in \mathsf{F} \subseteq \mathsf{F}[X]$. This means that $a \in \mathsf{J} \supseteq \mathsf{I}$. Since I, and therefore J, contains no constant polynomials other than $0_{\mathsf{F}[X]}$ (since we have $\mathsf{I} \subset \mathsf{F}[X]$), it follows that $a = 0_{\mathsf{F}}$. Thus we have F naturally regarded as a subfield of $\mathsf{K}(\mathsf{F})$, and so $\mathsf{K}(\mathsf{F})$ is an extension of F. We claim that if $A \in \mathsf{F}[\xi]$ then A splits in $\mathsf{K}(\mathsf{F})[\xi]$. It suffices to show (why?) that every polynomial $A \in \mathsf{F}[\xi]$ has a root in $\mathsf{K}(\mathsf{F})$. Let $A \in \mathsf{F}[\xi]$ be a nonconstant monic polynomial so that $A(\xi_A) \in \mathsf{I} \subseteq \mathsf{J}$. Let us denote by $\bar{A} \in \mathsf{K}(\mathsf{F})[\xi]$ the image of A under the inclusion of $\mathsf{F}[\xi]$ in $\mathsf{K}(\mathsf{F})[\xi]$. Then

$$\mathrm{Ev}_{\mathsf{K}(\mathsf{F})}(\bar{A})(\xi_A + \mathsf{J}) = A(\xi_A) + \mathsf{J} \quad \implies \mathrm{Ev}_{\mathsf{K}(\mathsf{F})}(\bar{A})(\xi_A + \mathsf{J}) = 0_{\mathsf{F}[X]} + \mathsf{J}.$$

Thus $\xi_A + \mathsf{J}$ is a root for A in $\mathsf{K}(\mathsf{F})$.

Now inductively define field extensions

$$\mathsf{K}_1 \subseteq \mathsf{K}_2 \subseteq \mathsf{K}_3 \subseteq \cdots$$

by $\mathsf{K}_1 = \mathsf{K}(\mathsf{F})$ and $\mathsf{K}_j = \mathsf{K}(\mathsf{K}_{j-1})$, using the construction above. Let $\mathsf{K} = \cup_{j \in \mathbb{Z}_{>0}} \mathsf{K}_j$. We claim that K is a field. To define the operations of addition and multiplication we let $a, b \in \mathsf{K}$. Then $a, b \in \mathsf{K}_n$ for some sufficiently large n, and so we merely define $a + b, ab \in \mathsf{K}_n \subseteq \mathsf{K}$ using the fact that K_n is already a field. We leave it to the reader to convince themselves that these definitions are independent of n and satisfy the properties of addition and multiplication for fields. We next claim that K is algebraically closed. Let $A \in \mathsf{K}[\xi]$ be irreducible. Since A has only finitely many nonzero coefficients in K, there exists $N \in \mathbb{Z}_{>0}$ such that all coefficients are in K_N. Thus

A splits in $\mathsf{K}_N[\xi]$ and so is of degree 1. Thus every irreducible polynomial in $\mathsf{K}[\xi]$ has degree 1, and so K is algebraically closed.

The above arguments show that there exists an algebraically closed field K containing F. Let $\bar{\mathsf{F}} \subseteq \mathsf{K}$ be the set of algebraic elements of K over F. This is a field by Proposition 4.6.20. Now let $A \in \bar{\mathsf{F}}[\xi]$ be irreducible. Since $\bar{\mathsf{F}} \subseteq \mathsf{K}$ there exists a root $a \in \mathsf{K}$ of $\bar{\mathsf{F}}$. Then $\bar{\mathsf{F}}(a)$ is an algebraic extension of $\bar{\mathsf{F}}$ and so an algebraic extension of F since $\bar{\mathsf{F}}$ is an algebraic extension of F. But this means that $a \in \bar{\mathsf{F}}$, and so A is of degree 1.

Finally, we show that two algebraic closures F_1 and F_2 are isomorphic via an F-isomorphism. Let

$$\mathscr{C} = \{(\mathsf{F}'_1, \mathsf{F}'_2, \phi) \mid \mathsf{F}'_1, \mathsf{F}'_2 \text{ extensions of } \mathsf{F},\ \mathsf{F} \subseteq \mathsf{F}'_1 \subseteq \bar{\mathsf{F}}_1,\ \mathsf{F} \subseteq \mathsf{F}'_2 \subseteq \bar{\mathsf{F}}_2,\\ \phi\colon \mathsf{F}'_1 \to \mathsf{F}'_2 \text{ an isomorphism}\}.$$

Since $(\mathsf{F}, \mathsf{F}, \mathrm{id}_{\mathsf{F}}) \in \mathscr{C}$, this set is nonempty. Let us partially order $\mathscr{C}$ by

$$(\mathsf{F}'_1, \mathsf{F}'_2, \phi) \preceq (\mathsf{E}'_1, \mathsf{E}'_2, \psi), \quad \iff \quad \mathsf{F}'_1 \subseteq \mathsf{E}'_1,\ \mathsf{F}'_2 \subseteq \mathsf{E}'_2,\ \phi(\mathsf{F}'_1) \subseteq \psi(\mathsf{E}'_1).$$

Let $\{(\mathsf{F}'_{1,a}, \mathsf{F}'_{2,a}, \phi_a)\}_{a \in A}$ be a totally ordered subset of $\mathscr{C}$. Define

$$\mathsf{E}_1 = \cup_{a \in A} \mathsf{F}'_{1,a}, \quad \mathsf{E}_2 = \cup_{a \in A} \mathsf{F}'_{2,a},$$

and define $\psi\colon \mathsf{E}_1 \to \mathsf{E}_2$ by asking that $\psi|\mathsf{F}_{1,a} = \phi_a$. It is easy to see that ψ is a bijection. Moreover, E_1 and E_2 are fields with addition and multiplication defined by using that in the sets $\mathsf{F}_{1,a}$ and $\mathsf{F}_{2,a}$, $a \in A$ (cf. the field operations on K defined above). Thus $(\mathsf{E}_1, \mathsf{E}_2, \psi)$ is an upper bound for $\{(\mathsf{F}'_{1,a}, \mathsf{F}'_{2,a}, \phi_a)\}_{a \in A}$. By Zorn's Lemma there exists a maximal element $(\mathsf{F}_1, \mathsf{F}_2, \phi)$ of $\mathscr{C}$. We claim that $\mathsf{F}_1 = \bar{\mathsf{F}}_1$. If not, there exists $a_1 \in \bar{\mathsf{F}}_1 \setminus \mathsf{F}_1$. Since $\bar{\mathsf{F}}_1$ is algebraic over F so is a_1. Let $M \in \mathsf{F}[\xi]$ be the minimal polynomial for a_1. We claim that M has a root in $\bar{\mathsf{F}}_2 \setminus \mathsf{F}_2$. If not, since ϕ is an isomorphism of F_1 and F_2, this would imply that M has all roots in F_1. But this cannot be, since a_1 is a root of M. Thus let $a_2 \in \bar{\mathsf{F}}_2 \setminus \mathsf{F}_2$ be a root of M. Note that M is the minimal polynomial of a_2. By Proposition 4.6.17 the vector spaces $\mathsf{F}_1(a_1)$ and $\mathsf{F}_2(a_2)$ over F_1 and F_2 have bases

$$\{1_{\mathsf{F}_1}, a_1, \dots, a_1^k\}, \quad \{1_{\mathsf{F}_2}, a_2, \dots, a_2^k\},$$

where $k = \deg(M)$. Thus the isomorphism ϕ extends in a natural way to an isomorphism ψ of $\mathsf{F}_1(a_1)$ and $\mathsf{F}_2(a_2)$. Then we have $(\mathsf{F}_1, \mathsf{F}_2, \phi) \prec (\mathsf{F}_1(a_1), \mathsf{F}_2(a_2), \psi)$, contradicting the maximality of $(\mathsf{F}_1, \mathsf{F}_2, \phi)$. Thus $\mathsf{F}_1 = \bar{\mathsf{F}}_1$. In a similar manner one shows that $\mathsf{F}_2 = \bar{\mathsf{F}}_2$. Since ϕ is an isomorphism of F_1 and F_2, the theorem follows. ■

4.6.5 Splitting fields

In the preceding section we saw that a distinguished rôle is played by extensions whose elements are roots of polynomials. In this section we study systematically the notion of constructing a field which contains *all* roots of a certain polynomial. The key idea in this construction is to quotient the polynomial ring $\mathsf{F}[\xi]$ by an ideal which was considered in Section 4.4.6.

To get started with the programme, the following definition provides the useful terminology.

4.6.24 Definition (Splits) If F is a field and if K is an extension of F, a polynomial $A \in \mathsf{F}[\xi]$ ***splits*** over K if there exists $a, r_1, \dots, r_k \in \mathsf{K}$ such that

$$A = a(\xi - r_1)\cdots(\xi - r_k). \qquad \bullet$$

A field K is a ***splitting field*** for $A \in \mathsf{F}[\xi]$ if A splits over K and but does not split over any proper subfield of K that is itself an extension of F. •

The idea of a splitting field for A is that it is the smallest field extension of F in which A is guaranteed to have all of its roots.

Theorem 4.4.44 gives us, for an irreducible monic polynomial $A \in \mathsf{F}[\xi]$ of positive degree, a means of constructing a field $\mathsf{F}[\xi]/(A)$ that contains a subfield isomorphic to F. If we identify F with its isomorphic image in $\mathsf{F}[\xi]/(A)$, then we can regard the field $\mathsf{F}[\xi]/(A)$ as an extension of F. We now turn to the properties of this extension, and ideas stemming from this. In the statement of the result we understand that, if K is a subfield of F, then the $\mathsf{K}[\xi]$ is naturally a subring of $\mathsf{F}[\xi]$ (cf. Proposition 4.4.8).

4.6.25 Proposition (A has a root in $\mathsf{F}[\xi]/(\mathrm{A})$) *Let F be a field, let $\mathrm{A} \in \mathsf{F}[\xi]$ be an irreducible monic polynomial of degree at least 1, and let $\mathsf{F}[\xi]/(\mathrm{A})$ be the corresponding quotient ring, thought of as a field extension of F. Then A has a root in $\mathsf{F}[\xi]/(\mathrm{A})$.*

Proof Define $r_0 = \xi + (A) \in \mathsf{F}[\xi]/(A)$. Then, writing $A = \xi^k - \sum_{j=0}^{k-1} a_j\xi^j$, we compute

$$r_0^k - \sum_{j=0}^{k-1} a_j r_0^j = \xi^k - \sum_{j=0}^{k-1} a_j\xi^j + (A) = A + (A) = 0_{\mathsf{F}[\xi]} + (A).$$

Thus, if we think of A as a polynomial in $(\mathsf{F}[\xi]/(A))[\eta]$, we have $\mathrm{Ev}_{\mathsf{F}[\xi]/(A)}(A)(r_0) = 0_{\mathsf{F}[\xi]/(A)}$, and so r_0 is a root of A in $\mathsf{F}[\xi]/(A)$, as desired. ■

Now we can prove that polynomials possess splitting fields. If a polynomial splits in some field, then it is irreducible in that field if and only if it has degree 1. With this idea, we now have the following important theorem.

4.6.26 Theorem (Existence and uniqueness of splitting fields) *If F is a field and if $\mathrm{A} \in \mathsf{F}[\xi]$ is a polynomial of degree at least 1, then there exists a field extension $\mathsf{F}(\mathrm{A})$ of F such that A splits in $\mathsf{F}(\mathrm{A})$. Moreover, A possesses a splitting field, and any two splitting fields for A are isomorphic via an F-isomorphism.*

Proof If A splits over F then the result follows trivially by taking $\mathsf{F}(A) = \mathsf{F}$. So suppose that A does not split in F, and denote $\mathsf{F}_0 = \mathsf{F}$ and $A_0 = A$. Construct a field F_1 and a polynomial $A_1 \in \mathsf{F}_1[\xi]$ as follows. Since A_0 does not split over A_0, we can write $A_0 = B_0 \cdot C_0$ where B_0 is monic, irreducible, and of degree at least 2. Let $\mathsf{F}_1 = \mathsf{F}_0[\xi]/(B_0)$, and note that B_0, and therefore A_0, has a root, say r_0, in F_1. Therefore, thinking of A_0 as a polynomial over the extension field F_1 of F_1, we can write $A_0 = (\xi - r_0) \cdot A_1$ for some $A_1 \in \mathsf{F}_1[\xi]$ by Proposition 4.4.25. If A_1 splits in F_1, then the result follows by taking $\mathsf{F}(A) = \mathsf{F}_1$. Otherwise, the construction can be repeated to define a field F_2 which is an extension of F_1, and so also of F_0. Note that A_1 has a root in F_2 so that we have

$A_0 = (\xi - r_0) \cdot (\xi - r_1) \cdot A_2$, so defining $A_2 \in \mathsf{F}_2[\xi]$. We claim that by continuing this process at most $k = \deg(A)$ times, we arrive at a field F_k in which A_0 splits. This is trivial since if the construction is carried out k times we arrive at an expression for A_0 of the form

$$A_0 = (\xi - r_0) \cdot (\xi - r_1) \cdots (\xi - r_k) \cdot A_{k+1},$$

which implies by Proposition 4.4.11 that $\deg(A_{k+1}) = 0$ and so A_{k+1} is a nonzero constant. Thus A_j splits over F_j for some $j \in \{0, 1, \dots, k\}$, and the result follows by taking $\mathsf{F}(A) = \mathsf{F}_j$.

Now let us prove the second assertion. The field $\mathsf{F}(A)$ constructed in the first part of the proof contains all roots of A. Let us denote the distinct roots by $r_1, \dots, r_k$. Now consider the subfield $\mathsf{K} = \mathsf{F}(r_1, \dots, r_k)$ of $\mathsf{F}(A)$. Clearly A splits in K. Moreover, any proper subfield of K that extends F cannot contain the roots $r_1, \dots, r_k$, since K is the smallest field extension containing these roots.

To show that two splitting fields are isomorphic, we first prove a lemma.

1 **Lemma** *Let F_1 and F_2 be fields with $\phi\colon \mathsf{F}_1 \to \mathsf{F}_2$ an isomorphism. Let $\mathrm{A} \in \mathsf{F}_1[\xi]$ be given by*

$$\mathrm{A} = \mathrm{a}_\mathrm{k}\xi^\mathrm{k} + \cdots + \mathrm{a}_1\xi + \mathrm{a}_0$$

and define $\phi_*\mathrm{A} \in \mathsf{F}_2[\xi]$ by

$$\phi_*\mathrm{A} = \phi(\mathrm{a}_\mathrm{k})\xi^\mathrm{k} + \cdots + \phi(\mathrm{a}_1)\xi + \phi(\mathrm{a}_0).$$

Let K_1 be a splitting field for A and K_2 be a splitting field for $\phi_*\mathrm{A}$. Then there exists an isomorphism $\psi\colon \mathsf{K}_1 \to \mathsf{K}_2$ such that $\psi|\mathsf{F} = \phi$.

Proof We prove the result by induction on $m = [\mathsf{K}_1 : \mathsf{F}_1]$. For $m = 1$ we have $\mathsf{K}_1 = \mathsf{F}_1$ and so A splits in F_1. It then holds that ϕ_*A splits in F_2 and so $\mathsf{K}_2 = \mathsf{F}_2$. The lemma then holds taking $\psi = \phi$. Now suppose that $[\mathsf{K}_1 : \mathsf{F}_1] \geq 2$. Thus there exists an irreducible polynomial P of degree at least 2 which divides A in $\mathsf{F}_1[\xi]$. Let $r_1 \in \mathsf{K}_1$ be a root of P. Note that ϕ_*P must divide ϕ_*A, and, therefore, ϕ_*P has a root $r_2 \in \mathsf{K}_2$. By Proposition 4.6.21 it follows that there exists an isomorphism $\psi'\colon \mathsf{F}_1(r_1) \to \mathsf{F}_2(r_2)$ mapping r_1 to r_2 and restricting to ϕ on F_1. Note that K_1 is a splitting field of A over $\mathsf{F}_1(r_1)$ and that K_2 is a splitting field of ϕ_*A over $\mathsf{F}_2(r_2)$. Since

$$[\mathsf{K}_1 : \mathsf{F}_1] = [\mathsf{K}_1 : \mathsf{F}_1(r_1)][\mathsf{F}_1(r_1) : \mathsf{F}_1]$$

by Proposition 4.6.4 and since $[\mathsf{F}_1(r_1) : \mathsf{F}_1] \geq 2$ by Proposition 4.6.17, it follows that $[\mathsf{K}_1 : \mathsf{F}_1(r_1)] < [\mathsf{K}_1 : \mathsf{F}_1]$. By the induction hypothesis there exists an isomorphism $\psi''\colon \mathsf{K}_1 \to \mathsf{K}_2$ which agrees with ψ' on $\mathsf{F}_1(r_1)$. Since ψ' agrees with ϕ on F_1, it follows that ψ'' agrees with ϕ on F_1. ▼

The theorem now follows by applying the lemma to the case when $\mathsf{F}_1 = \mathsf{F}_2 = \mathsf{F}$ and $\phi = \mathrm{id}_\mathsf{F}$. ■

In the first part of the theorem we construct a field which contains all roots of A. This could be achieved by merely taking the algebraic closure. However, we give a simpler, more direct, proof of this fact.

Let us look at some simple examples to illustrate the notion of a splitting field.

4.6.27 Examples (Splitting field)

1. We consider the field $\mathbb{Q}$ and the polynomial $A = \xi^2 - 2$. In Example 4.6.7–1 we constructed the field $\mathbb{Q}(\sqrt{2})$ which contains one of the roots, $\sqrt{2}$, of A. Moreover, because of (4.23) it also follows that $-\sqrt{2} \in \mathbb{Q}(\sqrt{2})$, and so $\mathbb{Q}(\sqrt{2})$ contains all roots of A. Moreover, because $\mathbb{Q}(\sqrt{2})$ is, by definition, the smallest field containing $\sqrt{2}$, it follows that $\mathbb{Q}(\sqrt{2})$ is a splitting field for A.
2. We again take the field $\mathbb{Q}$, but now consider the polynomial $A = \xi^4 - 2$. Any splitting field for A in $\mathbb{R}$ must contain $\sqrt[4]{2}$. Since A is irreducible in $\mathbb{Q}[\xi]$ it follows from Proposition 4.6.17 that

$$\mathbb{Q}(\sqrt[4]{2}) = \{\alpha_3 2^{3/4} + \alpha_2 2^{1/2} + \alpha_1 2^{1/4} + \alpha_0 \mid \alpha_0, \alpha_1, \alpha_2, \alpha_3 \in \mathbb{Q}\},$$

as asserted in Example 4.6.7–2. From this it follows that $-\sqrt[4]{2} \in \mathbb{Q}(\sqrt[4]{2})$. However, $\mathbb{Q}(\sqrt[4]{2})$ is *not* a splitting field for A since, in $\mathbb{Q}(\sqrt[4]{2})$ the prime factorisation of A is

$$A = (\xi - \sqrt[4]{2})(\xi + \sqrt[4]{2})(\xi^2 + \sqrt{2}),$$

i.e., A does not split in $\mathbb{Q}(\sqrt[4]{2})$. However, a splitting field for A as a subfield of $\mathbb{C}$ is $\mathbb{Q}(\sqrt[4]{2}, \mathrm{i}\sqrt[4]{2})$. •

Let us give a result concerning automorphisms of field extensions and roots of polynomials.

4.6.28 Proposition (F-automorphisms and roots) *Let F be a field, let $\mathrm{A} \in \mathsf{F}[\xi]$, and let K be an extension of F that contains a splitting field for A. Denote the distinct roots of A by $\{\mathrm{r}_1, \dots, \mathrm{r}_k\}$. If $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ then there exists $\sigma \in \mathfrak{S}_k$ such that $\phi(\mathrm{r}_j) = \mathrm{r}_{\sigma(j)}$, $j \in \{1, \dots, k\}$.*

Proof Since $\phi|\mathsf{F} = \mathrm{id}_{\mathsf{F}}$ and since ϕ is a field homomorphism we have

$$\mathrm{Ev}_{\mathsf{K}}(A)(\phi(r_j)) = \phi(\mathrm{Ev}_{\mathsf{K}}(r_j)) = 0_{\mathsf{K}},$$

so that $\phi(r_j)$ is a root of A for each $j \in \{1, \dots, k\}$. ■

One can also talk about extensions in which not one polynomial, but all polynomials in a family, split.

4.6.29 Definition (Splitting field for a family of polynomials) Let F be a field and let $(A_i)_{i \in I}$ be a family of polynomials in $\mathsf{F}[\xi]$. A ***splitting field*** for the family is a field extension K of F such that each polynomial A_i, $i \in I$, splits in K, but such that if K' is any proper subfield of K, there exists some polynomial A_{i_0} in the family that does not split in K'.

The following result essentially follows from the definitions of all the notions involved.

4.6.30 Theorem (Existence of splitting field for families of polynomials) *If* F *is a field and if* $(\mathrm{A_i})_{\mathrm{i\in I}}$ *is a family of polynomials in* $\mathsf{F}[\xi]$*, then there exists a splitting field for the family.*

Proof In the algebraic closure F let R denote the set of roots of all polynomials in $(A_i)_{i\in I}$. It is clear that $\mathsf{F}(R)$ is then a field in which all polynomials in $(A_i)_{i\in I}$ split. Moreover, it is also clear by definition of $\mathsf{F}(R)$ that any proper subfield of $\mathsf{F}(R)$ must omit some element of R, precluding all polynomials in $(A_i)_{i\in I}$ from splitting. ■

4.6.6 Normal field extensions

The notion of a normal field extension is a fairly simple one.

4.6.31 Definition (Normal extension) An extension K of a field F is ***normal*** if, for every irreducible polynomial $A \in \mathsf{F}[\xi]$ having a root in K, it holds that A splits in K. •

The idea is that if K is big enough to contain *some* root of a polynomial, to be normal it must be big enough to contain *all* roots of the polynomial. Let us illustrate this idea with some examples.

4.6.32 Examples (Normal extension)

1. The extension $\mathbb{Q}(\sqrt{2})$ of $\mathbb{Q}$ is normal, as we now show. We recall that
$$\mathbb{Q}(\sqrt{2}) = \{\alpha_1\sqrt{2} + \alpha_0 \mid \alpha_0, \alpha_1 \in \mathbb{Q}\}.$$
Let $a = \alpha_1\sqrt{2} + \alpha_0 \in \mathbb{Q}(\sqrt{2})$ and note that a is a root of the polynomial
$$\xi^2 - 2\alpha_0\xi + \alpha_0^2 - 2\alpha_1,$$
as can be verified directly. Note that the other root of this polynomial in $\mathbb{R}$ is $-\alpha_1\sqrt{2} + \alpha_0$. Since this number is in $\mathbb{Q}(\sqrt{2})$, it follows that $\mathbb{Q}(\sqrt{2})$ is normal.
While we have shown "by hand" that $\mathbb{Q}(\sqrt{2})$ is normal, we shall see in Proposition 4.6.36 that this actually follows since $\mathbb{Q}(\sqrt{2})$ is a splitting field for $\xi^2 - 2$.
2. The extension $\mathbb{Q}(\sqrt[4]{2})$ is not normal since the minimal polynomial of $\sqrt[4]{2}$ is $\xi^4 - 2$, which has a root in $\mathbb{Q}(\sqrt[4]{2})$ but does not split. •

To provide a nice characterisation of normal extensions we introduce the following notion which generalises complex conjugation.

4.6.33 Definition (Conjugate extension, conjugate elements) Let F be a field with algebraic closure $\bar{\mathsf{F}}$, let K and L be extensions of F contained in $\bar{\mathsf{F}}$, and let $a, b \in \bar{\mathsf{F}}$.

(i) The extensions K and L are ***conjugate*** if there exists $\phi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$ such that $\phi(\mathsf{K}) = \mathsf{L}$.

(ii) The elements a and b are ***conjugate*** if there exists $\phi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$ such that $\phi(a) = b$. •

The following result gives some useful and insightful characterisations of conjugate elements.

4.6.34 Proposition (Characterisations of conjugate elements) *Let* F *be a field with algebraic closure* $\bar{\mathsf{F}}$ *and let* $a, b \in \bar{\mathsf{F}}$. *Then the following statements are equivalent:*

(i) a *and* b *are conjugate;*

(ii) *there exists an* F*-isomorphism* $\phi\colon \mathsf{F}(a) \to \mathsf{F}(b)$ *such that* $\phi(a) = b$*;*

(iii) $M_a = M_b$.

Proof (i) $\implies$ (iii) Let $\psi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$ map a to b and note that

$$\mathrm{Ev}_{\bar{\mathsf{F}}}(M_a)(b) = \mathrm{Ev}_{\bar{\mathsf{F}}}(M_a)(\psi(a)) = \psi(\mathrm{Ev}_{\bar{\mathsf{F}}}(M_a)(a)) = 0_{\bar{\mathsf{F}}},$$

using the fact that ψ is an F-homomorphism, and so fixes F. Thus M_a is a monic irreducible polynomial possessing b as a root. In other words, $M_a = M_b$.

(iii) $\implies$ (ii) Here we apply Proposition 4.6.21 with $\mathsf{F}_1 = \mathsf{F}_2 = \mathsf{F}$ and $\phi = \mathrm{id}_{\mathsf{F}}$.

(ii) $\implies$ (i) Note that $\bar{\mathsf{F}}$ is an algebraic closure of $\mathsf{F}(a)$ and of $\phi(\mathsf{F}(a))$. Thus by the uniqueness part of Theorem 4.6.23 there exists an isomorphism $\Phi\colon \bar{\mathsf{F}} \to \bar{\mathsf{F}}$ such that $\Phi|\mathsf{F}(a) = \phi$. In particular, $\Phi|\mathsf{F} = \mathrm{id}_{\mathsf{F}}$, and so Φ is an F-isomorphism. ■

One of the important consequences of this characterisation of conjugate elements is the following.

4.6.35 Corollary (Conjugate elements and roots) *Let* F *be a field with algebraic closure* $\bar{\mathsf{F}}$ *and let* $a \in \bar{\mathsf{F}}$ *have minimal polynomial* M_a. *Then the set of conjugates of* a *is equal to the set of roots of* M_a.

Proof By Proposition 4.6.28 it follows that the set of roots of M_a are conjugates of a. Conversely, if b is a conjugate of a, then by Proposition 4.6.34 it holds that $M_b = M_a$, and so b is also a root of M_a. ■

The following result gives equivalent characterisations for normality of an extension, some of which involve the notion of conjugation.

4.6.36 Proposition (Characterisation of normal extensions) *Let* F *be a field with algebraic closure* $\bar{\mathsf{F}}$. *For an extension* $\mathsf{K} \subseteq \bar{\mathsf{F}}$ *of* F *the following statements are equivalent:*

(i) K *is normal;*

(ii) *if* $a \in \mathsf{K}$ *then every conjugate of* a *is also in* K*;*

(iii) *if* $\phi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$ *then* $\phi(\mathsf{K}) = \mathsf{K}$*;*

(iv) *if* $\psi\colon \mathsf{K} \to \bar{\mathsf{F}}$ *is an* F*-homomorphism then* $\psi(\mathsf{K}) = \mathsf{K}$*;*

(v) K *is the splitting field for a family of nonconstant polynomials in* $\mathsf{F}[\xi]$.

Proof (iii) $\implies$ (iv) Let $\psi\colon \mathsf{K} \to \bar{\mathsf{F}}$ be an F-homomorphism. Since $\bar{\mathsf{F}}$ is an algebraic closure of both K and $\psi(\mathsf{K})$, the uniqueness part of Theorem 4.6.23 gives the existence of an isomorphism $\Psi\colon \bar{\mathsf{F}} \to \bar{\mathsf{F}}$ such that $\Psi|\mathsf{K} = \psi$. Thus $\psi(\mathsf{K}) = \Psi(\mathsf{K}) = \mathsf{K}$.

(iv) $\implies$ (iii) Let $\phi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$ so that $\phi|\mathsf{K}$ is an F-homomorphism of K into $\bar{\mathsf{F}}$. Then $\phi(\mathsf{K}) = \mathsf{K}$, as desired.

(i) $\implies$ (v) Let $(M_a)_{a\in\mathsf{K}}$ be the family of minimal polynomials for elements of K. We claim that K is the splitting field in $\bar{\mathsf{F}}$ for $(M_a)_{a\in\mathsf{K}}$. Let us denote this splitting field by K'. If $a \in \mathsf{K}$ then a is a root of M_a and so $a \in \mathsf{K}'$. Conversely, let $a \in \mathsf{K}'$. Then a is a root of $M_{a'}$ for some $a' \in \mathsf{K}$. Since K is normal, it follows that $a \in \mathsf{K}$. Thus $\mathsf{K} = \mathsf{K}'$, as desired.

(v) $\implies$ (iii) Suppose that K is the splitting field for the family $(A_i)_{i\in I}$ of polynomials in $\mathsf{F}[\xi]$. Let $\phi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$. For $i \in I$, ϕ permutes the roots of A_i by Proposition 4.6.28. Therefore, since K is the smallest field containing F and all roots of all polynomials $(A_i)_{i\in I}$ by Theorem 4.6.30, it follows that $\phi(\mathsf{K}) = \mathsf{K}$.

(iii) $\implies$ (ii) This is simply the definition of what it means for elements to be conjugate.

(ii) $\implies$ (i) Let $A \in \mathsf{F}[\xi]$ be monic, irreducible, and with a root in K. In $\bar{\mathsf{F}}$ we have

$$A = (\xi - r_1)\cdots(\xi - r_k).$$

Since $M_{r_1} = \cdots = M_{r_k} = A$ it follows from Proposition 4.6.34 that $r_1, \dots, r_k$ are conjugate. Thus $r_1, \dots, r_k \in \mathsf{K}$, and A splits in K, as desired. ■

Let us illustrate all of this with some examples.

4.6.37 Examples (Characterisation of normal extensions)

1. The extension $\mathbb{Q}(\sqrt{2})$ of $\mathbb{Q}$ is normal. As we showed in Example 4.6.32–1 above (also recalling Example 4.6.10–1), the conjugate of $\alpha_1\sqrt{2} + \alpha_0 \in \mathbb{Q}(\sqrt{2})$ is $-\alpha_1\sqrt{2} + \alpha_0$, which is also in $\mathbb{Q}(\sqrt{2})$. In like manner we see that $\mathbb{Q}(\sqrt{2})$ is closed under the group of $\mathbb{Q}$-automorphisms of any algebraic closure of $\mathbb{Q}$. Also, the normality of $\mathbb{Q}(\sqrt{2})$ follows from the fact that it is a splitting field for $\xi^2 - 2$.
2. The extension $\mathbb{Q}(\sqrt[4]{2})$ of $\mathbb{Q}$ is not normal, and this is reflected, for example, by the fact that $\mathbb{Q}(\sqrt[4]{2})$ is not the splitting field for the polynomial $\xi^4 - 2$ (noting that this is the minimal polynomial for $\sqrt[4]{2} \in \mathbb{Q}(\sqrt[4]{2})$).

If one has a field extension that is not normal, it is possible to produce one that is normal in a natural way.

4.6.38 Definition (Normal closure) Let F be a field with algebraic closure $\bar{\mathsf{F}}$. If $\mathsf{K} \subseteq \bar{\mathsf{F}}$ is an extension of F, the ***normal closure*** of K is the extension $N(\mathsf{K})$ generated by

$$\{a \in \bar{\mathsf{F}} \mid a \text{ is conjugate to some element of } \mathsf{K}\}. \qquad \bullet$$

The normal closure is a field extension of F, as we show in the following result.

4.6.39 Proposition (Properties of the normal closure) *If F is a field with algebraic closure $\bar{\mathsf{F}}$ and if $\mathsf{K} \subseteq \bar{\mathsf{F}}$ is an extension of F, then the following statements hold:*

(i) $\mathrm{N}(\mathsf{K})$ *is a normal field extension of* F*;*

(ii) *if* $\mathsf{K}' \subset \mathrm{N}(\mathsf{K})$ *is a field extension then* K' *is not normal;*

(iii) $[\mathrm{N}(\mathsf{K}) : \mathsf{F}]$ *is finite if and only if* $[\mathsf{K} : \mathsf{F}]$ *is finite.*

Proof (i) Note that $N(\mathsf{K})$ is, by definition, the splitting field of $(M_a)_{a\in\mathsf{K}}$, and so is a normal field extension of F by Proposition 4.6.36.

(ii) As we indicated in the proof of part (i), $N(\mathsf{K})$ is the splitting field for the set of minimal polynomials of elements in K. Therefore, any proper subfield of $N(\mathsf{K})$ will necessarily exclude a root of some minimal polynomial M_a for $a \in \mathsf{K}$. Thus such a proper subfield cannot be normal.

(iii) It is clear from Proposition 4.6.4 that $[\mathsf{K} : \mathsf{F}]$ is finite if $[N(\mathsf{K}) : \mathsf{F}]$ is finite. Conversely, if $[\mathsf{K} : \mathsf{F}]$ is finite then let $\{a_1, \dots, a_k\}$ be a basis for K over F. Then $N(\mathsf{K})$ is the splitting field for $(M_{a_1}, \dots, M_{a_k})$. By Exercise 4.6.6 it follows that $[N(\mathsf{K}) : \mathsf{F}]$ is finite. ■

We can easily illustrate the notion of normal closure with an example.

4.6.40 Example (Normal closure) The normal closure of $\mathbb{Q}(\sqrt[4]{2})$ is $\mathbb{Q}(\sqrt[4]{2}, \mathrm{i}\sqrt[4]{2})$. Indeed, $\mathbb{Q}(\sqrt[4]{2}, \mathrm{i}\sqrt[4]{2})$ contains all roots of the minimal polynomial $M_{\sqrt[4]{2}}$. •

4.6.7 Separable polynomials and field extensions

In Sections 5.8.12 and 5.8.14 we shall encounter the need for polynomials with the property that, as polynomials over the algebraic closure of the field, they factor into a product of *distinct* degree one polynomials. This is the following definition.

4.6.41 Definition (Separable polynomial) Let F be a field with $\bar{\mathsf{F}}$ its algebraic closure. For $A \in \mathsf{F}[\xi]$, denote by $\bar{A} \in \bar{\mathsf{F}}[\xi]$ the image of A by the inclusion $\mathsf{F}[\xi] \subseteq \bar{\mathsf{F}}[\xi]$. A polynomial $A \in \mathsf{F}[\xi]$ is ***separable*** if we have

$$\bar{A} = a(\xi - b_1) \cdots (\xi - b_k)$$

for $a, b_1, \dots, b_k \in \bar{\mathsf{F}}$ with $r_1, \dots, r_k$ distinct. •

Let us agree that throughout this section we shall denote by $\bar{A} \in \bar{\mathsf{F}}[\xi]$ the image of $A \in \mathsf{F}[\xi]$ under the inclusion $\mathsf{F}[\xi] \subseteq \bar{\mathsf{F}}[\xi]$.

The following result characterises separable polynomials. We recall from Definition 4.4.28 the definition of the formal derivative A' of a polynomial A.

4.6.42 Proposition (Characterisation of separable polynomials) *Let* F *be a field. For* $\mathrm{A} \in \mathsf{F}[\xi]$ *the following statements are equivalent:*

(i) A *is separable;*

(ii) there exists an extension K *of* F *such that, as a polynomial in* $\mathsf{K}[\xi]$, A *has the form*

$$\mathrm{a}(\xi - \mathrm{b}_1) \cdots (\xi - \mathrm{b}_\mathrm{k})$$

for $\mathrm{a}, \mathrm{b}_1, \dots, \mathrm{b}_\mathrm{k} \in \mathsf{K}$ *with* $\mathrm{b}_1, \dots, \mathrm{b}_\mathrm{k}$ *distinct;*

(iii) A *and* A' *are coprime.*

Proof (i) $\implies$ (ii) Take $\mathsf{K} = \bar{\mathsf{F}}$, the algebraic closure of F.

(ii) $\implies$ (iii) Suppose that A and A' are not coprime and let $D \in \mathsf{F}[\xi]$ be a greatest common divisor of A and A' of positive degree. Since $\bar{D}|\bar{A}$ and $\bar{D}|\bar{A}'$, and since $\bar{\mathsf{F}}$ is algebraically closed, it follows that $\bar{A}$ and $\bar{A}'$ have a common root, and so $\bar{A}$ has a root of multiplicity greater than 1 by Proposition 4.4.29. If K is any extension in which A splits, then we have two cases: (1) $\mathsf{K} \subseteq \bar{\mathsf{F}}$ and (2) $\mathsf{K} \supseteq \bar{\mathsf{F}}$. In the first case we have A, as a polynomial in $\mathsf{K}[\xi]$, in the form

$$a(\xi - b_1) \cdots (\xi - b_k)$$

for $a, b_1, \dots, b_k \in \mathsf{K}$. Since $\mathsf{K} \subseteq \bar{\mathsf{F}}$ this is also the form of $\bar{A}$. Thus the roots of A in K have multiplicity greater than 1. Now, if $\mathsf{K} \supseteq \bar{\mathsf{F}}$, we write

$$\bar{A} = a(\xi - b_1) \cdots (\xi - b_k)$$

for $a, b_1, \dots, b_k \in \bar{\mathsf{F}}$. However, since $\bar{\mathsf{F}} \subseteq \mathsf{K}$ this is also the form of A in $\mathsf{K}[\xi]$. Thus the roots of A in K have multiplicity greater than 1. Thus there can be no extension of F in which A splits and the roots of A are of multiplicity 1.

(iii) $\Longrightarrow$ (i) If A and A' are coprime, then by Corollary 4.4.36 there exists $B, C \in \mathsf{F}[\xi]$ such that $BA + CA' = 1_{\mathsf{F}}$. Suppose that $b \in \bar{\mathsf{F}}$ is a root of $\bar{A}$. Then

$$1_{\bar{\mathsf{F}}} = \mathrm{Ev}_{\bar{\mathsf{F}}}(\bar{B}\bar{A} + \bar{C}\bar{A}')(b) = \mathrm{Ev}_{\bar{\mathsf{F}}}(\bar{C})(b)\mathrm{Ev}_{\bar{\mathsf{F}}}(\bar{A}')(b),$$

from which we deduce that b is not a root for $\bar{A}'$. Thus the roots of $\bar{A}$ have multiplicity 1 by Proposition 4.4.29. ■

In order to get a sufficiently fine understanding of separable polynomials, it is helpful to understand *irreducible* separable polynomials. The following result indicates that only exceptional irreducible polynomials are not separable.

4.6.43 Proposition (Irreducible separable polynomials) *Let F be a field and let $\mathrm{A} \in \mathsf{F}[\xi]$ be irreducible. Then the following statements are equivalent:*

(i) A is separable;

(ii) there exists an extension K of F in which A has a root of multiplicity 1;

(iii) $\mathrm{A}' \neq 0_{\mathsf{F}[\xi]}$;

(iv) either

(a) F has characteristic zero or

(b) F has characteristic $\mathrm{p} \in \mathbb{Z}_{>0}$ and A does not have the form

$$\mathrm{a_k}\xi^{\mathrm{kp}} + \cdots + \mathrm{a_1}\xi^{\mathrm{p}} + \mathrm{a_0},$$

for $\mathrm{a_0, a_1, \dots, a_k} \in \mathsf{F}$.

Proof (i) $\Longrightarrow$ (ii) Take K to be the algebraic closure of F and use Proposition 4.6.42.

(ii) $\Longrightarrow$ (iii) If A has a root $b \in \mathsf{K}$ of multiplicity 1 then $\mathrm{Ev}_{\mathsf{K}}(A')(b) \neq 0_{\mathsf{K}}$. In particular, $A' \neq 0_{\mathsf{F}[\xi]}$.

(iii) $\Longrightarrow$ (iv) We prove the contrapositive. Thus we assume that F does not have characteristic zero and that A is given by

$$A = a_k\xi^{kp} + \cdots + a_1\xi^p + a_0$$

with $p \in \mathbb{Z}_{>0}$ the characteristic of F. Then

$$A' = kpa_k\xi^{kp-1} + \cdots + 2pa_2\xi^{2p-1} + pa_1\xi^{p-1},$$

giving $A' = 0_{\mathsf{F}[\xi]}$ since $p\xi^j = 0_{\mathsf{F}[\xi]}$ for any $j \in \mathbb{Z}_{\geq 0}$.

(iv) $\implies$ (i) Suppose that F has characteristic zero. Since A is irreducible, $\deg(A) \geq 1$. Thus write

$$A = a_k\xi^k + \cdots + a_1\xi + a_0$$

with $k \in \mathbb{Z}_{>0}$ and with $a_k \neq 0_{\mathsf{F}}$. Then

$$A' = ka_k\xi^{k-1} + \cdots + 2a_2\xi + a_1.$$

Since F has characteristic zero, $ka_k\xi^{k-1} \neq 0_{\mathsf{F}[\xi]}$, and so $A' \neq 0_{\mathsf{F}[\xi]}$.

Now suppose that F has characteristic $p \in \mathbb{Z}_{>0}$ and write

$$A = a_k\xi^k + \cdots + a_1\xi + a_0$$

as above, supposing that $a_j \neq 0_{\mathsf{F}}$ for some j such that $p \nmid j$. Then, as above,

$$A' = ka_k\xi^{k-1} + \cdots + 2a_2\xi + a_1.$$

Since F has characteristic p, the expression $am\xi^j$ is zero if and only if $p \nmid m$. Thus $A' \neq 0_{\mathsf{F}[\xi]}$.

(iii) $\implies$ (i) Suppose that A is irreducible but not separable. Let K be an extension of F such that A splits in $\mathsf{K}[\xi]$. Since A is not separable there exists a root $b \in \mathsf{K}$ of A of multiplicity at least 2. Let M_b be the minimal polynomial of b over F. We claim that $M_b|A$ and $M_b|A'$. Since b is a root of multiplicity greater than 1 it follows from Proposition 4.4.29 that b is a root of both A and A'. From the definition of M_b, M_b divides A and A', as desired. Since $M_b|A$ and since A is irreducible, $A = \alpha M_b$ for $\alpha \in \mathsf{F}$. Therefore, $A|A'$. Since $\deg(A) > \deg(A')$, it follows that $A' = 0_{\mathsf{F}[\xi]}$. ■

Corresponding to the notion of a separable polynomial is that of a separable field extension.

4.6.44 Definition (Separable extension) An extension K of a field F is ***separable*** if, for every $a \in \mathsf{K}$, there exists a separable polynomial $A \in \mathsf{F}[\xi]$ with a as a root in K. •

Let us characterise separable field extensions.

4.6.45 Proposition (Characterisation of separable field extensions) *Let F be a field with K an algebraic extension of F. Then the following statements are equivalent:*

(i) K is separable;

(ii) the minimal polynomial $\mathrm{M_a}$ is separable for each $\mathrm{a} \in \mathsf{K}$;

(iii) for each $\mathrm{a} \in \mathsf{K}$, a is a root of multiplicity 1 of its minimal polynomial $\mathrm{M_a}$.

Proof (i) $\implies$ (ii) Let $a \in \mathsf{K}$. Then there exists a separable polynomial A possessing a as a root. Therefore, $M_a|A$ by definition of the minimal polynomial. If A is separable then the roots of A in $\bar{\mathsf{F}}$ all have multiplicity 1. This must then also be true of M_a, and so M_a is separable.

(ii) $\implies$ (iii) This is simply the definition of M_a being separable.

(iii) $\implies$ (i) Let $a \in \mathsf{K}$. If a is a root of multiplicity 1 of M_a then M_a is separable by Proposition 4.6.43. Therefore K is separable since a is a root of M_a. ■

The following relationship between separability and normality will be useful for us.

4.6.46 Corollary (Normal closure of a separable extension is separable) *If* K *is a separable extension of a field* F *then the normal closure of* K *is separable.*

Proof This follows since $N(\mathsf{K})$ is the splitting field for $(M_a)_{a\in\mathsf{K}}$ (by Proposition 4.6.36) and since each of the minimal polynomials M_a is separable (by Proposition 4.6.45). ■

The following result indicates that, in many cases of interest, separable extensions agree with algebraic extensions.

4.6.47 Proposition (Separable equals algebraic in characteristic zero) *If* F *is a field of characteristic zero, then an extension* K *of* F *is separable if and only if it is algebraic.*

Proof It is clear that separable extensions are algebraic. So suppose that K is an algebraic extension of F and let $a \in \mathsf{K}$. Then a is a root in K of the minimal polynomial M_a which is irreducible by Proposition 4.6.16. But M_a is then separable by Proposition 4.6.43. ■

There are also fields of nonzero characteristic where all algebraic extensions are separable. For example, it is possible to show that all finite fields, e.g., $\mathbb{Z}_p$ for p prime, have the property that all of their algebraic extensions are separable. Since this is not of much interest to us, we shall not prove it, but refer the reader to Section 4.6.10 for references. We shall, however, give an example of a polynomial that is not separable, and so which gives rise to a nonseparable field extension.

4.6.48 Example (Nonseparable field extension) Consider the field $\mathbb{Z}_2$ and its field $\mathsf{F} = \mathbb{Z}_2(\eta)$ of rational functions in indeterminate η. Consider the polynomial $A = \xi^2 - \eta \in \mathsf{F}[\xi]$. This polynomial is irreducible since it has no roots in $\mathsf{F}[\xi]$. Indeed, any root would have to be a rational function R satisfying $R^2 = \eta$. Thus R should be a polynomial, and its degree should be less than one, which is clearly absurd. The splitting field for A is denoted $\mathsf{F}(\sqrt{\eta})$, and in this field we have $A = (\xi + \sqrt{\eta})(\xi + \sqrt{\eta})$ since $2\sqrt{\eta} = 0_{\mathsf{F}(\sqrt{2})}$ and $-\eta = \eta$. Thus, while A is irreducible, it is not separable. Thus the extension $\mathsf{F}(\sqrt{\eta})$ is itself not separable. •

4.6.8 Galois extensions

Now we get to what for us is the main point of studying field extensions: the notion of a Galois extension. Let us first give the definition and explore some of its consequences. For the definition, one may wish to recall that $\mathsf{F} \subseteq \mathsf{K}^{\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})}$.

4.6.49 Definition (Galois extension) If F is a field, an algebraic extension K of F is ***Galois*** if $\mathsf{K}^{\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})} = \mathsf{F}$. •

The following characterisations of Galois extensions are useful.

4.6.50 Proposition (Characterisations of Galois extensions) *For a field* F *and an algebraic extension* K *of* F*, the following statements are equivalent:*

(i) K *is Galois;*

(ii) K *is normal and separable;*

(iii) for each a ∈ K, M_a *splits in* K *and each root has multiplicity* 1.

Proof (ii) ⟹ (iii) If K is normal then M_a splits in K since it is irreducible and has a root (a) in K. By Proposition 4.6.45 each of the roots of M_a have multiplicity 1 if K is separable.

(iii) ⟹ (ii) This follows from Propositions 4.6.36 and 4.6.45, after recalling from the proof of Proposition 4.6.36 that K is the splitting field for $(M_a)_{a\in\mathsf{K}}$.

(i) ⟹ (iii) Let $a \in \mathsf{K}$ and let $r_1, \dots, r_k$ be the distinct roots of the minimal polynomial M_a. Define

$$\tilde{M}_a = (\xi - r_1)\cdots(\xi - r_k).$$

Since $\tilde{M}_a$ and M_a have the same roots, but those of $\tilde{M}_a$ are of multiplicity 1, it follows that $\tilde{M}_a|M_a$. Let $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ and recall from Proposition 4.6.28 that there exists $\sigma \in \mathfrak{S}_k$ such that $\phi(r_j) = r_{\sigma(j)}$ for each $j \in \{1, \dots, k\}$. It follows that $\phi_*\tilde{M}_a = \tilde{M}_a$ (i.e., the coefficients of $\tilde{M}_a$ are fixed by ϕ). Since this holds for every $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$, by hypothesis it follows that $\tilde{M}_a \in \mathsf{F}[\xi]$. Moreover, $\mathrm{Ev}_{\mathsf{K}}(\tilde{M}_a)(a) = 0_{\mathsf{K}}$ since $a \in \{r_1, \dots, r_k\}$. Thus $M_a|\tilde{M}_a$ since M_a is the minimal polynomial. Since M_a and $\tilde{M}_a$ are both monic we must have $M_a = \tilde{M}_a$, giving (iii).

(iii) ⟹ (i) Let $\bar{\mathsf{F}}$ be an algebraic closure of F which is an extension of K. Let $a \in \mathsf{K}\setminus\mathsf{F}$. Let $\{r_1, \dots, r_k\}$ be the roots in K of the minimal polynomial M_a. By hypothesis we have

$$M_a = (\xi - r_1)\cdots(\xi - r_k).$$

Since $a \in \{r_1, \dots, r_k\}$ it follows from Corollary 4.6.35 that $\{r_1, \dots, r_k\}$ are the conjugates of a in $\bar{\mathsf{F}}$. Now, since $\deg(M_a) \geq 2$ (since $a \notin \mathsf{F}$) it follows that $k \geq 2$ and so there exists $b \in \{r_1, \dots, r_k\}$ such that $b \neq a$. Thus there exists $\phi \in \mathrm{Aut}_{\mathsf{F}}(\bar{\mathsf{F}})$ such that $\phi)a) = b$. Since $\phi(\mathsf{K}) = \mathsf{K}$ by hypothesis and by Proposition 4.6.36, it follows that ϕ defines $\psi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ such that $\psi(a) = b$. Thus we have shown that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ can only fix elements in F, as desired. ■

For us, the following easy consequence of the preceding result will be valuable.

4.6.51 Corollary (Separable extensions have Galois normal closures) *If* F *is a field with* K *a separable extension, the normal closure of* K *is a Galois extension of* F*.*

Proof Let $a \in N(\mathsf{K})$. Then a is a conjugate of an element $b \in \mathsf{K}$ and so $M_a = M_b$ by Proposition 4.6.34. Since M_a is separable, so is M_b. Thus $N(\mathsf{K})$ is normal and separable, and so Galois by Proposition 4.6.50. ■

4.6.9 Solvability of polynomials by radicals

One of the classical questions in algebra concerns the derivation of formulae for the roots of a polynomial. We shall study this is a little detail in Section 4.7.4 for polynomials with real coefficients, but here we say a few general things that will help us later on.

The idea is this. Given a field F consider a polynomial $A \in \mathsf{F}[\xi]$ given by

$$A = \xi^k + c_{k-1}\xi^{k-1} + \cdots + c_1\xi + c_0.$$

We can assume for the purpose of finding roots that A is monic. The objective is to derive a formula for the roots of A, possible in some appropriate extension, which involve iterative operations of addition, subtraction, multiplication, division, and taking rational powers of the coefficients $c_0, c_1, \dots, c_{k-1}$. Note that if we wish to ensure that we can perform every sum, difference, product, and quotient of these coefficients we must work in the extension $\mathsf{F}(c_0, c_1, \dots, c_{k-1})$ if the coefficients $c_0, c_1, \dots, c_{k-1}$ do not already live in F. By taking rational powers we mean expressions of the form $a^{k/l}$ for $k, l \in \mathbb{Z}_{>0}$. Noting that $a^{k/l} = (a^k)^{1/l}$ this amounts to being able to solve the equation $b^{1/l} = c$ for b in our extension.

With this as motivation we make the following definition which gives the sort of field extension we are interested in.

4.6.52 Definition (Radical extension) An extension K of a field F is ***radical*** if there exists a finite sequence

$$\mathsf{F} = \mathsf{K}_0 \subseteq \mathsf{K}_1 \subseteq \cdots \subseteq \mathsf{K}_n = \mathsf{K}$$

of extensions of F with the property that, for each $j \in \{1, \dots, n\}$, $\mathsf{K}_j = \mathsf{K}_{j-1}(a_j)$ where $a_j \in \mathsf{K}_j$ satisfies $a_j^{k_j} \in \mathsf{K}_{j-1}$ for some $k_j \in \mathbb{Z}_{>0}$. •

The idea then is that the extension from K_{j-1} to K_j is obtained by adjoining a root of the polynomial $\xi^{k_j} - a_j$, i.e., by adjoining the k_jth root of a_j.

The corresponding property of polynomials is the following.

4.6.53 Definition (Solvable by radicals) Let F be a field and let $A \in \mathsf{F}[\xi]$. The polynomial A is ***solvable by radicals*** if there exists a radical extension K of F in which A splits. •

This notion of solvability by radicals is easily illustrated by the usual quadratic equation.

4.6.54 Example (Quadratic equations and solvability by radicals) We consider the field $\mathbb{Q}$ with a polynomial $A = \xi^2 + c_1\xi + c_0$ with $c_0, c_1 \in \mathbb{C}$. We know that the roots of A are given by the quadratic formula:

$$r_1 = -\tfrac{1}{2}(-c_1 + \sqrt{c_1^2 - 4c_0}), \quad r_2 = -\tfrac{1}{2}(-c_1 - \sqrt{c_1^2 - 4c_0}).$$

Thus if we successively adjoint c_0, c_1, and $\sqrt{c_1^2 - 4c_0}$ to get the field $\mathbb{Q}(c_0, c_1, \sqrt{c_1^2 - 4c_0})$, then we are ensured to obtain a field where the roots exist. That A is solvable by radicals by taking the sequence of field extensions

$$\mathbb{Q} \subseteq \mathbb{Q}(c_0) \subseteq \mathbb{Q}(c_0, c_1) \subseteq \mathbb{Q}(c_0, c_1, \sqrt{c_1^2 - 4c_0}). \qquad \bullet$$

With this example we hope that the reader can see that any formula for the roots involving sums, differences, products, quotients, and rational powers starting with the coefficients will be equivalent to the polynomial being solvable by radicals.

The key result relates the notion of solvability by radicals to properties of the Galois group. The key property is the following which is valid for arbitrary groups.

4.6.55 Definition (Solvable group) A group G is ***solvable*** if there exists a finite sequence

$$\{e\} = \mathsf{N}_0 \subseteq \mathsf{N}_1 \subseteq \cdots \subseteq \mathsf{N}_n = \mathsf{G}$$

of normal subgroups such that the quotient group (cf. Proposition 4.1.20) is Abelian. •

The key, and rather surprising, result is then the following.

4.6.56 Theorem (Radical extensions and solvable Galois groups) *If* K *is a radical extension of a field* F *then* $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ *is solvable.*

Proof We suppose that we have

$$\mathsf{F} = \mathsf{K}_0 \subseteq \mathsf{K}_1 \subseteq \cdots \subseteq \mathsf{K}_n = \mathsf{K}$$

with $\mathsf{K}_j = \mathsf{K}_{j-1}(a_j)$ where $a_j^{k_j} \in \mathsf{K}_{j-1}$. We now make a few assumptions without loss of generality.

1. We assume that at each step we have either $a_j^{p_j} \in \mathsf{K}_{j-1}$ for a *prime* $p_j \in \mathbb{Z}_{>0}$. If this is not the case, and we instead have $a_j^{k_j} \in \mathsf{K}_{j-1}$ for a *nonprime* $k_j \in \mathbb{Z}_{>0}$, we write $k_j = p_{j1} \cdots p_{jm_j}$ as a product of primes and then add some terms to the sequence.
2. We assume that either
   (a) $\mathsf{F}(a_1, \dots, a_j)$ contains no roots of the polynomial $\xi^{p_j} - 1_{\mathsf{K}}$ that are not already contained in $\mathsf{F}(a_1, \dots, a_{j-1})$ or that
   (b) a_j is itself a root of the polynomial $\xi^{p_j} - 1_{\mathsf{K}}$.

   If this is not the case, then we take $b \neq 1_{\mathsf{K}}$ such that $b^{p_j} = 1_{\mathsf{K}}$ and add a term in the sequence:

$$\mathsf{F} = \mathsf{K}_0 \subseteq \mathsf{K}_1 \subseteq \cdots \subseteq \mathsf{K}_{j-1} \subseteq \mathsf{K}_{j-1}(b) \subseteq \mathsf{K}_j \subseteq \cdots \subseteq \mathsf{K}_n = \mathsf{K}.$$

   Note that this implies that the set $\{1_{\mathsf{K}}, b, b^2, \dots, b^{p_j-1}\}$ now contains all roots of $\xi^{p_j} = 1_{\mathsf{K}}$ (why?). Thus, for this extended sequence, our assumption holds.

Now we prove a lemma; note that the notation F and K in the lemma do not match that in the theorem.

1 Lemma *Let* F *be a field with extension* K *and let* $\mathsf{a} \in \mathsf{K}$ *satisfy* $\mathsf{a}^{\mathsf{p}} \in \mathsf{F}$ *for* p *prime. Further suppose that either*

(i) $\mathsf{F}(\mathsf{a})$ *contains no roots of* $\xi^{\mathsf{p}} - 1_{\mathsf{K}}$ *that are not already in* F *or that*

(ii) a *is itself a root of* $\xi^{\mathsf{p}} - 1_{\mathsf{K}}$.

Then $\mathrm{Aut}_{\mathsf{F}(\mathsf{a})}(\mathsf{K})$ *is a normal subgroup of* $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ *and* $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})/\mathrm{Aut}_{\mathsf{F}(\mathsf{a})}(\mathsf{K})$ *is Abelian.*

Proof We first claim that if $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ then $\phi(\mathsf{F}(a)) \subseteq \mathsf{F}(a)$, i.e., that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ leaves $\mathsf{F}(a)$ invariant. We have two cases.

1. If a is not a root of $\xi^p - 1_{\mathsf{K}}$ then

$$(\phi(a))^p = \phi(a^p) = a^p$$

since $a \in \mathsf{F}$. Thus $\phi(a)$ is a root of $\xi^p - a^p$, and so has the form $b^j a$ for some root b of $\xi^p - 1_{\mathsf{K}}$ (why?). Since our assumptions ensure that $b \in \mathsf{F}$ we have $\phi(a) \in \mathsf{F}(a)$.

2. If a is a root of $\xi^p - 1_{\mathsf{K}}$ then we have

$$(\phi(a))^p = \phi(a^p) = \phi(1_{\mathsf{K}}) = 1_{\mathsf{K}}$$

and so $\phi(a)$ is a root of $\xi^p - 1_{\mathsf{K}}$. Thus $\phi(a) = a^j$ for some $j \in \{1, \dots, p-1\}$. Thus $\phi(a) \in \mathsf{F}(a)$.

Thus we have $\phi(a) \in \mathsf{F}(a)$ in either case. Since $\phi|\mathsf{F} = \mathrm{id}_{\mathsf{F}}$ it follows that $\phi(\mathsf{F} \cup \{a\}) \subseteq \mathsf{F}(a)$. Since $\mathsf{F} \cup \{a\}$ generate $\mathsf{F}(a)$ and since ϕ is a homomorphism of fields, it follows that $\phi(\mathsf{F}(a)) \subseteq \mathsf{F}(a)$.

Now let us define $\Psi\colon \mathrm{Aut}_{\mathsf{F}}(\mathsf{K}) \to \mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ by $\Psi(\phi)(b) = \phi(b)$, this map making sense since above we showed that $\phi(\mathsf{F}(a)) \subseteq \mathsf{F}(a)$ and since $\phi|\mathsf{F} = \mathrm{id}_{\mathsf{F}}$. It is clear that $\ker(\Psi) = \mathrm{Aut}_{\mathsf{F}(a)}(\mathsf{K})$ since

$$\Psi(\phi) = \mathrm{id}_{\mathsf{F}(a)} \quad \Longleftrightarrow \quad \phi|\mathsf{F}(a) = \mathrm{id}_{\mathsf{F}(a)} \quad \Longleftrightarrow \quad \phi \in \mathrm{Aut}_{\mathsf{F}(a)}(\mathsf{K}).$$

To show that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})/\mathrm{Aut}_{\mathsf{F}(a)}(\mathsf{K})$ is Abelian we first show that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ is Abelian.

1. If a is not a root of unity then let us define $\phi_j \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$, $j \in \mathbb{Z}_{\geq 0}$, by $\phi_j(a) = b^j a$ where $b \in \mathsf{F}$ is a root of $\xi^p - 1_{\mathsf{K}}$. Note that

$$\phi_j \circ \phi_k(a) = b^{j+k} a = b^{k+j} a = \phi_k \circ \phi_j(a).$$

Thus $\phi_j \circ \phi_k(c) = \phi_k \circ \phi_j(c)$ for all $c \in \mathsf{F} \cup \{a\}$. Thus ϕ_j and ϕ_k commute on a set of generators for $\mathsf{F}(a)$, and so commute on $\mathsf{F}(a)$. Then, as above, if $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ we have $\phi = \phi_j$ for some $j \in \{0, 1, \dots, p-1\}$. This shows that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ is Abelian in this case.

2. If a is a root of $\xi^p - 1_{\mathsf{K}}$ then define $\phi_j \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$, $j \in \mathbb{Z}_{\geq 0}$, by $\phi_j(a) = a^j$. We then have

$$\phi_j \circ \phi_k(a) = a^{j+k} = a^{k+j} = \phi_j \circ \phi_k(a).$$

Now we argue as in the preceding case to see that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ is Abelian.

Thus we have a surjective homomorphism $\Psi\colon \mathrm{Aut}_{\mathsf{F}}(\mathsf{K}) \to \mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ into an Abelian group with kernel $\mathrm{Aut}_{\mathsf{F}(a)}(\mathsf{K})$. By Exercise 4.1.9 it follows that $\mathrm{Aut}_{\mathsf{F}}(\mathsf{K})/\mathrm{Aut}_{\mathsf{F}(a)}(\mathsf{K})$ is isomorphic to $\mathrm{Aut}_{\mathsf{F}}(\mathsf{F}(a))$ which we have shown to be Abelian. ▼

The theorem now follows by applying the lemma to each element in the sequence

$$\mathsf{F} = \mathsf{K}_0 \subseteq \mathsf{K}_1 \subseteq \cdots \subseteq \mathsf{K}_{j-1} \subseteq \mathsf{K}_{j-1}(b) \subseteq \mathsf{K}_j \subseteq \cdots \subseteq \mathsf{K}_n = \mathsf{K},$$

after making the assumptions given at the beginning of the proof. ■

The converse of the theorem is often true, but as we shall not need this, we stop with what we have.

4.6.10 Notes

We saw in Proposition 4.6.47 that the notions of algebraic and separable extensions agree for fields of characteristic zero. The same is true for the finite fields $\mathbb{Z}_p$ of prime characteristic p. However, there are nonfinite fields of nonzero characteristic which possess algebraic extensions that are not separable. Such matters as this are discussed, for example, in [Bourbaki 1990, A.V].

The proof we give of the existence of the algebraic closure (Theorem 4.6.23) is stated by [Lang 2005] as being due to Artin. The construction, it turns out, is more complicated than it needs to be, in some sense. In the proof, one constructs a nested sequence

$$\mathsf{K}_1 \subseteq \mathsf{K}_2 \subseteq \mathsf{K}_3 \subseteq \cdots$$

of field extensions of F and then takes their union to obtain a field which is guaranteed to be algebraically closed. The algebraic closure is then the set of algebraic elements in this big field extension. Gilmer [1968] shows that the algebraic closure is already contained in K_1.

We saw in Example 4.6.48 an example of a polynomial that is not separable. Such a polynomial must be defined over a field that is (1) of nonzero characteristic and (2) infinite. It turns out that the key property of a field F which ensures that its algebraic extensions are separable is that it be "perfect," by which we mean that either (1) it has characteristic zero or (2) it has nonzero characteristic p and the equation $a^p = b$ can be solved for a for every $b \in \mathsf{F}$. Sometimes one writes this property as $\mathsf{F}^p = \mathsf{F}$. We refer the reader to [Bourbaki 1990, A.V] for more discussion of this.

Fundamental Theorem of Galois Theory.

Exercises

4.6.1 For a field F denote by F_0 the prime field and let $S \subseteq \mathsf{F}$. Show that the smallest subfield of F containing S is $\mathsf{F}_0(S)$.

4.6.2 Prove Proposition 4.6.2.

4.6.3 Let K be an extension of a field F and let $a \in \mathsf{K}$ be algebraic so that $\mathsf{F}(a) = \mathsf{F}[a]$ by Proposition 4.6.17. If $A \in \mathsf{F}[\xi]$, show that the multiplicative inverse of $\mathrm{Ev}_{\mathsf{K}}(A)(a)$ in $\mathsf{F}(a)$, if it exists, is $\mathrm{Ev}_{\mathsf{K}}(B)(a)$ where B satisfies $AB + CM_a = 1_{\mathsf{F}[\xi]}$ for some $C \in \mathsf{F}[\xi]$.

4.6.4 Prove Proposition 4.6.9.

4.6.5 Let F be a field, let K be an extension of F, and let $a \in \mathsf{K}$ be algebraic over F. Show that

$$\mathsf{I}_a = \{A \in \mathsf{F}[\xi] \mid \mathrm{Ev}_{\mathsf{K}}(A)(a) = 0_{\mathsf{K}}\}$$

is a nonzero ideal of $\mathsf{F}[\xi]$.

4.6.6 Let F be a field and let $A_1, \dots, A_k \in \mathsf{F}[\xi]$. Show that K is a splitting field for $(A_j)_{j\in\{1,\dots,k\}}$ if and only if it is a splitting field for $A = A_1 \cdots A_k$.

Section 4.7

Construction of the complex numbers

In this section we give a well-motivated definition of the complex numbers. The motivation comes from some of our considerations about polynomials in Section 4.4. Specifically, we construct the complex numbers, denoted by $\mathbb{C}$, exactly to ensure that a certain polynomial in $\mathbb{R}[\xi]$ splits over $\mathbb{C}$. We then show that the complex numbers defined in this way are actually the algebraic closure of $\mathbb{R}$. Thus *every* polynomial in $\mathbb{R}[\xi]$ splits over $\mathbb{C}$.

Do I need to read this section? The usual presentation of the complex numbers is as points in a plane, with some specified rules of addition and multiplication. This presentation leaves much to be desired, but at the end of the day, this is all one really *needs*. For this reason, readers who are already familiar with complex arithmetic can skip this section. However, for such readers who also wish to have some understanding of what the complex numbers really are, this section is mandatory reading. •

4.7.1 What are the complex numbers?

Since many readers will likely be familiar with the complex numbers already, we provide a short summary of the constructions in this section since these constructions themselves may be unfamiliar even to readers who have had a course in complex analysis.

When one encounters polynomials in grade school, one learns the quadratic formula for degree two polynomials with real coefficients. Such a polynomial is typically given the form $P = a\xi^2 + b\xi + c$, where $a \neq 0$. To determine the roots, one uses a valuable trick called "completing the square:"

$$a\xi^2 + b\xi + c = (\sqrt{a}\xi + \tfrac{b}{2\sqrt{a}})^2 + \tfrac{4ac-b^2}{4a}.$$

The roots are then found by setting the expression equal to zero, and solving for ξ. In the expression on the right, this is straightforward, and produces the familiar formula

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

The interesting thing about this formula is that, when the ***discriminant*** $b^2 - 4ac$ is negative, the polynomial has no real roots, and so is therefore irreducible over $\mathbb{R}$. One can then ask, "What is the smallest field in which all quadratic polynomials in $\mathbb{R}[\xi]$ have roots?" An important simplification can be made with the following observation. When $b^2 - 4ac < 0$ we can write $b^2 - 4ac = (-1)(4ac - b^2)$ with $4ac - b^2 > 0$. Since the existence of the roots of the polynomial P turn on the existence of a real square root of $b^2 - 4ac$, this reduces matter of the existence of these roots to determining a square root of -1 (this assumes some nice properties of square roots, but let us put this discussion off until Section II-3.4). That is to say, we need only determine the field in which the polynomial $\xi^2 + 1$ has roots. However, we know how to do this by virtue of Theorem 4.4.44: since $\xi^2 + 1$ is irreducible in $\mathbb{R}[\xi]$, the quotient ring $\mathbb{R}[\xi]/(\xi^2 + 1)$ is a field. Moreover, by Proposition 4.6.25, we know that $\xi^2 + 1$ has a root in $\mathbb{R}[\xi]/(\xi^2 + 1)$, and therefore obviously splits in this field. It is this field that we defined to be $\mathbb{C}$.

With the construction to this point we know that every quadratic polynomial over $\mathbb{R}$ splits in $\mathbb{C}$. This, however, does not imply that *every* polynomial over $\mathbb{R}$ splits in $\mathbb{C}$. Nonetheless, this is indeed the case as we show in Theorem 4.7.6, the extremely important Fundamental Theorem of Algebra. It is interesting to observe that one must step out of the realm of the purely algebraic to prove the Fundamental Theorem of Algebra. However, it is fairly easily proved using the machinery developed in Chapter 3.

The definition of $\mathbb{C}$ as $\mathbb{R}[\xi]/(\xi^2 + 1)$ allows the easy derivation of the usual rules of complex arithmetic using Proposition 4.4.42. In this sense, this definition of $\mathbb{C}$ is entirely more satisfactory than the usual definition of $\mathbb{C}$ as being equal to $\mathbb{R}^2$, and with complex multiplication seemingly coming from nowhere. Of course, one pays a heavy price for this better understanding in that one needs to know some ring theory to appreciate the definition.

4.7.2 Definition and basic properties of complex numbers

Now let us proceed with the formalities.

4.7.1 Definition (Complex numbers, real part, imaginary part) The set of ***complex numbers***, denoted by $\mathbb{C}$, is the field $\mathbb{R}[\xi]/(\xi^2 + 1)$. Let $\mathrm{i} = \xi + (\xi^2 + 1) \in \mathbb{R}[\xi]/(\xi^2 + 1)$ and, following Proposition 4.4.42, denote a typical element of $\mathbb{C}$ by $z = x + \mathrm{i}y$ for $x, y \in \mathbb{R}$. Then x is the ***real part*** of z denoted by $\mathrm{Re}(z)$ and y is the ***imaginary part*** of z denoted by $\mathrm{Im}(z)$. A complex number of the form $x + \mathrm{i}0$, $x \in \mathbb{R}$, is called ***purely real*** and a complex number of the form $0 + \mathrm{i}y$, $y \in \mathbb{R}$, is called ***purely imaginary***. •

Let us give the properties of addition and multiplication in the field $\mathbb{C}$.

4.7.2 Proposition (Addition and multiplication in $\mathbb{C}$) *If* $\mathrm{z}_1 = \mathrm{x}_1 + \mathrm{i}\mathrm{y}_1, \mathrm{z}_2 = \mathrm{x}_2 + \mathrm{i}\mathrm{y}_2 \in \mathbb{C}$ *for* $\mathrm{x}_1, \mathrm{x}_2, \mathrm{y}_1, \mathrm{y}_2 \in \mathbb{R}$*, then*

$$\mathrm{z}_1 + \mathrm{z}_2 = (\mathrm{x}_1 + \mathrm{x}_2) + \mathrm{i}(\mathrm{y}_1 + \mathrm{y}_2), \quad \mathrm{z}_1\mathrm{z}_2 = (\mathrm{x}_1\mathrm{x}_2 - \mathrm{y}_1\mathrm{y}_2) + \mathrm{i}(\mathrm{x}_1\mathrm{y}_2 + \mathrm{x}_2\mathrm{y}_1).$$

Also, if $\mathrm{z} = \mathrm{x} + \mathrm{i}\mathrm{y} \in \mathbb{C} \setminus \{0 + \mathrm{i}0\}$, $\mathrm{x}, \mathrm{y} \in \mathbb{R}$*, then*

$$\mathrm{z}^{-1} = \frac{\mathrm{x}}{\mathrm{x}^2 + \mathrm{y}^2} - \mathrm{i}\frac{\mathrm{y}}{\mathrm{x}^2 + \mathrm{y}^2}.$$

Proof The formula for the sum follows directly from Proposition 4.4.42. The formula for the product does as well, but it is simpler, and more illustrative, to prove it directly.

Using the definition $\mathrm{i} = \xi + (\xi^2 + 1)$ we have

$$\begin{aligned}(x_1 + \mathrm{i}y_1)(x_2 + \mathrm{i}y_2) &= (x_1 + y_1\xi)(x_2 + y_2\xi) + (\xi^2 + 1)\\ &= x_1x_2 + (x_1y_2 + x_2y_1)\xi + y_1y_2\xi^2 + (\xi^2 + 1)\\ &= (x_1x_2 - y_1y_2) + (x_1y_2 + x_2y_1)\xi + (\xi^2 + 1)\\ &= (x_1x_2 - y_1y_2) + \mathrm{i}(x_1y_2 + x_2y_1),\end{aligned}$$

using the fact that $\xi^2 + (\xi^2 + 1) = -1 + (\xi^2 + 1)$. The final assertion is easily verified by checking that $z^{-1}z = zz^{-1} = 1 + \mathrm{i}0$. ■

Note that addition and multiplication are *consequences* of our definition of $\mathbb{C}$, and are not *included* in the definition of $\mathbb{C}$ as is often the case. Theorem 4.4.44 also eliminates the responsibility of showing that $\mathbb{C}$ is a field, as would be the case were the operations to be simply defined.

4.7.3 Remark ($\mathrm{i} = \sqrt{-1}$) Note that in the field $\mathbb{C}$ the element i satisfies $\mathrm{i}^2 + 1 = 0$. For this reason it is common to see the definition of i given as $\mathrm{i} = \sqrt{-1}$. By this it is really meant that, in the field $\mathbb{C}$, i is defined to be the number whose square is equal to -1. Again, in our construction this is a consequence of the manner in which $\mathbb{C}$ is built. •

Associated with the complex numbers are two useful operations.

4.7.4 Definition (Complex conjugate and complex modulus) If $z = x+\mathrm{i}y \in \mathbb{C}$, $x, y \in \mathbb{R}$, then the ***complex conjugate***, or simply the ***conjugate***, of z is the element $\bar{z} \in \mathbb{C}$ defined by $\bar{z} = x - \mathrm{i}y$. The ***complex modulus***, or simply the ***modulus***, of z is the element $|z| \in \mathbb{R}_{\geq 0}$ given by $|z| = x^2 + y^2$. •

Note that $\mathbb{R} \subseteq \mathbb{C}$ since $\mathbb{C}$ is a field extension of $\mathbb{R}$. More precisely, we have the injective map $i_{\mathbb{R}} \colon \mathbb{R} \to \mathbb{C}$ given by $i_{\mathbb{R}}(x) = x + \mathrm{i}0$. We shall simply think of $\mathbb{R}$ as being a subset of $\mathbb{C}$. We shall denote by 0 the complex zero, as well as the zero in $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$. We now have the following sequence of number systems that we have thus far introduced:

$$\mathbb{Z}_{>0} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C}.$$

We shall have no use for extending this idea any further. Note that the complex modulus generalises the absolute value first introduced for integers, then extended to rational and real numbers. One might also wonder whether the order on $\mathbb{Z}_{>0}$, which then led to orders on $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$, also extends to $\mathbb{C}$. It turns out that this is not possible. However, this is not of particular interest to us, so we say no more about it. In the references cited Section 4.7.5 the reader will find some discussion of this.

The following properties of complex conjugate and complex modulus are mostly easily proved, as the reader can verify in Exercise 4.7.2.

4.7.5 Proposition (Properties of conjugate and modulus) *For* $z, z_1, z_2 \in \mathbb{C}$*, the following statements hold:*

(i) $\bar{\bar{z}} = z$;

(ii) $\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2$;

(iii) $\overline{z_1 z_2} = \bar{z}_1 \bar{z}_2$;

(iv) if $z \neq 0 + i0$*, then* $\overline{z^{-1}} = \bar{z}^{-1}$;

(v) $\bar{z} = z$ *if and only if* z *is purely real;*

(vi) $\bar{z} = -z$ *if and only if* z *is purely imaginary*

(vii) $|\bar{z}| = |z|$;

(viii) $|z| = 0$ *if and only if* $z = 0 + i0$;

(ix) $|z_1 + z_2| \leq |z_1| + |z_2|$ *(**triangle inequality**);*

(x) $|z_1 - z_2| \geq \big||z_1| - |z_2|\big|$;

(xi) $|z_1 z_2| = |z_1||z_2|$;

(xii) if $z \neq 0 + i0$*, then* $|z^{-1}| = |z|^{-1}$.

That the triangle inequality holds for the complex modulus allows us to extend much of Chapter 3 to $\mathbb{C}$-valued functions. This will be discussed in detail in Section II-3.2, but we will use some of these ideas in our proof of Theorem 4.7.6.

4.7.3 Polynomials over $\mathbb{R}$ and $\mathbb{C}$

In this section we present some important properties of polynomials with coefficients in either $\mathbb{R}$ or $\mathbb{C}$. The most important part in this development is played by the following theorem. This theorem completes the story outlined in Section 4.7.1.

4.7.6 Theorem (Fundamental Theorem of Algebra) *The field* $\mathbb{C}$ *is algebraically closed.*

Proof Note that to show that $\mathbb{C}$ is algebraically closed, it suffices to show that if $A \in \mathbb{C}[\xi]$ is not constant, then A has a root. Indeed, if this is so, and if $A \in \mathbb{C}[\xi]$, then by Proposition 4.4.25 we can write $A = (\xi - z_1)A_2$ for $z_1 \in \mathbb{C}$ and for $A_1 \in \mathbb{C}[\xi]$. It then follows, since A_2 has a root, that $A_2 = (\xi - z_2)A_3$ for some $z_2 \in \mathbb{C}$ and $A_3 \in \mathbb{C}[\xi]$. This can clearly be continued until A is written as a product of degree one polynomials. Thus we shall show in the proof that every polynomial over $\mathbb{C}$ that is not constant has a root.

Fix $A = a_k\xi^k + \cdots + a_1\xi + a_0 \in \mathbb{C}[\xi]$ with $a_k \neq 0$. If $a_0 = 0$, then A immediately has zero as a root. Thus we suppose that $a_0 \neq 0$. For notational convenience let us agree to denote $A(z) = \mathrm{Ev}_{\mathbb{C}}(A)(z)$. The proof will consist largely of a series of lemmata.

1 Lemma *There exists* $R > 0$ *such that, if* $|z| \geq R$*, then* $|A(z)| \geq \frac{1}{2}|a_k||z|^k$.

Proof Let $M = |a_0| + |a_1| + \cdots + |a_{k-1}|$. If $M = 0$ then $A = a_k z^k$ so that $|A(z)| > \frac{1}{2}|a_k||z|^k$. Thus the lemma holds in this case. So we suppose that $M > 0$. If $|z| \geq 1$ then $|z|^k \geq |z|$. Therefore, if $|z| \geq 1$,

$$\frac{|a_{k-j}|}{|z|^j} \leq \frac{|a_{k-j}|}{|z|}.$$

Thus, using the triangle inequality,

$$\begin{aligned}\Big|-\frac{a_0}{z^k}-\frac{a_1}{z^{k-1}}-\cdots-\frac{a_{k-1}}{z}\Big| &\le \Big|\frac{a_0}{z^k}\Big|+\Big|\frac{a_1}{z^{k-1}}\Big|+\cdots+\Big|\frac{a_{k-1}}{z}\Big|\\ &\le \frac{M}{|z|}.\end{aligned}$$

Now using part (x) of Proposition 4.7.5,

$$\begin{aligned}|A(z)| &= \Big|z^k\Big(a_k-\Big(-\frac{a_0}{z^k}-\frac{a_1}{z^{k-1}}-\cdots-\frac{a_{k-1}}{z}\Big)\Big)\Big|\\ &\ge |z|^k\Big||a_k|-\frac{M}{|z|}\Big|,\end{aligned}$$

again provided that $|z| \ge 1$. If $|z| \ge \frac{2M}{|a_k|}$ then $\frac{M}{|z|} \le \frac{|z|}{2}$, and so the result follows by taking $R = \max\{1, \frac{2M}{|a_k|}\}$. ▼

2 Lemma *There exists* $r > 0$ *such that, if* $|z| \ge r$, *then* $|A(z)| \ge (|a_0| + 1)^2$.

Proof Let $R > 0$ be chosen as in Lemma 1. If $|z| \ge \max\{1, \frac{2(|a_0|+1)^2}{|a_k|}\}$ then

$$\begin{aligned}|A(z)| &\ge \tfrac{1}{2}|a_k||z|^k \ge \tfrac{1}{2}|a_k||z|\\ &\ge \frac{2(|a_0|+1)^2}{|a_k|}\frac{|a_k|}{2} = (|a_0|+1)^2.\end{aligned}$$

Thus the result follows by taking $r = \max\{1, \frac{2(|a_0|+1)^2}{|a_k|}\}$. ▼

3 Lemma *There exists* $z_0 \in \mathbb{C}$ *such that* $|A(z)| \ge |A(z_0)|$ *for every* $z \in \mathbb{C}$.

Proof Choose $r > 0$ as in Lemma 2. Then, for $|z| \ge r$ we have

$$|A(z)|^2 \ge |A(z)| \ge (|a_0|+1)^2 \ge |a_0|^2 = |A(0)|^2.$$

Thus $|A(0)| \le |A(z)|$ for every z satisfying $|z| \ge r$. Therefore, it suffices to find a point z_0 such that $|z_0| \le r$ and such that $|A(z_0)| \le |A(z)|$ for all $|z| \le r$. However, since the set $\{z \mid |z| \le r\}$ is compact, follows from the generalisation of Theorem 3.1.23 which holds for $\mathbb{R}$-valued functions on $\mathbb{R}$; the generalisation is stated as Theorem II-1.3.32. ▼

4 Lemma *If* $z_0 \in \mathbb{C}$ *satisfies* $|A(z_0)| \ne 0$, *then, for* $\epsilon > 0$, *there exists* $z \in \mathbb{C}$ *such that*

(i) $|z - z_0| < \epsilon$ *and*

(ii) $|A(z)| < |A(z_0)|$.

Proof For $w \in \mathbb{C}$ use the Binomial Theorem to write

$$A(z_0 + w) = b_k w^k + \cdots + b_1 w + b_0$$

for some $b_0, b_1, \dots, b_k \in \mathbb{C}$. Let $l \in \mathbb{Z}_{>0}$ be the smallest number such that $b_l \ne 0$; since $b_k = a_k$ (from the Binomial Theorem) this definition of l makes sense. Now define

$w_0 \in \mathbb{C}$ such that $w_0^l = -\frac{b_0}{b_l}$; such a number w_0 exists by Proposition II-3.1.1. Now define $g\colon \mathbb{C} \to \mathbb{R}_{\geq 0}$ by

$$g(h) = \left|h^{-l}(A(z_0 + hw_0) - b_0 - b_l w_0^l h^l)\right|^2 = \left|b_{l+1} w_0^{l+1} h + \cdots + b_k w_0^k h^{k-l}\right|^2.$$

Note that this is a continuous function of h and that $g(0 + \mathrm{i}0) = 0$. Now choose $h = x + \mathrm{i}0$ such that

1. $x \in (0, 1)$,
2. $|xw_0| < \epsilon$, and
3. $|g(x + \mathrm{i}0)| < |b_0|$.

Thus condition (i) is satisfied for $z = z_0 + hw_0$. Moreover,

$$\begin{aligned} |A(z_0)| = |b_0| = x^l|b_0| + (1 - x^l)|b_0| &= x^l|b_0| + |b_0 + b_l w_0^l x^l| \\ &\geq x^l|g(x + \mathrm{i}0)| + |b_0 + b_l w_0^l x^l| \\ &= |A(z_0 + xw_0) - b_0 - b_l w_0^l x^l| + |b_0 + b_l w_0^l x^l| \\ &\geq |A(z_0 + xw_0) - b_0 - b_l w_0^l x^l + b_0 + b_l w_0^l x^l| \\ &= |A(z_0 + xw_0)| = |A(z)|, \end{aligned}$$

giving the lemma. ▼

Now we complete the proof of the theorem. If A has no roots then it must hold that $|A(z)| > 0$ for every $z \in \mathbb{C}$. By Lemma 3 there exists z_0 such that

$$0 < |A(z_0)| = \inf\{|A(z)| \mid z \in \mathbb{C}\}. \tag{4.24}$$

By Lemma 4 it then follows that any neighbourhood of z_0 contains a point z such that $|A(z)| < |A(z_0)|$. But this contradicts (4.24). ■

As a consequence of the theorem we have the following useful result.

4.7.7 Corollary ($\mathbb{C}$ is the algebraic closure of $\mathbb{R}$) *The field of complex numbers is the algebraic closure of the field of real numbers.*

Proof Since we know that $\mathbb{C}$ is an algebraically closed extension of $\mathbb{R}$ by Theorem 4.7.6, it remains to show that all elements of $\mathbb{C}$ are algebraic over $\mathbb{R}$. To see this, let $a + ib \in \mathbb{C}$ for $a, b \in \mathbb{R}$. One readily checks that $a + ib$ is a root of the real polynomial $\xi^2 - 2a\xi + (a^2 + b^2)$, giving the result. ■

Now we know that the only irreducible polynomials over $\mathbb{C}$ are polynomials of degree one. Let us examine the consequences of this for polynomials over $\mathbb{R}$. A simple initial observation is the following.

4.7.8 Proposition (Structure of roots for elements of $\mathbb{R}[\xi]$) *If* $\mathrm{r} \in \mathbb{C}$ *is a root of* $\mathrm{A} \in \mathbb{R}[\xi] \subseteq \mathbb{C}[\xi]$*, then* $\bar{\mathrm{r}}$ *is also a root of* A*. Moreover, the multiplicities of the roots* r *and* $\bar{\mathrm{r}}$ *agree.*

Proof Write $A = a_k\xi^k + \cdots + a_1\xi + a_0$ with $a_k \neq 0$. Since r is a root for A we have

$$a_k r^k + \cdots + a_a r + a_0 = 0.$$

Taking the complex conjugate of this equation, and using Proposition 4.7.5, gives

$$a_k\bar{r}^k + \cdots + a_1\bar{r} + a_0 = 0,$$

thus showing that $\bar{r}$ is a root of A. To show that the multiplicities of r and $\bar{r}$ are the same, we proceed by induction on the multiplicity of r. If r has multiplicity 1 then we have $A = (\xi - r)(\xi - \bar{r})B$ where r is not a root of B. Therefore, $\bar{r}$ is also not a root of B since $\bar{\bar{r}} = r$. Now suppose that the whenever the multiplicity of r is equal to k, then so too is the multiplicity of $\bar{r}$ equal to k. Let A be a polynomial for which r is a root of multiplicity $k + 1$ and write

$$A = (\xi - r)(\xi - \bar{r})B.$$

Then r is a root of multiplicity k for B. We claim that $B \in \mathbb{R}[\xi]$. Indeed, using the Division Algorithm in $\mathbb{R}[\xi]$ we can write

$$A = Q(\xi - r)(\xi - \bar{r}) + R$$

for $Q, R \in \mathbb{R}[\xi]$ with $\deg(R) < 2$. We can also use the Division Algorithm in $\mathbb{C}[\xi]$ to write

$$A = Q'(\xi - r)(\xi - \bar{r}) + R'$$

for $Q', R' \in \mathbb{C}[\xi]$ with $\deg(R') < 2$. Since the quotient and remainder from the Division Algorithm in $\mathbb{C}[\xi]$ are unique by Corollary 4.4.15, we must have $Q' = Q$ and $R' = R$. It also follows that $B = Q$ and $R = 0_{\mathbb{R}[\xi]}$, so showing that $B \in \mathbb{R}[\xi]$. Now, by the induction hypothesis, $\bar{r}$ is a root of B of multiplicity k, so showing that $\bar{r}$ is a root of A of multiplicity $k + 1$. ■

A consequence of the root structure of polynomials over $\mathbb{R}$ is the following characterisation of irreducible polynomials over $\mathbb{R}$.

4.7.9 Theorem (Irreducible elements of $\mathbb{R}[\xi]$) *A polynomial* $\mathrm{A} \in \mathbb{R}[\xi]$ *is irreducible if and only if either*

(i) $\mathrm{A} = \mathrm{a}_1\xi + \mathrm{a}_0$ *for* $\mathrm{a}_0, \mathrm{a}_1 \in \mathbb{R}$ *with* $\mathrm{a}_1 \neq 0$*, or*

(ii) $\mathrm{A} = \mathrm{a}_2\xi^2 + \mathrm{a}_1\xi + \mathrm{a}_0$ *with* $\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2 \in \mathbb{R}$ *satisfying*

  (a) $\mathrm{a}_2 \neq 0$ *and*

  (b) $\mathrm{a}_1^2 - 4\mathrm{a}_2\mathrm{a}_0 < 0$.

Proof First note that any polynomials in $\mathbb{R}[\xi]$ described by one of the two conditions in the theorem statement is irreducible. This is trivial for condition (i), and for condition (ii) it follows from our discussion of the quadratic equation in Section 4.7.1.

Now suppose that A is irreducible. Using Proposition 4.7.8 and thinking of A as an element of $\mathbb{C}[\xi]$ we can write

$$A = a\prod_{j=1}^{k}((\xi - \rho_j)(\xi - \bar{\rho}_j))\prod_{l=1}^{m}(\xi - r_l)$$

for $a \in \mathbb{R}$, $\rho_j \in \mathbb{C} \setminus \mathbb{R}$, $j \in \{1, \dots, k\}$, and for $r_l \in \mathbb{R}$, $l \in \{1, \dots, m\}$. Note that

$$(\xi - \rho_j)(\xi - \bar{\rho}_j) = \xi^2 - 2\operatorname{Re}(\rho_j) + |\rho_j|^2 \in \mathbb{R}[\xi].$$

Thus A is irreducible if and only if either $k = 1$ and $l = 0$ or if $k = 0$ and $l = 1$. The latter case is condition (i) of the theorem statement. For the former case we have $a_0 = a|\rho_1|^2$, $a_1 = -2a\operatorname{Re}(\rho_1)$, and $a_2 = a$. Therefore

$$a_1^2 - 4a_2a_0 = 4a^2\operatorname{Re}(\rho_1)^2 - 4a^2|\rho_1|^2 < 0$$

since $\operatorname{Re}(\rho_1)^2 < |\rho_1|^2$. This then gives condition (ii). ■

4.7.4 Solving for roots

As we saw in Section 4.7.1, there is a formula for determining the roots of a quadratic polynomial in $\mathbb{R}[\xi]$. There are also similar formulae for cubic and quartic polynomials. Though these are messy, and are only to be used in emergencies, let us give these formulae in order to make more interesting the natural question to follow. Each of the following two theorems can be proved by simply substituting the expressions for the roots into the polynomial, and by brute force checking that they are indeed roots. In the theorems, we consider only monic polynomials, as this can be done without loss of generality.

4.7.10 Theorem (Roots of a cubic polynomial) *If* $\mathrm{A} = \xi^3 + \mathrm{a}_2\xi^2 + \mathrm{a}_1\xi + \mathrm{a}_0 \in \mathbb{R}[\xi]$, *then the roots of* A *are given by*

$$\begin{aligned} \mathrm{r}_1 &= -\tfrac{1}{3}\mathrm{a}_2 + (\mathrm{s} + \mathrm{t}), \\ \mathrm{r}_2 &= -\tfrac{1}{3}\mathrm{a}_2 - \tfrac{1}{2}(\mathrm{s} + \mathrm{t}) + \tfrac{1}{2}\mathrm{i}\sqrt{3}(\mathrm{s} - \mathrm{t}), \\ \mathrm{r}_3 &= -\tfrac{1}{3}\mathrm{a}_2 - \tfrac{1}{2}(\mathrm{s} + \mathrm{t}) - \tfrac{1}{2}\mathrm{i}\sqrt{3}(\mathrm{s} - \mathrm{t}), \end{aligned}$$

where

$$\mathrm{s} = (\mathrm{r} + \sqrt{\mathrm{d}})^{1/3}, \quad \mathrm{t} = (\mathrm{r} - \sqrt{\mathrm{d}})^{1/3},$$

and

$$\mathrm{d} = \mathrm{q}^3 + \mathrm{r}^2, \quad \mathrm{q} = \tfrac{1}{9}(3\mathrm{a}_1 - \mathrm{a}_2^2), \quad \mathrm{r} = \tfrac{1}{54}(9\mathrm{a}_1\mathrm{a}_2 - 27\mathrm{a}_0 - 2\mathrm{a}_3^3).$$

The formula for the quartic relies on finding a real root of a cubic polynomial. Note that any cubic polynomial in $\mathbb{R}[\xi]$ always has a real root (why?).

4.7.11 Theorem (Roots of a quartic polynomial) *If* $\mathrm{A} = \xi^4 + \mathrm{a}_3\xi^3 + \mathrm{a}_2\xi^2 + \mathrm{a}_1\xi + \mathrm{a}_0 \in \mathbb{R}[\xi]$, *then the roots of* A *are given by*

$$\begin{aligned} \mathrm{r}_1 &= -\tfrac{1}{4}\mathrm{a}_3 + \tfrac{1}{2}\mathrm{r} + \tfrac{1}{2}\mathrm{d}, \\ \mathrm{r}_2 &= -\tfrac{1}{4}\mathrm{a}_3 + \tfrac{1}{2}\mathrm{r} - \tfrac{1}{2}\mathrm{d}, \\ \mathrm{r}_3 &= -\tfrac{1}{4}\mathrm{a}_3 - \tfrac{1}{2}\mathrm{r} + \tfrac{1}{2}\mathrm{e}, \\ \mathrm{r}_4 &= -\tfrac{1}{4}\mathrm{a}_3 - \tfrac{1}{2}\mathrm{r} - \tfrac{1}{2}\mathrm{e}, \end{aligned}$$

where

$$\begin{aligned}
r &= (\tfrac{1}{4}a_3^2 - a_2 + s)^{1/2},\\
d &= \begin{cases} (\tfrac{3}{4}a_3^2 - r^2 - 2a_2 + \tfrac{1}{4}r^{-1}(4a_2a_3 - 8a_1 - a_3^3))^{1/2}, & r \neq 0,\\ (\tfrac{3}{4}a_3^2 - 2a_2 + 2(s^2 - 4a_0)^{1/2})^{1/2}, & r = 0,\end{cases}\\
e &= \begin{cases} (\tfrac{3}{4}a_3^2 - r^2 - 2a_2 - \tfrac{1}{4}r^{-1}(4a_2a_3 - 8a_1 - a_3^3))^{1/2}, & r \neq 0,\\ (\tfrac{3}{4}a_3^2 - 2a_2 - 2(s^2 - 4a_0)^{1/2})^{1/2}, & r = 0,\end{cases}
\end{aligned}$$

and where s *is any real root of the cubic polynomial*

$$\xi^3 - a_2\xi^2 + (a_1a_3 - 4a_0)\xi + 4(a_0a_2 - a_1^2 - a_0a_3^2).$$

The point of the above two theorems, other than that it must have been hard work to come up with the formulae for the roots, is that there are closed-form expressions for the roots of complex polynomials that are quadratic, cubic, and quartic.

In the preceding results we saw that there are formulae for the roots of quadratic, cubic, and quartic polynomials over $\mathbb{R}$. This then raises the question, "For a polynomial in $\mathbb{R}[\xi]$ of arbitrary degree, is there an expression for the roots involving addition, multiplication, and rational powers?" The answer is, "No, for there is no such formula even for quintic polynomials." In this section we give the proof of this fact. The main point of this is that finding the roots of polynomial equations is hard, and one must typically resort to numerical methods.

Let us state the theorem of Abel and Ruffini.[5] The theorem states that some quintic polynomials are not solvable by radicals. Obviously, some quintic polynomials *are* solvable by radicals (e.g., $\xi^5 - a$), so one must make sort of assumptions on the polynomial to ensure that it cannot be solved by radicals. The easiest of these is the following. Let $r_1 \in \mathbb{R}$ be transcendental. The existence of transcendental numbers is ensured since the set of algebraic numbers are countable and the set of real numbers are uncountable. Now let r_2 be transcendental over $\mathbb{Q}(r_1)$. Again, the existence of r_2 follows by a countability argument. Since $\mathbb{Q}(r_1)$ consists of rational functions in r_1, this is a countable $\mathbb{Q}$-vector space (why?). One can proceed in this way to get $r_1, r_2, r_3, r_4, r_5 \in \mathbb{R}$ such that r_j is transcendental over $\mathbb{Q}(r_1, \dots, r_{j-1})$ for $j \in \{2, 3, 4, 5\}$. Let us call five such real numbers ***sequentially transcendental***.

4.7.12 Theorem (Unsolvability of quintics by radicals) *Let* $a_0, a_1, a_2, a_3, a_4 \in \mathbb{R}$ *be such that the polynomial*

$$A = \xi^5 + a_4\xi^4 + a_3\xi^3 + a_3\xi^2 + a_1\xi + a_0$$

[5] Paolo Ruffini (1765–1822) was an Italian mathematician whose main mathematical contribution was the theorem given here. It can be argued that Ruffini was the first person to believe that the quintic equation cannot be solved by radicals. The mood in mathematics at the time of Ruffini's work was that quintics *could* be solved by radicals, so the work was largely disregarded.

has roots $\mathrm{r}_1, \mathrm{r}_2, \mathrm{r}_3, \mathrm{r}_4, \mathrm{r}_5 \in \mathbb{R}$ *that are sequentially transcendental. Then any radical extension of* $\mathbb{Q}(\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \mathrm{a}_4)$ *cannot contain these roots. In particular,* A *is not solvable by radicals.*

Proof We first claim that $\mathfrak{S}_5$ is not solvable. Let

$$\mathsf{G}_0 \subseteq \mathsf{G}_1 \cdots \subseteq \mathsf{G}_k = \mathfrak{S}_5$$

be a nested sequence of normal subgroups such that $\mathsf{G}_j/\mathsf{G}_{j-1}$ is Abelian for $j \in \{1, \dots, k\}$. By Exercise 4.1.9 this means that G_{j-1} is the kernel of the homomorphism $\phi_j \colon \mathsf{G}_j \to \mathsf{G}_j/\mathsf{G}_{j-1}$ onto an Abelian group. Therefore, if $a, b \in \mathsf{G}_j$ then $b^{-1}a^{-1}ba \in \mathsf{G}_{j-1}$ since

$$\begin{aligned}\phi_j(b^{-1}a^{-1}ba) &= \phi_j(b^{-1})\phi_j(a^{-1})\phi(b)\phi(a) = \phi_j(b^{-1})\phi_j(b)\phi_j(a^{-1})\phi_j(a)\\ &= \phi_j(a_{\mathsf{G}_j}) = e_{\mathsf{G}_j/\mathsf{G}_{j-1}}.\end{aligned}$$

A ***3-cycle*** in $\mathfrak{S}_5$ is a permutation that "cycles" three elements, say $\{j_1, j_2, j_3\}$ in the set $\{1, 2, 3, 4, 5\}$. Thus a 3-cycle maps (j_1, j_2, j_3) to (j_3, j_1, j_2) and leaves the other two elements in $\{1, 2, 3, 4, 5\}$ fixed. We shall simply denote such a 3-cycle by (j_1, j_2, j_3). We claim that each of the subgroups $\mathsf{G}_0, \dots, \mathsf{G}_k$ contains all 3-cycles. This is clear for G_k, so suppose it true for $\mathsf{G}_j, \dots, \mathsf{G}_k$. Let σ be a three cycle (j_1, j_2, j_3). Now note that

$$(j_1, j_2, j_3) = (j_5, j_1, j_3)^{-1} \circ (j_3, j_5, j_2)^{-1} \circ (j_5, j_1, j_3) \circ (j_3, j_5, j_2),$$

where j_4 and j_5 are distinct from j_1, j_2, j_3. One may verify this by direct computation. Since $(j_5, j_1, j_3), (j_3, j_5, j_2) \in \mathsf{G}_j$ by the induction hypothesis, it follows that $(j_1, j_2, j_3) \in \mathsf{G}_{j-1}$. This precludes G_0 from being the trivial group, and so precludes $\mathfrak{S}_5$ from being solvable.

Now let $\mathsf{K} = \mathbb{Q}(r_1, r_2, r_3, r_4, r_5)$. Note that

$$A = (\xi - r_1)(\xi - r_2)(\xi - r_3)(\xi - r_4)(\xi - r_5) \in \mathsf{K}[\xi],$$

and so we have

$$\begin{aligned}a_0 &= -r_1r_2r_3r_4r_5,\\ a_1 &= r_1r_2r_3r_4 + r_1r_2r_3r_5 + r_1r_2r_4r_5 + r_1r_3r_4r_5 + r_2r_3r_4r_5,\\ a_2 &= -r_1r_2r_3 - r_1r_2r_4 - r_1r_3r_4 - r_2r_3r_4 - r_1r_2r_5 - r_1r_3r_5\\ &\quad - r_2r_3r_5 - r_1r_4r_5 - r_2r_4r_5 - r_3r_4r_5,\\ a_3 &= r_1r_2 + r_1r_3 + r_2r_3 + r_1r_4 + r_2r_4 + r_3r_4 + r_1r_5 + r_2r_5 + r_3r_5 + r_4r_5,\\ a_4 &= -r_1 - r_2 - r_3 - r_4 - r_5.\end{aligned}$$

Let $\sigma \in \mathfrak{S}_5$ and note that σ permutes the roots $\{r_1, r_2, r_3, r_4, r_5\}$. Therefore, by Theorem 4.6.6we can think of σ as defining an element of $\mathrm{Aut}_{\mathbb{Q}}(\mathsf{K})$. Thus there are at least as many elements of $\mathrm{Aut}_{\mathbb{Q}}(\mathsf{K})$ as elements of $\mathfrak{S}_5$. However, since K is a splitting field for A it follows from Proposition 4.6.28 that $\mathrm{Aut}_{\mathbb{Q}}(\mathsf{K})$ has at most as many elements as $\mathfrak{S}_5$. This shows that $\mathrm{Aut}_{\mathbb{Q}}(\mathsf{K})$ is isomorphic to $\mathfrak{S}_5$ and so is not solvable. Thus K is not a radical extension, and so cannot be contained in any radical extension. ■

4.7.5 Notes

Reference for non-order on $\mathbb{C}$.
Solution by Hermite of quintic using theta functions.
Quintic proof follows [Stillwell 1994].

Exercises

4.7.1 Let $A \in \mathbb{R}[\xi]$ be an irreducible polynomial of degree 2. Show that $\mathbb{R}[\xi]/(A)$ is a field isomorphic to $\mathbb{C}$.

4.7.2 Prove Proposition 4.7.5.

4.7.3 Give the possible locations in $\mathbb{C}$ for the five roots of a quintic polynomial over $\mathbb{R}$.

4.7.4 Show that if $p \in \mathbb{R}[\xi]$ is irreducible, then either $p(\xi) = \xi - a$ for $a \in \mathbb{R}$ or $p(\xi) = (\xi - a)^2 + b^2$ where $a, b \in \mathbb{R}$ with $b \neq 0$.

Section 4.8

Modules

In this section we introduce a generalisation of vector spaces to allow scalars which lie in rings, and not just fields. The definitions in this section, for the most part, exactly mirror those for vector spaces. Indeed, a logically proper treatment would be to concern ourselves primarily with module theory, thinking of vector spaces merely as examples; this is in fact the way this material is typically treated at the graduate level in mathematics. However, since we wish to not put off the reader who only desires an understanding of the vector space theory, we cover vector space theory and module theory independently, with the cost merely being extra pages. We also, perhaps, benefit the more novice reader who wishes to learn elementary module theory by first presently a thorough account of the simpler, but related, vector space theory. We also assume that the reader of this section has read Section 4.5, so we omit some discussion and examples that might otherwise appear in our treatment.

While most of the constructions surrounding modules look rather the same as those for vector spaces, the reader should be aware: module theory is significantly more complicated than vector space theory, and it will not generally be true that the results from the Sections 4.5.1–4.5.6 will apply to modules. For example, modules do not necessarily possess bases, and so all results that rely on bases, and there are many of these, will not have direct generalisations to modules.

Do I need to read this section? This section can be bypassed on a first reading, and then read when the material is needed in Section 5.8. •

4.8.1 Definitions and basic properties

We proceed with the definitions.

4.8.1 Definition (Module) Let R be a ring. A ***left module*** (resp. ***right module***) over R, or a ***left*** **R*****-module*** (resp. ***right*** **R*****-module***), is a set M equipped with two operations: (1) **M*****-addition***, denoted by $\mathsf{M} \times \mathsf{M} \ni (x_1, x_2) \mapsto x_1 + x_2 \in \mathsf{M}$, and (2) **R*****-multiplication***, denoted by $\mathsf{R} \times \mathsf{M} \ni (r, x) \mapsto rx \in \mathsf{M}$ (resp. $\mathsf{M} \times \mathsf{R} \ni (x, r) \mapsto xr \in \mathsf{M}$). The operation of M-addition must satisfy the rules

(i) $x_1 + x_2 = x_2 + x_1$, $x_1, x_2 \in \mathsf{M}$ (***commutativity***),

(ii) $x_1 + (x_2 + x_3) = (x_1 + x_2) + x_3$, $x_1, x_2, x_3 \in \mathsf{M}$ (***associativity***),

(iii) there exists an element $0_{\mathsf{M}} \in \mathsf{M}$ with the property that $x + 0_{\mathsf{M}} = x$ for every $x \in \mathsf{M}$ (***zero element***), and

(iv) for every $x \in \mathsf{M}$, there exists an element $-x \in \mathsf{M}$ such that $x + (-x) = 0_{\mathsf{M}}$ (***negative element***), and R-multiplication must satisfy the rules

(v) $r_1(r_2x) = (r_1 \cdot r_2)x$ (resp. $(xr_1)r_2 = x(r_1 \cdot r_2)$), $r_1, r_2 \in \mathsf{R}$, $x \in \mathsf{M}$ (***associativity***),

(vi) $r(x_1+x_2) = rx_1+rx_2$ (resp. $(x_1+x_2)r = x_1r+x_2r$), $r \in \mathsf{R}$, $x_1, x_2 \in \mathsf{M}$ (***distributivity***), and

(vii) $(r_1 + r_2)x = r_1x + r_2x$ ($x(r_1 + r_2) = xr_1 + xr_2$), $r_1, r_2 \in \mathsf{R}$, $x \in \mathsf{M}$ (***distributivity*** again).

If R is a unit ring and if R-multiplication additionally satisfies

(viii) $1_{\mathsf{R}}x = x$ (resp. $x1_{\mathsf{R}} = x$), $x \in \mathsf{M}$,

then M is a ***left unity* R*-module*** (resp. ***right unity* R*-module***). •

We shall principally be interested in left modules, and will often refer to a "left module" as simply a "module," always being sure to explicitly say "right module" when this is what is meant. Moreover, if the ring R is commutative (as many of the rings we shall encounter are), then the notions of a left R-module and a right R-module are only distinguished by the notation of writing the ring element on the left or right side, respectively, of the module element. That is to say, if R is a commutative ring and M is a left R-module, then there is a natural right R-module defined simply by using R-multiplication $(x, r) \mapsto rx$.

Let us give some simple examples of modules.

4.8.2 Examples (Modules)

1. Let R be a ring. A ***trivial*** left R-module consists of one element: $\mathsf{M} = \{x\}$. the operations of M-addition and R multiplication are defined in the only way possible:
$$x + x = x, \quad rx = x.$$
One can verify that, if one takes the zero vector to be x, then this indeed gives the structure of a left R-module. Since there is essentially only one trivial R-module, it is often denoted by $\{0\}$.

2. We claim that if G is an Abelian group, then it has the structure of a module over $\mathbb{Z}$. Indeed, one can define addition and multiplication by
$$g_1 + g_2 = g_1 \cdot g_2, \quad jg = g^j$$
for $g, g_1, g_2 \in \mathsf{G}$ and for $j \in \mathbb{Z}$, and where we use the notation g^j introduced preceding Proposition 4.1.7. That this defines a (left or right, it matters not which since $\mathbb{Z}$ is commutative) $\mathbb{Z}$-module follows from the properties of the map $(j, g) \mapsto g^j$ as given in Proposition 4.1.7.

   Note that, since a $\mathbb{Z}$-module is an Abelian group using addition, it immediately follows that there is a 1–1 correspondence between Abelian groups and $\mathbb{Z}$-modules.

3. Let R be a ring and let $\mathsf{S} \subseteq \mathsf{R}$ be a subring. Then R is a left S-module if we define addition as addition in R and multiplication by $(s, r) \mapsto sr$ for $s \in \mathsf{S}$ and

$r \in \mathsf{R}$. If we define multiplication instead by $(r,s) \mapsto rs$, then this gives R the structure of a right S-module.

4. If R is a commutative ring and if I is an ideal of R, then one can verify that I is, in fact, an R-module.

   Note that this is an example having no interesting analogue for vector spaces. Indeed, if F is a field, then the only ideals of F are $\{0_{\mathsf{F}}\}$ and F. It is also true that, thinking of F as an F-vector space, the only subsets of F that are subspaces are $\{0_{\mathsf{F}}\}$ and F. In particular, there is no nontrivial strict subset of F that is a subspace of F.

   For rings, however, the story is indeed different. As a concrete example, consider the ideal $(\xi^2+1) \subseteq \mathbb{R}[\xi]$ generated by the polynomial ξ^2+1. This ideal is not equal to $\mathbb{R}[\xi]$ by Proposition 4.4.42. However, (ξ^2+1) is a module over the ring $\mathbb{R}[\xi]$.

5. Let M be an Abelian group and let R be a ring, and define R-multiplication by $(r,x) \mapsto e_{\mathsf{M}}$. One can then verify that, if M-addition is taken to be group multiplication in M, then this gives the structure of a left R-module. Note that, even if R is a unit ring, it is not the case that M is a unity module, unless $\mathsf{M} = \{e_{\mathsf{M}}\}$.

6. For a ring R and for $n \in \mathbb{Z}_{>0}$ we let R^n denote the n-fold Cartesian product of R with itself. We define R^n-addition and R-multiplication in R^n just as we did for the vector space F^n in Example 4.5.2–2:

$$(r_1,\dots,r_n) + (s_1,\dots,s_n) = (r_1+s_1,\dots,r_n+s_n), \quad r(r_1,\dots,r_n) = (rr_1,\dots,rr_n).$$

   This makes R^n into a left R-module. If we instead defined R-multiplication by

$$(r_1,\dots,r_n)r = (r_1r,\dots,r_nr),$$

   then we have a right R-module.

7. Let R be a ring, let S be a set, and let R^S be the set of maps from S to R. We define sum and multiplication in R^S by

$$(f+g)(x) = f(x) + g(x), \quad (rf)(x) = r(f(x))$$

   for $f,g \in \mathsf{R}^S$ and for $r \in \mathsf{R}$. These make R^S into a left R-module. One can also define multiplication by $(fr)(x) = (f(x))r$ to give a right R-module.

8. If R is a commutative ring and if M is a left R-module, then M is also naturally a right R-module with multiplication by ring elements defined by $xr = rx$ for $r \in \mathsf{R}$ and $x \in \mathsf{M}$. One can check, using commutativity of the ring, this does indeed define a right R-module structure. •

Certain of the properties of vector space operations also apply to modules. However, not all do, so we ask the reader to compare the following result with Proposition 4.5.3, and also to refer to Exercise 4.8.1 for further discussion.

4.8.3 Proposition (Properties of modules) *Let* R *be a field and let* M *be a left* R*-module. The following statements hold:*

(i) *there exists exactly one element* $0_M \in M$ *such that* $x + 0_M = x$ *for all* $x \in M$*;*

(ii) *for each* $x \in M$ *there exists exactly one element* $-x \in M$ *such that* $x + (-x) = 0_M$*;*

(iii) $r0_M = 0_M$ *for all* $r \in R$*;*

(iv) $0_R v = 0_M$ *for each* $x \in M$*;*

(v) $r(-x) = (-r)x = -(rx)$ *for all* $r \in R$ *and* $x \in M$*.*

Proof This follows, *mutatis mutandis*, the proof of Proposition 4.5.3. ■

For modules we have the analogue of linear maps between vector spaces.

4.8.4 Definition (Module homomorphism) Let R be a ring and let M and N be a left R-modules. An ***R-homomorphism*** of M and N is a map $L: M \to N$ having the properties that

(i) $L(x_1 + x_2) = L(x_1) + L(x_2)$ for every $x_1, x_2 \in M$ and

(ii) $L(rx) = rL(x)$ for every $r \in R$ and $x \in M$.

An R-homomorphism L is an ***R-monomorphism*** (resp. ***R-epimorphism***, ***R-isomorphism***) if L is injective (resp. surjective, bijective). If there exists an isomorphism between left R-modules M and N, then M and N are ***R-isomorphic***. An R-homomorphism from M to itself is called an ***R-endomorphism*** of M. The set of R-homomorphisms from M to N is denoted by $\mathrm{Hom}_R(M; N)$, and the set of R-endomorphisms of M is denoted by $\mathrm{End}_R(M)$. •

4.8.5 Notation ("Linear map" versus "homomorphism") We shall reserve the term "linear map" for a homomorphism of vector spaces. The term "homomorphism" may be used for both vector spaces and modules. •

4.8.2 Submodules

Associated with the concept of a module are various concepts that mirror those for vector spaces. The following is one such.

4.8.6 Definition (Submodule) Let R be a ring. A subset N of a left (resp. right) R-module M is a ***submodule*** if $x_1 + x_2 \in N$ for all $x_1, x_2 \in N$, and if $rx \in N$ (resp. $xr \in N$) for every $r \in R$ and for every $x \in N$. •

Of course, it holds that a submodule is itself a module.

4.8.7 Proposition (A submodule is a module) *Let* R *be a ring. A nonempty subset* $N \subseteq M$ *of an* R*-module* M *is a (left or right) submodule if and only if* N *is a (left or right) module using the operations of* M*-addition and* R*-multiplication in* M*, restricted to* N*.*

Proof The proof follows that for vector spaces, and the proof in this case is Exercise 4.5.11. ■

As with linear maps, one can define the notions of kernel and image of a homomorphism, and these are submodules.

4.8.8 Definition (Kernel and image of a module homomorphism) Let R be a ring, let N and M be (left or right) R-modules, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N};\mathsf{M})$.

(i) The ***image*** of L is $\mathrm{image}(\mathsf{L}) = \{\mathsf{L}(y) \mid y \in \mathsf{N}\}$.

(ii) The ***kernel*** of L is $\ker(\mathsf{L}) = \{y \in \mathsf{N} \mid \mathsf{L}(y) = 0_{\mathsf{M}}\}$. •

4.8.9 Proposition (Kernel and image are submodules) *Let R be a ring, let N and M be (left or right) R-modules, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N};\mathsf{M})$. Then $\mathrm{image}(\mathsf{L})$ and $\ker(\mathsf{L})$ are submodules of M and N, respectively.*

Proof This follows as in the vector space case, and the proof in this case is Exercise 4.5.16. ■

One can also form linear combinations from subsets of a module, and the resulting set of linear combinations is a submodule, in complete analogy to the situation with vector spaces.

4.8.10 Definition (Linear combination) Let R be a ring and let M be an R-module. If $S \subseteq \mathsf{M}$ is nonempty, a ***linear combination*** from S is an element of M of the form

$$c_1x_1 + \cdots + c_kx_k,$$

where $c_1,\dots,c_k \in \mathsf{R}$ and $x_1,\dots,x_k \in \mathsf{M}$. We call $c_1,\dots,c_k$ the ***coefficients*** in the linear combination. •

4.8.11 Proposition (The set of linear combinations is a submodule) *If R is a ring, if M is a (left or right) R-module, and if $\mathsf{S} \subseteq \mathsf{M}$ is nonempty, then the set of linear combinations from S is a submodule of M.*

Proof Follows the proof of Proposition 4.5.11, *mutatis mutandis.* ■

Note that we did *not* state in the preceding result that the set of linear combinations from a set S is the smallest submodule containing S, as we did with vector spaces in Proposition 4.5.11. Indeed, this is not true. What is true is the following result, which recalls Notation 4.1.8.

4.8.12 Proposition *If R is a ring, if M is a left (resp. right) R-module, and if $\mathsf{S} \subseteq \mathsf{M}$ is nonempty, then the smallest submodule of M containing S is the subset given by*

$$\left\{\sum_{j=1}^{k} r_jx_j + \sum_{l=1}^{m} k_ly_l \;\middle|\; k,m \in \mathbb{Z}_{>0},\ r_1,\dots,r_k \in \mathsf{R},\right.$$
$$\left. y_1,\dots,y_m \in \mathbb{Z},\ x_1,\dots,x_k,y_1,\dots,y_m \in \mathsf{S}\right\} \quad (4.25)$$

(resp.

$$\left\{\sum_{j=1}^{k} \mathrm{x_j r_j} + \sum_{l=1}^{m} \mathrm{k_l y_l} \;\middle|\; \mathrm{k, m} \in \mathbb{Z}_{>0},\ \mathrm{r_1, \dots, r_k} \in \mathsf{R}, \right.$$
$$\left. \mathrm{y_1, \dots, y_m} \in \mathbb{Z},\ \mathrm{x_1, \dots, x_k, y_1, \dots, y_m} \in \mathsf{S} \right\}.\Bigg)$$

Moreover, if R *is a unit ring and if* M *is a unity module, then the smallest submodule of* M *containing* S *is equal to the set of linear combinations from* S*.*

Proof We shall give the proof only for left modules, the proof for right modules differing only in notation.

Let N_S be the set defined in (4.25). We leave it to the reader to straightforwardly check that N_S is a submodule. Then suppose that N is a submodule of M containing S. Since N is a submodule, for $k \in \mathbb{Z}$ and $x \in S$ we have $kx \in \mathsf{N}$, and for $r \in \mathsf{R}$ and $x \in S$ we have $rx \in \mathsf{N}$. It then immediately follows, again since N is a submodule, that $\mathsf{N}_S \subseteq \mathsf{N}$. Thus N_S is contained in any submodule containing S, and so is the smallest such submodule.

If R is a unit ring and M is unitary, then, for $k \in \mathbb{Z}$ and $x \in S$, we have $kx = k(1_{\mathsf{R}}x) = (k \cdot 1_{\mathsf{R}})x$. This shows that N_S is contained in the set of linear combinations from S. ■

With this characterisation, the following definition admits a ore or less explicit description.

4.8.13 Definition (Submodule generated by a set) If R is a ring, if M is a (left or right) R-module, and if $S \subseteq \mathsf{M}$ is nonempty, then the ***submodule generated by* S** is the smallest submodule of M containing S. This submodule is denoted by $\mathrm{span}_{\mathsf{R}}(S)$. •

Let us give an example that illustrates how vector spaces and modules can differ with regard to the notion of subspaces and submodules generated by sets.

4.8.14 Example (Submodule generated by a set) Consider the ring $\mathsf{R} = 2\mathbb{Z}$ of even integers, and note that $\mathsf{M} = \mathbb{Z}$ is a left R-module (see Example 4.8.2–3). Let N be the submodule of M generated by $\{1\}$. Then we see that the odd integers are elements of N, but are not of form rx for $r \in \mathsf{R}$ and $x \in S$. That is, there are elements in the submodule generated by S that are not linear combinations of elements of S. •

4.8.3 Linear independence, basis, and rank

To this point, the definitions and simple results, with minor and essentially uninteresting exceptions, have exactly mirrored the development of vector spaces in Section 4.5. In this section we see one of the principal ways in which modules differ from vector spaces: they do not generally possess bases. This lies at the heart of a great deal of why module theory is more difficult than vector space theory.

Nonetheless, one can *define* the concept of a basis for a module, so let us do this. We first define the notion of linear independence.

4.8.15 Definition (Linearly independent) Let R be a ring and let M be a left (resp. right) R-module.

(i) A finite family $(x_1, \dots, x_k)$ of elements of M is ***linearly independent*** if the equality

$$c_1x_1 + \cdots + c_kx_k = 0_{\mathsf{M}} \text{ (resp. } x_1c_1 + \cdots + x_kc_k = 0_{\mathsf{M}}), \qquad c_1, \dots, c_k \in \mathsf{R},$$

is satisfied only if $c_1 = \cdots = c_k = 0_{\mathsf{R}}$.

(ii) A finite set $S = \{x_j \mid j \in \{1, \dots, k\}\}$ is linearly independent if the finite family corresponding to the set is linearly independent.

(iii) An nonempty family $(v_a)_{a\in A}$ of vectors in V is ***linearly independent*** if every finite subfamily of $(v_a)_{a\in A}$ is linearly independent.

(iv) A nonempty subset $S \subseteq \mathsf{M}$ is linearly independent if every nonempty finite subset of S is linearly independent.

(v) A nonempty subset $S \subseteq \mathsf{M}$ is ***linearly dependent*** if it is not linearly independent. •

As we did for vector spaces in Proposition 4.5.17, one can show that the two possibly conflicting notions of linear independence of finite sets of vectors are actually not in conflict.

We have the following properties of linearly independent and linearly dependent sets.

4.8.16 Proposition (Properties of linearly (in)dependent sets) *Let R be a ring, let M be a left (resp. right) R-module, and let $\mathsf{S} \subseteq \mathsf{M}$ be nonempty. Then the following statements hold:*

(i) if $\mathsf{S} = \{\mathsf{x}\}$ for some $\mathsf{x} \in \mathsf{M}$, then S is linearly independent if and only if $\mathsf{x} \neq 0_{\mathsf{M}}$;

(ii) if $0_{\mathsf{M}} \in \mathsf{S}$ then S is linearly dependent;

(iii) if S is linearly independent and if $\mathsf{T} \subseteq \mathsf{S}$ is nonempty, then T is linearly independent;

(iv) if S is linearly dependent and if $\mathsf{T} \subseteq \mathsf{M}$, then $\mathsf{S} \cup \mathsf{T}$ is linearly dependent;

(v) if S is linearly independent, if $\{\mathsf{x}_1, \dots, \mathsf{x}_k\} \subseteq \mathsf{S}$, and if

$$\mathsf{r}_1\mathsf{x}_1 + \cdots + \mathsf{r}_k\mathsf{x}_k = \mathsf{s}_1\mathsf{x}_1 + \cdots + \mathsf{s}_k\mathsf{x}_k \qquad (\textit{resp. } \mathsf{x}_1\mathsf{r}_1 + \cdots + \mathsf{x}_k\mathsf{r}_k = \mathsf{x}_1\mathsf{s}_1 + \cdots + \mathsf{x}_k\mathsf{s}_k)$$

for $\mathsf{r}_,1 \dots, \mathsf{r}_k, \mathsf{s}_1, \dots, \mathsf{s}_k \in \mathsf{R}$, then $\mathsf{r}_j = \mathsf{s}_j$, $\mathsf{j} \in \{1, \dots, \mathsf{k}\}$.

Proof This follows along the same lines as the proof of Proposition 4.5.19. ■

Note that we did not include part (vi) from Proposition 4.5.19 in the preceding result. This is because it is false, as the following example shows.

4.8.17 Example (A counterexample on linear independence) Let $\mathsf{R} = \mathbb{Z}$ and let $\mathsf{M} = \mathbb{Z}^2$. Let $S = \{(1,0),(0,2)\}$ and take $\boldsymbol{x} = (0,1)$. Then $\boldsymbol{x} \notin \operatorname{span}_{\mathbb{Z}}(S)$, but we have

$$(0,2) - 2(0,1) = (0,0),$$

giving a linear combination in $S \cup \{\boldsymbol{x}\}$ that sums to the zero vector, but whose coefficients are nonzero. Thus S is linearly independent, $\boldsymbol{x} \notin \operatorname{span}_{\mathsf{R}}(S)$, but $S \cup \{\boldsymbol{x}\}$ is linearly dependent. This situation cannot happen for vector spaces. •

We may now define the notion of a basis for a module in exactly the same way as we did for a vector space.

4.8.18 Definition (Basis for a module) Let R be a ring and let M be a left module over R. A ***basis*** for M is a subset $\mathscr{B}$ of M with the properties that

(i) $\mathscr{B}$ is linearly independent and

(ii) $\operatorname{span}_{\mathsf{R}}(\mathscr{B}) = \mathsf{M}$.

A module which possesses a basis is a ***free module***. •

First let us settle that the analogue of Theorem 4.5.22 for vector spaces does not hold for modules.

4.8.19 Example (A module that is not free) Let $\mathsf{R} = \mathbb{Z}_6 = \mathbb{Z}/6\mathbb{Z}$ and take M to be the ideal generated by $2 + 6\mathbb{Z}$. As we indicated in Example 4.8.2–4, an ideal in a ring is a module. In this case we may easily verify that

$$\mathsf{M} = \{0 + 6\mathbb{Z}, 2 + 6\mathbb{Z}, 4 + 6\mathbb{Z}\}.$$

We claim that M has no basis. Indeed, we have

$$\begin{aligned}(3 + 6\mathbb{Z})(2 + 6\mathbb{Z}) &= 0 + 6\mathbb{Z},\\ (3 + 6\mathbb{Z})(2 + 6\mathbb{Z}) &= 0 + 6\mathbb{Z},\\ (1 + 6\mathbb{Z})(2 + 6\mathbb{Z}) + (2 + 6\mathbb{Z})(4 + 6\mathbb{Z}) &= 0 + 6\mathbb{Z},\end{aligned}$$

which shows that, for any finite subset of M not containing the zero element, there exists a set of nonzero coefficients in $\mathbb{Z}_6$ for which the corresponding linear combination is zero. In particular, any finite subset of M is linearly dependent, and so M cannot be free.

Furthermore, note that $\mathbb{Z}_6$ is itself a $\mathbb{Z}_6$-module with basis $1 + 6\mathbb{Z}$. Thus our example also shows that, even when a module possesses a basis, it can have submodules that do not possess a basis. •

Another pitfall of which to be wary is that submodules of a free module may not be free. An example illustrates this.

4.8.20 Example (A free module with a nonfree submodule) We let $\mathsf{R} = \mathbb{Z} \times \mathbb{Z}$ and define a ring structure on R by defining addition and scalar multiplication by

$$(j_1, k_1) + (j_2, k_2) = (j_1 + j_2, k_1 + k_2), \qquad (j_1, k_1) \cdot (j_2, k_2) = (j_1 j_2, k_1 k_2),$$

respectively. We leave it to the reader to verify straightforwardly that these operations do indeed make R into a commutative unit ring with identity $(1, 1)$. Thus R is a module over itself (Example 4.8.2–3). Moreover, the module is free as $\{(1, 1)\}$ is easily verified to be a basis.

Now consider the subset $\mathsf{N} = \mathbb{Z} \times \{0\} \subseteq \mathsf{R}$. One can easily verify that this is a submodule. However, if $(j, 0) \in \mathsf{N}$ then $(0, 1) \cdot (j, 0) = (0, 0)$, whence the set $\{(j, 0)\}$ is linearly dependent. Therefore, *every* subset of N is linearly dependent, and this prohibits N from being free. •

Since modules do not always possess a basis, one might try to relax the notion of basis by not requiring the set to be linearly independent. Doing so gives the following notion.

4.8.21 Definition (Generators for a module) If R is a ring and if M is a (left or right) R-module, a set of ***generators*** for M is a subset $\mathscr{G} \subseteq \mathsf{M}$ such that $\mathrm{span}_{\mathsf{R}}(\mathscr{G}) = \mathsf{M}$. The module M is ***finitely generated*** if it possesses a finite set of generators. •

While not every module is free, it is true that every module possesses a set of generators: for example, the module itself is obviously a set of generators. However, one typically wants to choose a smaller set of generators. It is typically nontrivial to determine the minimum number of generators needed for a given module.

Having established that not all modules are free, one can then ask, “Do any modules, apart from vector spaces, possess bases?” The answer to this question is, “Yes,” and we refer to the notion of direct sum in Example 4.8.36 for an important example of a module over a general ring having a basis. Next one can ask, “If a module possesses a basis, does it hold that all bases have the same cardinality?” The answer here is, “No,” as the following example shows.

4.8.22 Example (A module having bases of different size) This example uses some concepts from Section 5.2.1, in particular Theorem 5.2.11.

Let R be a unit ring and let R_0^∞ be the set of maps from $\mathbb{Z}_{>0}$ into R having the property that if $f \in \mathsf{R}_0^\infty$ then the set

$$\{j \in \mathbb{Z}_{>0} \mid f(j) \neq 0_{\mathsf{R}}\}$$

is finite (cf. Example 4.5.2–4 and Example 4.8.36 below). As we shall see in Corollary 5.5.6, the set $\mathsf{S} = \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^\infty; \mathsf{R}_0^\infty)$ has the structure of a ring, and therefore the structure of a left module over itself. We claim that, for any $k \in \mathbb{Z}_{>0}$, there exists a basis for S as a left S-module with k elements. To see this, let $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}}$ be the standard basis for R_0^∞ defined by

$$\boldsymbol{e}_j(k) = \begin{cases} 1_{\mathsf{R}}, & j = k, \\ 0_{\mathsf{R}}, & j \neq k. \end{cases}$$

To represent an element of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^\infty; \mathsf{R}_0^\infty)$ we use matrices with countably infinite rows and columns and with entries in R, following Theorem 5.2.11. Let us denote by $\boldsymbol{\varepsilon}(i, j)$ the infinite matrix all of whose entries are 0_{R} except that in the ith row and jth column, which is 1_{R}. In terms of the standard basis, $\boldsymbol{\varepsilon}(i, j)$ is uniquely defined by its mapping the ith standard basis element to the jth standard basis element. For an arbitrary $\boldsymbol{A} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^\infty; \mathsf{R}_0^\infty)$ we note that $\boldsymbol{A} \circ \boldsymbol{\varepsilon}(i, j)$ is homomorphism all of whose columns are zero except the jth column which is equal to the ith column of $\boldsymbol{A}$. For $k \in \mathbb{Z}_{>0}$ define k elements $\boldsymbol{E}_1^k, \dots, \boldsymbol{E}_k^k$ of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^\infty; \mathsf{R}_0^\infty)$ as follows:

1. if $k = 1$ take $\boldsymbol{E}_1^1 = \mathrm{id}_{\mathsf{R}_0^\infty}$;
2. otherwise define

$$\boldsymbol{E}_l^k = \sum_{j=0}^{\infty}\left(\boldsymbol{\varepsilon}(jk+1, (j+1)k) + \sum_{m=1}^{k-1} \boldsymbol{\varepsilon}(jk+l+1, jk+l)\right).$$

By parsing the definition, one can see that, for $k > 1$, for $l \in \{1, \dots, k\}$, and for $\boldsymbol{A} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^\infty; \mathsf{R}_0^\infty)$, the homomorphism $\boldsymbol{A} \circ \boldsymbol{E}_l^k$ has the property that, for $j \in \mathbb{Z}_{>0}$, its $((j-1)k + l)$th column is the jth column of $\boldsymbol{A}$, and all other columns of $\boldsymbol{A} \circ \boldsymbol{E}_l^k$ are zero. Thus the effect of $\boldsymbol{E}_k^l$ is to "expand" the columns of $\boldsymbol{A}$ so that they lie in the $jk + l$th columns, $j \in \mathbb{Z}_{>0}$, of $\boldsymbol{A} \circ \boldsymbol{E}_k^l$.

It is clear that $\mathscr{B}_1 = \{\boldsymbol{E}_1^1\}$ is a basis for S as a left S-module. Indeed, the set $\mathscr{B}_1$ is clearly linearly independent, and also, for any $\boldsymbol{A} \in \mathsf{S}$, we have $\boldsymbol{A} = \boldsymbol{A} \circ \boldsymbol{E}_1^1$, so $\mathsf{S} = \mathrm{span}_{\mathsf{S}}(\mathscr{B}_1)$.

We claim that, for each $k \in \mathbb{Z}_{>0}$, the set $\mathscr{B}_k = \{\boldsymbol{E}_1^k, \dots, \boldsymbol{E}_k^k\}$ is a basis for the S-module S. First suppose that

$$\boldsymbol{A}_1\boldsymbol{E}_1^k + \cdots + \boldsymbol{A}_k\boldsymbol{E}_k^k = \boldsymbol{0}_{\infty\times\infty}, \tag{4.26}$$

where by $\boldsymbol{0}_{\infty\times\infty}$ we mean the zero element of S. By construction of the homomorphisms $\boldsymbol{E}_l^k$, $l \in \{1, \dots, k\}$, it follows that all columns of all matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ appear as some column of the left-hand side of (4.26). From this it follows that all columns of all of the matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are zero. This gives linear independence of $\mathscr{B}_k$. Now let $\boldsymbol{A} \in \mathsf{S}$. Define $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k \in \mathsf{S}$ as follows. For $l \in \{1, \dots, k\}$ and for $j \in \mathbb{Z}_{>0}$, the jth column of $\boldsymbol{A}_l$ is the $((j-1)k + l)$th column of $\boldsymbol{A}$. Thus $\boldsymbol{A}_l$ "collapses" some of the columns of $\boldsymbol{A}$. One can readily verify that, if $\boldsymbol{A}_1, \dots, \boldsymbol{A}_k$ are defined in this way, we have

$$\boldsymbol{A} = \boldsymbol{A}_1\boldsymbol{E}_1^k + \cdots + \boldsymbol{A}_k\boldsymbol{E}_k^k.$$

Thus $\mathsf{S} = \operatorname{span}_{\mathsf{S}}(\mathscr{B}_k)$.

Thus we have an example of a module which has finite bases of all possible cardinality. •

Note that the preceding example was one where we constructed collections of *finite* bases of arbitrary size. One might wonder whether a module with an infinite basis can possess bases with different cardinality. It turns out that this is not possible.

4.8.23 Theorem (All bases have the same size if there is an infinite basis) *If* R *is a unit ring and if* M *is a (left or right)* R*-module possessing a basis* $\mathscr{B}$ *such that* $\operatorname{card}(\mathscr{B}) \geq \operatorname{card}(\mathbb{Z}_{>0})$*, then any other basis for* M *has the same cardinality as* $\mathscr{B}$*.*

Proof First we show that if M possesses an infinite basis, then any other basis cannot be finite. Thus let $\mathscr{B}$ be an infinite basis and suppose that $\mathscr{B}'$ is a finite basis. Since $\mathscr{B}'$ generates M, and since every one of the finite elements of $\mathscr{B}'$ is itself a finite linear combination of elements of $\mathscr{B}$, this means that there exists a finite subset $\{x_1, \dots, x_k\} \subseteq \mathscr{B}$ which generates M. In particular, if $x \in \mathscr{B} \setminus \{x_1, \dots, x_m\}$, it follows that x is a linear combination of the elements $\{x_1, \dots, x_k\}$. This contradicts linear independence of $\mathscr{B}$, and so we conclude that if one basis is finite, all bases must be finite.

The matter of showing that two infinite bases have the same cardinality now goes just like that part of the proof of Theorem 4.5.25 where we showed that two infinite bases for a vector space have the same cardinality. We leave to the reader the matter of checking that the argument indeed carries through to the more general case. ■

Based on this, one can make the following definition.

4.8.24 Definition (Invariant rank property) A ring R has the ***invariant rank property*** if, whenever M is a (left or right) R-module possessing a basis, the cardinality of any two bases of M agree. •

The content of Theorem 4.5.25 is then that every field has the invariant rank property. In fact, commutative unit rings have the invariant rank property as the following result shows.

4.8.25 Theorem (Commutative unit rings have the invariant rank property) *A commutative unit ring* R *has the invariant rank property.*

Proof Let M be an R-module and let $\mathsf{I} \subseteq \mathsf{R}$ be a maximal ideal, the existence of which follows from Theorem 4.2.19. Note that R/I is then a field by Theorem 4.3.9. Denote by IM the submodule

$$\mathsf{IM} = \{r_1x_1 + \dots + r_kx_k \mid r_1, \dots, r_k \in \mathsf{I},\ x_1, \dots, x_k \in \mathsf{M},\ k \in \mathbb{Z}_{>0}\}.$$

We claim that M/IM is a vector space over R/I using the natural addition and the scalar multiplication defined by

$$(r + \mathsf{I})(x + \mathsf{IM}) = rx + \mathsf{IM}.$$

It should be verified that this definition makes sense. Thus suppose that

$$r + \mathsf{I} = r' + \mathsf{I}, \quad x + \mathsf{IM} = x' + \mathsf{IM}$$

and compute

$$r'x' + \mathsf{IM} = (r' - r)x' - r(x - x') + rx + \mathsf{IM} = rx + \mathsf{IM},$$

as desired. One should also verify that the module axioms are satisfied for this definition, but this is elementary and we leave it for the reader to supply the details.

Let $\mathscr{B} \subseteq \mathsf{M}$ and denote

$$\mathscr{B} + \mathsf{IM} = \{y + \mathsf{IM} \mid\ y \in \mathscr{B}\}.$$

We claim that $\mathscr{B} + \mathsf{IM}$ generates M/IM as an R/I-vector space if $\mathscr{B}$ generates M as an R-module. Indeed, let $x + \mathsf{IM} \in \mathsf{M}/\mathsf{IM}$ and write $x = \sum_{j=1}^{k} c_j x_j$ for some $x_1, \dots, x_k \in \mathscr{B}$ and $c_1, \dots, c_k \in \mathsf{R}$. Then

$$x + \mathsf{IM} = \left(\sum_{j=1}^{k} c_j x_j\right) + \mathsf{IM} = \sum_{j=1}^{k} c_j(x_j + \mathsf{IM}) = \sum_{j=1}(c_j + \mathsf{I})(x_j + \mathsf{IM}),$$

as desired.

We also claim that if $\mathscr{B}$ is linearly independent over R then $\mathscr{B} + \mathsf{IM}$ is linearly independent over R/I. Suppose that

$$\sum_{j=1}^{k}(c_j + \mathsf{I})(x_j + \mathsf{IM}) = 0_{\mathsf{M}/\mathsf{IM}}$$

for $x_j + \mathsf{IM} \in \mathscr{B} + \mathsf{IM}$, $j \in \{1, \dots, k\}$. Then $\sum_{j=1}^{k} c_j x_j \in \mathsf{IM}$ which implies that

$$\sum_{j=1}^{k} c_j x_j = \sum_{j=1}^{m} c'_j x'_j$$

for some $c'_j \in \mathsf{I} \setminus \{0_{\mathsf{R}}\}$, $x'_j \in \mathscr{B}$, $j \in \{1, \dots, m\}$. Since $\mathscr{B}$ is linearly independent it follows that $m = k$ and $c_j = c'_j$, $j \in \{1, \dots, k\}$. Thus $c_j \in \mathsf{I}$ and so $c_j + \mathsf{I} = 0_{\mathsf{R}/\mathsf{I}}$, $j \in \{1, \dots, k\}$, as desired.

The above argument shows that if $\mathscr{B}$ is a basis for M then $\mathscr{B} + \mathsf{IM}$ is a basis for M/IM. It remains to show that $\operatorname{card}(\mathscr{B}) = \operatorname{card}(\mathscr{B} + \mathsf{IM})$, which will follow if we can show that the elements of the set $\mathscr{B} + \mathsf{IM}$ are distinct. Suppose that $x + \mathsf{IM} = x' + \mathsf{IM}$ for $x, x' \in \mathscr{B}$. Then $x - x' = \sum_{j=1}^{k} c_j x_j$ for $c_j \in \mathsf{I}$ and $x_j \in \mathscr{B}$, $j \in \{1, \dots, k\}$. If $x \neq x'$ then it follows there exists $j \in \{1, \dots, k\}$ such that $x_j = x$ and $a_j = 1_{\mathsf{R}}$, contradicting the fact that I is maximal. ■

Despite the fact that two bases, should a basis even exist, may not have the same cardinality, it still holds that any element of a module is uniquely expressed as a linear combination of basis elements.

4.8.26 Proposition (Unique representation of elements in bases) *If* R *is a ring, if* M *is a left (resp. right)* R*-module, and if* $\mathscr{B}$ *is a basis for* M*, then, for* $\mathsf{x} \in \mathsf{M}$ *there exists a unique finite subset* $\{\mathsf{x}_1, \dots, \mathsf{x}_k\} \subseteq \mathscr{B}$ *and unique nonzero coefficients* $\mathsf{c}_1, \dots, \mathsf{c}_k \in \mathsf{R}$ *such that*

$$\mathsf{x} = \mathsf{c}_1\mathsf{x}_1 + \dots + \mathsf{c}_k\mathsf{x}_k \ (\textit{resp. } \mathsf{x} = \mathsf{x}_1\mathsf{c}_1 + \dots + \mathsf{x}_k\mathsf{c}_k).$$

Proof This follows as does the proof of Proposition 4.5.23. ■

Finally, let us define the analogue of dimension for modules, in the cases when it can be defined.

4.8.27 Definition (Rank of a module) Let R be a ring, let M be a (left or right) R-module having the property that, if $\mathscr{B}_1$ and $\mathscr{B}_2$ are bases for M, then $\operatorname{card}(\mathscr{B}_1) = \operatorname{card}(\mathscr{B}_2)$. Then, if $\mathscr{B}$ is a basis for M, $\operatorname{card}(\mathscr{B})$ is the ***rank*** of M, denoted by $\operatorname{rank}(\mathsf{M})$. •

4.8.4 Intersections, sums, and products

Next, following what we did for vector spaces, we discuss ways of manipulating submodules of a module. The definitions here mirror those for vector spaces. However, not all of the statements we make in Section 4.5.5 have analogues for modules, with the obstruction typically being that not all modules are free.

We first proceed with the definitions.

4.8.28 Definition (Sum and intersection) Let R be a ring, let M be a (left or right) R-module, and let $(\mathsf{N}_j)_{j\in J}$ be a family of submodules of M indexed by a set J.

(i) The ***sum*** of $(\mathsf{N}_j)_{j\in J}$ is the submodule generated by $\cup_{j\in J}\mathsf{N}_j$, and is denoted by $\sum_{j\in J}\mathsf{N}_j$.

(ii) The ***intersection*** of $(\mathsf{N}_j)_{j\in J}$ is the set $\cap_{j\in J}\mathsf{N}_j$ (i.e., the set theoretic intersection). •

As with vector spaces, we will often write finite sums of submodules as

$$\sum_{j=1}^{k}\mathsf{N}_j = \mathsf{N}_1 + \dots + \mathsf{N}_k.$$

It is also true that the intersection of a family of submodules is a submodule, the proof going just like that in the vector space case.

As with vector spaces, a special rôle is played by so-called direct sums. Let us state the definitions and results; the proofs follow the vector space case, *mutatis mutandis*.

4.8.29 Definition (Internal direct sum of submodules) Let R be a ring, let M be a (left or right) R-module, and let $(\mathsf{N}_j)_{j\in J}$ be a collection of submodules of M. The module M is the ***internal direct sum*** of the submodules $(\mathsf{N}_j)_{j\in J}$, and we write $\mathsf{M} = \bigoplus_{j\in J}\mathsf{N}_j$, if, for any $x \in \mathsf{M} \setminus \{0_{\mathsf{M}}\}$, there exists unique indices $\{j_1, \dots, j_k\} \subseteq J$ and unique nonzero members $y_{j_l} \in \mathsf{N}_{j_l}$, $l \in \{1, \dots, k\}$, such that $x = y_{j_1} + \dots + y_{j_k}$. Each of the submodules N_j, $j \in J$, is a ***summand*** in the internal direct sum. •

4.8.30 Proposition (Representation of the zero vector in an internal direct sum of submodules) *Let* R *be a ring, let* M *be a (left or right)* R*-module, and suppose that* M *is the internal direct sum of the submodules* $(\mathsf{N}_j)_{j\in J}$*. If* $j_1,\dots,j_k \in J$ *are distinct and if* $y_{j_l} \in \mathsf{N}_{j_l}$*,* $l \in \{1,\dots,k\}$*, satisfy*

$$y_{j_1} + \dots + y_{j_k} = 0_{\mathsf{M}},$$

then $y_{j_l} = 0_{\mathsf{M}}$*,* $l \in \{1,\dots,k\}$.

4.8.31 Proposition (Characterisation of internal direct sum for modules) *Let* R *be a ring, let* M *be a (left or right)* R*-module, and let* $(\mathsf{N}_j)_{j\in J}$ *be a collection of submodules of* M*. Then* $\mathsf{M} = \bigoplus_{j\in J} \mathsf{N}_j$ *if and only if*

(i) $\mathsf{N} = \sum_{j\in J} \mathsf{N}_j$ *and*

(ii) if, for any $j_0 \in J$*, we have* $\mathsf{N}_{j_0} \cap \left(\sum_{j\in J\setminus\{j_0\}} \mathsf{N}_j\right) = \{0_{\mathsf{M}}\}$.

While it is not the case that modules always possess bases, there still holds the analogy between bases and internal direct sums, when the former do exist. Specifically, we have the following result, whose proof follows that of Theorem 4.5.38.

4.8.32 Theorem (Bases and internal direct sums for modules) *Let* R *be a ring, let* M *be a (left or right)* R*-module, and let* $\mathscr{B}$ *be a basis for* M*, and define a family* $(\mathsf{N}_y)_{y\in\mathscr{B}}$ *of submodules by* $\mathsf{N}_y = \operatorname{span}_{\mathsf{R}}(y)$*. Then* $\mathsf{M} = \bigoplus_{y\in\mathscr{B}} \mathsf{N}_y$*.*

Note that the result holds even when the cardinality of different bases are not the same.

4.8.33 Example (Example 4.8.22 cont'd) In Example 4.8.22 we gave an example of a ring S such that the S-module S had bases of any finite cardinality. From this it follows that the S-modules S, $\mathsf{S}\oplus\mathsf{S}$, $\mathsf{S}\oplus\mathsf{S}\oplus\mathsf{S}$, etc., are all isomorphic! •

The following definition of direct product and direct sum mirrors the situation for vector spaces. We make use here of the notion of the general Cartesian product from Section 1.6.2.

4.8.34 Definition (Direct product and direct sum of modules) Let R be a ring and let $(\mathsf{M}_j)_{j\in J}$ be a family of (left or right) R-modules.

(i) The ***direct product*** of the family $(\mathsf{M}_j)_{j\in J}$ is the R-module $\prod_{j\in J} \mathsf{M}_j$ with addition and multiplication defined by

$$(f_1 + f_2)(j) = f_1(j) + f_2(j), \quad (rf)(j) = r(f(j))$$

for $f, f_1, f_2 \in \prod_{j\in J} \mathsf{M}_j$ and for $r \in \mathsf{R}$.

(ii) The ***direct sum*** of the family $(\mathsf{M}_j)_{j\in J}$ is the submodule $\bigoplus_{j\in J} \mathsf{M}_j$ of $\prod_{j\in J} \mathsf{M}_j$ consisting of those elements $f\colon J \to \cup_{j\in J}\mathsf{M}_j$ for which the set $\{j \in J \mid f(j) \neq 0_{\mathsf{M}_j}\}$ is finite. Each of the modules M_j, $j \in J$, is a ***summand*** in the direct sum. •

As with vector spaces, the direct product and direct sum agree for finite index sets, and we shall often write

$$\prod_{j=1}^{k} \mathsf{M}_j = \bigoplus_{j=1}^{k} \mathsf{M}_j = \mathsf{M}_1 + \cdots + \mathsf{M}_k$$

in this case. We also have a connection, as with vector spaces, with internal direct sums and direct sums as follows.

4.8.35 Proposition (Internal direct sum and direct sum of modules) *Let* R *be a ring, let* M *be a (left or right)* R*-module, and let* $(\mathsf{N}_\mathsf{j})_{\mathsf{j}\in\mathsf{J}}$ *be a family of submodules of* M *such that* M *is the internal direct sum of these submodules. Let* $\mathrm{i}_{\mathsf{N}_\mathsf{j}} : \mathsf{N}_\mathsf{j} \to \mathsf{M}$ *be the inclusion. Then the map from the direct sum* $\bigoplus_{\mathsf{j}\in\mathsf{J}} \mathsf{N}_\mathsf{j}$ *to* M *defined by*

$$\mathrm{f} \mapsto \sum_{\mathrm{j}\in\mathrm{J}} \mathrm{i}_{\mathsf{N}_\mathrm{j}} \mathrm{f}(\mathrm{j})$$

(noting that the sum is finite) is an isomorphism.

As with vector spaces, the direct sum as we have defined it is often called the external direct sum. Since the preceding result indicates that the two notions of direct sum are essentially the same, we shall, again as with the vector space case, often omit explicit reference to whether we are using the external or internal direct sum; the precise situation will be clear from the context.

An important example of a direct sum module is the following which mirrors Example 4.5.43 for vector spaces.

4.8.36 Example (The direct sum of copies of R) The construction of Example 4.5.43 also holds for modules. Thus let J be an arbitrary index set and let $\bigoplus_{j\in J} \mathsf{R}$ be the direct sum of "J copies" of the ring R. In the case when $J = \{1, \dots, n\}$ we have $\bigoplus_{j\in J} \mathsf{R} = \mathsf{R}^n$ and in the case when $J = \mathbb{Z}_{>0}$ we use the notation $\bigoplus_{j\in J} \mathsf{R} = \mathsf{R}_0^\infty$. If R is a unit ring, for $j \in J$ define $\boldsymbol{e}_j : J \to \mathsf{R}$ by

$$\boldsymbol{e}_j(j') = \begin{cases} 1_\mathsf{R}, & j' = j, \\ 0_\mathsf{R}, & j' \neq j. \end{cases}$$

One can show that $\{\boldsymbol{e}_j\}_{j\in J}$ is a basis for $\bigoplus_{j\in J} \mathsf{R}$, just as in the case for the direct sum of J copies of a field. We call $\{\boldsymbol{e}_j\}_{j\in J}$ the ***standard basis*** for $\bigoplus_{j\in J} \mathsf{R}$. •

4.8.37 Notation (Alternative notation for direct sums and direct products of copies of R) Following Notation 4.5.44, we shall find it sometimes convenient to denote

$$\prod_{j\in J} \mathsf{R} = \mathsf{R}^J, \quad \bigoplus_{j\in J} \mathsf{R} = \mathsf{R}_0^J$$

for the direct product and direct sum, respectively, of J copies of R. •

Up to this point in this section, everything we have said has been essentially transcripted from the corresponding Section 4.5.5 for vector spaces. However, some of the material from Section 4.5.5 has no analogue for modules. Specifically, for general rings R, it does not hold that every R-module is isomorphic to a direct sum of copies of R. Correspondingly, all of the nice characterisations one has for vector spaces being characterised essentially by their dimension have no analogue for rings.

4.8.5 Complements and quotients

An examination of our discussion in Section 4.5.6 of complements and quotients for vector spaces reveals that bases played a significant rôle. Therefore, we expect the discussion to be more complicated for modules, and indeed this is the case.

However, things start out benignly enough.

4.8.38 Definition (Complement of a submodule) If R is a ring, if M is a (left or right) R-module, and if N is a submodule of M, a ***complement*** of N in M is a submodule P of M such that $M = N \oplus P$. •

But right away we run into difficulties since submodules do not necessarily possess complements.

4.8.39 Example (A submodule without a complement) We take the ring $\mathbb{Z}$ thought of as a $\mathbb{Z}$-module, so that the submodules of $\mathbb{Z}$ are the ideals. We claim that the only submodules of $\mathbb{Z}$ that have complements are $\{0\}$ and $\mathbb{Z}$. Indeed, suppose that $I_1, I_2 \subseteq \mathbb{Z}$ are submodules so that, for example, $I_1 \cap I_2 = \{0\}$. Since $\mathbb{Z}$ is a principal ideal domain, we have $I_1 = (k_1)$ and $I_2 = (k_2)$ for $k_1, k_2 \in \mathbb{Z}$. If neither of k_1 and k_2 are nonzero, then, if k is the greatest common denominator for k_1 and k_2, we have $k \in (k_1) \cap (k_2)$. Thus we must have either $k_1 = 0$ or $k_2 = 0$. It follows that if $I \subseteq \mathbb{Z}$ is a submodule with a complement, then either $I = \{0\}$ or $I = \mathbb{Z}$. •

Despite the fact that submodules do not always have complements, it *is* the case that the quotient module can be constructed, just as can be constructed the quotient space by a subspace.

4.8.40 Definition (Quotient by a submodule) Let R be a ring, let M be a (left or right) R-module, and let N be a submodule of M. The ***quotient*** of M by N is the set of equivalence classes in M under the equivalence relation

$$x_1 \sim x_2 \quad \Longleftrightarrow \quad x_1 - x_2 \in N.$$

We denote by M/N the quotient of M by N, and we denote by $\pi_{M/N} \colon M \to M/N$ the map assigning to $x \in M$ its equivalence class. •

As with vector spaces, we denote by $x + N$ the equivalence class of $x \in M$ under the equivalence relation defined by the submodule N. The set M/N is naturally an Abelian group using addition in M, just as the quotient by a subspace is an Abelian group. However, M/N also has the structure of an R-module, as the following result indicates.

4.8.41 Proposition (The quotient by a submodule is a module) *Let* R *be a ring, let* M *be a left (resp. right)* R*-module, and let* N *be a submodule of* M*. The operations of addition and multiplication in* M/N *defined by*

$$(x_1 + \mathsf{N}) + (x_2 + \mathsf{N}) = (x_1 + x_2) + \mathsf{N},$$
$$r(x + \mathsf{N}) = (rx) + \mathsf{N} \ (\textit{resp.}\ (x + \mathsf{N})r = (xr) + \mathsf{N}), \qquad x, x_1, x_2 \in \mathsf{M},\ r \in \mathsf{R},$$

respectively, satisfy the axioms for a left (resp. right) R*-module. Moreover, if* R *is a unit ring and if* M *is unitary, then* M/N *is also unitary.*

Proof The proof that M/N is a left (or right) R-module follows, *mutatis mutandis*, along the lines of the proof of Proposition 4.5.54. The last statement is straightforward:

$$(1_{\mathsf{R}}(x + \mathsf{N})) = (1_{\mathsf{R}}x) + \mathsf{N} = x + \mathsf{N}$$

for all $x \in \mathsf{M}$. ■

Let us illustrate the character of the quotient module when a complement does not exist.

4.8.42 Example (Quotient module (Example 4.8.39 cont'd)) We take $\mathsf{R} = \mathbb{Z}$, $\mathsf{M} = \mathbb{Z}^2$, and N to be the submodule generated by $\{(1,0),(0,2)\}$. Let us understand the structure of the quotient module M/N by establishing an isomorphism from it to something somewhat familiar.

To do this we first note that $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$ has the structure of a $\mathbb{Z}$-module if we define addition and multiplication by

$$(k_1 + 2\mathbb{Z}) + (k_2 + 2\mathbb{Z}) = (k_1 + k_2) + 2\mathbb{Z}, \quad j(k + 2\mathbb{Z}) = jk + 2\mathbb{Z}$$

for $k, k_1, k_2 \in \mathbb{Z}_2$ and for $j \in \mathbb{Z}$ (this is a special case of Example 4.8.2–2). We now claim that the $\mathbb{Z}$-modules M/N and $\mathbb{Z}_2$ are isomorphic. Indeed, the map

$$\mathsf{M}/\mathsf{N} \ni (k_1, k_2) + \mathsf{N} \mapsto k_2 + 2\mathbb{Z} \in \mathbb{Z}_2 \tag{4.27}$$

is an isomorphism. That the map is a homomorphism of $\mathbb{Z}$-modules is easily shown; we leave the details to the reader. To see that the map is injective, suppose that $k_2 + 2\mathbb{Z} = 0 + 2\mathbb{Z}$. This implies that k_2 is even and so $(k_1, k_2) \in \mathsf{N}$ for any $k_1 \in \mathbb{Z}$. Thus the preimage of the zero element under the map (4.27) consists only of the zero element, showing that the map is injective by Exercise 4.8.3. That the map (4.27) is surjective is clear, and so establishes the fact that the map is a $\mathbb{Z}$-module isomorphism. •

While it is not true that complements exist in general, it is still true that when they *do* exist, they are isomorphic to the quotient.

4.8.43 Theorem (Relationship between complements and quotients) *Let* R *be a ring, let* M *be a (left or right)* R*-module, and let* N *be a submodule of* M *with a complement* P*. Then the map* $\iota_{\mathsf{N},\mathsf{P}} \colon \mathsf{P} \to \mathsf{M}/\mathsf{N}$ *defined by*

$$\iota_{\mathsf{N},\mathsf{P}}(z) = z + \mathsf{N}$$

is an isomorphism of R*-modules.*

Proof This follows along the lines of the proof of Theorem 4.5.56, using Proposition 4.8.31. ■

Let us now address the important problem of understanding when a submodule possesses a complement. Here it turns out that one's intuition can be enhanced by the use of commutative diagrams, the precise background for which we refer to Section 1.3.3.

4.8.44 Definition (Exact sequence) Let R be a ring. A ***sequence*** of R-modules is a commutative diagram on a directed graph (V, E) with the following properties:

(i) V is a subset of $\mathbb{Z}$;

(ii) $E = \begin{cases} \{(j, j+1) \mid j \in V \setminus \sup(V)\}, & V \text{ has an upper bound}, \\ \{(j, j+1) \mid j \in V\}, & V \text{ has no upper bound}; \end{cases}$

(iii) for each $j \in V$, the set assigned to the vertex j is an R-module M_j;

(iv) for each $(j, j+1) \in E$, $f_{(j,j+1)} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_j; \mathsf{M}_{j+1})$.

If additionally it holds that

(v) $\mathrm{image}(f_{(j-1,j)}) = \ker(f_{(j,j+1)})$ for each $j \in V \setminus \{\inf(V), \sup(V)\}$,

then the sequence is ***exact***. •

The definition is complicated by rigour. Stripping away the rigour one is left with a simple idea. Namely, a sequence of R-modules is a commutative diagram that can be represented as

$$\begin{tikzcd} \cdots \arrow[r, "f_{(j-2,j-1)}"] & \mathsf{M}_{j-1} \arrow[r, "f_{(j-1,j)}"] & \mathsf{M}_j \arrow[r, "f_{(j,j+1)}"] & \mathsf{M}_{j+1} \arrow[r, "f_{(j+1,j+2)}"] & \cdots \end{tikzcd}$$

where all maps are R-module homomorphisms. The diagram may be finite in length, unbounded on the left, unbounded on the right, or unbounded on both the left and the right. If the homomorphism with codomain M_j has image equal to the kernel of the homomorphism with domain M_j, then the sequence is exact.

The following properties of exact sequences of modules are useful.

4.8.45 Proposition (Properties of exact sequences) *Let* algR *be a ring and consider an exact sequence of* R*-modules on a directed graph* (V, E)*. If for some* $\mathrm{j} \in \mathrm{V}$ *it holds that* $\mathsf{M}_{\mathrm{j}} = \{0\}$*, then the following statements hold whenever the maps appearing are defined:*

(i) $\mathrm{f}_{(\mathrm{j}-2,\mathrm{j}-1)} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_{\mathrm{j}-1}; \mathsf{M}_{\mathrm{j}})$ *is surjective;*

(ii) $\mathrm{f}_{(\mathrm{j}+1,\mathrm{j}+2)} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_{\mathrm{j}}; \mathsf{M}_{\mathrm{j}+1})$ *is injective.*

Proof (i) The part of the diagram of concern looks like

$$\begin{tikzcd}\cdots \arrow[r, "f_{(j-3,j-2)}"] & \mathsf{M}_{j-2} \arrow[r, "f_{(j-2,j-1)}"] & \mathsf{M}_{j-1} \arrow[r, "f_{(j-1,j)}"] & \{0\} \arrow[r, "f_{(j,j+1)}"] & \cdots\end{tikzcd}$$

and we know that $\text{image}(f_{(j-2,j-1)}) = \ker(f_{(j-1,j)}) = \mathsf{M}_{j-1}$ since $f_{(j-1,j)}(x) = 0$ for every $x \in \mathsf{M}_{j-1}$.

(ii) Here the part of the diagram of interest looks like

$$\begin{tikzcd}\cdots \arrow[r, "f_{(j-1,j)}"] & \{0\} \arrow[r, "f_{(j,j+1)}"] & \mathsf{M}_{j+1} \arrow[r, "f_{(j+1,j+2)}"] & \mathsf{M}_{j+2} \arrow[r, "f_{(j+2,j+3)}"] & \cdots\end{tikzcd}$$

and we know that $\ker(f_{(j+1,j+2)}) = \text{image}(f_{j,j+1}) = \{0\}$ since $f_{(j,j+1)}$ maps the only element of $\{0\}$ to the zero element of M_{j+1}. ■

We shall be primarily interested in a particular sort of exact sequence.

4.8.46 Definition (Short exact sequence) Let R be a ring. A ***short exact sequence*** of R-modules is an exact sequence of R-modules on a directed graph (V, E) with the following properties:

(i) $V = \{1, 2, 3, 4, 5\}$;

(ii) $\mathsf{M}_1 = \mathsf{M}_5 = \{0\}$. •

In terms of our diagrammatic representation of an exact sequence, a short exact sequence is an exact sequence of the form

$$\begin{tikzcd}\{0\} \arrow[r, "f_1"] & \mathsf{N} \arrow[r, "f_2"] & \mathsf{M} \arrow[r, "f_3"] & \mathsf{P} \arrow[r, "f_4"] & \{0\}\end{tikzcd}$$

Note that the only possibility for the map f_1 is that it map the single element in the trivial module $\{0\}$ to the zero element of N, and that the only possibility for the map f_5 is that it map all elements of P to the zero element of the trivial module $\{0\}$. For this reason, these maps are sometimes omitted from the diagram. From Proposition 4.8.45 note that f_2 is injective and f_3 is surjective.

There are two standard examples of short exact sequences which we give in the following result.

4.8.47 Proposition (Some short exact sequences) *Let* R *be a ring, let* M *be a (left or right)* R*-module, and let* N *be a submodule. Then the statements hold:*

(i) the sequence

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "\mathrm{i}_{\mathsf{N}}"] & \mathsf{M} \arrow[r, "\pi_{\mathsf{M}/\mathsf{N}}"] & \mathsf{M}/\mathsf{N} \arrow[r] & \{0\}\end{tikzcd}$$

is exact, where i_{N} *is the inclusion map;*

(ii) the sequence

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "\mathrm{i}_{\mathsf{N}}"] & \mathsf{M} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

is exact, where P *is a complement to* N*, where* i_{N} *is the inclusion map, and where* pr_{P} *is the projection from* $\mathsf{M} = \mathsf{N} \oplus \mathsf{P}$ *onto* P*.*

Proof (i) Since the map i_{N} is injective and since the map $\pi_{\mathsf{M}/\mathsf{N}}$ is surjective, the only thing to check is whether $\operatorname{image}(i_{\mathsf{N}}) = \ker(\pi_{\mathsf{M}/\mathsf{N}})$. However, this is clear since $x + \mathsf{N} = 0_{\mathsf{M}} + \mathsf{N}$ if and only if $x \in \mathsf{N}$.

(ii) Again, since i_{N} is injective and pr_{P} is surjective, we only have to verify that $\operatorname{image}(i_{\mathsf{N}}) = \ker(\mathrm{pr}_{\mathsf{N}})$. This, however, follows from the definition of direct sum. Indeed, if $x \in \mathsf{N}$ then $x = x + 0_{\mathsf{M}}$ is the unique decomposition of x as a sum of an element of N and an element of P. Thus $\mathrm{pr}_2(x) = 0_{\mathsf{M}}$. Conversely, if $\mathrm{pr}_2(x) = 0_{\mathsf{M}}$, then it follows that $x = x + 0_{\mathsf{M}}$ must be the unique decomposition of x as a sum of an element of N and an element of P. In particular, $x \in \mathsf{N}$. ■

Since we know that quotient modules always exist, but complements to submodules do not, we know that the two parts of the preceding proposition are not in exact correspondence. We wish to establish conditions under which the diagrams *are* in correspondence. In order to do this, we first need to be clear about what "correspondence" means.

4.8.48 Definition (Isomorphic short exact sequences) Let R be a ring. Two exact sequences of R-modules represented by the diagrams

$$\begin{tikzcd} \{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] & \mathsf{M} \arrow[r, "g"] & \mathsf{P} \arrow[r] & \{0\} \end{tikzcd}$$

and

$$\begin{tikzcd} \{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] & \mathsf{M}' \arrow[r, "g'"] & \mathsf{P}' \arrow[r] & \{0\} \end{tikzcd}$$

are isomorphic if there exist isomorphisms $\phi\colon \mathsf{N} \to \mathsf{N}'$, $\psi\colon \mathsf{M} \to \mathsf{M}'$, and $\chi\colon \mathsf{P} \to \mathsf{P}'$ such that the diagram

$$\begin{tikzcd} \{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] \arrow[d, "\phi"'] & \mathsf{M} \arrow[r, "g"] \arrow[d, "\psi"'] & \mathsf{P} \arrow[r] \arrow[d, "\chi"'] & \{0\} \\ \{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"'] & \mathsf{M}' \arrow[r, "g'"'] & \mathsf{P}' \arrow[r] & \{0\} \end{tikzcd}$$

commutes. •

The idea behind isomorphic short exact sequences, like that of all isomorphic things, is that two isomorphic short exact sequences are essentially the same. It is this notion that we shall use to connect the two sequences of Proposition 4.8.47 in cases when this is possible. Indeed, we have the following result.

4.8.49 Theorem (Complements and short exact sequences) *Let R be a ring, let M be a (left or right) R-module, and let N be a submodule of M. Then the following statements are equivalent:*

(i) there exists a complement P to N in M;

(ii) there exists a submodule P of M and $f \in \operatorname{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{P})$ such that

(a) $f(z) = z$ for all $z \in \mathsf{P}$ and such that

(b) *the short exact sequences*

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "\mathrm{i}_{\mathsf{N}}"] & \mathsf{M} \arrow[r, "\pi_{\mathsf{M}/\mathsf{N}}"] & \mathsf{M}/\mathsf{N} \arrow[r] & \{0\}
\end{tikzcd}$$

and

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "\mathrm{i}_{\mathsf{N}}"] & \mathsf{M} \arrow[r, "\mathrm{f}"] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

are isomorphic with the corresponding isomorphisms of N *and* M *being the identity maps.*

Moreover, any complement P *to* N *has the property of* P *from part (ii), and if a submodule* P *and a homomorphism* $\mathrm{f} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{P})$ *have the properties from part (ii), then* P *is a complement to* N.

Proof Suppose that there exists a complement P to N. We claim that the diagrams

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"] & \mathsf{M} \arrow[r, "\pi_{\mathsf{M}/\mathsf{N}}"] & \mathsf{M}/\mathsf{N} \arrow[r] & \{0\}
\end{tikzcd}$$

and

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"] & \mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

are isomorphic, where we use the notation from the statement of Proposition 4.8.47. To see this, we consider the isomorphisms $\phi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N};\mathsf{N})$, $\psi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N}\oplus\mathsf{P})$, and $\chi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}/\mathsf{N};\mathsf{P})$ defined by

$$\phi(y) = y, \quad \psi(x) = \mathrm{pr}_{\mathsf{N}}(x) + \mathrm{pr}_{\mathsf{P}}(x), \quad \chi(x+\mathsf{N}) = \mathrm{pr}_{\mathsf{P}}(x),$$

respectively, where $\mathrm{pr}_{\mathsf{N}} \colon \mathsf{M} \to \mathsf{N}$ is the projection onto the first component in the direct sum $\mathsf{M} = \mathsf{N} \oplus \mathsf{P}$. The map ψ is an isomorphism by the definition of the internal direct sum, and the map χ is an isomorphism by Theorem 4.8.43. We must verify that the diagram

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"] \arrow[d, "\phi"'] & \mathsf{M} \arrow[r, "\pi_{\mathsf{M}/\mathsf{N}}"] \arrow[d, "\psi"'] & \mathsf{M}/\mathsf{N} \arrow[r] \arrow[d, "\chi"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"'] & \mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"'] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

commutes. One can easily see (verify this!) that it is sufficient to show that the two diagrams

$$\begin{tikzcd}
\mathsf{N} \arrow[r, "i_{\mathsf{N}}"] \arrow[d, "\phi"'] & \mathsf{M} \arrow[d, "\psi"] \\
\mathsf{N} \arrow[r, "i_{\mathsf{N}}"'] & \mathsf{N} \oplus \mathsf{P}
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathsf{M} \arrow[r, "\pi_{\mathsf{M}/\mathsf{N}}"] \arrow[d, "\psi"'] & \mathsf{M}/\mathsf{N} \arrow[d, "\chi"] \\
\mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"'] & \mathsf{P}
\end{tikzcd}$$

commute. For the left diagram, let $y \in \mathsf{N}$ and compute

$$\psi \circ i_{\mathsf{N}}(y) = i_{\mathsf{N}}(y)$$

since $\mathrm{pr}_{\mathsf{N}}(i_{\mathsf{N}}(y)) = i_{\mathsf{N}}(y)$ and since $\mathrm{pr}_{\mathsf{P}}(x) = 0_{\mathsf{M}}$, and also compute

$$i_{\mathsf{N}} \circ \phi(y) = i_{\mathsf{N}}(y),$$

which gives the commutativity of the left of the preceding two diagrams. For the rightmost diagram, let $x \in \mathsf{M}$ and write $x = y + z$ for $y \in \mathsf{N}$ and $z \in \mathsf{P}$. Then

$$\chi \circ \pi_{\mathsf{M}/\mathsf{N}}(x) = \chi(x + \mathsf{N}) = \chi(z + \mathsf{N}) = z$$

and

$$\mathrm{pr}_{\mathsf{P}} \circ \psi(x) = z,$$

giving the commutativity of the right diagram as well.

Now suppose that P and f satisfy the conditions of part (ii). Define $g_{\mathsf{P}} = i_{\mathsf{P}} \circ f \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{M})$, and note that for $x \in \mathsf{M}$

$$g_{\mathsf{P}} \circ g_{\mathsf{P}}(x) = (i_{\mathsf{P}} \circ f) \circ \underbrace{(i_{\mathsf{P}} \circ f)(x)}_{\in \mathsf{P}} = i_{\mathsf{P}} \circ f(x) = g_{\mathsf{P}}(x),$$

using the fact that $g_{\mathsf{P}}(z) = z$ for $z \in \mathsf{P}$. We now use a lemma.

1 **Lemma** *If* R *is a ring, if* M *is a (left or right)* R*-module, and if* $\mathrm{g} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{M})$ *satisfies* $\mathrm{g} \circ \mathrm{g} = \mathrm{g}$*, then* $\mathsf{M} = \mathrm{image}(\mathrm{g}) \oplus \ker(\mathrm{g})$.

Proof Suppose that $x \in \ker(f) \cap \mathrm{image}(f)$. Then there exists $y \in \mathsf{M}$ such that $x = f(y)$. Also

$$f(x) = f(f(y)) = f(y) = x = 0_{\mathsf{M}}$$

which implies that $\ker(f) \cap \mathrm{image}(f) = \{0_{\mathsf{M}}\}$. If $x \in \mathsf{M}$ we can write $x = x - f(x) + f(x)$ with, clearly, $f(x) \in \mathrm{image}(f)$. Since

$$f(x - f(x)) = f(x) - f(f(x)) = f(x) - f(x) = 0_{\mathsf{M}},$$

it also holds that $\mathsf{M} = \ker(f) + \mathrm{image}(f)$, and so the result follows from Proposition 4.8.31. ▼

By the lemma we have $\mathsf{M} = \ker(g_{\mathsf{P}}) \oplus \mathrm{image}(g_{\mathsf{P}})$. We claim that $\ker(g_{\mathsf{P}}) = \mathsf{N}$ and that $\mathrm{image}(g_{\mathsf{P}}) = \mathsf{P}$. By exactness of the sequence

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"] & \mathsf{M} \arrow[r, "f"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

if $y \in \mathsf{N}$ then $y \in \ker(f)$ and so we immediately have $g_{\mathsf{P}}(y) = 0_{\mathsf{M}}$; thus $\mathsf{N} \subseteq \ker(g_{\mathsf{P}})$. If $g_{\mathsf{P}}(x) = 0_{\mathsf{M}}$ then, since i_{P} is injective, we have $f(x) = 0_{\mathsf{M}}$ by Exercise 4.8.3. Thus $x \in \ker(f) = \mathsf{N}$, so giving $\mathsf{N} = \ker(g_{\mathsf{P}})$. If $z \in \mathsf{P}$ then $f(z) = z$ which gives $g_{\mathsf{P}}(z) = z$, and so $z \in \mathrm{image}(g_{\mathsf{P}})$. If $z \in \mathrm{image}(g_{\mathsf{P}})$ then $z \in \mathrm{image}(i_{\mathsf{P}})$, immediately giving $\mathsf{P} = \mathrm{image}(g_{\mathsf{P}})$. Thus P is a complement to N as desired.

The last statement of the theorem is a direct consequence of the proof above. ■

We next give some conditions one can use to test when a given submodule has a complement. This requires some additional language.

4.8.50 Definition (Splitting of short exact sequences) Let R be a ring and consider a sort exact sequence represented by the diagram

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] & \mathsf{M} \arrow[r, "g"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

The short exact sequence

(i) ***splits on the left*** if there exists $f' \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ such that $f' \circ f = \mathrm{id}_{\mathsf{N}}$ and

(ii) ***splits on the right*** if there exists $g' \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{P};\mathsf{M})$ such that $g \circ g' = \mathrm{id}_{\mathsf{P}}$. •

Note that the map f' in the definition is a left-inverse of f and the map g' is a right-inverse of g. By Proposition 1.3.9 we know that f possesses a left-inverse by virtue of being injective, and that g possesses a right-inverse by virtue of being surjective. Thus the central idea of a short exact sequence that splits on the left (resp. right) is that the left-inverse (resp. right-inverse) be an R-module homomorphism.

The next result we give relates split exact sequences to direct sums.

4.8.51 Theorem (Direct sums and split exact sequences) *Let R be a ring, let M, N, and P be (left or right) R-modules, and consider a short exact sequence*

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "\mathsf{f}"] & \mathsf{M} \arrow[r, "g"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

of R-modules. Then the following statements are equivalent:

(i) the short exact sequence splits on the left;

(ii) the short exact sequence splits on the right;

(iii) the short exact sequence is isomorphic to the short exact sequence

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "\mathrm{i}_{\mathsf{N}}"] & \mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

where the corresponding isomorphisms of N and P are the identity maps.

Proof (i) $\Longrightarrow$ (iii) Let $f' \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ be a left-inverse of f and define $\psi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N}\oplus\mathsf{P})$ by $\psi(x) = (f'(x), g(x))$. It is then a direct computation to check that the diagram

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] \arrow[d, "\mathrm{id}_{\mathsf{N}}"'] & \mathsf{M} \arrow[r, "g"] \arrow[d, "\psi"'] & \mathsf{P} \arrow[r] \arrow[d, "\mathrm{id}_{\mathsf{P}}"'] & \{0\} \\ \{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"'] & \mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"'] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

commutes. From this we see that Exercise 4.8.7 gives ψ an isomorphism, and so gives this part of the proof.

(ii) $\implies$ (iii) Let $g' \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{P};\mathsf{M})$ be a right-inverse of g and define $\phi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N} \oplus \mathsf{P};\mathsf{M})$ by $\phi(y,z) = (f(y), g'(z))$. One can then verify directly that the diagram

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"] \arrow[d, "\mathrm{id}_{\mathsf{N}}"'] & \mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"] \arrow[d, "\phi"'] & \mathsf{P} \arrow[r] \arrow[d, "f"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"'] & \mathsf{M} \arrow[r, "g"'] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

commutes, and then, by Exercise 4.8.7, ϕ is an isomorphism, which gives the result.

(iii) $\implies$ (i) and (ii) We are given the commutative diagram

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "i_{\mathsf{N}}"] \arrow[d, "\mathrm{id}_{\mathsf{N}}"'] & \mathsf{N} \oplus \mathsf{P} \arrow[r, "\mathrm{pr}_{\mathsf{P}}"] \arrow[d, "\phi"'] & \mathsf{P} \arrow[r] \arrow[d, "\mathrm{id}_{\mathsf{P}}"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"'] & \mathsf{M} \arrow[r, "g"'] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

with ϕ an isomorphism. If we define $f' \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ and $g' \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{P};\mathsf{M})$ by $f' = \mathrm{pr}_{\mathsf{N}} \circ \phi^{-1}$ and $g' = \phi \circ i_{\mathsf{P}}$, respectively, then we leave as an exercise to reader the verification that f' is a left-inverse of f and that g' is a right inverse of g. ■

4.8.6 Torsion

One of the features that a module may possess that a vector space does not possess is "torsion." For non-mathematicians, torsion refers to a twisting motion. And in the mathematical context of this section, the reader might try to see how this non-mathematical notion carries over to the mathematical setting.

First the definition.

4.8.52 Definition (Order ideal, torsion element, torsion submodule, torsion module) Let R be a commutative ring and let M be an R-module.

(i) For $x \in \mathsf{M}$ the set
$$\mathrm{ann}(x) = \{r \in \mathsf{R} \mid rx = 0_{\mathsf{M}}\}$$
is the ***annihilator*** of x.

(ii) If N is a submodule of M, the set
$$\mathrm{ann}(\mathsf{N}) = \{r \in \mathsf{R} \mid ry = 0_{\mathsf{M}} \text{ for all } y \in \mathsf{N}\}$$
is the ***annihilator*** of N.

(iii) An element $x \in \mathsf{M}$ is a ***torsion element*** if $\mathrm{ann}(x)$ contains a nonzerodivisor.

(iv) The set of torsion elements of M is denoted by $\mathrm{Tor}(\mathsf{M})$ and is called the ***torsion submodule***.

(v) If $\mathrm{Tor}(\mathsf{M}) = \mathsf{M}$ then M is a ***torsion module***.

(vi) If $\mathrm{Tor}(\mathsf{M}) = \{0_{\mathsf{M}}\}$ then M is ***torsion-free***. •

First we should be sure that the name torsion submodule is deserved, and that some other natural properties hold.

4.8.53 Proposition (The set of torsion elements is a submodule, etc.) *If* R *is a commutative ring and if* M *is an* R*-module, then the following statements hold:*

(i) $\operatorname{ann}(\mathsf{x})$ *is an ideal for each* $\mathsf{x} \in \mathsf{M}$*;*

(ii) $\operatorname{Tor}(\mathsf{M})$ *is a submodule;*

(iii) $\mathsf{M}/\operatorname{Tor}(\mathsf{M})$ *is a torsion-free module.*

Proof (i) That $\operatorname{ann}(x)$ is an ideal is directly checked.

(ii) Let $x, y \in \operatorname{Tor}(\mathsf{M})$ and let $r, s \in \mathsf{R}$ be nonzerodivisors such that $rx = sy = 0_{\mathsf{M}}$. Then $-sy = 0_{\mathsf{M}}$ and so

$$rx = -sy = 0_{\mathsf{M}} \quad\Longrightarrow\quad rsx = -rsy = 0_{\mathsf{M}} \quad\Longrightarrow\quad rs(x+y) = 0_{\mathsf{M}}.$$

Since rs is a nonzerodivisor by Exercise 4.2.12, it follows that $x + y \in \operatorname{Tor}(\mathsf{M})$. Also, if $x \in \operatorname{Tor}(\mathsf{M})$ and if $r \in \mathsf{R}$ is a nonzerodivisor such that $rx = 0_{\mathsf{M}}$, then, for any $r' \in \mathsf{R}$, we have $rr'x = r'rx = 0_{\mathsf{M}}$ so that $rx \in \operatorname{Tor}(\mathsf{M})$.

(iii) For $x \in \mathsf{M}$ we have

$$\operatorname{ann}(x + \operatorname{Tor}(\mathsf{M})) = \{r \in \mathsf{R} \mid rx \in \operatorname{Tor}(\mathsf{M})\}.$$

We must show that $\operatorname{ann}(x + \operatorname{Tor}(\mathsf{M}))$ consists of only zerodivisors for every $x \notin \operatorname{Tor}(\mathsf{M})$. If $r \in \operatorname{ann}(x + \operatorname{Tor}(\mathsf{M}))$ there exists a nonzerodivisor $r' \in \mathsf{R}$ such that $r'rx = 0_{\mathsf{M}}$. The following cases may occur.

1. *rr' is a nonzerodivisor:* In this case we have $x \in \operatorname{Tor}(\mathsf{M})$.
2. *rr' is a zerodivisor:* This implies that r is a zerodivisor since the product of two nonzerodivisors is a nonzerodivisor by Exercise 4.2.12.

Therefore, if $x \notin \operatorname{Tor}(\mathsf{M})$ and if $r \in \operatorname{ann}(x + \operatorname{Tor}(\mathsf{M}))$, it follows that r is a zero divisor, as desired. ■

It is also true that $\operatorname{ann}(\mathsf{N})$ is an ideal of R for every submodule N of M. This is easily checked.

Let us give some examples of torsion elements.

4.8.54 Examples (Torsion elements)

1. Note that $0_{\mathsf{M}} \in \operatorname{Tor}(\mathsf{M})$ for any module M, so $\operatorname{Tor}(\mathsf{M}) \neq \emptyset$.
2. By Proposition 4.5.3(vi), for a vector space V over a field F, $\operatorname{Tor}(\mathsf{V}) = \{0_{\mathsf{V}}\}$.
3. For an integral domain R and an R-module M, $x \in \operatorname{Tor}(\mathsf{M})$ if and only if $\operatorname{ann}(x) \neq \{0_{\mathsf{R}}\}$.
4. Consider the $\mathbb{Z}$-module $\mathbb{Z}_4 = \mathbb{Z}/4\mathbb{Z}$. We note that $\operatorname{ann}(2 + 4\mathbb{Z}) = \{2j \mid j \in \mathbb{Z}_{\geq 0}\}$, and so $2 + 4\mathbb{Z}$ is a torsion element. •

4.8.7 Algebras

It is sometimes of interest to consider additional structure on a vector space, namely a product between vectors. In this section we provide the definitions for such a structure. The main examples will only arise later in Sections 5.1 and 5.4.

4.8.55 Definition (Algebra) If R is a ring, a ***left*** **R-*algebra*** (resp. ***right*** **R-*algebra***) is a left (resp. right) R-module A equipped with a binary operation, denoted by $(x_1, x_2) \mapsto x_1 \cdot x_2$, satisfying the following conditions:

(i) $x_1 \cdot (x_2 + x_3) = (x_1 \cdot x_2) + (x_1 \cdot x_3)$, $x_1, x_2, x_3 \in \mathsf{A}$ (***left distributivity***);

(ii) $(x_1 + x_2) \cdot x_3 = (x_1 \cdot x_3) + (x_2 \cdot x_3)$, $x_1, x_2, x_3 \in \mathsf{R}$ (***right distributivity***);

(iii) $(x_1 \cdot x_2) \cdot x_3 = x_1 \cdot (x_2 \cdot x_3)$, $x_1, x_2, x_3 \in \mathsf{A}$ (***associativity*** of multiplication);

(iv) $r(x_1 \cdot x_2) = (rx_1) \cdot x_2 = x_1 \cdot (rx_2)$ (resp. $(x_1 \cdot x_2)r = (x_1 r) \cdot x_2 = x_1 \cdot (x_2 r)$) (***distributivity*** of product). A left (resp. right) R-algebra is a ***left unity*** **R-*algebra*** (resp. ***right unity*** **R-*algebra***) if it is a left (resp. right) unity R-module. •

Note that an algebra combines three operations: vector addition, scalar multiplication, and a product. Most authors assume that an algebra is associative. In this case one can think of an algebra in one of two ways: (1) it is a vector space with the addition of a product which gives a ring structure; (2) it is a ring with the addition of a scalar product which gives a vector space structure.

Let us give some simple examples of algebras.

4.8.56 Examples (Algebras)

1. If M is any R-module, and if we define a product on M by $x_1 \cdot x_2 = 0_{\mathsf{M}}$ for all $x_1, x_2 \in \mathsf{M}$, then this defines the structure of an algebra.
2. If K is a field extension of F, then we have already seen that K is an F-vector space. If we take the product between vectors in K to be the product coming from the ring structure, then one immediately verifies that this makes K an associative F-algebra.
3. Let us define a product on F^3 by

$$(u_1, u_2, u_3) \cdot (v_1, v_2, v_3) = (u_2 v_3 - u_3 v_2, u_3 v_1 - u_1 v_3, u_1 v_2 - u_2 v_1).$$

In F^3 readers will recognise this as the vector- or cross-product. One can verify that this product satisfies all properties for an F-algebra except for associativity of the product. Sometimes nonassociative algebras are of interest, and this is one such instance. •

As with all algebraic objects we have dealt with in this chapter, there is a notion of a map between algebras that preserves the algebra structure.

4.8.57 Definition (Algebra homomorphism and antihomomorphism) Let R be a ring and let A and B be left (resp. right) R-algebras. An **R-*homomorphism*** of A and B is a map $\phi\colon \mathsf{A} \to \mathsf{B}$ with the following properties:

(i) $\phi(x_1 + x_2) = \phi(x_1) + \phi(x_2)$ for $x_1, x_2 \in \mathsf{A}$;

(ii) $\phi(rx) = r\phi(x)$ (resp. $\phi(xr) = \phi(x)r$) for $r \in \mathsf{R}$ and $x \in \mathsf{A}$;

(iii) $\phi(x_1 \cdot x_2) = \phi(x_1) \cdot \phi(x_2)$ for $x_1, x_2 \in \mathsf{A}$.

If the last property is replaced with

(iv) $\phi(x_1 \cdot x_2) = \phi(x_2) \cdot \phi(x_1)$ for $x_1, x_2 \in \mathsf{A}$,

then ϕ is an **R-*antihomomorphism***. An R-homomorphism ϕ is an **R-*monomorphism*** (resp. **R-*epimorphism*, R-*isomorphism***) if ϕ is injective (resp. surjective, bijective). If there exists an isomorphism between left R-algebras A and B, then A and B are **R-*isomorphic***. •

4.8.8 Notes

Exercises

4.8.1 Give an example of a commutative unit ring R and a left unity R-module M for which there exists $r \in \mathsf{R} \setminus \{0_\mathsf{R}\}$ and $x \in \mathsf{M} \setminus \{0_\mathsf{M}\}$ satisfying $rx = 0_\mathsf{M}$.

4.8.2 Give an example of a commutative ring R, an R-module M, and submodules N_1 and N_2 of M such that (1) $\mathsf{N}_1 \neq \mathsf{N}_2$ and (2) $\mathsf{N}_1 \cap \mathsf{N}_2 \neq \{0_\mathsf{M}\}$. Is this phenomenon possible for vector spaces?

4.8.3 Let R be a ring, let M and N be (left or right) R-modules, and let $\mathsf{L} \in \mathrm{Hom}_\mathsf{R}(\mathsf{M};\mathsf{N})$. Show that L is injective if and only if $\ker(\mathsf{L}) = \{0_\mathsf{M}\}$.

4.8.4 Let R be a ring, let M and N be (left or right) R-modules, and let $\mathsf{A} \subseteq \mathsf{M}$ and $\mathsf{B} \subseteq \mathsf{N}$ be submodules. Show that the modules

$$(\mathsf{M} \oplus \mathsf{N})/(\mathsf{A} \oplus \mathsf{B}), \quad (\mathsf{M}/\mathsf{A}) \oplus (\mathsf{N}/\mathsf{B})$$

are isomorphic.

4.8.5 Let R be a ring, let M and N be R-modules, and let $\mathsf{L} \in \mathrm{Hom}_\mathsf{R}(\mathsf{M};\mathsf{N})$ be an epimorphism. Show that the map $\mathsf{L}_0 \in \mathrm{Hom}_\mathsf{R}(\mathsf{M}/\ker(\mathsf{L});\mathsf{N})$ defined by $\mathsf{L}_0(x + \ker(\mathsf{L})) = \mathsf{L}(x)$ is a well-defined isomorphism of R-modules.

4.8.6 Let R be a ring.

(a) Suppose that two short exact sequences of R-modules

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] & \mathsf{M} \arrow[r, "g"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

and

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] & \mathsf{M}' \arrow[r, "g'"] & \mathsf{P}' \arrow[r] & \{0\}\end{tikzcd}$$

are isomorphic so that there exist isomorphisms $\phi\colon \mathsf{N} \to \mathsf{N}'$, $\psi\colon \mathsf{M} \to \mathsf{M}'$, and $\chi\colon \mathsf{P} \to \mathsf{P}'$ such that the diagram

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] \arrow[d, "\phi"'] & \mathsf{M} \arrow[r, "g"] \arrow[d, "\psi"'] & \mathsf{P} \arrow[r] \arrow[d, "\chi"'] & \{0\} \\ \{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"'] & \mathsf{M}' \arrow[r, "g'"'] & \mathsf{P}' \arrow[r] & \{0\}\end{tikzcd}$$

commutes. Show that the diagram

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] \arrow[d, "\phi^{-1}"'] & \mathsf{M}' \arrow[r, "g'"] \arrow[d, "\psi^{-1}"'] & \mathsf{P}' \arrow[r] \arrow[d, "\chi^{-1}"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"'] & \mathsf{M} \arrow[r, "g"'] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

commutes.

(b) Consider three short exact sequences of R-modules

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] & \mathsf{M} \arrow[r, "g"] & \mathsf{P} \arrow[r] & \{0\}
\end{tikzcd}$$

and

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] & \mathsf{M}' \arrow[r, "g'"] & \mathsf{P}' \arrow[r] & \{0\}
\end{tikzcd}$$

and

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N}'' \arrow[r, "f''"] & \mathsf{M}'' \arrow[r, "g''"] & \mathsf{P}'' \arrow[r] & \{0\}
\end{tikzcd}$$

Suppose that the first two are isomorphic and that the second two are isomorphic so that there exist isomorphisms $\phi\colon \mathsf{N} \to \mathsf{N}'$, $\psi\colon \mathsf{M} \to \mathsf{M}'$, $\chi\colon \mathsf{P} \to \mathsf{P}'$, $\phi'\colon \mathsf{N}' \to \mathsf{N}''$, $\psi'\colon \mathsf{M}' \to \mathsf{M}''$, and $\chi'\colon \mathsf{P}' \to \mathsf{P}''$ such that the diagrams

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] \arrow[d, "\phi"'] & \mathsf{M} \arrow[r, "g"] \arrow[d, "\psi"'] & \mathsf{P} \arrow[r] \arrow[d, "\chi"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"'] & \mathsf{M}' \arrow[r, "g'"'] & \mathsf{P}' \arrow[r] & \{0\}
\end{tikzcd}$$

and

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] \arrow[d, "\phi'"'] & \mathsf{M}' \arrow[r, "g'"] \arrow[d, "\psi'"'] & \mathsf{P}' \arrow[r] \arrow[d, "\chi'"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N}'' \arrow[r, "f''"'] & \mathsf{M}'' \arrow[r, "g''"'] & \mathsf{P}'' \arrow[r] & \{0\}
\end{tikzcd}$$

commute. Show that the diagram

$$\begin{tikzcd}
\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] \arrow[d, "\phi'\circ\phi"'] & \mathsf{M} \arrow[r, "g"] \arrow[d, "\psi'\circ\psi"'] & \mathsf{P} \arrow[r] \arrow[d, "\chi'\circ\chi"'] & \{0\} \\
\{0\} \arrow[r] & \mathsf{N}'' \arrow[r, "f''"'] & \mathsf{M}'' \arrow[r, "g''"'] & \mathsf{P}'' \arrow[r] & \{0\}
\end{tikzcd}$$

commutes.

(c) Conclude that the relation in the set of exact sequences given by "two sequences are equivalent if they are isomorphic" is an equivalence relation.

4.8.7 Let R be a ring. For two exact sequences

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] & \mathsf{M} \arrow[r, "g"] & \mathsf{P} \arrow[r] & \{0\}\end{tikzcd}$$

and

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] & \mathsf{M}' \arrow[r, "g'"] & \mathsf{P}' \arrow[r] & \{0\}\end{tikzcd}$$

of R-modules, suppose that the diagram

$$\begin{tikzcd}\{0\} \arrow[r] & \mathsf{N} \arrow[r, "f"] \arrow[d, "\phi"'] & \mathsf{M} \arrow[r, "g"] \arrow[d, "\psi"'] & \mathsf{P} \arrow[r] \arrow[d, "\chi"'] & \{0\} \\ \{0\} \arrow[r] & \mathsf{N}' \arrow[r, "f'"] & \mathsf{M}' \arrow[r, "g'"] & \mathsf{P}' \arrow[r] & \{0\}\end{tikzcd}$$

is commutative.

(a) Show that if ϕ and χ are surjective, then ψ is surjective.

(b) Show that if ϕ and χ are injective, then ψ is injective.

(c) Show that if ϕ and χ are bijective, then ψ is bijective.

(This result is called the ***Short Five Lemma***.)

4.8.8 If R is an integral domain, show that a free R-module is torsion free.

Section 4.9

Modules over principal ideal domains

In this section we develop a rather special area of module theory: the structure of modules over a principal ideal domain. In this case, we shall see that certain of the bad features of general modules do not arise. Moreover, for finitely generated modules over principal ideal domains, one can be quite explicit about their structure.

Do I need to read this section? The material here is needed to form a complete understanding of (1) the equivalence of homomorphisms of modules over principal ideal domains in Section 5.5.6 and (2) the structure theory for endomorphisms of vector spaces in Section 5.8. Readers only needing the statements of these results can skip the details needed from this section. Certainly this section can be bypassed on a first reading.

It might be pointed out, however, that this section serves quite well the purpose of showing the ways in which module theory differs from vector space theory. So for readers feeling as if they might benefit from a better understanding of this, perhaps this section is a good one to read. •

4.9.1 Free modules over principal ideal domains

We begin our study of modules over principal ideal domains by studying free modules, i.e., those possessing bases. For a general free module, it may be the case that submodules are themselves not free (see Example 4.8.19). However, for principal ideal domains this cannot happen.

4.9.1 Theorem (Submodules of free modules over principal ideal domains) *Let* R *be a principal ideal domain and let* M *be a free* R*-module. If* N *is a submodule of* M *then* N *is free and* $\operatorname{rank}(\mathsf{N}) \leq \operatorname{rank}(\mathsf{M})$.

Proof Let $\{e_j\}_{j\in J}$ be a basis for M and, for $J' \subseteq J$, define

$$\mathsf{N}_{J'} = \mathsf{N} \cap \bigoplus_{j'\in J'} \operatorname{span}_{\mathsf{R}}(e_{j'}).$$

One may verify, exactly as Proposition 4.5.34 for vector spaces, that $\mathsf{N}_{J'}$ is a submodule for each $J' \subseteq J$. Now define

$$P_J = \{(J', \mathscr{B}_{J'}) \mid J' \subseteq J, \mathscr{B}_{J'} \text{ is a basis for } \mathsf{N}_{J'} \text{ with } \operatorname{card}(\mathscr{B}_{J'}) \leq \operatorname{card}(J')\}$$

First let us show that P_J is nonempty. Let $J' = \{j\} \subseteq J$. One can directly check that

$$\mathsf{I}_j = \{r \in \mathsf{R} \mid re_j \in \mathsf{N}\}$$

is an ideal of R. Therefore, since R is a principal ideal domain, this ideal is of the form (r_j) for some $r_j \in \mathsf{R}$. If $r_j = 0_{\mathsf{R}}$ then $N_{J'} = \{0_{\mathsf{M}}\}$ and so $\varnothing$ is a basis for $\mathsf{N}_{J'}$. If $r_j \neq 0_{\mathsf{R}}$ then we claim that $\{r_j e_j\}$ is a basis for $\mathsf{N}_{J'}$. Indeed, since $\mathsf{N}_{J'} = \operatorname{span}_{\mathsf{R}}(e_j)$ in this case, if $x \in \mathsf{N}_{J'}$ we must have $x = re_j$ for some $r \in \mathsf{R}$. Thus $r \in \mathsf{I}_j$ and so $r = sr_j$ for some $s \in \mathsf{R}$ since $\mathsf{I}_j = (r_j)$. Therefore, $x = s(r_j e_j)$ and so $x \in \operatorname{span}_{\mathsf{R}}(r_j e_j)$ and so $\mathsf{N}_{J'} \subseteq \operatorname{span}_{\mathsf{R}}(r_j e_j)$. If $x \in \operatorname{span}_{\mathsf{R}}(r_j e_j)$ then clearly $x \in \operatorname{span}_{\mathsf{R}}(e_j)$, and so we have $\mathsf{N}_{J'} = \operatorname{span}_{\mathsf{R}}(r_j e_j)$, as desired. Therefore, in either of the cases $r_j = 0_{\mathsf{R}}$ or $r_j \neq 0_{\mathsf{R}}$, we have $\operatorname{card}(\{\mathscr{B}_j\}) \leq \operatorname{card}(J')$ where $\mathscr{B}_j$ is any basis for $\mathsf{N}_{J'}$. Thus $(\{j\}, \mathsf{N}_{\{j'\}}) \in P_J$.

Now let us place a partial order $\preceq$ on P_J by

$$(J_1, \mathscr{B}_{J_1}) \preceq (J_2, \mathscr{B}_{J_2}) \quad \Longleftrightarrow \quad J_1 \subseteq J_2,\ \mathscr{B}_{J_1} \subseteq \mathscr{B}_{J_2}.$$

We claim that this partial order on P_J satisfies the hypotheses of Zorn's Lemma. Let $\{(J_a, \mathscr{B}_{J_a}) \mid a \in A\}$ be a totally ordered subset of P_J and define $\bar{J} = \cup_{a\in A} J_a$ and $\bar{\mathscr{B}} = \cup_{a\in A} \mathscr{B}_{J_a}$. We shall show that $(\bar{J}, \bar{\mathscr{B}})$ is an upper bound for $\{(J_a, \mathscr{B}_{J_a}) \mid a \in A\}$. The total order on $\{(J_a, \mathscr{B}_{J_a}) \mid a \in A\}$ induces a total order on the index set A in the obvious way:

$$a_1 \leq a_2, \quad \Longleftrightarrow \quad (J_{a_1}, \mathscr{B}_{J_{a_1}}) \preceq (J_{a_2}, \mathscr{B}_{J_{a_2}}).$$

We shall tacitly use this partial order. Note that the families

$$(J_a)_{a\in A}, \quad \left(\bigoplus_{j\in J_a} \operatorname{span}_{\mathsf{R}}(e_j)\right)$$

of subsets of J and M, respectively, are totally ordered by inclusion since $\{(J_a, \mathscr{B}_{J_a}) \mid a \in A\}$ is totally ordered. That is to say,

$$\begin{aligned} J_{a_1} \subseteq J_{a_2} \quad &\Longleftrightarrow \quad a_1 \leq a_2, \\ \bigoplus_{j\in J_{a_1}} \operatorname{span}_{\mathsf{R}}(e_j) \subseteq \bigoplus_{j\in J_{a_2}} \operatorname{span}_{\mathsf{R}}(e_j) \quad &\Longleftrightarrow \quad a_1 \leq a_2 \end{aligned}$$

(why?). It, therefore, follows directly that the family $(\mathsf{N}_{J_a})_{a\in A}$ of subsets of M is also totally ordered by inclusion, i.e.,

$$\mathsf{N}_{J_{a_1}} \subseteq \mathsf{N}_{J_{a_2}} \quad \Longleftrightarrow \quad a_1 \leq a_2$$

(again, why?). Therefore, it holds that

$$\bigcup_{a\in A} \mathsf{N}_{J_a} = \sum_{a\in A} \mathsf{N}_{J_a}$$

(cf. Exercise 4.5.17). Thus we compute

$$\begin{aligned}\sum_{a\in A}\mathsf{N}_{J_a}=\bigcup_{a\in A}\mathsf{N}_{J_a}=\bigcup_{a\in A}\left(\mathsf{N}\cap\bigoplus_{j\in J_a}\operatorname{span}_{\mathsf{R}}(e_j)\right)\\ =\mathsf{N}\cap\left(\bigcup_{a\in A}\bigoplus_{j\in J_a}\operatorname{span}_{\mathsf{R}}(e_j)\right)\\ =\mathsf{N}\cap\left(\sum_{a\in A}\bigoplus_{j\in J_a}\operatorname{span}_{\mathsf{R}}(e_j)\right)\\ =\mathsf{N}\cap(\bigoplus_{j\in J_a}\operatorname{span}_{\mathsf{R}}(e_j)=\mathsf{N}_{\bar{J}},\end{aligned}$$

using Proposition 1.1.7. We claim that $\bar{\mathscr{B}}$ is a basis for $\mathsf{N}_{\bar{J}}$. Let $\{e_{j_1},\dots,e_{j_k}\}\subseteq\bar{\mathscr{B}}$ and let $a_0\in A$ be sufficiently large (with respect to the total order on A) that $\{e_{j_1},\dots,e_{j_k}\}\subseteq\mathscr{B}_{J_{a_0}}$. It follows immediately that $\{e_{j_1},\dots,e_{j_k}\}$ is linearly independent, and so $\bar{\mathscr{B}}$ is linearly independent. If $x\in\mathsf{N}_{\bar{J}}$ then there exists $a_0\in A$ such that $x\in\mathsf{N}_{J_{a_0}}$. In particular, x is a finite linear combination of elements of $\mathscr{B}_{J_{a_0}}$ since $\mathscr{B}_{J_{a_0}}$ is a basis for $\mathsf{N}_{J_{a_0}}$. Thus x is a finite linear combination of elements of $\bar{\mathscr{B}}$, and so $\bar{\mathscr{B}}$ is indeed a basis for $\mathsf{N}_{\bar{J}}$. Since $\operatorname{card}(\mathscr{B}_{J_a})\le\operatorname{card}(J_a)$ it follows that $\operatorname{card}(\bar{\mathscr{B}})\le\operatorname{card}(\bar{J})$ (why?). Therefore, this shows that $(\bar{J},\bar{\mathscr{B}})\subseteq P_J$. It is clear that $(\bar{J},\bar{\mathscr{B}})$ is an upper bound for $\{(J_a,\mathscr{B}_{J_a})\mid a\in A\}$, and so the partially ordered set P_J does indeed satisfy the hypotheses of Zorn's Lemma.

Now let $(K,\mathscr{B}_K)\in P_J$ be a maximal element. We shall show that $K=J$. Suppose that $K\subset J$ and let $j\in J\setminus K$. Take $K'=K\cup\{j\}$ so that $\mathsf{N}_K\subseteq\mathsf{N}_{K'}$. If $\mathsf{N}_K=\mathsf{N}_{K'}$ then $(K',\mathscr{B}_K)\in P_J$ has the property that $(K,\mathscr{B}_K)\prec(K',\mathscr{B}_K)$, thus contradicting the maximality of $(K,\mathscr{B}_K)$. We may, therefore, suppose that $\mathsf{N}_K\subset\mathsf{N}_{K'}$. If $x\in\mathsf{N}_{K'}\setminus\mathsf{N}_K$ then one can write $x=x'+re_j$ for $x'\in\mathsf{N}_K$ and where $r\neq 0_{\mathsf{R}}$. One can then directly check that

$$\mathsf{I}_{K'}=\{r\in\mathsf{R}\mid\ x-re_j\in\mathsf{N}_K\text{ for some }x\in\mathsf{N}_{K'}\}$$

is an ideal of R and so we have $\mathsf{I}_{K'}=(r_0)$ for some $r_0\in\mathsf{R}\setminus\{0_{\mathsf{R}}\}$. Let $x_0\in\mathsf{N}_{K'}$ have the property that $x_0-r_0e_j\in\mathsf{N}_K$.

We claim that $\mathscr{B}_K\cup\{x_0\}$ is a basis for $\mathsf{N}_{K'}$. Indeed, if $x\in\mathsf{N}_{K'}$ then let $r\in\mathsf{R}$ have the property that $x-re_j\in\mathsf{N}_K$. We then have $r=sr_0$ for some $s\in\mathsf{R}$, and so $x-sx_0\in\mathsf{N}_K$. Thus $y\in\operatorname{span}_{\mathsf{R}}(\mathsf{N}_K\cup\{x_0\})$ or, equivalently, $y\in\operatorname{span}_{\mathsf{R}}(\mathscr{B}_K\cup\{x_0\})$. Therefore, $\mathscr{B}_K\cup\{x_0\}$ generates $\mathsf{N}_{K'}$. Moreover, since $x_0\notin\mathsf{N}_K$, it follows that $\mathscr{B}_K\cup\{x_0\}$ is linearly independent.

Now we have

$$\operatorname{card}(\mathscr{B}_K\cup\{x_0\})=\operatorname{card}(\mathscr{B}_K)+1\le\operatorname{card}(K)+1=\operatorname{card}(K').$$

Thus $(K',\mathscr{B}_K\cup\{x_0\})\in P_J$ and $(K,\mathscr{B}_K)\prec(K',\mathscr{B}_{K'})$, contradicting the maximality of $(K,\mathscr{B}_K)$. From this we conclude that $K=J$. Therefore,

$$\mathsf{N}_K=\mathsf{N}_J=\mathsf{N}\cap\bigoplus_{j\in J}\operatorname{span}_{\mathsf{R}}(e_j)=\mathsf{N}\cap\mathsf{M}=\mathsf{N}.$$

Thus $\mathscr{B}_K$ is a basis for $\mathsf{N}_K = \mathsf{N}$. Moreover,

$$\operatorname{rank}(\mathsf{N}) = \operatorname{card}(\mathscr{B}_K) \le \operatorname{card}(K) \le \operatorname{card}(I) = \operatorname{rank}(\mathsf{M}),$$

which gives the theorem. ■

Let us present an example that shows the manner in which modules over principal ideal domains differ from vector spaces, even though Theorem 4.9.1 tells us that there are some similarities in the two situations.

4.9.2 Example (Submodule of a module over a principal ideal domain) Let us consider the principal ideal domain $\mathbb{Z}$ and the $\mathbb{Z}$-module $\mathbb{Z}^2$ of ordered pairs of elements of $\mathbb{Z}$. Note that $\{(1,0),(0,1)\}$ is a basis for $\mathbb{Z}^2$, and so, by Theorem 4.9.1, every submodule of $\mathbb{Z}^2$ possesses a basis with at most two elements. Let us consider the submodule N of $\mathbb{Z}^2$ generated by the vectors $\{(1,0),(0,2)\}$. Note that $\mathsf{N} \subset \mathbb{Z}^2$ since, for example, $(0,1) \notin \mathsf{N}$. However, it still holds that $\operatorname{rank}(\mathsf{N}) = 2 = \operatorname{rank}(\mathbb{Z}^2)$. Note that this is a situation that cannot happen for finite-dimensional vector spaces. That is, if U is a *strict* subspace of a finite-dimensional vector space V, then it always holds that $\dim(\mathsf{U}) < \dim(\mathsf{V})$. (For infinite-dimensional vector spaces, this is not necessarily the case; see Exercise 4.5.24.) •

Let us next examine the interplay between freeness and torsion for modules over principal ideal domains.

4.9.3 Theorem (Torsion-free modules over principal ideal domains are free and vice versa) *If* R *is a principal ideal domain and if* M *is a finitely generated* R*-module, the following statements are equivalent:*

(i) M *is free;*

(ii) $\operatorname{Tor}(\mathsf{M}) = \{0_{\mathsf{M}}\}$.

Proof It is evident (and is Exercise 4.8.8) that a free R-module is torsion-free provided that R is simply an integral domain.

So suppose that R is a principal ideal domain and that M is a torsion-free R-module generated by $\{x_1,\dots,x_k\}$. For $k = 1$ we have $\{x_1\}$ as linearly independent since M is torsion-free (why?). Now suppose that the generating set $\{x_1,\dots,x_k\}$ is ordered such that the set $\{x_1,\dots,x_m\}$ is the largest linearly independent subset for $m \in \{1,\dots,k\}$. For $j \in \{m+1,\dots,k\}$ the set $\{x_1,\dots,x_m,x_j\}$ is linearly dependent so there exists $c_1,\dots,c_m,c_j \in \mathsf{R}$, not all zero, such that

$$c_1x_1 + \cdots + c_mx_m + c_jx_j = 0_{\mathsf{M}}.$$

Moreover, it is evident that $c_j \neq 0_{\mathsf{R}}$ since the set $\{x_1,\dots,x_m\}$ is linearly independent. Define $C = c_{m+1}\cdots c_k$ (note that $C \neq 0_{\mathsf{R}}$) so that

$$\{Cx \mid x \in \operatorname{span}_{\mathsf{R}}(x_1,\dots,x_k)\} \subseteq \operatorname{span}_{\mathsf{R}}(x_1,\dots,x_m).$$

Thus the submodule on the left is a submodule of the free module on the right. By Theorem 4.9.1 it follows that the module on the left is free. However,

$$\{Cx \mid x \in \operatorname{span}_{\mathsf{R}}(x_1,\dots,x_k)\} = \{Cx \mid x \in \mathsf{M}\}.$$

We claim that freeness of the module on the right implies freeness of M. Indeed, if

$$\{Cx \mid x \in \mathsf{M}\} = \operatorname{span}_{\mathsf{R}}(Ce_1, \dots, Ce_n)$$

then it is easy to see that $\mathsf{M} = \operatorname{span}_{\mathsf{R}}(e_1, \dots, e_n)$. ■

Let us illustrate the theorem with an example.

4.9.4 Example (Nonfree module over $\mathbb{Z}$) The $\mathbb{Z}$-module $\mathbb{Z}_4 = \mathbb{Z}/4\mathbb{Z}$ is not torsion-free as we saw in Example 4.8.54–4. The module is, however, finitely generated (indeed, it is consists of only four elements). Thus it should also not be free. Indeed, note that $(2 + 4\mathbb{Z}) + (2 + 4\mathbb{Z}) = 0 + 4\mathbb{Z}$. Thus there are two elements in $\mathbb{Z}_4$ which sum to zero. This cannot happen in any free, finitely generated $\mathbb{Z}$-module, since such modules are isomorphic to $\mathbb{Z}^n$ for some $n \in \mathbb{Z}_{>0}$. •

The next situation we consider is the case when a finitely generated module over a principal ideal domain is not torsion-free, and hence not free. We see that in this case one can do the best that could be expected, i.e., "separate" the free and nonfree parts. This result constitutes the first decomposition theorem that we will encounter for finitely generated modules over principal ideal domains.

4.9.5 Theorem (Nonfree modules over principal ideal domains) *If R is a principal ideal domain and if M is a finitely generated R-module, then the R-module $\mathsf{M}/\mathrm{Tor}(\mathsf{M})$ is free. Moreover, there exists a free submodule $\mathsf{M}_{\mathrm{free}}$ of M such that $\mathsf{M} = \mathrm{Tor}(\mathsf{M}) \oplus \mathsf{M}_{\mathrm{free}}$. Finally, if T and N are torsion and free submodules of M such that $\mathsf{M} = \mathsf{T} \oplus \mathsf{N}$, then $\mathsf{T} = \mathrm{Tor}(\mathsf{M})$.*

Proof Since $\mathsf{M}/\mathrm{Tor}(\mathsf{M})$ is torsion-free by Proposition 4.8.53(iii), it follows from Theorem 4.9.3 that it is also free. To show that there exists a complement to $\mathrm{Tor}(\mathsf{M})$ in M we use the following lemma.

1 **Lemma** *Let R be a commutative unit ring, let M and N be R-modules, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{N})$ be an epimorphism. If N is free then there exists a complement to $\ker(\mathsf{L})$ in M.*

Proof Let $\{f_j\}_{j\in J}$ be a basis for N and for $j \in J$ let $x_j \in \mathsf{L}^{-1}(f_j)$ (by the Axiom of Choice). Take $\mathsf{P} = \operatorname{span}_{\mathsf{R}}(x_j \mid\ j \in J)$. We claim that P is a complement to $\ker(\mathsf{L})$. Let $x \in \mathsf{M}$ and write

$$\mathsf{L}(x) = \sum_{j\in J} c_j f_j,$$

for $c_j \in \mathsf{R}$, $j \in J$, only finitely many of which are nonzero. Then define $x' = x - \sum_{j\in J} c_j x_j$. We then have

$$\mathsf{L}(x') = \mathsf{L}(x) - \sum_{j\in J} c_j \mathsf{L}(x_j) = 0_{\mathsf{N}}$$

so that $x' \in \ker(\mathsf{L})$. Now let $y \in \ker(\mathsf{L}) \cap \mathsf{P}$. Since $y \in \mathsf{P}$ we can write $y = \sum_{j\in J} c_j x_j$ for some $c_j \in \mathsf{R}$, $j \in J$, only finitely many of which are nonzero. Since $y \in \ker(\mathsf{L})$ we have

$$\mathsf{L}(y) = \sum_{j\in J} c_j \mathsf{L}(x_j) = \sum_{j\in J} c_j f_j = 0_{\mathsf{N}}.$$

Since $\{f_j\}_{j\in J}$ is a basis, it follows that $c_j = 0_{\mathsf{R}}$ for each $j \in J$ and so $\ker(\mathsf{L}) \cap \mathsf{P} = \{0_{\mathsf{M}}\}$, giving the result. ▼

Applying this lemma to the case of the projection from M to the free module $\mathsf{M}/\mathrm{Tor}(\mathsf{M})$ we see that $\mathrm{Tor}(\mathsf{M})$ does indeed possess a complement.

Now we prove the final assertion of the theorem. Clearly we must have $\mathsf{T} \subseteq \mathrm{Tor}(\mathsf{M})$. Now suppose that $x \in \mathrm{Tor}(\mathsf{M})$ and write $x = x_1 + x_2$ for $x_1 \in \mathsf{T}$ and $x_2 \in \mathsf{N}$. Then there exists nonzero $a, a_1 \in \mathsf{R}$ such that $ax = 0_{\mathsf{M}}$ and $a_1 x_1 = 0_{\mathsf{M}}$. Thus $aa_1x_2 = 0_{\mathsf{M}}$ which gives $x_2 = 0_{\mathsf{M}}$ since N is free, and so torsion-free. Thus $x \in \mathsf{T}$, giving $\mathsf{T} = \mathrm{Tor}(\mathsf{M})$, as desired. ■

Note that the choice of free complement to $\mathrm{Tor}(\mathsf{M})$ in Theorem 4.9.5 is not necessarily unique. However, since any complement is isomorphic to the quotient $\mathsf{M}/\mathrm{Tor}(\mathsf{M})$, it follows that any two complements are isomorphic.

4.9.2 Cyclic modules over principal ideal domains

In Theorem 4.9.5 we saw that a finitely generated module over a principal ideal domain can be decomposed as a direct sum of a torsion module and a free module. Since free modules are easy to understand, we should focus on understanding finitely generated torsion modules. It turns out that the way to do this is through cyclic modules and primary modules. In this section we study cyclic modules.

Let us first give the definition.

4.9.6 Definition (Cyclic module) For a commutative ring R, an R-module M is a ***cyclic module*** if $\mathsf{M} = \{rx \mid r \in \mathsf{R}\}$ for some $x \in \mathsf{M}$. •

For fields, cyclic submodules are isomorphic to the field thought of as a vector space. For general modules, cyclic modules can be more complicated, although one has the following useful characterisation.

4.9.7 Proposition (Characterisation of cyclic modules) *If R is a commutative ring M is a cyclic R-module with $\mathsf{M} = \mathrm{span}_{\mathsf{R}}(\mathrm{x})$, then M is isomorphic to the R-module $\mathsf{R}/\mathrm{ann}(\mathrm{x})$.*

Proof Consider the map $\sigma \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}; \mathsf{M})$ given by $\sigma(r) = rx$. It is clear that this is an epimorphism of R-modules. It is then easy to show that the map $\sigma_0 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}/\ker(\sigma); \mathsf{M})$ is an isomorphism (see Exercise 4.8.5). Since $\mathrm{ann}(x) = \ker(\sigma)$ the result follows. ■

For modules over principal ideal domains, there are useful facts that can be exploited to better understand their structure. We being with a definition that makes sense since ideals in principal ideal domains are generated by single ring elements.

4.9.8 Definition (Order) If R is a principal ideal domain, M is an R-module, and if N is a submodule of M, an ***order*** of N is a generator of the ideal $\mathrm{ann}(\mathsf{N})$. An ***order*** of $x \in \mathsf{M}$ is an order of the submodule $\mathrm{span}_{\mathsf{R}}(x)$. •

Note that any two orders for a submodule are associates by Propositions 4.2.60 and 4.2.61.

Now let us give some of the basic properties of cyclic modules over principal ideal domains.

4.9.9 Proposition (Properties of cyclic modules over principal ideal domains) *Let* R *be a principal ideal domain and let* $\mathsf{M} = \mathrm{span}_{\mathsf{R}}(\mathrm{x})$ *be an* R*-module with order* r*. Then the following statements hold:*

(i) if M *is cyclic then any submodule of* M *is cyclic;*

(ii) if $\mathrm{r} = \mathrm{p}_1^{\mathrm{k}_1} \cdots \mathrm{p}_{\mathrm{m}}^{\mathrm{k}_{\mathrm{m}}}$ *is a prime factorisation with* $\mathrm{p}_1, \dots, \mathrm{p}_{\mathrm{m}}$ *nonassociate primes, then* M *is isomorphic to the* R*-module*

$$\mathsf{R}/(\mathrm{p}_1^{\mathrm{k}_1}) \oplus \cdots \oplus \mathsf{R}/(\mathrm{p}_{\mathrm{m}}^{\mathrm{k}_{\mathrm{m}}}).$$

Proof (i) Suppose that $\mathsf{M} = \mathrm{span}_{\mathsf{R}}(x)$ and let $\mathrm{ann}(x) = (a)$. Then let $\sigma \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}; \mathsf{M})$ be the epimorphism defined by $\sigma(r) = rx$. Let N be a submodule of M and note that $\sigma^{-1}(\mathsf{N})$ is then a submodule of R; that is, $\sigma^{-1}(\mathsf{N})$ is an ideal of R. Since R is a principal ideal domain, $\sigma^{-1}(\mathsf{N}) = (r)$ for some $r \in \mathsf{R}$. Then $\sigma(r)$ is a generator for N, giving the result.

(ii) By Proposition 4.9.7 it suffices to show that $\mathsf{R}/(r)$ is isomorphic to

$$\mathsf{R}/(p_1^{k_1}) \oplus \cdots \oplus \mathsf{R}/(p_m^{k_m}).$$

It will, therefore, suffice (by induction) to show that $\mathsf{R}/(r_1 r_2)$ is isomorphic to $\mathsf{R}/(r_1) \oplus \mathsf{R}/(r_2)$ when r_1 and r_2 are relatively prime. Let $\phi_1 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}; \mathsf{R})$ be defined by $\phi_1(r) = r r_1$ and note that ϕ_1 maps (r_2) onto $(r_1 r_2)$. One can then easily check that this induces a well-defined homomorphism $\bar{\phi}_1 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}/(r_2); \mathsf{R}/(r_1 r_2))$ defined by

$$\bar{\phi}_1(r + (r_2)) = \phi_1(r) + (r_1 r_2) = r r_1 + (r_1 r_2).$$

In like manner one defines a homomorphism $\bar{\phi}_2 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}/(r_1); \mathsf{R}/(r_1 r_2))$ given by $\bar{\phi}_2(r + (r_1)) = r r_2 + (r_1 r_2)$. Now define

$$\begin{aligned} \bar{\phi} \colon & \mathsf{R}/(r_1) \oplus \mathsf{R}/(r_2) \to \mathsf{R}/(r_1 r_2) \\ & (r + (r_1), s + (r_2)) \mapsto \bar{\phi}_2(r) + \bar{\phi}_1(s) = (r r_2 + s r_1) + (r_1 r_2). \end{aligned}$$

We leave it to the reader to verify that this map is well-defined. Since r_1 and r_2 are relatively prime, by Proposition 4.2.77(ii) there exists $s_1, s_2 \in \mathsf{R}$ such that $s_1 r_1 + s_2 r_2 = 1_{\mathsf{R}}$. Therefore, for $a \in \mathsf{R}$ we compute

$$\bar{\phi}(a s_2 + (r_1), a s_1 + (r_2)) = a s_2 r_2 + a s_2 r_1 + (r_1 r_2) = a + (r_1 r_2),$$

showing that $\bar{\phi}$ is an epimorphism. Now suppose that

$$\bar{\phi}(r + (r_1), s + (r_2)) = 0_{\mathsf{R}} + (r_1 r_2) \quad \Longrightarrow \quad r r_2 + s r_1 \in (r_1 r_2).$$

Thus $r r_2 + s r_1 = a r_1 r_2$ for some $a \in \mathsf{R}$. Therefore,

$$s_1 r r_2 + s_1 s r_1 = s_1 a r_1 r_2.$$

Now we use the relation

$$s_1 r_1 + s_2 r_2 = 1_{\mathsf{R}} \quad \Longrightarrow \quad s r_1 s_1 = s - s r_2 s_2$$

to give

$$s_1rr_2 + s - sr_2s_2 = s_1ar_1r_2 \quad \Longrightarrow \quad s = s_1ar_1r_2 + sr_2s_2 - s_1rr_2,$$

giving $s \in (r_2)$. A similar argument gives $r \in (r_1)$. Therefore,

$$(r + (r_1), s + (r_2)) = (0_{\mathsf{R}} + (r_1), 0_{\mathsf{R}} + (r_2))$$

showing that $\bar{\phi}$ is injective by Exercise 4.8.3. ■

The following example illustrates the preceding result.

4.9.10 Example (Cyclic module over $\mathbb{Z}$) We consider $\mathbb{Z}_6 = \mathbb{Z}/6\mathbb{Z}$ as a $\mathbb{Z}$-module, noting that $\mathbb{Z}$ is a principal ideal domain (by Theorems 4.2.45 and 4.2.55). Clearly $\mathbb{Z}_6$ is cyclic and $6 \in \mathbb{Z}$ is an order. Since the prime factorisation of 6 is $6 = 2 \cdot 3$, Proposition 4.9.9 tells us that $\mathbb{Z}_6$ is isomorphic to $\mathbb{Z}_2 \oplus \mathbb{Z}_3$. The proof of Proposition 4.9.9 moreover gives the isomorphism as

$$(j + 2\mathbb{Z}, k + 3\mathbb{Z}) \mapsto (3j + 2k) + 6\mathbb{Z}.$$ •

4.9.3 Primary modules

In a cyclic module all elements have the same orders. We next consider a generalisation where not all elements have the same orders, but where all orders are (up to multiplication by a unit) powers of the same prime ring element.

4.9.11 Definition (Primary module) For a principal ideal domain R, a ***primary module*** over R is an R-module M such that there exists a prime $p \in \mathsf{R}$ such that for each $x \in \mathsf{M}$ there exists $k \in \mathbb{Z}_{>0}$ for which $\operatorname{ann}(x) = (p^k)$. •

For modules that are not primary, one can still define submodules that are primary.

4.9.12 Definition (Primary submodule) For a principal ideal domain R, a prime $p \in \mathsf{R}$, and an R-module M, the ***p-primary submodule*** is the subset

$$\mathsf{M}(p) = \{x \in \mathsf{M} \mid \operatorname{ann}(x) = (p^k) \text{ for some } k \in \mathbb{Z}_{>0}\}.$$ •

We should verify that $\mathsf{M}(p)$ is indeed a submodule.

4.9.13 Proposition (Primary submodules are submodules) *For a principal ideal domain R, a prime $\mathsf{p} \in \mathsf{R}$, and an R-module M, $\mathsf{M}(\mathsf{p})$ is a submodule.*

Proof We begin with a simple lemma which uses the notation from the statement of the result.

1 Lemma *If* $\mathrm{p}^{\mathrm{k}} \in \operatorname{ann}(\mathrm{x})$ *then* $\operatorname{ann}(\mathrm{x}) = (\mathrm{p}^{\mathrm{j}})$ *for some* $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$.

Proof We have $\operatorname{ann}(x) = (r)$ for some $r \in \mathsf{R}$ and so $r|p^k$. Since R is a unique factorisation domain we must then have $r = up^j$ for some $j \in \{1, \dots, k\}$. ▼

Let $x_1, x_2 \in \mathsf{M}(p)$ with $\operatorname{ann}(x_1) = (p^{k_1})$ and $\operatorname{ann}(x_2) = (p^{k_2})$. Let $k = \max\{k_1, k_2\}$ and note that $p^k(x_1 + x_2) = 0_{\mathsf{M}}$ so that $(p^k) \subseteq \operatorname{ann}(x_1 + x_2)$. Then $\operatorname{ann}(x_1 + x_2) = (p^j)$ for some $j \in \{1, \dots, k\}$ by the lemma and so $x_1 + x_2 \in \mathsf{M}(p)$. Also let $x \in \mathsf{M}(p)$ and let $r \in \mathsf{R}$, supposing that $\operatorname{ann}(x) = (p^k)$. Then $p^k rx = rp^k x = 0_{\mathsf{M}}$ and so $(p^k) \subseteq \operatorname{ann}(rx)$. The lemma gives $\operatorname{ann}(rx) = (p^j)$ for some $j \in \{1, \dots, k\}$ and so $rx \in \mathsf{M}(p)$. ■

In order to explain the importance of the notion of primary modules in the theory of modules over a principal ideal domain, the following result tells us that any torsion module is a direct sum of primary modules.

4.9.14 Theorem (Torsion modules and primary modules over principal ideal domains) *If* R *is a principal ideal domain and if* M *is a torsion* R*-module, then*

$$\mathsf{M} = \bigoplus_{\substack{\mathrm{p} \in \mathsf{R}, \\ \mathrm{p}\ \textit{prime}}} \mathsf{M}(\mathrm{p}).$$

Proof Let $x \in \mathsf{M}$ be nonzero and let $\operatorname{ann}(x) = (r)$. Write $r = p_1^{k_1} \cdots p_m^{k_m}$ for nonassociate primes $p_1, \dots, p_m$ and for $k_1, \dots, k_m \in \mathbb{Z}_{>0}$. For $j \in \{1, \dots, m\}$ define

$$r_j = p_1^{k_1} \cdots p_{j-1}^{k_{j-1}} p_{j+1}^{k_{j+1}} \cdots p_m^{k_m}.$$

Since $\{r_1, \dots, r_m\}$ are relatively prime, by Proposition 4.2.77(ii) there exists $s_1, \dots, s_k \in \mathsf{R}$ such that

$$s_1 r_1 + \cdots + s_k r_k = 1_{\mathsf{R}}.$$

Since $p_j^{k_j} s_j r_j x = s_j r_s p_j^{k_j} x = 0_{\mathsf{M}}$, $s_j r_j x \in \mathsf{M}(p_j)$ for $j \in \{1, \dots, m\}$. Therefore, since

$$x = 1_{\mathsf{R}} x = s_1 r_1 x + \cdots + s_m r_m x \in \mathsf{M}(p_1) + \cdots + \mathsf{M}(p_m).$$

Now we show that the sum is direct. Let $p \in \mathsf{R}$ be a prime and denote

$$\mathsf{M}(!p) = \sum_{\substack{q \in \mathsf{R}, \\ q \text{ prime}, \\ q \neq p}} \mathsf{M}(q).$$

Suppose that $x \in \mathsf{M}(p) \cap \mathsf{M}(!p)$ so that $p^k x = 0_{\mathsf{M}}$ for some $k \in \mathbb{Z}_{>0}$ and so that $x = x_1 + \cdots + x_m$ for $x_j \in \mathsf{M}(q_j)$ for some collection $q_1, \dots, q_m$ of nonassociate primes none of which equals p. Thus we also have $q_j^{k_j} x = 0_{\mathsf{M}}$ for some $k_j \in \mathbb{Z}_{>0}$, $j \in \{1, \dots, m\}$. If $d = q_1^{k_1} \cdots q_m^{k_m}$ then $x \in \mathsf{M}(d)$ and, since d and p are relatively prime (by unique factorisation), by Proposition 4.2.77(ii) there exists $r, s \in \mathsf{R}$ such that $rp^k + sd = 1_{\mathsf{R}}$. Therefore,

$$x = 1_{\mathsf{R}} x = (rp^k + sd)x = 0_{\mathsf{M}},$$

showing that $\mathsf{M}(p) \cap \mathsf{M}(!p) = \{0_{\mathsf{M}}\}$, as desired. ■

The decomposition of a torsion module given in the preceding theorem has a name.

4.9.15 Definition (Primary decomposition) Let R be a principal ideal domain and let M be a torsion R-module. The direct sum decomposition

$$\mathsf{M} = \bigoplus_{\substack{p \in \mathsf{R}, \\ p \text{ prime}}} \mathsf{M}(p)$$

is the ***primary decomposition*** of M. •

The preceding theorem gives us an important decomposition for torsion modules, and so by virtue of Theorem 4.9.5, of finitely generated modules over principal ideal domains. Our next result gives a refinement of this theorem by showing that each of the direct summands in Theorem 4.9.14 admits a decomposition into cyclic modules in a specific manner. This theorem contains the bulk of the effort required to understand the structure of modules over principal ideal domains, and the proof is quite technical in places.

4.9.16 Theorem (Primary modules and cyclic modules over principal ideal domains) *Let* R *be a principal ideal domain and let* M *be a finitely generated primary* R*-module with* $\mathsf{M} = \mathsf{M}(\mathrm{p})$ *for some prime* $\mathrm{p} \in \mathsf{R}$*. Then there exists* $\mathrm{k}_1, \dots, \mathrm{k}_\mathrm{m} \in \mathbb{Z}_{>0}$ *such that*

(i) $\mathrm{k}_1 \geq \cdots \geq \mathrm{k}_\mathrm{m}$ *and*

(ii) M *is isomorphic as an* R*-module to the direct sum*

$$\mathsf{R}/(\mathrm{p}^{\mathrm{k}_1}) \oplus \cdots \oplus \mathsf{R}/(\mathrm{p}^{\mathrm{k}_\mathrm{m}}).$$

Moreover, if $\mathrm{q} \in \mathsf{R}$ *is a prime such that* M *is isomorphic to the* R*-module*

$$\mathsf{R}/(\mathrm{q}^{\mathrm{l}_1}) \oplus \cdots \oplus \mathsf{R}/(\mathrm{q}^{\mathrm{l}_\mathrm{n}}),$$

where $\mathrm{l}_1, \dots, \mathrm{l}_\mathrm{n} \in \mathbb{Z}_{>0}$ *satisfy* $\mathrm{l}_1 \geq \cdots \geq \mathrm{l}_\mathrm{n}$*, then* q *and* p *are associates,* $\mathrm{n} = \mathrm{m}$*, and* $\mathrm{l}_\mathrm{j} = \mathrm{k}_\mathrm{j}$*,* $\mathrm{j} \in \{1, \dots, \mathrm{m}\}$*.*

Proof For the existence part of the proof we employ a technical lemma. The proof of technical lemma relies on a simple fact which we prove first.

1 Lemma *If* R *is a principal ideal domain, if* M *is an* R*-module, if* $\mathrm{x} \in \mathsf{M}$*, and if* $\mathrm{ann}(\mathrm{x}) = (\mathrm{p}^\mathrm{k})$ *for some prime* $\mathrm{p} \in \mathsf{R}$ *and some* $\mathrm{k} \in \mathbb{Z}_{>0}$*, then* $\mathrm{p}^\mathrm{j}\mathrm{x} \neq 0_\mathsf{M}$ *for* $\mathrm{j} \in \{0, 1, \dots, \mathrm{k} - 1\}$*.*

Proof Suppose that $p^j x = 0_\mathsf{M}$ for some $j \in \mathbb{Z}_{\geq 0}$. Then $p^j \in \mathrm{ann}(x)$ and so $p^k | p^j$. Thus, by unique factorisation, $j \geq k$. ▼

Now we state and prove the technical lemma.

2 Lemma *Let* R *be a principal ideal domain, let* $\mathrm{p} \in \mathsf{R}$ *be a prime, and let* M *be an* R*-module such that, for some* $\mathrm{k} \in \mathbb{Z}_{>0}$*,* $\mathrm{p}^\mathrm{k}\mathrm{x} = 0_\mathsf{M}$ *for each* $\mathrm{x} \in \mathsf{M}$ *and* $\mathrm{p}^{\mathrm{k}-1}\mathrm{x} \neq 0_\mathsf{M}$ *for some* $\mathrm{x} \in \mathsf{M}$*. If* $\mathrm{x} \in \mathsf{M}$ *satisfies* $\mathrm{ann}(\mathrm{x}) = (\mathrm{p}^\mathrm{k})$*, then* $\mathrm{span}_\mathsf{R}(\mathrm{x})$ *possesses a complement in* M*.*

Proof The result is vacuous when $\mathsf{M} = \operatorname{span}_{\mathsf{R}}(x)$ so let us suppose otherwise. Denote by $\mathscr{C}_x$ the set of all submodules P of M satisfying $\operatorname{span}_{\mathsf{R}}(x) \cap \mathsf{P} = \{0_{\mathsf{M}}\}$.

Let us first show that $\mathscr{C}_x$ is not empty. Let $z \in \mathsf{M} \setminus \operatorname{span}_{\mathsf{R}}(x)$. Since $p^k z = 0_{\mathsf{M}} \in \operatorname{span}_{\mathsf{R}}(x)$ there exists a least $j \in \mathbb{Z}_{>0}$ such that $p^j z \in \operatorname{span}_{\mathsf{R}}(x)$. Thus $p^{j-1}z \notin \operatorname{span}_{\mathsf{R}}(x)$ and $p^j z = sx$ for some $s \in \mathsf{R}$. We may then factor s as $s = rp^m$ for some $r \in \mathsf{R}$ and some $m \in \mathbb{Z}_{\geq 0}$ such that $p \nmid r$. Therefore,

$$0_{\mathsf{M}} = p^k z = p^{k-j} p^j z = p^{k-j} r p^m x.$$

Since $p \nmid r$ and $p^{k-1}x \neq 0_{\mathsf{M}}$ (by the lemma above), $k - j + m \geq k$; thus $m \geq j \geq 1$. Now take $y = p^{j-1}z - rp^{m-1}x$. Since $p^{j-1}z \notin \operatorname{span}_{\mathsf{R}}(x)$ it follows that $y \neq 0_{\mathsf{M}}$. Also,

$$py = p^j z - rp^m x = p^j z - sx = p^j z - p^j z = 0_{\mathsf{M}}.$$

Let $sy \in \operatorname{span}_{\mathsf{R}}(x) \cap \operatorname{span}_{\mathsf{R}}(y)$ so that $sy \in \operatorname{span}_{\mathsf{R}}(x)$ for some $s \in \mathsf{R}$. Suppose that $sy \neq 0_{\mathsf{m}}$. Then $p \nmid s$ since $py = 0_{\mathsf{M}}$. Therefore, s and p^k are relatively prime and, by Proposition 4.2.77(ii), there exists $a, b \in \mathsf{R}$ such that $as + bp^k = 1_{\mathsf{R}}$. Then

$$y = 1_{\mathsf{R}} y = asy + bp^k y = a(sy) \in \operatorname{span}_{\mathsf{R}}(x).$$

Therefore,

$$p^{j-1}z = y + rp^{m-1}x \in \operatorname{span}_{\mathsf{R}}(x).$$

If $j - 1 = 0$ then this contradicts the choice that $z \notin \operatorname{span}_{\mathsf{R}}(x)$. If $j - 1 \neq 0$ then this contradicts the fact that j is the least positive integer satisfying $p^j z \in \operatorname{span}_{\mathsf{R}}(x)$. Either way, we conclude that $sy = 0_{\mathsf{M}}$ and so $\operatorname{span}_{\mathsf{R}}(y) \cap \operatorname{span}_{\mathsf{R}}(x) = \{0_{\mathsf{M}}\}$. Thus $\mathscr{C}_x$ is nonempty.

Now place a partial order on $\mathscr{C}_x$ by set inclusion. Let $\{\mathsf{P}_j \mid j \in J\}$ be a totally ordered subset of $\mathscr{C}_x$. Then $\cup_{j \in J} \mathsf{P}_j$ is an upper bound for this totally ordered subset since

$$\operatorname{span}_{\mathsf{R}}(x) \cap \left(\cup_{j \in J} \mathsf{P}_j\right) = \cup_{j \in J}\left(\operatorname{span}_{\mathsf{R}}(x) \cap \mathsf{P}_j\right) = \{0_{\mathsf{M}}\}$$

by Proposition 1.1.7. By Zorn's Lemma we may conclude the existence of a maximal element N of $\mathscr{C}_x$.

We now work with the quotient module M/N. Note that $p^k(x + \mathsf{N}) = 0_{\mathsf{M}/\mathsf{N}}$. Since $p^{k-1}x \neq 0_{\mathsf{M}}$ by Lemma 1 and since $\operatorname{span}_{\mathsf{R}}(x) \cap \mathsf{N} = \{0_{\mathsf{M}}\}$ we have $p^{k-1}x \notin \mathsf{N}$. Thus $p^{k-1}(x + \mathsf{N}) \neq 0_{\mathsf{M}/\mathsf{N}}$ and so $\operatorname{ann}(x + \mathsf{N}) = (p^k)$. This also shows that $p^{k-1} \notin \operatorname{ann}(\mathsf{M}/\mathsf{N})$. Thus the R-module M/N shares with M the property that $p^k(x' + \mathsf{N}) = 0_{\mathsf{M}/\mathsf{N}}$ for every $x' + \mathsf{N} \in \mathsf{M}/\mathsf{N}$ and that there exists $x' + \mathsf{N} \in \mathsf{M}/\mathsf{N}$ such that $p^{k-1}(x' + \mathsf{N}) \neq \mathsf{M}/\mathsf{N}$. Now we claim that $\mathsf{M}/\mathsf{N} = \operatorname{span}_{\mathsf{R}}(x + \mathsf{N})$. Suppose not. Then, as we saw during the course of proving that $\mathscr{C}_x$ is nonempty, there exists a nonzero $y + \mathsf{N} \in \mathsf{M}/\mathsf{N}$ such that

$$\operatorname{span}_{\mathsf{R}}(x + \mathsf{N}) \cap \operatorname{span}_{\mathsf{R}}(y + \mathsf{N}) = \{0_{\mathsf{M}/\mathsf{N}}\}.$$

Since $\operatorname{span}_{\mathsf{R}}(x) \cap \mathsf{N} = \{0_{\mathsf{M}}\}$, this implies that

$$\operatorname{span}_{\mathsf{R}}(x) \cap \left(\operatorname{span}_{\mathsf{R}}(y) + \mathsf{N}\right) = \{0_{\mathsf{M}}\}.$$

Thus we see that $\operatorname{span}_{\mathsf{R}}(y) + \mathsf{N} \in \mathscr{C}_x$. Since $y \notin \mathsf{N}$ this contradicts the maximality of N, and so we conclude that N is the cyclic module generated by $x + \mathsf{N}$. Thus N is a complement to $\operatorname{span}_{\mathsf{R}}(x)$, as desired. ▼

Now we proceed with the existence part of the proof of the theorem. The proof is by induction on the number of generators of M. If M is generated by $x \in \mathsf{M}$ then $\mathsf{M} = \mathsf{R}/(p^k)$ for some $k \in \mathbb{Z}_{>0}$, and the result is clear. So suppose the result is true for modules possessing $m-1$ generators and let M be generated by $x_1, \dots, x_m$. Suppose that

$$\operatorname{ann}(x_1) = (p^{k_1}),\ \operatorname{ann}(x_j) = (p^{l_j}), \qquad j \in \{2, \dots, m\}.$$

Without loss of generality suppose that $k_1 = \max\{k_1, l_2, \dots, l_m\}$. From this assumption and from Lemma 1 it follows that $p^{k_1}x = 0_{\mathsf{M}}$ for every $x \in \mathsf{M}$ and that $p^{k_1-1}x \neq 0_{\mathsf{M}}$ for some $x \in \mathsf{M}$. By Lemma 2 there exists a complement N to $\operatorname{span}_{\mathsf{R}}(x_1)$ in M. Denote by $\pi\colon \mathsf{M} \to \mathsf{N}$ the projection defined by the direct sum decomposition $\mathsf{M} = \operatorname{span}_{\mathsf{R}}(x_1) \oplus \mathsf{N}$. Note that since M is generated by $\{x_1, \dots, x_m\}$, N is generated by $\{\pi(x_1), \dots, \pi(x_m)\}$. Moreover $\pi(x_1) = 0_{\mathsf{N}}$ and so N is finitely generated by at most $m-1$ elements. By the induction hypothesis N is isomorphic to

$$\mathsf{R}/(p^{k_2}) \oplus \dots \oplus \mathsf{R}/(p^{k_m})$$

with $k_2 \geq \dots \geq k_m$. Since $p^{k_1}x = 0_{\mathsf{M}}$ for every $x \in \mathsf{N}$ it follows that $k_1 \geq k_2$. Since $\operatorname{span}_{\mathsf{R}}(x_1)$ is isomorphic to $\mathsf{R}/(p^{k_1})$ the existence part of the theorem follows.

Now we prove the uniqueness. To do so we introduce some notation and an accompanying lemma. For the moment, let M be a general R-module and define

$$\mathsf{M}[p] = \{x \in \mathsf{M} \mid\ px = 0_{\mathsf{M}}\}$$

for a prime $p \in \mathsf{R}$. Let us adopt the notation

$$r\mathsf{M} = \{rx \mid\ x \in \mathsf{M}\}$$

for convenience. Note that $\mathsf{R}/(p)$ is a field by Theorem 4.3.9 since (p) is a maximal ideal by Theorem 4.2.64 and Proposition 4.2.70. Therefore, the following lemma makes sense.

3 Lemma *Let R be a principal ideal domain, let $\mathrm{p} \in \mathsf{R}$ be prime, and let M be an R-module. Then the following statements hold:*

(i) $\mathsf{M}[\mathrm{p}]$ *is a submodule of* M;

(ii) $\mathsf{M}[\mathrm{p}]$ *is a vector space over* $\mathsf{R}/(\mathrm{p})$;

(iii) *if* $\mathsf{M} = \mathsf{N} \oplus \mathsf{P}$ *then* $\mathsf{M}[\mathrm{p}] = \mathsf{N}[\mathrm{p}] \oplus \mathsf{P}[\mathrm{p}]$.

Proof (i) This is a straightforward computation.

(ii) We define scalar multiplication in the vector space by $(r + (p))x = rx$. Let us show that this operation is well-defined. If $r + (p) = r' + (p)$ then $r - r' = sp$ for some $s \in \mathsf{R}$. Then

$$(r' + (p))x = r'x = (r - sp)x = rx = (r + (p))x,$$

as desired. To show that this definition of scalar multiplication, combined with vector addition as inherited from the module structure, makes $\mathsf{M}[p]$ into a $\mathsf{R}/(p)$-vector space is now straightforward.

(iii) It is clean that $\mathsf{M}[p] \subseteq \mathsf{N}[p] \oplus \mathsf{P}[p]$. Moreover, since $\mathsf{N}[p] \subseteq \mathsf{N}$ and $\mathsf{P}[p] \subseteq \mathsf{P}$ it follows that $\mathsf{N}[p] \cap \mathsf{P}[p] = \{0_{\mathsf{M}}\}$. Now let $x \in \mathsf{M}[p]$ and write $x = y + z$ for $y \in \mathsf{N}$ and $z \in \mathsf{P}$. Since $px = 0_{\mathsf{M}}$ we have $py + pz = 0_{\mathsf{M}}$. Since $py \in \mathsf{N}$ and $pz \in \mathsf{P}$ it follows that $py = pz = 0_{\mathsf{M}}$, and so $y \in \mathsf{N}[p]$ and $z \in \mathsf{P}[p]$. Thus $\mathsf{M}[p] = \mathsf{N}[p] \oplus \mathsf{P}[p]$. ▼

Now we proceed to the uniqueness part of the proof of the theorem, returning to the case where M satisfies the hypotheses of the theorem. We have M isomorphic to both R-modules

$$\mathsf{R}/(p^{k_1}) \oplus \cdots \oplus \mathsf{R}/(p^{k_m}), \quad \mathsf{R}/(q^{l_1}) \oplus \cdots \oplus \mathsf{R}/(q^{l_n}),$$

with $k_1 e.g., \cdots \geq k_m$ and $l_1 \geq \cdots \geq l_n$. It follows that $\operatorname{ann}(\mathsf{M}) = (p^{k_1}) = (q^{l_1})$. Thus p_{k_1} and q^{l_1} are associates, and by unique factorisation it follows that q and p are associates and that $k_1 = l_1$.

Let us next show that $n = m$. Let $x_1, \dots, x_m$ and $y_1, \dots, y_n$ have the properties that $\operatorname{ann}(x_j)$ is isomorphic to (p^{k_j}), $j \in \{1, \dots, m\}$, and that $\operatorname{ann}(y_j)$ is isomorphic to (p^{l_j}), $j \in \{1, \dots, n\}$. Then, by Lemma 3, the $\mathsf{R}/(p)$-vector spaces

$$\operatorname{span}_{\mathsf{R}}(x_1)[p] \oplus \cdots \oplus \operatorname{span}_{\mathsf{R}}(x_m)[p]$$

and

$$\operatorname{span}_{\mathsf{R}}(y_1)[p] \oplus \cdots \oplus \operatorname{span}_{\mathsf{R}}(y_n)[p]$$

are isomorphic to $\mathsf{M}[p]$. Moreover, each of the summands is nontrivial given that they are each cyclic modules whose order is a power of p. Moreover, given the definition of scalar multiplication in these $\mathsf{R}/(p)$-vector spaces, each of the summands is one-dimensional. Thus $\mathsf{M}[p]$ is isomorphic to two direct sums of one-dimensional vector spaces, one with m components and the other with n components. Therefore, $m = n$.

Finally we show that $k_j = l_j$, $j \in \{1, \dots, m\}$. We do this by induction on k_1. If $k_1 = 1$ then $l_1 = 1$ (as we showed above) and so $l_j = k_j = 1$, $j \in \{1, \dots, m\}$. Now suppose that $k_j = l_j$, $j \in \{1, \dots, m\}$, whenever $k_j \in \{1, \dots, a-1\}$ for some $a \geq 2$. Then we can write

$$(k_1, \dots, k_m) = (k_1, \dots, k_r, 1, \dots, 1), \qquad k_r > 1$$

and

$$(l_1, \dots, l_m) = (l_1, \dots, l_s, 1, \dots, 1), \qquad l_s > 1,$$

for some $r, s \in \{1, \dots, m\}$. This then gives $p\mathsf{M}$ as being isomorphic to both of the submodules

$$p \operatorname{span}_{\mathsf{R}}(x_1) \oplus \cdots \oplus p \operatorname{span}_{\mathsf{R}}(x_m) = p \operatorname{span}_{\mathsf{R}}(x_1) \oplus \cdots \oplus p \operatorname{span}_{\mathsf{R}}(x_r)$$

and

$$p \operatorname{span}_{\mathsf{R}}(y_1) \oplus \cdots \oplus p \operatorname{span}_{\mathsf{R}}(y_m) = p \operatorname{span}_{\mathsf{R}}(y_1) \oplus \cdots \oplus p \operatorname{span}_{\mathsf{R}}(y_s).$$

Now note that $p \operatorname{span}_{\mathsf{R}}(x_1) = \operatorname{span}_{\mathsf{R}}(px_1)$ is a cyclic R-module with order $p^{k_1 - 1}$. Thus, by the induction hypotheses, $r = s$ and $k_j = l_j$, $j \in \{1, \dots, r\}$, giving the result. ■

The decomposition of a primary module as a direct sum of cyclic modules as in the theorem has a name.

4.9.17 Definition (Cyclic decomposition) Let R be a principal ideal domain and let M be a finitely generated primary R-module with $\mathsf{M} = \mathsf{M}(p)$ for some prime $p \in \mathsf{R}$. The decomposition, as in Theorem 4.9.16,

$$\mathsf{M} = \mathsf{C}_1 \oplus \cdots \oplus \mathsf{C}_m$$

with C_j isomorphic to $\mathsf{R}/(p^{k_j})$, $j \in \{1, \ldots, m\}$. is the ***cyclic decomposition*** of M. •

4.9.4 The primary-cyclic decomposition

With the results of the preceding three sections we can now fairly easily state and prove two important decomposition theorems for finitely generated modules over principal ideal domains. As we shall see in Section 5.8, these decomposition theorems are useful in linear algebra in distinct ways.

Let us simply state the theorem.

4.9.18 Theorem (Primary-cyclic decomposition) *If* R *is a principal ideal domain and if* M *is a finitely generated* R*-module, then there exists*

(i) $\mathrm{n} \in \mathbb{Z}_{\geq 0}$,

(ii) *nonassociate primes* $\mathrm{p}_1, \ldots, \mathrm{p}_\mathrm{n} \in \mathsf{R}$,

(iii) $\mathrm{m_j} \in \mathbb{Z}_{>0}$ *and* $\mathrm{l_{j1}}, \ldots, \mathrm{l_{jm_j}} \in \mathbb{Z}_{>0}$ *for each* $\mathrm{j} \in \{1, \ldots, \mathrm{n}\}$,

(iv) *primary submodules* $\mathsf{P}_1, \ldots, \mathsf{P}_\mathrm{n}$ *of* M, *and*

(v) *a free submodule* F *of* M

such that

(vi) $\mathrm{l_{j_1}} \geq \cdots \geq \mathrm{l_{jm_j}}$ *for each* $\mathrm{j} \in \{1, \ldots, \mathrm{n}\}$,

(vii) P_j *is isomorphic to*

$$\mathsf{R}/(\mathrm{p}_\mathrm{j}^{\mathrm{l_{j1}}}) \oplus \cdots \oplus \mathsf{R}/(\mathrm{p}_\mathrm{j}^{\mathrm{l_{jm_j}}})$$

for each $\mathrm{j} \in \{1, \ldots, \mathrm{n}\}$, *and*

(viii) $\mathsf{M} = \mathsf{P}_1 \oplus \cdots \oplus \mathsf{P}_\mathrm{n} \oplus \mathsf{F}$.

Moreover, if M *is a direct sum*

$$\mathsf{M} = \mathsf{Q}_1 \oplus \cdots \oplus \mathsf{Q}_\mathrm{m} \oplus \mathsf{E}$$

with Q_j *a* $\mathrm{q_j}$*-primary submodule isomorphic to*

$$\mathsf{R}/(\mathrm{q}_\mathrm{j}^{\mathrm{s_{j1}}}) \oplus \cdots \oplus \mathsf{R}/(\mathrm{q}_\mathrm{j}^{\mathrm{s_{jn_j}}})$$

for a prime $\mathrm{q_j} \in \mathsf{R}$, *for* $\mathrm{n_j} \in \mathbb{Z}_{>0}$, *and for* $\mathrm{s_{j1}}, \ldots, \mathrm{s_{jn_j}} \in \mathbb{Z}_{>0}$ *satisfying* $\mathrm{s_{j1}} \geq \cdots \geq \mathrm{s_{jn_j}}$, $\mathrm{j} \in \{1, \ldots, \mathrm{m}\}$, *and with* E *free, then* E *and* F *are isomorphic,* $\mathrm{m} = \mathrm{n}$, *and there exists a bijection* $\sigma\colon \{1, \ldots, \mathrm{n}\} \to \{1, \ldots, \mathrm{n}\}$ *such that, for each* $\mathrm{j} \in \{1, \ldots, \mathrm{n}\}$, $\mathsf{Q}_\mathrm{j} = \mathsf{P}_{\sigma(\mathrm{j})}$, $\mathrm{q_j}$ *and* $\mathrm{p}_{\sigma(\mathrm{j})}$ *are associates,* $\mathrm{n_j} = \mathrm{m}_{\sigma(\mathrm{j})}$, *and* $\mathrm{s_{ja}} = \mathrm{l}_{\sigma(\mathrm{j})\mathrm{a}}$ *for* $\mathrm{a} \in \{1, \ldots, \mathrm{n_j}\}$.

Proof By Theorem 4.9.5 we can write $\mathsf{M} = \mathsf{T} \oplus \mathsf{F}$ with T a torsion module and with F free. Moreover, the torsion part of this decomposition is unique and any two choices of free complement will be isomorphic since all complements are isomorphic to M/T and since R has the invariant rank property by Theorem 4.8.25. By Theorem 4.9.14 we know that T is a direct sum over a finite number of nonassociate primes of primary modules associated with these primes. The theorem then follows from Theorem 4.9.16. ■

The idea is that first M is a finite direct sum of it the primary submodules $\mathsf{P}_1, \dots, \mathsf{P}_n$, along with a free module F. This decomposition follows, as explained in the proof of the preceding theorem, from Theorems 4.9.5 and 4.9.14. Each of the primary modules $\mathsf{P}_j = \mathsf{M}(p_j)$, $j \in \{1, \dots, n\}$, can then be decomposed using the cyclic decomposition as a finite direct sum of cyclic submodules with orders $p_j^{l_{j1}}, \dots, p_j^{l_{jm_j}}$ by Theorem 4.9.16. The prime powers

$$p_1^{l_{11}}, \dots, p_1^{l_{1m_1}}, \dots, p_n^{l_{n1}}, \dots, p_n^{l_{m_n}}$$

are called the ***elementary divisors***, and are unique up to multiplication by units.

The decomposition of the theorem has a name.

4.9.19 Definition (Primary-cyclic decomposition) Let R be a principal ideal domain and let M be a finitely generated R-module. The decomposition $\mathsf{M} = \mathsf{P}_1 \oplus \cdots \oplus \mathsf{P}_n \oplus \mathsf{F}$ of Theorem 4.9.18 is the ***primary-cyclic decomposition*** of M. •

The simplest example of a principal ideal domain is $\mathbb{Z}$, and modules over $\mathbb{Z}$ are simply Abelian groups. Thus, applying Theorem 4.9.18 in this case gives a classification of finitely generated Abelian groups. Let us record this.

4.9.20 Example (Primary-cyclic decomposition for $\mathbb{Z}$) We let G be a finitely generated Abelian group, i.e., a finitely generated $\mathbb{Z}$-module by Example 4.8.2–2. Then Theorem 4.9.18 tells us that G is isomorphic to the group

$$\mathbb{Z}_{p_1^{k_1}} \oplus \cdots \oplus \mathbb{Z}_{p_n^{k_n}} \oplus \mathbb{Z}^k$$

for some $n, k \in \mathbb{Z}_{\geq 0}$ and $k_1, \dots, k_n \in \mathbb{Z}_{>0}$ and for some primes $p_1, \dots, p_n \in \mathbb{Z}_{>0}$. •

4.9.5 The invariant factor decomposition

The next decomposition we give is simply a rearrangement of the primary cyclic decomposition. However, it is a rearrangement done in such a way that (1) it is interesting and (2) it is uniquely defined in the same sense that the primary-cyclic decomposition is uniquely defined.

Let us simply state the theorem.

4.9.21 Theorem (Invariant factor decomposition) *If* R *is a principal ideal domain and if* M *is a finitely generated* R*-module, then there exists*

(i) $\mathrm{m} \in \mathbb{Z}_{\geq 0}$,

(ii) *(not necessarily distinct) nonzero nonunits* $\mathrm{r}_1, \dots, \mathrm{r}_\mathrm{m} \in \mathsf{R}$,

(iii) *cyclic submodules* $\mathsf{N}_1, \dots, \mathsf{N}_\mathrm{m}$ *of* M *with orders* $\mathrm{r}_1, \dots, \mathrm{r}_\mathrm{m}$, *respectively, and*

(iv) *a free submodule* F *of* M *such that*

$\mathrm{r}_1 | \cdots | \mathrm{r}_\mathrm{m}$ *and such that* $\mathsf{M} = \mathsf{N}_1 \oplus \dots \mathsf{N}_\mathrm{m} \oplus \mathsf{F}$.

Moreover, if M *is a direct sum*

$$\mathsf{M} = \mathsf{P}_1 \oplus \cdots \oplus \mathsf{P}_\mathrm{n} \oplus \mathsf{E}$$

with P_j *cyclic with nonzero nonunit orders* s_j, $\mathrm{j} \in \{1, \dots, \mathrm{n}\}$, *satisfying* $\mathrm{s}_1 | \cdots | \mathrm{s}_\mathrm{n}$ *and with* E *free, then* E *and* F *are isomorphic,* $\mathrm{n} = \mathrm{m}$, *and* $(\mathrm{s}_\mathrm{j}) = (\mathrm{r}_\mathrm{j})$, $\mathrm{j} \in \{1, \dots, \mathrm{n}\}$.

Proof The existence part of the proof consists of constructing the decomposition from the primary-cyclic decomposition of Theorem 4.9.18. We do this by arranging the data from the primary-cyclic decomposition in a particular way. Since the free module F is "along for the ride," let us simply suppose that M is a torsion module. We then know by Theorem 4.9.14 that there are nonassociate primes $p_1, \dots, p_k \in \mathsf{R}$ such that

$$\mathsf{M} = \mathsf{M}(p_1) \oplus \cdots \oplus \mathsf{M}(p_k).$$

For each $j \in \{1, \dots, k\}$ we then have $l_{j1}, \dots, l_{jm_j} \in \mathbb{Z}_{>0}$ such that $l_{j1} \geq \cdots \geq l_{jm_j}$ and such that $\mathsf{M}(p_j)$ is isomorphic to

$$\mathsf{R}/(p_j^{l_{j1}}) \oplus \cdots \oplus \mathsf{R}/(p_j^{l_{jm_j}}).$$

Let us take $m = \max\{m_1, \dots, m_k\}$. Let us then use this notation to write the following table for arranging the powers of primes:

$$\begin{matrix} p_1^{l_{1m}} & p_2^{l_{2m}} & \cdots & p_k^{l_{km}} \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{l_{12}} & p_2^{l_{22}} & \cdots & p_k^{l_{k2}} \\ p_1^{l_{11}} & p_2^{l_{21}} & \cdots & p_k^{l_{k1}} \end{matrix}$$

We adopt the convention that $l_{js} = 0$ if $s > m_j$. We then define

$$r_{m-j+1} = p_1^{l_{1j}} \cdots p_k^{l_{kj}}, \qquad j \in \{1, \dots, m\},$$

i.e., r_j is the product of the terms in the jth row in the table. Let us also take N_j to be the cyclic module with order r_j obtained by taking the direct sum of the cyclic submodules A_{sj}, of $\mathsf{M}(p_s)$ of order $p_j^{l_{sj}}$, $s \in \{1, \dots, k\}$. It is clear from our table that $r_1 | \cdots | r_m$, and that the $\mathsf{M} = \mathsf{N}_1 \oplus \cdots \oplus \mathsf{N}_m$.

Now we prove the uniqueness of the decomposition. Two decompositions will involve only the primes $p_1, \dots, p_k$ and so will give rise to two tables

$$\begin{array}{cccc} p_1^{l_{1m}} & p_2^{l_{2m}} & \cdots & p_k^{l_{km}} \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{l_{12}} & p_2^{l_{22}} & \cdots & p_k^{l_{k2}} \\ p_1^{l_{11}} & p_2^{l_{21}} & \cdots & p_k^{l_{k1}} \end{array}, \qquad \begin{array}{cccc} p_1^{q_{1n}} & p_2^{q_{2n}} & \cdots & p_k^{q_{kn}} \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{q_{12}} & p_2^{q_{22}} & \cdots & p_k^{q_{k2}} \\ p_1^{q_{11}} & p_2^{q_{21}} & \cdots & p_k^{q_{k1}}, \end{array}$$

for which the products of the rows are the ring elements $r_1, \dots, r_m$ and $s_1, \dots, s_n$. The requirement that $r_1 | \cdots | r_m$ and $s_1 | \cdots | s_n$ ensures that $l_{j1} \geq \cdots \geq l_{jm}$, $j \in \{1, \dots, k\}$ and $q_{j1} \geq \cdots \geq q_{jn}$, $j \in \{1, \dots, k\}$. Thus $\mathsf{M}(p_j)$ is isomorphic to both R-modules

$$\mathsf{R}/(p_j^{l_{j1}}) \oplus \cdots \oplus \mathsf{R}/(p_j^{l_{jm}})$$

and

$$\mathsf{R}/(p_j^{q_{j1}}) \oplus \cdots \oplus \mathsf{R}/(p_j^{q_{jn}})$$

(allowing that some components in the direct sum are zero). The uniqueness part of the result then follows from the uniqueness part of Theorem 4.9.16. ■

The ring elements $r_1, \dots, r_m$ in the statement of the theorem are called ***invariant factors*** of M. These are uniquely defined, up to multiplication by a unit, by the module M.

The decomposition of the preceding theorem has a name.

4.9.22 Definition (Invariant factor decomposition) Let R be a principal ideal domain and let M be a finitely generated R-module. The decomposition $\mathsf{M} = \mathsf{N}_1 \oplus \dots \mathsf{N}_m \oplus \mathsf{F}$ of Theorem 4.9.21 is the ***invariant factor decomposition*** of M. •

4.9.23 Example (Invariant factor decomposition for $\mathbb{Z}$) Let us give a concrete example of the structure of a finitely generated $\mathbb{Z}$-module to illustrate how one goes from the primary-cyclic decomposition to the invariant factor decomposition. As in Example 4.9.20 we let G be a finitely generated Abelian group and, without loss of generality, take

$$\mathsf{G} = \mathbb{Z}_{p_1^{k_1}} \oplus \cdots \oplus \mathbb{Z}_{p_n^{k_n}} \oplus \mathbb{Z}^k.$$

For concreteness we take the $n = 3$ and the following data:

j	p_j	k_j
1	2	1
2	2	3
3	3	2
4	3	2
5	3	5
6	7	2

The primes that will participate in our invariant factors are then 2, 3, and 7. There will be three invariant factors since the most appearances by any prime (in this case the prime 3) is three. The rule for constructing the invariant factors is to start by constructing the largest one as the product of the highest powers of participating primes, in this case 2^3, 3^5, and 7^2. The next to largest invariant factor then takes the products of the next highest powers of the participating primes, in this case 2^1, 3^2, and 7^0. This procedure is most easily illustrated in tabular form, just as it was in the proof of the theorem:

$$\begin{array}{ccc} 2^0 & 3^2 & 7^0 \\ 2^1 & 3^2 & 7^0 \\ 2^3 & 3^5 & 7^2 \end{array}$$

The invariant factors are then

$$r_1 = 2^0 \cdot 3^2 \cdot 7^0 = 9, \quad r_2 = 2^1 \cdot 3^2 \cdot 7^0 = 18, \quad r_3 = 2^3 \cdot 3^5 \cdot 7^2 = 95256. \qquad \bullet$$

Let us give an entirely related characterisation of the submodules of a free module over a principal ideal domain. Such submodules are necessarily free by Theorem 4.9.1 and have a well-defined rank by Theorem 4.8.25.

4.9.24 Theorem (Submodules of free modules over a principal ideal domain) *Let* R *be a principal ideal domain and let* M *be a finitely generated, free* R*-module with* $\mathsf{N} \subseteq \mathsf{M}$ *a submodule of finite rank* k*. Then there exists a basis* $\mathscr{B}$ *for* M, $\{\mathrm{e}_1, \dots, \mathrm{e}_k\} \subseteq \mathscr{B}$*, and* $\mathrm{r}_1, \dots, \mathrm{r}_k \in \mathsf{R}$ *such that*

(i) $\mathrm{r}_1 | \cdots | \mathrm{r}_k$ *and*

(ii) $\{\mathrm{r}_1\mathrm{e}_1, \dots, \mathrm{r}_k\mathrm{e}_k\}$ *is a basis for* N.

Proof Our proof makes use of Theorem 5.5.20 which itself follows from Theorem 5.2.43.

We let $i_{\mathsf{N}} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N}; \mathsf{M})$ be the inclusion of N into M. If $\mathrm{rank}(\mathsf{N}) = k$ then $\mathrm{rank}(i_{\mathsf{N}}) = k$. Therefore, by Theorem 5.5.20 there exists bases $\{f_1, \dots, f_k\}$ and $\{e_1, \dots, e_n\}$ for N and M, respectively, and $r_1, \dots, r_k \in \mathsf{R}$ such that

1. $r_1 | \cdots | r_k$ and
2. $i_{\mathsf{N}}(f_j) = r_j e_j$ for $j \in \{1, \dots, k\}$.

This implies that $f_j = r_j e_j$ (an elements of M) and so the result follows. ■

Thus a finite rank submodule of a free module over a principal ideal domain comes equipped with invariant factors as well. An example illustrates this.

4.9.25 Example (Subgroups of free, finitely generated Abelian groups) A free, finitely generated Abelian group is a free, finitely generated $\mathbb{Z}$-module. A free, finitely generated $\mathbb{Z}$-module is isomorphic to $\mathbb{Z}^n$. Let $\{e_1, \dots, e_n\}$ be the standard basis for $\mathbb{Z}^n$. The "simplest" submodules of $\mathbb{Z}^n$ are those generated by $\{j_1 e_1, \dots, j_k e_k\}$. Theorem 4.9.24 says that after a change of basis, *any* submodule of $\mathbb{Z}^n$ has this form. Moreover, by properly rearranging the prime factors of $j_1, \dots, j_k$, cf. Example 4.9.23, one can ensure that $j_1 | \cdots | j_k$. This rearrangement of prime factors is uniquely determined by this divisibility condition. •

Exercises

4.9.1 Let R be a principal ideal domain.

(a) Show that if M is a finitely generated torsion R-module then $\operatorname{ann}(\mathsf{M}) \neq \{0_{\mathsf{R}}\}$.

We wish to now see that the preceding statement is not generally true when M is not finitely generated. To see this, let $P \subseteq \mathbb{Z}$ be the set of prime numbers (these are taken to be positive, as usual) and define the $\mathbb{Z}$-module

$$\mathsf{M} = \oplus_{p \in P} \mathbb{Z}_p.$$

For this module answer the following questions.

(b) Show that M is a torsion $\mathbb{Z}$-module, but is not finitely generated.

(c) What is the primary decomposition of M?

(d) What is $\operatorname{ann}(\mathsf{M})$?

4.9.2 Let R be a principal ideal domain and let M be a finitely generated R-module. Suppose that we have two invariant factor decompositions for M:

$$\mathsf{M} = \mathsf{N}_{11} \oplus \cdots \oplus \mathsf{N}_{1m} \oplus \mathsf{F}_1 = \mathsf{N}_{21} \oplus \cdots \oplus \mathsf{N}_{2m} \oplus \mathsf{F}_2,$$

with F_1 and F_2 free, and with N_{1j} and N_{2j} cyclic of order r_j with $r_1 | \cdots | r_m$. Is it the case that $\mathsf{N}_{1j} = \mathsf{N}_{2j}$ for $j \in \{1, \dots, m\}$?

Chapter 5

Linear algebra

Linear algebra is one of the most important branches of mathematics as concerns its widespread applications. Indeed, the number of applications of mathematics which rely on linear algebra is quite astonishing. Moreover, the abstract viewpoint offered by linear algebra makes disparate topics look “the same,” and it is precisely to this that linear algebra owes much of its power. In this chapter we give a fairly thorough description of the algebraic aspects of linear algebra, paying special attention to both special cases of finite-dimensional linear algebra, and infinite-dimensional linear algebra. We begin in Sections 5.1, 5.2, and 5.3 by discussing matrices, which can be seen as the concrete side of linear algebra. Then, in Sections 5.4 and 5.5 we discuss abstract linear algebra. In Section 5.7 we encounter for the first time the important topic of duality. Particularly in Sections 5.1, 5.4, and 5.7 we give special consideration to understanding linear algebra where infinite-dimensional vector spaces are involved. The reader will begin to see here why infinite-dimensional and finite-dimensional linear algebra are actually quite different. In Section 5.8 we focus on some of the special structure of linear transformations of finite-dimensional vector spaces.

Do I need to read this chapter? The ideas in this chapter are integral to understanding the transform methods in Chapters IV-5–IV-7. More precisely, the material in Chapter III-6 is essential to understanding transform methods, and the material in Chapter III-6 is built on the material in this chapter. Also, our discussion in Chapter V-3 involves some deep knowledge of linear algebra, such as we develop in this chapter. Thus this chapter is one of the central background chapters in the volume, and should be comprehended to a large degree before moving on. Exception can be made for Sections 5.2 and 5.5 which are concerned with linear algebra over rings, and with the material in Section 5.8, all of which can be omitted until it is needed. •

Contents

Section 5.1

Matrices over fields

A matrix is a deceptively simple object when one is merely confronted with its definition. Its simplicity belies its breadth of use, and the comparative complexity of the operations that can be performed on and with matrices. In this section we give a comparatively simple presentation of matrices, one that more or less mirrors the standard presentation, albeit with perhaps more rigour and generality. In particular, we treat matrices with arbitrary numbers of rows and columns, the advantage of this being that in Theorem 5.4.21 we will be able to draw a very general, indeed the most general possible, connection between matrices and linear maps between arbitrary vector spaces. At present, however, the generality of arbitrary numbers of rows and columns may seem unnecessarily complicated. In Section 5.2 we shall add another layer of generality by considering matrices over rings. Parts of this more general development are similar to that for fields, so we suffer from some redundancy. However, to complicate the simpler development of this section, which is all that will be needed by many readers, would be poor pedagogy.

Do I need to read this section? Readers having had a basic course in linear algebra where basic matrix manipulations are presented can perhaps bypass this section until some of the results are needed subsequently. Readers having had a more applied course in linear algebra where the basic ideas are presented, but perhaps not proved, might be interested in seeing following some of the proofs concerning topics like row and column rank and row reduction. •

5.1.1 Matrices over fields: definitions and notation

Let us begin with the definition of matrix with entries in a field. In this section and in the next we shall use the symbol I for an index set, as it is extremely convenient to do so. We shall also use the letter i as an index. This is at odds with our use in Chapter 2 of the symbol I to exclusively stand for an interval, and our use in Section 4.7 and Chapter II-3 of the letter i to represent $\sqrt{-1}$ (although this latter notational imperfection is less alarming because of the font difference).

5.1.1 Definition (Matrix over a field) Let F be a field and let I and J be index sets. A ***matrix over*** F in $I \times J$ is a map $A\colon I \times J \to \mathsf{F}$. The expression $A(i, j)$, $i \in I$, $j \in J$, is the ***(i, j)th component*** of the matrix A, and is said to lie in the ***ith row*** and the ***jth column*** of A. If, for each $i_0 \in I$ the set

$$\{j \in J \mid A(i_0, j) \neq 0_{\mathsf{F}}\}$$

is finite, then $\boldsymbol{A}$ is ***row finite***, and, if for each $j_0 \in J$ the set

$$\{i \in I \mid \boldsymbol{A}(i, j_0) \neq 0_{\mathsf{F}}\}$$

is finite, then $\boldsymbol{A}$ is ***column finite***.

The set of matrices over F in $I \times J$ is denoted by $\mathrm{Mat}_{I\times J}(\mathsf{F})$. If $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$, then a matrix over F in $I \times J$ is an $\mathbf{m \times n}$ ***matrix***, and the set of $m \times n$ matrices is denoted by $\mathrm{Mat}_{m\times n}(\mathsf{F})$. •

The case that will be by far the most interesting for us will be the case when I and J are finite. In this case it is customary to represent an $m \times n$ matrix as an array of elements of F:

$$\boldsymbol{A} = \begin{bmatrix} A(1,1) & \cdots & A(1,n) \\ \vdots & \ddots & \vdots \\ A(m,1) & \cdots & A(m,n) \end{bmatrix}. \tag{5.1}$$

Also, the most commonly encountered matrices will have components from $\mathbb{R}$ or $\mathbb{C}$. However, we shall have occasion to use the more general notion of a matrix whose components lie in a polynomial ring, and this discussion we postpone until Section 5.2.

Let us give some specific examples of matrices.

5.1.2 Examples (Matrices over fields)

1. For general index sets I and J, the ***zero matrix*** is the element $\boldsymbol{0}_{I\times J}$ of $\mathrm{Mat}_{I\times J}(\mathsf{F})$ defined by $\boldsymbol{0}_{I\times J}(i, j) = 0_{\mathsf{F}}$ for each $(i, j) \in I \times J$. If $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$ then we denote $\boldsymbol{0}_{m\times n} = \boldsymbol{0}_{I\times J}$.
2. A ***square*** matrix is a matrix in $I \times I$ for some index set I.
3. A square matrix $\boldsymbol{A}\colon I \times I \to \mathsf{F}$ is ***diagonal*** if $\boldsymbol{A}(i_1, i_2) = 0_{\mathsf{F}}$ whenever $i_1 \neq i_2$.
4. The diagonal matrix $\boldsymbol{I}_I\colon I \to \mathsf{F}$ defined by
$$\boldsymbol{I}_I(i_1, i_2) = \begin{cases} 1_{\mathsf{F}}, & i_1 = i_2, \\ 0_{\mathsf{F}}, & i_1 \neq i_2 \end{cases}$$
is called the ***identity matrix***. If $I = \{i, \dots, n\}$ then we denote $\boldsymbol{I}_n = \boldsymbol{I}_I$. Note that the identity matrix is both row and column finite.
5. A square matrix $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ is ***upper triangular*** if $\boldsymbol{A}(i_1, i_2) = 0_{\mathsf{F}}$ for $i_1 > i_2$ and is ***lower triangular*** if $\boldsymbol{A}(i_1, i_2) = 0_{\mathsf{F}}$ for $i_1 < i_2$. •

A useful, but simple, operation on matrices involves "swapping" rows and columns.

5.1.3 Definition (Matrix transpose) Let F be a field, let I and J be index sets, and let $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$. The ***transpose*** of $\boldsymbol{A}$ is the matrix $\boldsymbol{A}^T \in \mathrm{Mat}_{J\times I}(\mathsf{F})$ defined by $\boldsymbol{A}^T(j,i) = \boldsymbol{A}(i,j)$, $i \in I$, $j \in J$. •

A nice linear algebraic interpretation of the transpose comes when one studies duality, as we shall see in Theorem 5.7.22. One can easily verify that the transpose satisfies the following equality:

$$(\boldsymbol{A}^T)^T = \boldsymbol{A}.$$

Sometimes it is useful to partition the sets of row and column indices as, say

$$I = \overset{\circ}{\bigcup_{a\in A}} I_a, \quad J = \overset{\circ}{\bigcup_{b\in B}} J_b.$$

In this case, if $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$, then for each $(a,b) \in A \times B$ we can define $\boldsymbol{A}_{ab} \in \mathrm{Mat}_{I_a\times J_b}(\mathsf{F})$ by $\boldsymbol{A}_{ab}(i,j) = \boldsymbol{A}(i,j)$ for $(i,j) \in I_a \times J_b$. This collection of matrices associated to $\boldsymbol{A}$ and the partitions of I and J is called a ***partition*** of $\boldsymbol{A}$. If the sets A and B that index the partition of rows and columns, respectively, are countable and well ordered as

$$a_1 < a_2 < a_3 < \cdots, \quad b_1 < b_2 < b_3 < \cdots,$$

then it is convenient to represent the partition of $\boldsymbol{A}$ as

$$\boldsymbol{A} = \left[\begin{array}{c|c|c|c} \boldsymbol{A}_{a_1b_1} & \boldsymbol{A}_{a_1b_2} & \boldsymbol{A}_{a_1b_3} & \cdots \\ \hline \boldsymbol{A}_{a_2b_1} & \boldsymbol{A}_{a_2b_2} & \boldsymbol{A}_{a_2b_3} & \cdots \\ \hline \boldsymbol{A}_{a_3b_1} & \boldsymbol{A}_{a_3b_2} & \boldsymbol{A}_{a_3b_3} & \cdots \\ \hline \vdots & \vdots & \vdots & \ddots \end{array}\right].$$

In most cases we will encounter the sets A and B indexing the partitions will be finite, even though the partition itself may be comprised of infinite sets. This looks somewhat simpler in the case of finite index sets for the rows and columns, where one can use the natural order on the a finite index set obtained by enumerating it. Here it is convenient to use the representation (5.1) of $\boldsymbol{A}$ as an array of elements of F. Doing so allows us to denote a partition of $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$ by

$$\boldsymbol{A} = \left[\begin{array}{c|c|c} \boldsymbol{A}_{11} & \cdots & \boldsymbol{A}_{1s} \\ \hline \vdots & \ddots & \vdots \\ \hline \boldsymbol{A}_{r1} & \cdots & \boldsymbol{A}_{rs} \end{array}\right],$$

where $\boldsymbol{A}_{ab} \in \mathrm{Mat}_{m_a\times n_b}(\mathsf{F})$, $a \in \{1,\dots,r\}$, $b \in \{1,\dots,s\}$. Therefore, we must have

$$\sum_{a=1}^{r} m_a = m, \quad \sum_{b=1}^{s} n_b = n.$$

A special case of this partitioning for square matrices will be of particular interest. Thus we let $A \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ and let $n_1, \dots, n_k \in \mathbb{Z}_{>0}$ sum to n. If

$$A = \left[\begin{array}{c|c|c|c} A_1 & 0_{n_1\times n_2} & \cdots & 0_{n_1\times n_k} \\ \hline 0_{n_2\times n_1} & A_2 & \cdots & 0_{n_2\times n_k} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline 0_{n_k\times n_1} & 0_{n_k\times n_2} & \cdots & A_k \end{array}\right]$$

for $A_j \in \mathrm{Mat}_{n_j\times n_j}(\mathsf{F})$, then we say that A is ***block diagonal***. We will sometimes write

$$A = \mathrm{diag}(A_1, \dots, A_k) \tag{5.2}$$

in this case. Note that a diagonal matrix is a special case of a block diagonal matrix with $n_1 = \cdots = n_k = 1$.

Another specific partition is by rows and columns of A. Thus we might write A in such a way as to distinguish its columns by

$$A = \left[\begin{array}{c|c|c} c_1 & \cdots & c_n \end{array}\right]$$

for $c_j \in \mathrm{Mat}_{m\times 1}(\mathsf{F})$, $j \in \{1, \dots, n\}$. To distinguish the rows of A we may write

$$A = \left[\begin{array}{c} r_1 \\ \hline \vdots \\ \hline r_m \end{array}\right]$$

for $r_i \in \mathrm{Mat}_{1\times n}(\mathsf{F})$, $i \in \{1, \dots, m\}$.

We have been a little vague about how one should think about the rows and columns of a matrix, i.e., what set do columns and rows live in? It will be useful to be clear about this, so let us do this. Above, when in the finite index set case we wrote a matrix as partitioned into its rows and columns, we tacitly thought of columns as being themselves matrices with one column, and rows as themselves being matrices with one row. However, it is often most useful to think of rows and columns as being vectors. Let us do this precisely, establishing the notation for this at the same time.

5.1.4 Definition (Column and row vectors) Let F be a field, let I and J be index sets, and let $A \in \mathrm{Mat}_{I\times J}(\mathsf{F})$. For $i \in I$ define $r(A, i) \in \mathsf{F}^J$ and for $j \in J$ define $c(A, j) \in \mathsf{F}^I$ by

$$r(A, i)(j) = A(i, j), \quad c(A, j)(i) = A(i, j),$$

respectively. The ***i*th *row vector*** of A is the element of F^J defined by $j \mapsto r(A, i)(j)$ and the ***j*th *column vector*** of A is the element of F^I defined by $i \mapsto c(A, j)(i)$. •

5.1.2 The algebra of matrices over fields

In this section we show how one can add and multiply matrices, and give some interesting properties of how these operations fit together. As we shall see, matrices give us nontrivial examples of vector spaces and algebras.

Next we indicate how to add and multiply matrices.

5.1.5 Definition (Sum and product of matrices over fields) Let F be a field and let I, J, and K be index sets.

(i) If $\boldsymbol{A}, \boldsymbol{B} \in \operatorname{Mat}_{I\times J}(\mathsf{F})$ then the ***sum*** of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{A} + \boldsymbol{B} \in \operatorname{Mat}_{I\times J}(\mathsf{F})$ defined by

$$(\boldsymbol{A} + \boldsymbol{B})(i, j) = \boldsymbol{A}(i, j) + \boldsymbol{B}(i, j).$$

(ii) If $\boldsymbol{A} \in \operatorname{Mat}_{I\times J}(\mathsf{F})$ and $\boldsymbol{B} \in \operatorname{Mat}_{J\times K}(\mathsf{F})$ then the ***product*** of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{AB} \in \operatorname{Mat}_{I\times K}(\mathsf{F})$ defined by

$$(\boldsymbol{AB})(i, k) = \sum_{j\in J} \boldsymbol{A}(i, j)\boldsymbol{B}(j, k),$$

and is defined whenever the sum is finite.

(iii) If $\boldsymbol{A} \in \operatorname{Mat}_{I\times J}(\mathsf{F})$ and $a \in \mathsf{F}$ then ***multiplication*** of $\boldsymbol{A}$ by a is the matrix $a\boldsymbol{A} \in \operatorname{Mat}_{I\times J}(\mathsf{F})$ defined by $(a\boldsymbol{A})(i, j) = a(\boldsymbol{A}(i, j))$. •

Note that the product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is always defined when $I = \{1, \dots, m\}$, $J = \{1, \dots, n\}$, and $K = \{1, \dots, r\}$. In this case we obtain the usual matrix product with which the majority of readers will be familiar. The following result gives more general conditions under which the product can be defined. We ask the reader to prove this result, and explore some related matters, in Exercise 5.1.1.

5.1.6 Proposition (Definability of the product of matrices over fields) *If F is a field and if I, J, and K are index sets, then the following statements hold for $\mathbf{A} \in \operatorname{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ and $\mathbf{B} \in \operatorname{Mat}_{\mathrm{J}\times\mathrm{K}}(\mathsf{F})$:*

(i) the product $\mathbf{AB}$ is defined if $\mathbf{A}$ is row finite;

(ii) the product $\mathbf{AB}$ is defined if $\mathbf{B}$ is column finite.

Moreover, if both $\mathbf{A}$ and $\mathbf{B}$ are column (resp. row) finite, then $\mathbf{AB}$ is column (resp. row) finite.

The sum and product of matrices have the following properties.

5.1.7 Proposition (Properties of sum and product of matrices over fields) *Let F be a field, let I, J, K, and L be index sets, and let $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \in \operatorname{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$, $\mathbf{B}_1, \mathbf{B}_2 \in \operatorname{Mat}_{\mathrm{J}\times\mathrm{K}}(\mathsf{F})$, $\mathbf{C}_1 \in \operatorname{Mat}_{\mathrm{K}\times\mathrm{L}}(\mathsf{F})$, and $\mathrm{a}_1, \mathrm{a}_2 \in \mathsf{F}$. Then the following equalities hold:*

(i) $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{A}_2 + \mathbf{A}_1$;

(ii) $(\mathbf{A}_1 + \mathbf{A}_2) + \mathbf{A}_3 = \mathbf{A}_1 + (\mathbf{A}_2 + \mathbf{A}_3)$;

(iii) $\mathbf{A}_1 + \mathbf{0}_{\mathrm{I}\times\mathrm{J}} = \mathbf{A}_1$;

(iv) if $-\mathbf{A}_1 \in \operatorname{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ is defined by $(-\mathbf{A}_1)(\mathrm{i}, \mathrm{j}) = -(\mathbf{A}_1(\mathrm{i}, \mathrm{j}))$, $\mathrm{i} \in \mathrm{I}$, $\mathrm{j} \in \mathrm{J}$, then $\mathbf{A}_1 + (-\mathbf{A}_1) = \mathbf{0}_{\mathrm{I}\times\mathrm{J}}$;

(v) if $\mathbf{A}_1$ is row finite, or if $\mathbf{B}_1$ and $\mathbf{B}_2$ are column finite, then $\mathbf{A}_1(\mathbf{B}_1 + \mathbf{B}_2) = \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_1\mathbf{B}_2$;

(vi) if $\mathbf{A}_1$ and $\mathbf{A}_2$ are row finite, or if $\mathbf{B}_1$ is column finite, then $(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{B}_1 = \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_1$;

(vii) if $\mathbf{A}_1$ and $\mathbf{B}_1$ are row finite, or if $\mathbf{B}_1$ and $\mathbf{C}_1$ are column finite, then $(\mathbf{A}_1\mathbf{B}_1)\mathbf{C}_1 = \mathbf{A}_1(\mathbf{B}_1\mathbf{C}_1)$;

(viii) $\mathbf{I}_J\mathbf{A}_1 = \mathbf{A}_1\mathbf{I}_I = \mathbf{A}_1$;

(ix) $a_1(a_2\mathbf{A}_1) = (a_1a_2)\mathbf{A}_1$;

(x) $(a_1 + a_2)\mathbf{A}_1 = a_1\mathbf{A}_1 + a_2\mathbf{A}_1$;

(xi) $a_1(\mathbf{A}_1 + \mathbf{A}_2) = a_1\mathbf{A}_1 + a_1\mathbf{A}_2$;

(xii) $1_F\mathbf{A}_1 = \mathbf{A}_1$.

Proof This is Exercise 5.1.2. ■

The immediately gives the following result concerning the structure of sets of matrices.

5.1.8 Corollary (Matrices over fields as elements of a vector space) *If* F *is a field and if* I *and* J *are index sets, then* $\mathrm{Mat}_{I\times J}(F)$ *is an* F*-vector space with addition given by the sum of matrices and with scalar multiplication being given by the multiplication of a matrix by a scalar.*

Of special interest is the case for matrices in $\mathrm{Mat}_{I\times I}(F)$. In this case there is useful additional structure.

5.1.9 Corollary (Matrices over fields as elements of an algebra) *If* F *is a field and if* I *is an index set, then the set of column finite matrices in* $\mathrm{Mat}_{I\times I}(F)$ *is an* F*-algebra with the vector space structure of Corollary 5.1.8 and with the product given by the product of matrices. Moreover,* $\mathbf{I}_I$ *is a unity element for the ring structure of* $\mathrm{Mat}_{I\times I}(F)$.

The transpose interacts with the algebraic operations on matrices in the following manner.

5.1.10 Proposition (Matrix algebra and matrix transpose) *Let* F *be a field, let* I *and* J *be index sets, and let* $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{I\times J}(F)$. *Then the following statements hold:*

(i) $(\mathbf{A}_1 + \mathbf{A}_2)^T = \mathbf{A}_1^T + \mathbf{A}_2^T$;

(ii) the product $\mathbf{A}_1\mathbf{A}_2$ *is defined if and only if the product* $\mathbf{A}_2^T\mathbf{A}_1^T$ *is defined, and when these are defined we have* $(\mathbf{A}_1\mathbf{A}_2)^T = \mathbf{A}_2^T\mathbf{A}_1^T$.

Proof This is Exercise 5.1.7. ■

5.1.3 Matrices as linear maps

An important interpretation of a matrix is as a linear map between vector spaces. We shall establish the generality of this interpretation for vector spaces in Theorem 5.4.21. For the moment we merely indicate how a matrix can be regarded as a linear map of certain vector spaces. We do this for general row and column index sets, although most of our interest will be in the case of finite index sets.

Let F be a field and let I be an index set. Following Definition 1.3.1, we denote by F^I the set of maps from a set I to F. As in Definition 4.5.39, and following Notation 4.5.44, we think of F^I as the direct product of I copies of F, which is an

F-vector space. The direct sum of I copies of F (see Example 4.5.43) we denote by F_0^I. If $I = \{1, \dots, n\}$ (the case of most interest to us) then we furthermore have

$$\mathsf{F}^I = \mathsf{F}_0^I = \mathsf{F}^n.$$

With this notation we have the following definition.

5.1.11 Definition (Matrix-vector product) Let F be a field and let I and J be index sets. For a matrix $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ and $\boldsymbol{x} \in \mathsf{F}^J$, the ***product*** of $\boldsymbol{A}$ and $\boldsymbol{x}$ is the element $\boldsymbol{A}\boldsymbol{x}$ of F^I given by

$$\boldsymbol{A}\boldsymbol{x}(i) = \sum_{j\in J} \boldsymbol{A}(i,j)\boldsymbol{x}(j),$$

and is defined whenever the sum is finite. •

For matrices with finite numbers of rows and columns, the matrix-vector product is *always* defined. Specifically, if $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$ then the product $\boldsymbol{A}\boldsymbol{x}$ gives the matrix-vector product with which most readers are familiar. This is most easily visualised by writing elements of F^n as "column vectors" as follows:

$$\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{A}(1,1) & \cdots & \boldsymbol{A}(1,n) \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}(m,1) & \cdots & \boldsymbol{A}(m,n) \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(1) \\ \vdots \\ \boldsymbol{x}(n) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}(1,1)\boldsymbol{x}(1) + \cdots + \boldsymbol{A}(1,n)\boldsymbol{x}(n) \\ \vdots \\ \boldsymbol{A}(m,1)\boldsymbol{x}(1) + \cdots + \boldsymbol{A}(m,n)\boldsymbol{x}(n) \end{bmatrix}. \tag{5.3}$$

However, the case of arbitrary row and column index sets will be of interest to us, so let us consider sufficient conditions under which the matrix-vector product is defined.

5.1.12 Proposition (Definability and properties of the matrix-vector product) *Let F be a field, let I and J be index sets, and let $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ and $\mathbf{x} \in \mathsf{F}^{\mathrm{J}}$. Then the following statements hold:*

(i) if $\mathbf{x} \in \mathsf{F}_0^{\mathrm{J}}$ then $\mathbf{Ax}$ is defined;

(ii) if $\mathbf{x} \in \mathsf{F}_0^{\mathrm{J}}$ and if $\mathbf{A}$ is column finite then $\mathbf{Ax}$ is defined and is an element of $\mathsf{F}_0^{\mathrm{I}}$;

(iii) if $\mathbf{A}$ is row finite then $\mathbf{Ax}$ is defined.

Proof The definedness of the matrix-vector product in each case is a simple matter of checking that the sum involved in finite, and this follows easily in each case from the definition of the matrix-vector product. The additional conclusion in part (ii) that the matrix-vector product lies in F_0^I follows from the definition of the matrix-vector product and of column finiteness. ■

The following result gives an interpretation of the matrix-vector product that introduces an important object: a linear map associated to a column finite matrix. We also indicate the relationship between row finite matrices and linear maps.

5.1.13 Theorem (Matrices as linear maps) *Let* F *be a field, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$*. Then the following statements hold:*

(i) if $\mathbf{A}$ *is column finite, then the map* $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ *is an* F*-linear map from* $\mathsf{F}_0^{\mathrm{J}}$ *to* $\mathsf{F}_0^{\mathrm{I}}$*;*

(ii) if $\mathbf{A}$ *is row finite, then the map* $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ *is an* F*-linear map from* F^{J} *to* F^{I}*.*

Moreover, the following statements also hold:

(iii) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^{\mathrm{J}}; \mathsf{F}_0^{\mathrm{I}})$*, then there exists a unique column finite matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ *for all* $\mathbf{x} \in \mathsf{F}^{\mathrm{J}}$*;*

(iv) if J *is finite, then, if* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{J}}; \mathsf{F}^{\mathrm{I}})$*, then there exists a unique (necessarily row finite) matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ *for all* $\mathbf{x} \in \mathsf{F}_0^{\mathrm{J}}$*;*

(v) if J *is infinite, then there exists* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{J}}; \mathsf{F}^{\mathrm{I}})$ *for which there is no row finite matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ *for each* $\mathbf{x} \in \mathsf{F}^{\mathrm{J}}$*.*

Proof (i) Denote the map from F_0^J to F_0^I by L_A. That the map is well-defined and takes values in F_0^I is a consequence of part (ii) of Proposition 5.1.12. For $x, x_1, x_2 \in \mathsf{F}_0^J$ and for $a \in \mathsf{F}$ we have

$$\begin{aligned}\mathsf{L}_A(x_1 + x_2)(i) &= \sum_{j\in J} A(i,j)(x_1(j) + x_2(j)) = \sum_{j\in J} A(i,j)x_1(j) + \sum_{j\in J} A(i,j)x_2(j)\\ &= \mathsf{L}_A(x_1)(i) + \mathsf{L}_A(x_2)(i)\end{aligned}$$

and

$$\mathsf{L}_A(ax)(i) = \sum_{j\in J} A(i,j)(ax)(j) = a\sum_{j\in J} A(i,j)x(j) = a\mathsf{L}_A(x),$$

using the fact that all sums are finite, and using the vector space structure of F_0^I and F_0^J. This gives linearity of L_A.

(ii) This follows in the same manner as part (i).

(iii) Let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^J; \mathsf{F}_0^I)$. Let $\{e_i\}_{i\in I}$ and $\{f_j\}_{j\in J}$ be the standard bases for F_0^I and F_0^J, respectively. For $j \in J$ we have

$$\mathsf{L}(f_j) = a_{i_1 j}e_{i_1} + \cdots + a_{i_{k_j} j}e_{i_{k_j}} \tag{5.4}$$

for some unique $k_j \in \mathbb{Z}_{\geq 0}$, some unique basis elements $\{e_{i_1}, \dots, e_{i_{k_j}}\}$, and some unique nonzero $a_{i_1 j}, \dots, a_{i_{k_j} j} \in \mathsf{F}$ since $\{e_i\}_{i\in I}$ is a basis for F_0^I. We then define $A \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ by

$$A(i,j) = \begin{cases} a_{ij}, & i \in \{i_1, \dots, i_{k_j}\},\\ 0_{\mathsf{F}}, & \text{otherwise.}\end{cases}$$

It is now a straightforward matter to show that $\mathsf{L}(x) = Ax$. Indeed, if $x \in \mathsf{F}_0^J$ then we can write

$$x = x_1 f_{j_1} + \cdots + x_k f_{j_k}$$

for some $x_1, \dots, x_k \in \mathsf{F}$. Now let $e_{i_1}, \dots, e_{i_m}$ be standard basis elements with the property that

$$\mathsf{L}(f_{j_l}) \in \mathrm{span}_{\mathsf{F}}(e_{i_1}, \dots, e_{i_m})$$

for each $l \in \{1, \dots, k\}$. Then

$$\begin{aligned}\mathsf{L}(\boldsymbol{x}) &= \mathsf{L}(x_1\boldsymbol{f}_{j_1} + \dots + x_k\boldsymbol{f}_{j_k})\\ &= x_1\mathsf{L}(\boldsymbol{f}_{j_1}) + \dots + x_k\mathsf{L}(\boldsymbol{f}_{j_k})\\ &= x_1 a_{i_1 j_1}\boldsymbol{e}_{i_1} + \dots + x_1 a_{i_m j_1}\boldsymbol{e}_{i_m} + \dots + x_k a_{i_1 j_k}\boldsymbol{e}_{i_1} + \dots + x_k a_{i_m j_k}\boldsymbol{e}_{i_m}\\ &= \boldsymbol{A}\boldsymbol{x},\end{aligned}$$

as desired. The uniqueness of $\boldsymbol{A}$ follows from the fact that, for each $j \in J$, the choice of $\{\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_{k_j}}\}$ and $a_{i_1 j}, \dots, a_{i_{k_j} j}$ in (5.4) are unique.

(iv) Suppose that $J = \{1, \dots, n\}$ so that $\mathsf{F}^J = \mathsf{F}^n$, and let $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n\}$ be the standard basis for F^n. Let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^n; \mathsf{F}^I)$ and define $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ by asking that, for $(i, j) \in I \times J$, $\mathsf{L}(\boldsymbol{f}_j)(i) = \boldsymbol{A}(i, j)$. We claim that $\mathsf{L}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathsf{F}^n$. Indeed, for $\boldsymbol{x} \in \mathsf{F}^n$ write

$$\boldsymbol{x} = \boldsymbol{x}(1)\boldsymbol{f}_1 + \dots + \boldsymbol{x}(n)\boldsymbol{f}_n,$$

and then compute, for $i \in I$,

$$\begin{aligned}\mathsf{L}(\boldsymbol{x})(i) &= \mathsf{L}(\boldsymbol{x}(1)\boldsymbol{f}_1 + \dots + \boldsymbol{x}(n)\boldsymbol{f}_n)(i)\\ &= \boldsymbol{x}(1)\mathsf{L}(\boldsymbol{f}_1)(i) + \dots + \boldsymbol{x}(n)\mathsf{L}(\boldsymbol{f}_n)(i)\\ &= \boldsymbol{A}(i, 1)\boldsymbol{x}(1) + \dots + \boldsymbol{A}(i, n)\boldsymbol{x}(n) = (\boldsymbol{A}\boldsymbol{x})(i),\end{aligned}$$

as desired.

(v) We let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^J; \mathsf{F}^I)$, and we again let $\{\boldsymbol{f}_j\}_{j\in J}$ be the standard basis for F_0^J. Note that this is not a basis for F^J, but nonetheless is a linearly independent subset. Now define $\boldsymbol{x}_0 \in \mathsf{F}^J$ by $\boldsymbol{x}_0(j) = 1_{\mathsf{F}}$ for $j \in J$. If J is infinite then the set $\{\boldsymbol{f}_j\}_{j\in J} \cup \{\boldsymbol{x}_0\}$ is linearly independent (why?). Therefore, by Theorem 4.5.26, there exists a basis $\mathscr{B}$ for F^J such that $\{\boldsymbol{f}_j\}_{j\in J} \cup \{\boldsymbol{x}_0\} \subseteq \mathscr{B}$. Define $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^J; \mathsf{F}^I)$ by asking that $\mathsf{L}(\boldsymbol{x}_0) = \boldsymbol{y}_0$ for some nonzero $\boldsymbol{y}_0 \in \mathsf{F}^I$, and that $\mathsf{L}(\boldsymbol{u}) = \boldsymbol{0}_{\mathsf{F}^I}$ for $\boldsymbol{u} \in \mathscr{B} \setminus \{\boldsymbol{x}_0\}$. To then define $\mathsf{L}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathsf{F}^J$ we note that we can write

$$\boldsymbol{x} = c_1\boldsymbol{u}_1 + \dots + c_k\boldsymbol{u}_k$$

for some unique $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k \in \mathscr{B}$ and nonzero $c_1, \dots, c_k \in \mathsf{F}$. We claim that there exists no row finite matrix $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ such that $\mathsf{L}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ for every $\boldsymbol{x} \in \mathsf{F}^J$. We demonstrate this by a contradiction. Suppose that $\boldsymbol{A}$ is a row finite matrix such that $\mathsf{L}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ for every $\boldsymbol{x} \in \mathsf{F}^J$. Then, for each $j \in J$, $\boldsymbol{A}\boldsymbol{f}_j = \boldsymbol{0}_{\mathsf{F}^I}$ by the definition of L. However, as can easily be seen by the definition of matrix-vector multiplication, $\boldsymbol{A}\boldsymbol{f}_j$ is exactly the jth column vector of $\boldsymbol{A}$. Thus all columns of $\boldsymbol{A}$ are zero, and so $\boldsymbol{A}$ is the zero matrix. But this contradicts the fact that L is nonzero. ■

Note the lack of general symmetry in the relationship between column finite matrices and elements of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^J; \mathsf{F}_0^I)$ (where there is an exact correspondence) and in the relationship between row finite matrices and elements of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^J; \mathsf{F}^I)$ (where a row finite matrix defines a homomorphism, but not necessarily the converse). This will become clear in Theorem 5.1.13 when we indicate exactly which linear maps from $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^J; \mathsf{F}^I)$ are characterised by row finite matrices. The issue, roughly speaking, is that the vector space F^J is "too big" to have all of linear maps characterised by row finite matrices.

5.1.14 Notation (Matrices as linear maps) Accepting an abuse of notation, we shall denote by $A \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^J; \mathsf{F}_0^I)$ (resp. $A \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^J; \mathsf{F}^I)$) the homomorphism associated to the column finite (resp. row finite) matrix $A\colon I \times J \to \mathsf{F}$. Also, we shall also write Ax instead of $A(x)$, even when we are thinking of A as a map. •

The next result relates the matrix product to its corresponding concept in terms of linear maps.

5.1.15 Proposition (Matrix product and composition of linear maps) *Let* F *be a field, let* I, J, *and* K *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *and* $\mathbf{B} \in \mathrm{Mat}_{\mathrm{J}\times\mathrm{K}}(\mathsf{F})$*. Then the following statements hold:*

(i) if $\mathbf{A}$ *and* $\mathbf{B}$ *are column finite then the matrix corresponding to the composition of the homomorphisms* $\mathbf{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^{\mathrm{J}}; \mathsf{F}_0^{\mathrm{I}})$ *and* $\mathbf{B} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^{\mathrm{K}}; \mathsf{F}_0^{\mathrm{J}})$ *is* $\mathbf{AB}$*;*

(ii) if $\mathbf{A}$ *and* $\mathbf{B}$ *are row finite then the matrix corresponding to the composition of the homomorphisms* $\mathbf{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{J}}; \mathsf{F}^{\mathrm{I}})$ *and* $\mathbf{B} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{K}}; \mathsf{F}^{\mathrm{J}})$ *is* $\mathbf{AB}$*.*

Proof (i) That AB is column finite follows from Proposition 5.1.6. Let $\{e_i\}_{i\in I}$, $\{f_j\}_{j\in J}$, and $\{g_k\}_{k\in K}$ be the standard bases for F_0^I, F_0^J, and F_0^K, respectively. We compute

$$A \circ B(g_k) = A\left(\sum_{j\in J} B(j,k) f_j\right) = \sum_{j\in J} B(j,k) A(f_j) = \sum_{j\in J}\sum_{i\in I} B(j,k)A(i,j)e_i,$$

where all sums are finite since A and B are column finite. This directly gives

$$A \circ B(g_k) = \sum_{i\in I} (AB)(i,k)e_i,$$

using the definition of matrix product. A reference to the proof of Theorem 5.1.13 shows that AB is the matrix associated to the homomorphism $A \circ B$, as desired.

(ii) That AB is row finite follows from Proposition 5.1.6. Let $z \in \mathsf{F}^K$ and for $i \in I$ compute

$$\begin{aligned}(A \circ B(x))(i) &= \sum_{j\in J} A(i,j)((Bx)(j)) = \sum_{j\in J} A(i,j)\left(\sum_{k\in K} B(j,k)x(k)\right)\\ &= \sum_{k\in K} (AB)(i,k)x(k) = (ABx)(i),\end{aligned}$$

giving $A \circ B(x) = ABx$, as desired. ■

Next we consider the character of the transpose of a matrix as a linear map. If $I = \{1,\dots,m\}$ and $J = \{1,\dots,n\}$, then, if $A \in \mathrm{Mat}\,A_{I\times J}(\mathsf{F})$, we can think of A^T as a linear map from F^m to F^n. For arbitrary index sets I and J, the story is more complicated since A^T is *not* a linear map from F_0^I to F_0^J, even if A is column finite, and so itself a linear map from F_0^J to F_0^I. However, the transpose *is* a still a homomorphism, as we shall now show. In order to relate the homomorphisms A and A^T in a revealing way, it is useful to observe that elements of F^I can be regarded as elements of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F})$. Indeed, to $\boldsymbol{y} \in \mathsf{F}_0^I$ we associate the element $\mathsf{L}_{\boldsymbol{y}} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F})$ defined by

$$\mathsf{L}_{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{i \in I} \boldsymbol{y}(i)\boldsymbol{x}(i),$$

the sum being finite. Following Theorem 5.1.13, one can think of $\mathsf{L}_{\boldsymbol{y}}$ as a matrix in $\mathrm{Mat}_{\{1\}\times I}(\mathsf{F})$ given by $\mathsf{L}_{\boldsymbol{y}}(1, i) = \boldsymbol{y}(i)$. We shall discuss this relationship between F^I and $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F})$ in more detail in Proposition 5.7.5. There we will see that, in fact, F^I and $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F})$ are isomorphic as F-vector spaces.

5.1.16 Theorem (Transpose as a linear map) *Let* F *be a field and let* I *and* J *be index sets. If* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *is column finite (and so defines a linear map from* $\mathsf{F}_0^{\mathrm{J}}$ *to* $\mathsf{F}_0^{\mathrm{I}}$ *by Theorem 5.1.13), then the map* $\mathbf{y} \mapsto \mathbf{A}^{\mathrm{T}}\mathbf{y}$ *is a linear map from* F^{I} *to* F^{J}*. Moreover, the relation*

$$\mathsf{L}_{\mathbf{A}^{\mathrm{T}}(\mathbf{y})}(\mathbf{x}) = \mathsf{L}_{\mathbf{y}}(\mathbf{A}\mathbf{x})$$

holds for every $\mathbf{x} \in \mathsf{F}_0^{\mathrm{J}}$ *and* $\mathbf{y} \in \mathsf{F}^{\mathrm{I}}$*.*

Proof Since $\boldsymbol{A}$ is column finite if and only if $\boldsymbol{A}^T$ is row finite, it follows from part (ii) of Theorem 5.1.13 that $\boldsymbol{A}^T$ defines an F-linear map from F^I to F^J, and the form of this linear map is just as stated. The second assertion in the theorem is simply the straightforward observation that both $\mathsf{L}_{\boldsymbol{A}^T(\boldsymbol{y})}(\boldsymbol{x})$ and $\mathsf{L}_{\boldsymbol{y}}(\boldsymbol{A}\boldsymbol{x})$ are given by

$$\sum_{i \in I} \sum_{j \in J} \boldsymbol{A}(i, j)\boldsymbol{x}(j)\boldsymbol{y}(i),$$

the sums both being finite since $\boldsymbol{x} \in \mathsf{F}_0^J$ and since $\boldsymbol{A}$ is column finite. ■

For readers only familiar with linear algebra in finite-dimensions, the lack of symmetry in the character of the linear maps $\boldsymbol{A}$ and $\boldsymbol{A}^T$ will seem strange. However, as we shall see in our development in Section 5.7.2, the lack of symmetry is explained by the fact that, for infinite-dimensional vector spaces, a vector space and its algebraic dual are not isomorphic. Indeed, this is one of the important distinctions that arises between finite- and infinite-dimensional linear algebra.

Let us continue our discussion of the transpose as a linear map by understanding some of its properties in relation to those of the linear map itself.

5.1.17 Proposition (Properties of transpose as a linear map) *Let* F *be a field, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *be column finite. Then the following statements hold:*

(i) $\mathbf{A}$ *is surjective if and only if* $\mathbf{A}^{\mathrm{T}}$ *is injective;*

(ii) $\mathbf{A}$ *is injective if and only if* $\mathbf{A}^{\mathrm{T}}$ *is surjective;*

(iii) $\mathbf{A}$ *is an isomorphism if and only if* $\mathbf{A}^{\mathrm{T}}$ *is an isomorphism.*

Proof For the proof we will rely on the facts, proved as Proposition 5.4.46, that a linear map $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ is injective (resp. surjective) if and only if there exists $\mathsf{M} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}; \mathsf{U})$ such that $\mathsf{M} \circ \mathsf{L} = \mathrm{id}_{\mathsf{U}}$ (resp. $\mathsf{L} \circ \mathsf{M} = \mathrm{id}_{\mathsf{V}}$). This is not an especially difficult thing to prove, but we postpone the proof until we are dealing with abstract linear maps, rather than matrices.

(i) Note that $\boldsymbol{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F}_0^I)$ is surjective if and only if there exists $\boldsymbol{B} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F}_0^I)$ such that $\boldsymbol{AB} = \boldsymbol{I}_I$, noting that $\boldsymbol{I}_I$, as an element of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^I; \mathsf{F}_0^I)$ is the identity map. Since $\boldsymbol{A}$ and $\boldsymbol{B}$ are both column finite matrices by Theorem 5.1.13, by Proposition 5.1.10 the product $\boldsymbol{B}^T\boldsymbol{A}^T$ is well-defined, and so we have

$$(\boldsymbol{AB})^T = \boldsymbol{I}_I^T \quad\Longleftrightarrow\quad \boldsymbol{B}^T\boldsymbol{A}^T = \boldsymbol{I}_I.$$

Note that $\boldsymbol{I}_I$, thought of as a row finite matrix, and so an element of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^I; \mathsf{F}^I)$, corresponds to the identity map. Thus $\boldsymbol{A}$ is surjective if and only $\boldsymbol{A}^T$ possesses a left inverse, i.e., if and only if $\boldsymbol{A}^T$ is injective, as per Proposition 1.3.9.

(ii) This follows along the same lines as part (i), except that $\boldsymbol{A}$ in injective if and only if there exists $\boldsymbol{B}$ such that $\boldsymbol{BA} = \boldsymbol{I}_I$.

(iii) This is an immediate consequence of parts (i) and (ii). ■

Every matrix in $I \times J$ defines two subspaces, one of F^I and one of F^J.

5.1.18 Definition (Columnspace and rowspace) Let F be a field, let I and J be index sets, and let $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$.

(i) The ***columnspace*** of $\boldsymbol{A}$ is the subspace of F^I generated by the column vectors of $\boldsymbol{A}$, and is denoted by $\mathrm{colspace}(\boldsymbol{A})$.

(ii) The ***rowspace*** of $\boldsymbol{A}$ is the subspace of F^J generated by the row vectors of $\boldsymbol{A}$, and is denoted by $\mathrm{rowspace}(\boldsymbol{A})$. •

The following characterisations of the columnspace and rowspace follow immediately from the definition of a matrix as a linear map, and we leave the proof to the reader as Exercise 5.1.8.

5.1.19 Proposition (Interpretation of columnspace and rowspace) *Let* F *be a field, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$. *Then the following statements hold:*

(i) if $\mathbf{A}$ *is column finite then* $\mathrm{colspace}(\mathbf{A}) = \mathrm{image}(\mathbf{A})$*;*

(ii) if $\mathbf{A}$ *is row finite then* $\mathrm{rowspace}(\mathbf{A}) = \mathrm{image}(\mathbf{A}^{\mathrm{T}})$*.*

5.1.4 Invertible matrices over fields

The notion of an invertible matrix, or equivalently an invertible linear map, is an important one, and we shall encounter this in various places, even in this section (see Theorems 5.1.33 and 5.1.42). In this section we simply introduce the notion of an invertible matrix, and give some of its more elementary properties.

When a linear map $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^J; \mathsf{F}_0^I)$ (or $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^J; \mathsf{F}_0^I)$) is invertible, we have the following result which mirrors similar conclusions for groups and rings (see Exercises 4.1.4 and 4.2.6).

5.1.20 Proposition (The inverse of an isomorphism is linear) *If* F *is a field, if* I *and* J *are index sets, and if* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{J}}_0; \mathsf{F}^{\mathrm{I}}_0)$ *(resp.* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{J}}; \mathsf{F}^{\mathrm{I}})$*) is an isomorphism, then the inverse of* L *is an element of* $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{I}}_0, \mathsf{F}^{\mathrm{J}}_0)$ *(resp.* $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{I}}, \mathsf{F}^{\mathrm{J}})$*).*

Proof For $y, y_1, y_2 \in \mathsf{F}^I_0$ (resp. $y, y_1, y_2 \in \mathsf{F}^I$) let $x = \mathsf{L}^{-1}(x)$, $x_1 = \mathsf{L}^{-1}(y_1)$, and $x_2 = \mathsf{L}^{-1}(y_2)$. Then compute

$$\mathsf{L}^{-1}(y_1 + y_2) = \mathsf{L}^{-1}(\mathsf{L}(x_1) + \mathsf{L}(x_2)) = \mathsf{L}^{-1} \circ \mathsf{L}(x_1 + x_2) = x_1 + x_2 = \mathsf{L}^{-1}(y_1) + \mathsf{L}^{-1}(x_2)$$

and, for $a \in \mathsf{F}$, compute

$$\mathsf{L}^{-1}(ay) = \mathsf{L}^{-1}(a\mathsf{L}(x)) = \mathsf{L}^{-1} \circ \mathsf{L}(ax) = ax = a\mathsf{L}^{-1}(y),$$

which gives the result. ■

If $\boldsymbol{A} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ is column finite and is an isomorphism of F^I_0, it follows from Theorem 5.1.13 that the inverse of the linear map $\boldsymbol{A}$ has associated with it a column finite matrix. However, if $\boldsymbol{A} \in \mathrm{Hom}_{I\times I}(\mathsf{F})$ is row finite and an isomorphism of F^I, it may not be the case that the inverse of this isomorphism is represented by a row finite matrix (cf. part (v) of Theorem 5.1.13). However, it actually *is* in fact the case that this homomorphism is represented by a row finite matrix, and the following result demonstrates this.

5.1.21 Proposition (The inverse of a row finite matrix is a row finite matrix) *Let* F *be a field, let* I *be an index set, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{F})$ *be row finite and an isomorphism of* F^{I}*. Then there exists a row finite matrix* $\mathbf{A}^{-1} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{F})$ *such that* $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_{\mathrm{I}}$*.*

Proof Since $\boldsymbol{A}^T$ is column finite, it is to be regarded as a linear map from F^I_0 to F^I_0 by Theorem 5.1.13. Moreover, this linear map is invertible by Proposition 5.1.17. By Theorem 5.1.13, to the inverse of the linear map $\boldsymbol{A}^T$ we can associate a column finite matrix $\boldsymbol{B} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ such that $\boldsymbol{A}^T\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}^T = \boldsymbol{I}_I$, noting that $\boldsymbol{I}_I$ is the matrix associated with the identity map on F^I_0. By Proposition 5.1.10 we then have

$$(\boldsymbol{A}^T\boldsymbol{B})^T = \boldsymbol{B}^T\boldsymbol{A} = \boldsymbol{I}_I^T = \boldsymbol{I}_I$$

and

$$(\boldsymbol{B}\boldsymbol{A}^T)^T = \boldsymbol{A}\boldsymbol{B}^T = \boldsymbol{I}_I^T = \boldsymbol{I}_I.$$

The result follows by taking $\boldsymbol{A}^{-1} = \boldsymbol{B}^T$ since $\boldsymbol{B}^T$ is row finite. ■

Theorem 5.1.13 and Propositions 5.1.20 and 5.1.21 make possible the following definition.

5.1.22 Definition (Invertible matrix over a field) Let F be a field, let I be an index set, and let $\boldsymbol{A} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$.

(i) If $\boldsymbol{A}$ is column finite, then it is ***invertible*** if it is an isomorphism from F^I_0 to F^I_0.

(ii) If $\boldsymbol{A}$ is row finite, then it is ***invertible*** if it is an isomorphism from F^I to F^I.

The ***inverse*** of an invertible matrix $\boldsymbol{A}$ is the matrix $\boldsymbol{A}^{-1} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ associated to the inverse of the isomorphism from F^I_0 to F^I_0 (or from F^I to F^I) associated with $\boldsymbol{A}$. •

Here is a simple example of an invertible matrix. We shall encounter large classes of invertible matrices as we go along in this section.

5.1.23 Example (The identity matrix is invertible) Let F be a field and let I be an index set. One may verify that, the identity matrix $\boldsymbol{I}_I$ is invertible in both senses of Definition 5.1.22, since it is both row and column finite. Indeed, the linear map associated to it is simply the identity map. Its inverse is then also the identity matrix: $\boldsymbol{I}_I^{-1} = \boldsymbol{I}_I$. •

Let us next consider some of the simpler properties of invertible matrices.

5.1.24 Proposition (Properties of the matrix inverse) *Let* F *be a field, let* I *be an index set, and let* $\mathbf{A}, \mathbf{B} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{F})$*. Then the following statements hold:*

(i) if $\mathbf{A}$ *is invertible, then so is* $\mathbf{A}^{-1}$*, and* $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$*;*

(ii) $\mathbf{A}$ *is invertible if and only if* $\mathbf{A}^{\mathrm{T}}$ *is invertible and, if* $\mathbf{A}$ *is invertible, then* $(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}$*;*

(iii) if $\mathbf{A}$ *and* $\mathbf{B}$ *are invertible, then* $\mathbf{AB}$ *is invertible and* $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$*.*

Proof (i) We assume that $\boldsymbol{A}$ is column finite. The case where $\boldsymbol{A}$ is row finite follows in a similar manner. As a map of sets, we know that $\boldsymbol{A}\colon \mathsf{F}_0^I \to \mathsf{F}_0^I$ being invertible implies the existence of a unique map $\boldsymbol{A}^{-1}\colon \mathsf{F}_0^I \to \mathsf{F}_0^I$ such that $\boldsymbol{A} \circ \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \circ \boldsymbol{A} = \mathrm{id}_{\mathsf{F}_0^I}$ by Proposition 1.3.9. Therefore, by Proposition 5.1.20 we know that this set map $\boldsymbol{A}^{-1}$ must be exactly the linear map associated to the inverse matrix for $\boldsymbol{A}$. Note that $\boldsymbol{A} \circ \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \circ \boldsymbol{A} = \mathrm{id}_{\mathsf{F}_0^I}$ then implies, by Proposition 1.3.9, that $\boldsymbol{A}^{-1}$ is invertible with inverse equal to $\boldsymbol{A}$.

(ii) That $\boldsymbol{A}$ is invertible if and only if $\boldsymbol{A}^T$ is invertible is part (iii) of Proposition 5.1.17. Since $\boldsymbol{I}_I$ is the matrix associated with both the identity map on both F^I and F_0^I, if $\boldsymbol{A}$ is invertible then its inverse matrix $\boldsymbol{A}^{-1}$ satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_I.$$

By Proposition 5.1.10 it makes sense to take the transpose of this equation to get

$$(\boldsymbol{A}^{-1})^T\boldsymbol{A}^T = \boldsymbol{A}^T(\boldsymbol{A}^{-1})^T = \boldsymbol{I}_I.$$

It follows from Proposition 1.3.9 that $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1}$.

(iii) We suppose that $\boldsymbol{A}$ and $\boldsymbol{B}$ are column finite. The case where they are row finite follows along the same lines. We note that, thinking of matrices as linear maps and using Proposition 5.1.15,

$$(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1})(\boldsymbol{A}\boldsymbol{B}) = (\boldsymbol{A}\boldsymbol{B})(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) = \mathrm{id}_{\mathsf{F}_0^I}.$$

By Proposition 1.3.9 this implies that $\boldsymbol{AB}$ is invertible with inverse $\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$. ■

5.1.25 Notation (Inverse of transpose) In cases where the equality $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$ makes sense (e.g., when $\boldsymbol{A}$ is column and row finite) then one often writes

$$\boldsymbol{A}^{-T} = (\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T.$$ •

5.1.5 Elementary operations and elementary matrices

In this section we deal exclusively with matrices in $\mathrm{Mat}_{m\times n}(\mathsf{F})$, i.e., matrices with finite numbers of rows and columns.

In this section we describe operations that can be performed on the rows and columns of matrices to produce new matrices. The best way to motivate these operations is via the use of systems of linear equations. We postpone a systematic discussion of this to Section 5.1.8. A reader who (justifiably) thinks that the row and column operations we discuss here come from thin air may wish to refer to the discussion in that section before proceeding. For now, we simply give the definition.

5.1.26 Definition (Elementary row operation) Let F be a field, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{F})$. The matrix $\boldsymbol{A}_2$ is obtained by an ***elementary row operation*** from $\boldsymbol{A}_1$ if one of the following hold:

(i) there exists distinct $i_1, i_2 \in \{1, \dots, n\}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & i \notin \{i_1, i_2\}, \\ \boldsymbol{A}_1(i_2,j), & i = i_1, \\ \boldsymbol{A}_1(i_1,j), & i = i_2, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree except that the i_1st and i_2nd rows are interchanged;

(ii) there exists $i_0 \in \{1, \dots, n\}$ and a nonzero $u \in \mathsf{F}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & i \neq i_0, \\ u\boldsymbol{A}_1(i,j), & i = i_0, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree, except that the i_0th row of $\boldsymbol{A}_2$ is the i_0th row of $\boldsymbol{A}_1$ multiplied by u;

(iii) there exists distinct $i_1, i_2 \in \{1, \dots, n\}$ and $a \in \mathsf{F}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & i \neq i_1, \\ \boldsymbol{A}_1(i,j) + a\boldsymbol{A}_1(i_2,j), & i = i_1, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree except that the i_1st row of $\boldsymbol{A}_2$ is obtained by adding a times the i_2nd row of $\boldsymbol{A}_1$ to the i_1st row of $\boldsymbol{A}_1$.

The matrix $\boldsymbol{A}_2$ is ***row equivalent*** to $\boldsymbol{A}_1$ if there exists $k \in \mathbb{Z}_{>0}$ and matrices $\boldsymbol{A}'_1, \dots, \boldsymbol{A}'_k \in \mathrm{Mat}_{m\times n}(\mathsf{F})$ such that $\boldsymbol{A}'_1 = \boldsymbol{A}_1$, $\boldsymbol{A}'_k = \boldsymbol{A}_2$, and $\boldsymbol{A}'_{j+1}$ is obtained by an elementary row operation from $\boldsymbol{A}'_j$ for each $j \in \{1, \dots, k-1\}$. •

We may analogously define operations on columns of a matrix. Let us record the definition formally for clarity.

5.1.27 Definition (Elementary column operation) Let F be a field, let $m, n \in \mathbb{Z}_{>0}$, and let $A_1, A_2 \in \mathrm{Mat}_{m\times n}(\mathsf{F})$. The matrix A_2 is obtained by an ***elementary column operation*** from A_1 if one of the following hold:

(i) there exists distinct $j_1, j_2 \in \{1, \dots, m\}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$A_2(i, j) = \begin{cases} A_1(i, j), & j \notin \{j_1, j_2\}, \\ A_1(i, j_2), & j = j_1, \\ A_1(i, j_1), & j = j_2, \end{cases}$$

i.e., A_1 and A_2 agree except that the j_1st and j_2nd columns are interchanged;

(ii) there exists $j_0 \in \{1, \dots, m\}$ and a nonzero $u \in \mathsf{F}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$A_2(i, j) = \begin{cases} A_1(i, j), & j \neq j_0, \\ uA_1(i, j), & j = j_0, \end{cases}$$

i.e., A_1 and A_2 agree, except that the j_0th column of A_2 is the j_0th column of A_1 multiplied by u;

(iii) there exists distinct $j_1, j_2 \in \{1, \dots, m\}$ and $a \in \mathsf{F}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$A_2(i, j) = \begin{cases} A_1(i, j), & j \neq j_1, \\ A_1(i, j) + aA_1(i, j_2), & j = j_1, \end{cases}$$

i.e., A_1 and A_2 agree except that the j_1st column of A_2 is obtained by adding a times the j_2nd column of A_1 to the j_1st column of A_1.

The matrix A_2 is ***column equivalent*** to A_1 if there exists $k \in \mathbb{Z}_{>0}$ and matrices $A'_1, \dots, A'_k \in \mathrm{Mat}_{m\times n}(\mathsf{F})$ such that $A'_1 = A_1$, $A'_k = A_2$, and A'_{j+1} is obtained by an elementary column operation from A'_j for each $j \in \{1, \dots, k-1\}$. •

We leave to the reader as Exercise 5.1.9 the verification of the following result.

5.1.28 Proposition (Row and column equivalence are equivalence relations) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and define relations* $\sim_{\mathrm{r}}$ *and* $\sim_{\mathrm{c}}$ *in* $\mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$ *by*

$$\begin{aligned} \mathbf{A}_1 \sim_{\mathrm{r}} \mathbf{A}_2 \quad &\iff \quad \mathbf{A}_1 \textit{ and } \mathbf{A}_2 \textit{ are row equivalent}, \\ \mathbf{A}_1 \sim_{\mathrm{c}} \mathbf{A}_2 \quad &\iff \quad \mathbf{A}_1 \textit{ and } \mathbf{A}_2 \textit{ are column equivalent}, \end{aligned}$$

respectively. Then $\sim_{\mathrm{r}}$ *and* $\sim_{\mathrm{c}}$ *are equivalence relations.*

The following result is more or less obvious, given the definition of the transpose of a matrix.

5.1.29 Proposition (Row and column operations and transpose) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row equivalent;*

(ii) $\mathbf{A}_1^T$ *and* $\mathbf{A}_2^T$ *are column equivalent.*

Associated to elementary row and column operations are particular matrices whose properties we now begin to explore.

5.1.30 Definition (Elementary row matrix, elementary column matrix) Let F be a field and let $n \in \mathbb{Z}_{>0}$. A matrix $A \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ is

(i) an ***elementary row matrix*** if A is obtained by an elementary row operation from I_n and is

(ii) an ***elementary column matrix*** if A is obtained by an elementary column operation from I_n. •

The following result then relates the notions of elementary row matrices, elementary column matrices, and matrix transpose.

5.1.31 Proposition (Elementary row and column operations and transpose) *If* F *is a field and if* $\mathrm{n} \in \mathbb{Z}_{>0}$*, then the following statements for* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{F})$ *are equivalent:*

(i) $\mathbf{A}$ *is an elementary row matrix;*

(ii) $\mathbf{A}$ *is an elementary column matrix;*

(iii) $\mathbf{A}^T$ *is an elementary row matrix;*

(iv) $\mathbf{A}^T$ *is an elementary column matrix.*

Proof In the proof it will be convenient to let $\boldsymbol{\varepsilon}(i, j) \in \mathrm{Mat}_{n\times n}(\mathsf{F})$, $i, j \in \{1, \dots, n\}$, be the matrix all of whose components are zero, except for the (i, j)th component which is 1_{F}. One can verify directly from the definition of matrix multiplication that the following lemma holds.

1 Lemma *For* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{F})$ *and*

(i) for $\mathrm{i}, \mathrm{j} \in \{1, \dots, \mathrm{n}\}$*,* $\boldsymbol{\varepsilon}(\mathrm{i}, \mathrm{j})\mathbf{A}$ *is the matrix all of whose rows are zero except the* i*th row, which is equal to the* j*th row of* $\mathbf{A}$*, and*

(ii) for $\mathrm{i}, \mathrm{j} \in \{1, \dots, \mathrm{n}\}$*,* $\mathbf{A}\boldsymbol{\varepsilon}(\mathrm{i}, \mathrm{j})$ *is the matrix all of whose columns are zero, except the* j*th column, which is equal to the* i*th column of* $\mathbf{A}$.

Using this notation, we can write down the elementary matrix associated with the three types of elementary row and column operations.

1. The swapping of rows i_1 and i_2: Here we have the elementary matrix

$$I_n - \boldsymbol{\varepsilon}(i_1, i_1) - \boldsymbol{\varepsilon}(i_2, i_2) + \boldsymbol{\varepsilon}(i_1, i_2) + \boldsymbol{\varepsilon}(i_2, i_1).$$

2. The swapping of the columns j_1 and j_2: Here we have the elementary matrix

$$I_n - \boldsymbol{\varepsilon}(j_1, j_1) - \boldsymbol{\varepsilon}(j_2, j_2) + \boldsymbol{\varepsilon}(j_1, j_2) + \boldsymbol{\varepsilon}(j_2, j_1).$$

3. The multiplication of row i_0 by u: Here we have the elementary matrix

$$\boldsymbol{I}_n - \boldsymbol{\varepsilon}(i_0, i_0) + u\boldsymbol{\varepsilon}(i_0, i_0).$$

4. The multiplication of column j_0 by u: Here we have the elementary matrix

$$\boldsymbol{I}_n - \boldsymbol{\varepsilon}(j_0, j_0) + u\boldsymbol{\varepsilon}(j_0, j_0).$$

5. Adding a times row i_2 to row i_1: Here the elementary matrix is

$$\boldsymbol{I}_n + a\boldsymbol{\varepsilon}(i_2, i_1).$$

6. Adding a times column j_2 to column j_1: The corresponding elementary matrix is

$$\boldsymbol{I}_n + a\boldsymbol{\varepsilon}(j_2, j_1).$$

The proof of the proposition is now a simple matter of verification, with the aid of Lemma 1, of the conclusions for each of the previous types of elementary matrices. We leave the trivial verification of this to the reader. ■

Since elementary row and elementary column matrices amount to the same thing, we shall simply call such matrices ***elementary matrices***.

The following result explains the value of elementary matrices in terms of row and column operations.

5.1.32 Proposition (Elementary matrices and elementary row and column operations) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*. Then the following statements hold:*

(i) if $\mathbf{A}_2$ *is obtained from* $\mathbf{A}_1$ *by an elementary row operation and if* $\mathbf{E}_\mathrm{m} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{m}}(\mathsf{F})$ *is the elementary matrix obtained by applying the same row operation to* $\mathbf{I}_\mathrm{m}$*, then* $\mathbf{A}_2 = \mathbf{E}_\mathrm{m}\mathbf{A}_1$*;*

(ii) if $\mathbf{A}_2$ *is obtained from* $\mathbf{A}_1$ *by an elementary column operation and if* $\mathbf{E}_\mathrm{n} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{F})$ *is the elementary matrix obtained by applying the same column operation to* $\mathbf{I}_\mathrm{n}$*, then* $\mathbf{A}_2 = \mathbf{A}_1\mathbf{E}_\mathrm{n}$*.*

Proof By Propositions 5.1.10 and 5.1.29, it suffices to prove the proposition for row operations.

We now consider the three types of row operations in succession, using the corresponding elementary matrices obtained during the course of the proof of Proposition 5.1.31.

1. Suppose that $\boldsymbol{A}_2$ is obtained from $\boldsymbol{A}_1$ by swapping the i_1st and i_2nd rows. Suppose that $\boldsymbol{E}_m$ is the matrix obtained by applying the same row operation to $\boldsymbol{I}_m$. Then

$$\boldsymbol{E}_m = \boldsymbol{I}_m - \boldsymbol{\varepsilon}(i_1, i_1) - \boldsymbol{\varepsilon}(i_2, i_2) + \boldsymbol{\varepsilon}(i_1, i_2) + \boldsymbol{\varepsilon}(i_2, i_1).$$

One then computes directly, using Lemma 1 from the proof of Proposition 5.1.31, that $\boldsymbol{A}_2 = \boldsymbol{E}_m\boldsymbol{A}_1$.

2. Suppose that A_2 is obtained from A_1 by multiplying the i_0th row of A_1 by a nonzero $u \in \mathsf{F}$, but leaving all other rows unchanged. Let E_m be the matrix obtained by applying the same row operation to I_m. Then we have

$$E_m = I_m - \varepsilon(i_0, i_0) + u\varepsilon(i_0, i_0).$$

An application of Lemma 1 from the proof of Proposition 5.1.31 gives $A_2 = E_m A_1$.

3. Suppose that A_2 agrees with A_1 except that the i_1st row of A_2 is obtained by adding a times the i_2nd row of A_1 to the i_1st row of A_1, and let E_m be the matrix obtained by applying the same row operation to I_m. Thus

$$E_m = I_m + a\varepsilon(i_2, i_1)$$

It then immediately from Lemma 1 of the proof of Proposition 5.1.31 that $A_2 = E_m A_1$.

This completes the proof. ■

We now establish an important link between invertible matrices and elementary matrices.

5.1.33 Theorem (Invertible matrices are products of elementary matrices and vice versa) *Let F be a field and let $n \in \mathbb{Z}_{>0}$. The following statements for $\mathbf{A} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ are equivalent:*

(i) $\mathbf{A}$ is invertible;

(ii) $\mathbf{A}$ is a product of a finite number of elementary matrices.

Proof We first prove a lemma.

1 Lemma *If $\mathbf{A}$ is an elementary matrix, then $\mathbf{A}$ is invertible, and its inverse is an elementary matrix.*

Proof We consider the three types of elementary row operations that may be used in forming A. We use the notation introduced in the proof of Proposition 5.1.32 of $\varepsilon(i, j) \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ for $i, j \in \{1, \dots, n\}$.

1. Suppose that A is obtained from I_n by swapping the i_1st and i_2nd rows. Then

$$A = I_n - \varepsilon(i_1, i_1) - \varepsilon(i_2, i_2) + \varepsilon(i_1, i_2) + \varepsilon(i_2, i_2).$$

If we define $B \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ by $B = A$ then one directly computes, using Lemma 1 from the proof of Proposition 5.1.31, that $AB = BA = I_n$. It is clear that B is an elementary matrix.

2. Suppose that A is obtained from I_n by multiplying the i_0th row of I_n by $u \in \mathsf{F}^*$, but leaving all other rows unchanged. In this case we have

$$A = I_n - \varepsilon(i_0, j_0) + u\varepsilon(i_0, i_0).$$

Then define

$$B = I_n - \varepsilon(i_0, j_0) + u^{-1}\varepsilon(i_0, i_0),$$

and check directly, using Lemma 1 from the proof of Proposition 5.1.31, that $AB = BA = I_n$. One can see that B is, as per Proposition 5.1.32, the elementary matrix corresponding to the elementary row operation of multiplying the i_0th row by u^{-1}.

3. Suppose that $\boldsymbol{A}$ agrees with $\boldsymbol{I}_n$ except that the i_1st row of $\boldsymbol{A}$ is obtained by adding r times the i_2nd row of $\boldsymbol{I}_n$ to the i_1st row of $\boldsymbol{I}_n$. As in the proof of Proposition 5.1.32 we have
$$\boldsymbol{A} = \boldsymbol{I}_n + r\boldsymbol{\varepsilon}(i_2, i_1).$$
By taking
$$\boldsymbol{B} = \boldsymbol{I}_n - r\boldsymbol{\varepsilon}(i_2, i_1),$$
one can check that $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}_n$, showing that $\boldsymbol{A}$ is invertible. Moreover, $\boldsymbol{B}$ is the elementary matrix corresponding to the elementary row operation of subtracting r times the i_2nd row from the i_1st row. ▼

We now proceed with the proof.

(i) $\Longrightarrow$ (ii) We first claim that, since $\boldsymbol{A}$ is invertible, the first column of $\boldsymbol{A}$ must have at least one nonzero element. Indeed, if the first column of $\boldsymbol{A}$ were comprised of all zeros, then $\boldsymbol{A}\boldsymbol{e}_1 = \boldsymbol{0}_{\mathsf{F}^n}$, implying that $\boldsymbol{A}$ is not injective by Exercise 4.5.23. So we may assume that $\boldsymbol{A}$ has a nonzero element in its first column. By an elementary row operation of swapping rows, arrive at a matrix $\boldsymbol{A}_1'$ whose $(1,1)$-component is nonzero. Now by the elementary row operation of multiplying the first row of $\boldsymbol{A}_1'$ by the inverse of the $(1,1)$-component, arrive at a matrix $\boldsymbol{A}_1''$ whose $(1,1)$-component is 1_{F}. Now, for each $j \in \{2, \dots, n\}$, perform the elementary row operation of subtracting from the jth row the product of the first row with the $(j,1)$-component of $\boldsymbol{A}_1''$. Upon doing these row operations one arrives at a matrix of the form
$$\left[\begin{array}{c|c} 1_{\mathsf{F}} & \boldsymbol{a}_{11} \\ \hline \boldsymbol{0}_{(n-1)\times 1} & \boldsymbol{A}_1 \end{array}\right] \tag{5.5}$$
for some $\boldsymbol{a}_{11} \in \mathrm{Mat}_{1\times(n-1)}(\mathsf{F})$ and for some $\boldsymbol{A}_1 \in \mathrm{Mat}_{(n-1)\times(n-1)}(\mathsf{F})$. Moreover, by Proposition 5.1.32, the matrix in (5.5) is the product of $\boldsymbol{A}$ with a finite number of elementary matrices. By the lemma above, an elementary matrix is invertible. Since the product of invertible matrices is invertible (see Proposition 5.1.24), it then follows that the matrix in (5.5) is invertible. We claim that this implies that $\boldsymbol{A}_1$ is invertible. Indeed, suppose that $\boldsymbol{A}_1$ is not invertible. Then by Exercise 4.5.23 and Corollary 5.4.44 there exists $\boldsymbol{x}_1 \in \mathsf{F}^{n-1} \setminus \{\boldsymbol{0}_{\mathsf{F}^{n-1}}\}$ such that $\boldsymbol{A}_1\boldsymbol{x}_1 = \boldsymbol{0}_{\mathsf{F}^{n-1}}$. But then, if we define $\boldsymbol{x} \in \mathsf{F}^n$ to have a zero first component with the remaining $n-1$ components equal to those of $\boldsymbol{x}_1$, it follows that $\boldsymbol{x}$ is nonzero and in the kernel of the matrix (5.5). Thus we conclude that $\boldsymbol{A}_1$ is invertible. We can then apply the above sequence of row operations to the invertible matrix $\boldsymbol{A}_1$, or equivalently to the last $n-1$ rows of the matrix (5.5), to arrive at a matrix in the form
$$\left[\begin{array}{c|c|c} 1_{\mathsf{F}} & a_{21} & \boldsymbol{a}_{21} \\ \hline 0_{\mathsf{F}} & 1_{\mathsf{F}} & \boldsymbol{a}_{22} \\ \hline \boldsymbol{0}_{(n-2)\times 1} & \boldsymbol{0}_{(n-2)\times 1} & \boldsymbol{A}_2 \end{array}\right]$$
for $a_{21} \in \mathsf{F}$, $\boldsymbol{a}_{21}, \boldsymbol{a}_{22} \in \mathrm{Mat}_{1\times(n-2)}(\mathsf{F})$, and $\boldsymbol{A}_2 \in \mathrm{Mat}_{(n-2)\times(n-2)}(\mathsf{F})$. This process can now be repeated $n-2$ more times to arrive at a matrix $\boldsymbol{B}$ whose diagonal entries are 1_{F} and whose entries below the diagonal are zero. Then, for $j \in \{1, \dots, n-1\}$ one performs the row operations of subtracting from the jth row the nth row multiplied by the (j,n)-component of $\boldsymbol{B}$. After these row operations, we arrive at a matrix whose last column is zero, except for the nth row which is 1_{F}. This process can be repeated for the column

j, $j \in \{2, \dots, n-1\}$, and what results is the $n \times n$ identity matrix. Thus we have shown that an invertible matrix can be transformed by a finite sequence of row operations to the $n \times n$ identity matrix. By Proposition 5.1.32 this means that

$$\boldsymbol{E}_1 \cdots \boldsymbol{E}_k \boldsymbol{A} = \boldsymbol{I}_n$$

for elementary matrices $\boldsymbol{E}_1, \dots, \boldsymbol{E}_k$. It, therefore, follows that, since elementary matrices are invertible,

$$\boldsymbol{A} = \boldsymbol{E}_k^{-1} \cdots \boldsymbol{E}_1^{-1}.$$

Since the inverse of an elementary matrix is also an elementary matrix by the lemma above, this then gives $\boldsymbol{A}$ as a product of elementary matrices.

(ii) $\implies$ (i) This follows since the product of two (and so any finite number, by induction) invertible matrices is invertible (see Proposition 5.1.24). ■

5.1.6 Rank and equivalence for matrices over fields

In this section, with the exception of the definition of equivalence in Definition 5.1.38 and of Example 5.1.44, we deal exclusively with matrices in $\mathrm{Mat}_{m\times n}(\mathsf{F})$ where F is a field.

The notion of rank is an important one in linear algebra. In this section we introduce two possible variants of the notion of rank, and show that they are, in fact, the same for matrices over fields. In Definition 5.4.1 we shall give another definition of rank, and will show that this one is also equivalent to the one's we give here.

We begin with two notions of rank.

5.1.34 Definition (Row rank and column rank for matrices over fields) Let F be a field, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$.

(i) The ***row rank*** of $\boldsymbol{A}$ is the dimension of the rowspace of $\boldsymbol{A}$.

(ii) The ***column rank*** of $\boldsymbol{A}$ is the dimension of the columnspace of $\boldsymbol{A}$. •

If one restricts to matrices with entries in a field, then it does in fact hold that row rank and column rank agree. The proof of the theorem is a little clumsy since it develops in an *ad hoc* way certain concepts that we will deal with more systematically in Sections 5.4 and 5.7.

5.1.35 Theorem (Row rank and column rank agree for matrices over fields) *If F is a field, if $m, n \in \mathbb{Z}_{>0}$, and if $\mathbf{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$, then $\dim_{\mathsf{F}}(\mathrm{image}(\mathbf{A})) = \dim_{\mathsf{F}}(\mathrm{image}(\mathbf{A}^T))$. In particular, the row rank and the column rank of $\mathbf{A}$ are the same.*

Proof Let us begin by proving a lemma.

1 Lemma *Let F be a field, let $m, n \in \mathbb{Z}_{>0}$, and let $\mathbf{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$. Then there exists a basis $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ for F^n such that*

(i) $\{\mathbf{A}\mathbf{u}_1, \dots, \mathbf{A}\mathbf{u}_r\}$ is a basis for $\mathrm{image}(\mathbf{A})$ and

(ii) $\{\mathbf{u}_{r+1}, \dots, \mathbf{u}_n\}$ is a basis for $\ker(\mathbf{A})$.

Proof Let $\{\boldsymbol{u}_{r+1},\dots,\boldsymbol{u}_n\}$ be any basis for $\ker(\boldsymbol{A})$, and by Theorem 4.5.26 extend this to a basis $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n\}$ for F^n. We claim that $\{\boldsymbol{A}\boldsymbol{u}_1,\dots,\boldsymbol{A}\boldsymbol{u}_r\}$ is a basis for $\operatorname{image}(\boldsymbol{A})$. To show that the set is linearly independent suppose that

$$c_1\boldsymbol{A}\boldsymbol{u}_1+\dots+c_r\boldsymbol{A}\boldsymbol{u}_r=\boldsymbol{0}_{\mathsf{F}^m}$$

for $c_1,\dots,c_r\in\mathsf{F}$. Then, by linearity,

$$\boldsymbol{A}(c_1\boldsymbol{u}_1+\dots+c_r\boldsymbol{u}_r)=\boldsymbol{0}_{\mathsf{F}^m},$$

implying that $c_1\boldsymbol{u}_1+\dots+c_r\boldsymbol{u}_r=\boldsymbol{0}_{\mathsf{F}^n}$ since $\mathsf{F}^n=\ker(\boldsymbol{A})\oplus\operatorname{span}_{\mathsf{F}}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_r)$. Therefore, $c_1=\dots=c_r=0_{\mathsf{F}}$, giving linear independence of $\{\boldsymbol{A}\boldsymbol{u}_1,\dots,\boldsymbol{A}\boldsymbol{u}_r\}$. Next suppose that $\boldsymbol{y}\in\operatorname{image}(\boldsymbol{A})$. Then there exists $\boldsymbol{x}\in\mathsf{F}^n$ such that $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$. Therefore, if

$$\boldsymbol{x}=c_1\boldsymbol{u}_1+\dots+c_n\boldsymbol{u}_n,$$

we have

$$\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}=\boldsymbol{A}(c_1\boldsymbol{u}_1+\dots+c_n\boldsymbol{u}_n)=c_1\boldsymbol{A}\boldsymbol{u}_1+\dots+c_r\boldsymbol{A}\boldsymbol{u}_r,$$

using the fact that $\boldsymbol{A}$ is linear and that $\boldsymbol{u}_{r+1},\dots,\boldsymbol{u}_n\in\ker(\boldsymbol{A})$. Thus $\operatorname{image}(\boldsymbol{A})=\operatorname{span}_{\mathsf{F}}(\boldsymbol{A}\boldsymbol{u}_1,\dots,\boldsymbol{A}\boldsymbol{u}_r)$. This gives the lemma. ▼

Now let $\{\boldsymbol{u}_1,\dots,\boldsymbol{u}_n\}$ be a basis as in the lemma, and, if necessary, extend $\{\boldsymbol{A}\boldsymbol{u}_{r+1},\dots,\boldsymbol{A}\boldsymbol{u}_n\}$ to a basis of F^m which we denote by

$$\{\boldsymbol{v}_1=\boldsymbol{A}\boldsymbol{u}_1,\dots,\boldsymbol{v}_r=\boldsymbol{A}\boldsymbol{u}_r,\boldsymbol{v}_{r+1},\dots,\boldsymbol{v}_n\}.$$

For $\boldsymbol{y}\in\mathsf{F}^m$ define a linear map $\mathsf{L}_{\boldsymbol{y}}\in\operatorname{Hom}(\mathsf{F}^m;\mathsf{F})$ by

$$\mathsf{L}_{\boldsymbol{y}}(\boldsymbol{z})=\boldsymbol{y}(1)\boldsymbol{z}(1)+\dots+\boldsymbol{y}(m)\boldsymbol{z}(m).$$

The following lemma shows that $\mathsf{L}_{\boldsymbol{y}}$ exactly determines $\boldsymbol{y}$.

2 Lemma *The map* $\mathbf{y}\mapsto\mathsf{L}_{\mathbf{y}}$ *is a bijection from* F^{m} *to* $\operatorname{Hom}(\mathsf{F}^{\mathrm{m}};\mathsf{F})$*.*

Proof Suppose that $\mathsf{L}_{\boldsymbol{y}_1}=\mathsf{L}_{\boldsymbol{y}_2}$. Then, if $\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_m\}$ is the standard basis, for each $j\in\{1,\dots,m\}$ we have

$$\mathsf{L}_{\boldsymbol{y}_1}(\boldsymbol{e}_j)=\mathsf{L}_{\boldsymbol{y}_2}(\boldsymbol{e}_j)\quad\Longrightarrow\quad\boldsymbol{y}_1(j)=\boldsymbol{y}_2(j).$$

Thus $\boldsymbol{y}_1=\boldsymbol{y}_2$, and so the map in question is surjective. Now let $\mathsf{L}\in\operatorname{Hom}(\mathsf{F}^m;\mathsf{F})$. By Theorem 5.1.13 there exists $\boldsymbol{A}\in\operatorname{Mat}_{1\times m}(\mathsf{F})$ such that $\boldsymbol{A}\boldsymbol{z}=\mathsf{L}(\boldsymbol{z})$ for every $\boldsymbol{z}\in\mathsf{F}^m$. If one takes $\boldsymbol{y}\in\mathsf{F}^m$ to be the sole row vector of $\boldsymbol{A}$, one directly verifies that $\boldsymbol{A}\boldsymbol{y}=\mathsf{L}_{\boldsymbol{y}}(\boldsymbol{z})$ for every $\boldsymbol{z}\in\mathsf{F}^m$, and so the map in question is surjective. ▼

Next we prove a lemma asserting the existence of homomorphisms of F^m and F having certain properties.

3 Lemma *If* $\{\mathbf{v}_1,\dots,\mathbf{v}_{\mathrm{m}}\}$ *is a basis for* F^{m} *then there exists* $\mathsf{L}_1,\dots,\mathsf{L}_{\mathrm{m}} \in \mathrm{Hom}(\mathsf{F}^{\mathrm{m}};\mathsf{F})$ *such that, for* $\mathrm{j},\mathrm{k} \in \{1,\dots,\mathrm{m}\}$*,*

$$\mathsf{L}_{\mathrm{j}}(\mathbf{v}_{\mathrm{k}}) = \begin{cases} 1_{\mathsf{F}}, & \mathrm{j} = \mathrm{k}, \\ 0_{\mathsf{F}}, & \mathrm{j} \neq \mathrm{k}. \end{cases}$$

Proof Fix $j \in \{1,\dots,m\}$. To define L_j we merely note that, since $\{v_1,\dots,v_m\}$ is a basis, for $\boldsymbol{y} \in \mathsf{F}^m$ we have

$$\boldsymbol{y} = c_1 v_1 + \dots + c_m v_m.$$

Therefore, since L_j is linear,

$$\mathsf{L}_j(\boldsymbol{y}) = c_1 \boldsymbol{L}_j v_1 + \dots + c_m \boldsymbol{L}_j v_m = c_j.$$

That $\mathsf{L}_1,\dots,\mathsf{L}_m$ have the desired property follows immediately. ▼

Next, for $j \in \{1,\dots,m\}$, let $\boldsymbol{v}_j^* \in \mathsf{F}^m$ have the property that

$$\mathsf{L}_{\boldsymbol{v}_j^*}(\boldsymbol{v}_k) = \begin{cases} 1_{\mathsf{F}}, & j = k, \\ 0_{\mathsf{F}}, & j \neq k, \end{cases}$$

this being possible by Lemmas 2 and 3. We claim that the set $\{\boldsymbol{v}_1^*,\dots,\boldsymbol{v}_m^*\}$ is linearly independent. Indeed, suppose that

$$c_1 \boldsymbol{v}_1^* + \dots + c_m \boldsymbol{v}_m^* = \boldsymbol{0}_{\mathsf{F}^m}.$$

Then we have

$$(c_1 \mathsf{L}_{\boldsymbol{v}_1^*} + \dots + c_m \mathsf{L}_{\boldsymbol{v}_m^*})(\boldsymbol{y}) = 0_{\mathsf{F}}$$

for every $\boldsymbol{y} \in \mathsf{F}^m$. For $j \in \{1,\dots,m\}$, one then directly computes

$$0_{\mathsf{F}} = c_1 \mathsf{L}_{\boldsymbol{v}_1^*}(\boldsymbol{v}_j) + \dots + c_m \mathsf{L}_{\boldsymbol{v}_m^*}(\boldsymbol{v}_j) = c_j,$$

giving linear independence as desired. Similarly define linearly independent vectors $\boldsymbol{u}_1^*,\dots,\boldsymbol{u}_n^* \in \mathsf{F}^n$ by

$$\mathsf{L}_{\boldsymbol{v}_j^*}(\boldsymbol{v}_k) = \begin{cases} 1_{\mathsf{F}}, & j = k, \\ 0_{\mathsf{F}}, & j \neq k. \end{cases}$$

Next, for $j \in \{1,\dots,r\}$ and for $k \in \{1,\dots,n\}$, compute

$$\begin{aligned} \mathsf{L}_{\boldsymbol{A}^T\boldsymbol{v}_j^*}(\boldsymbol{u}_k) &= \sum_{l=1}^{n} (\boldsymbol{A}^T\boldsymbol{v}_j^*)(l)\boldsymbol{u}_k(l) = \sum_{s=1}^{m}\sum_{l=1}^{n} \boldsymbol{A}(s,l)\boldsymbol{v}_j^*(s)\boldsymbol{u}_k(l) \\ &= \sum_{s=1}^{m} (\boldsymbol{A}\boldsymbol{u}_k)(s)\boldsymbol{v}_j^*(s) = \sum_{s=1}^{m} \boldsymbol{v}_k(s)\boldsymbol{v}_j^*(s) = \mathsf{L}_{\boldsymbol{v}_j^*}(\boldsymbol{v}_k) \\ &= \begin{cases} 1_{\mathsf{F}}, & j = k, \\ 0_{\mathsf{F}}, & j \neq k. \end{cases} \end{aligned}$$

A similar computation shows that $\mathsf{L}_{\boldsymbol{A}^T\boldsymbol{v}_j^*}(\boldsymbol{u}_k) = 0_{\mathsf{F}}$ for $k \in \{1,\dots,n\}$ and $j \in \{r+1,\dots,m\}$.

Using these computations we now show that $\{\boldsymbol{A}^T\boldsymbol{v}_1^*,\dots,\boldsymbol{A}^T\boldsymbol{v}_r^*\}$ is a basis for $\operatorname{image}(\boldsymbol{A}^T)$. To show linear independence suppose that

$$c_1\boldsymbol{A}^T\boldsymbol{v}_1^* + \cdots + c_r\boldsymbol{A}^T\boldsymbol{v}_r^* = \boldsymbol{0}_{\mathsf{F}^n}.$$

Then

$$\boldsymbol{A}^T(c_1\boldsymbol{v}_1^* + \cdots + c_r\boldsymbol{v}^*) = \boldsymbol{0}_{\mathsf{F}^n},$$

which gives $c_1 = \cdots = c_r = 0_{\mathsf{F}}$ since $\mathsf{F}^n = \ker(\boldsymbol{A}^T) \oplus \operatorname{span}_{\mathsf{F}}(\boldsymbol{v}_1^*,\dots,\boldsymbol{v}_r^*)$. Also, if $\boldsymbol{x} \in \operatorname{image}(\boldsymbol{A}^T)$, then there exists $\boldsymbol{y} \in \mathsf{F}^m$ such that $\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{x}$. Since $\{\boldsymbol{v}_1^*,\dots,\boldsymbol{v}_m^*\}$ is linearly independent, it is a basis for F^m, and so

$$\boldsymbol{y} = c_1\boldsymbol{v}_1^* + \cdots + c_m\boldsymbol{v}_m^*.$$

Therefore,

$$\boldsymbol{A}^T\boldsymbol{y} = c_1\boldsymbol{A}^T\boldsymbol{v}_1^* + \cdots + c_rA^T\boldsymbol{v}_r^*,$$

using linearity of A^T and the fact that $\boldsymbol{A}^T\boldsymbol{v}_j^* = \boldsymbol{0}_{\mathsf{F}^n}$ for $j \in \{r+1,\dots,m\}$. Thus $\{\boldsymbol{A}^T\boldsymbol{v}_1^*,\dots,\boldsymbol{A}^T\boldsymbol{v}_r^*\}$ spans $\operatorname{image}(\boldsymbol{A}^T)$. Thus $\dim_{\mathsf{F}}(\operatorname{image}(\boldsymbol{A}^T)) = r = \dim_{\mathsf{F}}(\operatorname{image}(\boldsymbol{A}))$.

That the row and column ranks of $\boldsymbol{A}$ agree follows from Proposition 5.1.19. ■

Based on the theorem we now make the following definition.

5.1.36 Definition (Rank of matrices over fields) If F is a field, if $m,n \in \mathbb{Z}_{>0}$, and if $\boldsymbol{A} \in \operatorname{Mat}_{m\times n}(\mathsf{F})$, then the ***rank*** of $\boldsymbol{A}$ is equal to the column or row rank of $\boldsymbol{A}$, and is denoted by $\operatorname{rank}(\boldsymbol{A})$. •

From Theorem 5.1.35 we have the following result.

5.1.37 Corollary (Rank and rank of transpose agree for matrices over fields) *If* F *is a field, if* $\mathrm{m,n} \in \mathbb{Z}_{>0}$, *and if* $\mathbf{A} \in \operatorname{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$, *then* $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^{\mathrm{T}})$.

Next we turn to the important notion of equivalence of matrices. This idea may seem somewhat contrived at the present time. However, we shall see in Section 5.4.7 that equivalence has a rather natural interpretation in terms of linear maps between vector spaces.

5.1.38 Definition (Equivalence of matrices over fields) Let F be a field, let I and J be index sets, and let $\boldsymbol{A}_1,\boldsymbol{A}_2 \in \operatorname{Mat}_{I\times J}(\mathsf{F})$ be column finite. The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are ***equivalent*** if there exists column finite invertible matrices $\boldsymbol{P} \in \operatorname{Mat}_{I\times I}(\mathsf{F})$ and $\boldsymbol{Q} \in \operatorname{Mat}_{J\times J}(\mathsf{F})$ such that $\boldsymbol{A}_2 = \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{Q}$. •

Of course, we have the following result, whose simple proof we leave as an exercise for the reader (Exercise 5.1.10).

5.1.39 Proposition (Equivalence of matrices is an equivalence relation) *If* F *is a field and if* $m, n \in \mathbb{Z}_{>0}$*, then the relation in* $\mathrm{Mat}_{m\times n}(\mathsf{F})$ *defined by*

$$\mathbf{A}_1 \sim \mathbf{A}_2 \qquad \Longleftrightarrow \qquad \mathbf{A}_1 \textit{ and } \mathbf{A}_2 \textit{ are equivalent}$$

is an equivalence relation.

An important part of our approach to understanding equivalence is the following property of rank in terms of elementary row and column operations.

5.1.40 Proposition (Elementary operations and rank) *Let* F *be a field, let* $m, n \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{F})$*. Then the following statements hold:*

(i) if $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row equivalent then* $\mathrm{rank}(\mathbf{A}_1) = \mathrm{rank}(\mathbf{A}_2)$*;*

(ii) if $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are column equivalent then* $\mathrm{rank}(\mathbf{A}_1) = \mathrm{rank}(\mathbf{A}_2)$*.*

Proof We shall only prove the first statement, the second following from Proposition 5.1.19 and Theorem 5.1.35.

It is clear that, if A_2 is obtained from A_1 by a single elementary row operation, then the row vectors of A_2 are contained in the rowspace of A_1. Moreover, since row equivalence is an equivalence relation by Proposition 5.1.28, it also holds that the row vectors of A_1 are contained in the rowspace of A_2. Thus $\mathrm{rank}(A_1) = \mathrm{rank}(A_2)$ if A_2 is obtained from A_1 by an elementary row operation. The result follows since row equivalence of A_1 and A_2 means that A_2 is obtained from A_1 by a finite sequence of elementary row operations. ■

The following result characterises equivalence for matrices over fields. For an extension of the following result to matrices with infinite rows and/or columns, we refer to Corollary 5.4.43.

5.1.41 Theorem (Characterisation of equivalence for matrices over fields) *Let* F *be a field, let* $m, n \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{F})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are equivalent;*

(ii) $\mathrm{rank}(\mathbf{A}_1) = \mathrm{rank}(\mathbf{A}_2)$*;*

(iii) there exists $r \in \mathbb{Z}_{\geq 0}$ *such that* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are equivalent to a matrix of the form*

$$\left[\begin{array}{c|c} \mathbf{I}_r & \mathbf{0}_{r\times(n-r)} \\ \hline \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{array}\right].$$

Moreover, the number r *in part (iii) is the rank of* $\mathbf{A}_1$ *and* $\mathbf{A}_2$*.*

Proof The following lemma is the key to our proof.

1 Lemma *If* F *is a field, if* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*, then* $\mathbf{A}$ *is equivalent to the matrix*

$$\left[\begin{array}{c|c} \mathbf{I}_{\mathrm{r}} & \mathbf{0}_{\mathrm{r}\times(\mathrm{n}-\mathrm{r})} \\ \hline \mathbf{0}_{(\mathrm{m}-\mathrm{r})\times\mathrm{r}} & \mathbf{0}_{(\mathrm{m}-\mathrm{r})\times(\mathrm{n}-\mathrm{r})} \end{array}\right],$$

where $\mathrm{r} = \mathrm{rank}(\mathbf{A})$.

Proof If $A = \mathbf{0}_{m\times n}$ then the result is immediate with $r = 0$. If $A \neq \mathbf{0}_{m\times n}$ then, by swapping rows and columns, arrive at a matrix A'_1 with the property that $A'_1(1,1) \neq 0_{\mathsf{F}}$. By multiplying the first row of A'_1 by $A'(1,1)^{-1}$ one arrives at a matrix A''_1 whose $(1,1)$-component is 1_{F}. Now, for $i \in \{2,\dots,m\}$, subtract $A''(i,1)$ times the first row of A''_1 from the ith row. The resulting matrix has zeros in the first column, except for the first row. Similarly perform elementary column operations to produce a matrix whose first row is zero, except for the first column. In summary, by a finite sequence of elementary row and column operations, we have arrived at a matrix of the form

$$\left[\begin{array}{c|c} 1_{\mathsf{F}} & \mathbf{0}_{1\times(n-1)} \\ \hline \mathbf{0}_{(m-1)\times 1} & A_1. \end{array}\right].$$

By Proposition 5.1.40 this matrix has rank r. Therefore, it must be the case that $\mathrm{rank}(A_1) = r-1$. If $r-1 = 0$ then $A_1 = \mathbf{0}_{(m-1)\times(n-1)}$ and the lemma is proved. Otherwise, one can continue the process performed on A on the matrix A_1 to arrive at a matrix of the form

$$\left[\begin{array}{c|c|c} 1_{\mathsf{F}} & 0_{\mathsf{F}} & \mathbf{0}_{1\times(n-2)} \\ \hline 0_{\mathsf{F}} & 1_{\mathsf{F}} & \mathbf{0}_{1\times(m-2)} \\ \hline \mathbf{0}_{(m-2)\times 1} & \mathbf{0}_{(m-2)\times 1} & A_2. \end{array}\right].$$

This process can be continued r times to arrive, after a finite sequence of elementary row and column operations, at a matrix in the desired form. The lemma follows from Propositions 5.1.32 and 5.1.33, along with the fact that the product of invertible matrices is invertible by virtue that the matrix product corresponds to composition of maps by Proposition 5.1.15. ▼

Now we can proceed easily with the proof of the theorem.

(i) $\Longrightarrow$ (ii) This follows from Lemma 1 since equivalence is an equivalence relation.

(ii) $\Longrightarrow$ (iii) This follows immediately from Lemma 1.

(iii) $\Longrightarrow$ (i) This follows since equivalence of matrices is an equivalence relation. ■

We next specialise the preceding result to the case of invertible matrices, and here we see that the additional structure of invertibility allows one to say more.

5.1.42 Theorem (Equivalence for invertible matrices) *Let* F *be a field, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{F})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}$ *is invertible;*

(ii) $\mathrm{rank}(\mathbf{A}) = \mathrm{n}$*;*

(iii) $\mathbf{A}$ *is equivalent to* $\mathbf{I}_{\mathrm{n}}$*;*

(iv) $\mathbf{A}$ *is row equivalent to* $\mathbf{I}_{\mathrm{n}}$*;*

(v) *there exists a unique invertible matrix* $\mathbf{P}$ *such that* $\mathbf{PA} = \mathbf{I}_{\mathrm{n}}$*;*

(vi) $\mathbf{A}$ *is column equivalent to* $\mathbf{I}_n$*;*

(vii) *there exists a unique invertible matrix* $\mathbf{Q}$ *such that* $\mathbf{AQ} = \mathbf{I}_n$*.*

Moreover, the matrices $\mathbf{P}$ *and* $\mathbf{Q}$ *in parts (v) and (vii) are equal.*

Proof (i) $\Longrightarrow$ (ii) If $\boldsymbol{A}$ is invertible then $\operatorname{rank}(\boldsymbol{A}) = n$ by the definition of rank and by Proposition 5.1.19.

(ii) $\Longrightarrow$ (iii) By Theorem 5.1.41, if $\operatorname{rank}(\boldsymbol{A}) = n$ then $\boldsymbol{A}$ can be transformed into $\boldsymbol{I}_n$ by a finite number of combined elementary row and column operations. However, by Proposition 5.1.32 and Theorem 5.1.33 this means that $\boldsymbol{I}_n = \boldsymbol{PAQ}$ for invertible matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$.

(iii) $\Longrightarrow$ (iv) We make use here of the reduced row echelon form introduced in Definition 5.1.45. By Theorem 5.1.41 we know that $\operatorname{rank}(\boldsymbol{A}) = n$. Therefore, by Propositions 5.1.40, 5.1.46, and Theorem 5.1.47, we know that $\boldsymbol{A}$ is row equivalent to a matrix in reduced row echelon form with n leading ones. The only such matrix is the $n \times n$ identity matrix.

(iv) $\Longrightarrow$ (v) If $\boldsymbol{A}$ can be transformed into $\boldsymbol{I}_n$ by a finite sequence of elementary row operations, then by Proposition 5.1.32 and Theorem 5.1.33 this means that $\boldsymbol{I}_n = \boldsymbol{PA}$ for some invertible matrix $\boldsymbol{P}$. Since $\boldsymbol{A}$ is row equivalent to $\boldsymbol{I}_n$, $\operatorname{rank}(\boldsymbol{A}) = n$ by Proposition 5.1.40. Thus $\boldsymbol{A}$ is invertible so that $\boldsymbol{P}$ is uniquely defined by the requirement that $\boldsymbol{PA} = \boldsymbol{I}_n$, since $\boldsymbol{P} = \boldsymbol{A}^{-1}$ in this event.

(v) $\Longrightarrow$ (vi) Since $\boldsymbol{PA} = \boldsymbol{I}_n$ with $\boldsymbol{P}$ invertible, $\boldsymbol{A} = \boldsymbol{P}^{-1}$ and so $\boldsymbol{A}$ is invertible by Proposition 5.1.24. Thus $\boldsymbol{A}$ has rank n by definition of rank and by Proposition 5.1.19. By Theorem 5.1.35 this means that $\boldsymbol{A}^T$ also has rank n. By the above proved implications, (ii) $\Longrightarrow$ (v), and therefore there exists an invertible matrix $\boldsymbol{P}$ such that $\boldsymbol{PA}^T = \boldsymbol{I}_n$. Taking the transpose of this equation and using Proposition 5.1.10 gives $\boldsymbol{AP}^T = \boldsymbol{I}_n$. Since $\boldsymbol{P}^T$ is invertible by Proposition 5.1.24 we conclude that (vi) holds.

(vi) $\Longrightarrow$ (vii) If $\boldsymbol{A}$ can be transformed into $\boldsymbol{I}_n$ by a finite sequence of elementary column operations, then by Proposition 5.1.32 and Theorem 5.1.33 this means that $\boldsymbol{I}_n = \boldsymbol{AQ}$ for some invertible matrix $\boldsymbol{Q}$. Since $\boldsymbol{A}$ is column equivalent to $\boldsymbol{I}_n$, $\operatorname{rank}(\boldsymbol{A}) = n$ by Proposition 5.1.40. Thus $\boldsymbol{A}$ is invertible, so uniqueness of $\boldsymbol{Q}$ follows since $\boldsymbol{Q} = \boldsymbol{A}^{-1}$.

(vii) $\Longrightarrow$ (i) If $\boldsymbol{AQ} = \boldsymbol{I}_n$ with $\boldsymbol{Q}$ invertible, then $\boldsymbol{A} = \boldsymbol{Q}^{-1}$. Since $\boldsymbol{Q}^{-1}$ is invertible by Proposition 5.1.24, it follows that $\boldsymbol{A}$ is invertible.

The final statement in the theorem follows from Proposition 1.3.9. ■

The preceding result, along with general properties of maps and inverses as given in Proposition 1.3.9, has the following immediate corollary.

5.1.43 Corollary (Left and right inverses for finite-dimensional matrices) *Let* F *be a field, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{F})$*. The following statements are equivalent:*

(i) $\mathbf{A}$ *possesses an inverse;*

(ii) $\mathbf{A}$ *possesses a left inverse;*

(iii) $\mathbf{A}$ *possesses a unique left inverse;*

(iv) $\mathbf{A}$ *possesses a right inverse;*

(v) $\mathbf{A}$ *possesses a unique right inverse;*

(vi) $\mathbf{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{\mathrm{n}}; \mathsf{F}^{\mathrm{n}})$ *is injective;*

(vii) $\mathbf{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^n; \mathsf{F}^n)$ *is surjective;*

(viii) $\mathbf{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^n; \mathsf{F}^n)$ *is bijective.*

Note that the preceding result is not generally true for square matrices defined using infinite index sets, as the following example shows.

5.1.44 Example (An infinite-dimensional counterexample) Let F be a field and take the index set $I = \mathbb{Z}_{>0}$. Let $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}}$ be the standard basis for $\mathsf{F}_0^{\mathbb{Z}_{>0}}$ and define $\boldsymbol{A} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}_0^{\mathbb{Z}_{>0}}; \mathsf{F}_0^{\mathbb{Z}_{>0}})$ by asking that $\boldsymbol{A}\boldsymbol{e}_j = \boldsymbol{e}_{2j}$ for each $j \in \mathbb{Z}_{>0}$. For a general $\boldsymbol{x} = c_1\boldsymbol{e}_1 + \cdots + c_k\boldsymbol{e}_k \in \mathsf{F}_0^{\mathbb{Z}_{>0}}$, then define

$$\boldsymbol{A}\boldsymbol{x} = c_1\boldsymbol{A}\boldsymbol{e}_1 + \cdots + c_k\boldsymbol{A}\boldsymbol{e}_k.$$

The matrix associated to this linear map is

$$\boldsymbol{A} = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

With $\boldsymbol{A}$ so defined, we claim that $\boldsymbol{A}$ is injective but not surjective, and possesses multiple left-inverses but no right inverse. To see that $\boldsymbol{A}$ is injective, suppose that $\boldsymbol{x} = c_1\boldsymbol{e}_1 + \cdots + c_k\boldsymbol{e}_k$, and that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}_{\mathsf{F}_0^{\mathbb{Z}_{>0}}}$. Then

$$c_1\boldsymbol{A}\boldsymbol{e}_1 + \cdots + c_k\boldsymbol{A}\boldsymbol{e}_k = c_1\boldsymbol{e}_2 + \cdots + c_k\boldsymbol{e}_{2k} = \boldsymbol{0}_{\mathsf{F}_0^{\mathbb{Z}_{>0}}}.$$

By linear independence of the standard basis, $c_1 = \cdots = c_k = 0_{\mathsf{F}}$, and so $\boldsymbol{A}$ is injective by Exercise 4.5.23. That $\boldsymbol{A}$ is not surjective follows since, for example, $\boldsymbol{e}_j \notin \mathrm{image}(\boldsymbol{A})$ for j odd. Since $\boldsymbol{A}$ is injective, it has a left-inverse by Proposition 1.3.9. In fact, $\boldsymbol{A}$ has many left-inverses, two of which are given by the matrices

$$\boldsymbol{B}_1 = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \boldsymbol{B}_2 = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Since $\boldsymbol{A}$ is not surjective, it has no right inverse.

This example illustrates another of the important differences between finite-dimensional linear algebra and infinite-dimensional linear algebra. •

5.1.7 Characterisations of row and column equivalence for matrices over fields

In this section we deal exclusively with matrices in $\mathrm{Mat}_{m\times n}(\mathsf{F})$.

Next we turn to a description for row and column equivalent matrices. One of our objectives will be to produce analogues of Theorems 5.1.41 and 5.1.42 (these being for equivalence) for row equivalence. In part (iii) of Theorem 5.1.41 we gave a particular simple form for a representative of the equivalence class of equivalent matrices. The corresponding form for row equivalence is more complicated, and is as given in the following definition.

5.1.45 Definition (Row Hermite matrix, reduced row echelon matrix) Let F be a field, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$. For $i \in \{1, \dots, m\}$ denote

$$E(i) = \begin{cases} \min\{j \in \{1, \dots, n\} \mid \boldsymbol{A}(i, j) \neq 0_{\mathsf{F}}\}, & \text{the } i\text{th row of } \boldsymbol{A} \text{ is nonzero}, \\ \infty, & \text{the } i\text{th row of } \boldsymbol{A} \text{ is zero}. \end{cases}$$

Then

(i) $\boldsymbol{A}$ is in ***row Hermite form*** if there exists $k \in \{1, \dots, m\}$ such that

(a) the first k rows of $\boldsymbol{A}$ are nonzero and the last $n - k$ rows of $\boldsymbol{A}$ are zero and such that

(b) $i_1 < i_2$, $i_1, i_2 \in \{1, \dots, k\}$, implies that $E(i_1) < E(i_2)$,

and

(ii) $\boldsymbol{A}$ is in ***reduced row echelon form*** if it is in row Hermite form, and if additionally

(a) $\boldsymbol{A}(i, E(i)) = 1_{\mathsf{F}}$ for all $i \in \{1, \dots, m\}$ such that $E(i) \neq \infty$ and

(b) $\boldsymbol{A}(i, j) = 0_{\mathsf{F}}$ for $j \neq E(i)$ and for all $i \in \{1, \dots, m\}$. •

Sometimes what we call "row Hermite form" is called "row echelon form." As we shall see in Section 5.2.7, Hermite form is also relevant for matrices whose entries are elements of a principal ideal domain.

By parsing the definitions one can easily deduce that a matrix in row Hermite form looks like

$$\begin{bmatrix}
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & \# & \cdots & * & * & \cdots & * & * & \cdots & * & * & \cdots & * & * & \cdots & * \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & \# & \cdots & * & * & \cdots & * & * & \cdots & * & * & \cdots & * \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & \# & \cdots & * & * & \cdots & * & * & \cdots & * \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & \# & \cdots & * & * & \cdots & * \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}}
\end{bmatrix}$$

where an entry denoted by # stands for any nonzero element of F and an entry denoted by $*$ stands for any element of F. Similarly, a matrix in reduced row echelon form looks like

$$\begin{bmatrix}
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & * & 0_{\mathsf{F}} & \cdots & * & 0_{\mathsf{F}} & \cdots & * & 0_{\mathsf{F}} & \cdots & * & * & \cdots & * \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & * & 0_{\mathsf{F}} & \cdots & * & 0_{\mathsf{F}} & \cdots & * & * & \cdots & * \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & * & 0_{\mathsf{F}} & \cdots & * & * & \cdots & * \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & * & * & \cdots & * \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \cdots & \vdots \\
0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}}
\end{bmatrix}$$

where, again, an entry denoted by $*$ stands for any element of F. For a matrix in reduced row echelon form, in a nonzero row the first nonzero element is always 1_{F}. These entries are called ***leading ones***.

The following simple result characterises the rowspace of a matrix in row Hermite form.

5.1.46 Proposition (Rowspace of a matrix in row Hermite form) *If* F *is a field, if* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$ *is in row Hermite form, then the nonzero rows of* $\mathbf{A}$ *form a basis for* $\mathrm{rowspace}(\mathbf{A})$.

Proof Let us denote the row vectors of $\boldsymbol{A}$ by $\{\boldsymbol{r}_1, \dots, \boldsymbol{r}_m\}$, thinking of these as vectors in F^n. Suppose that the first k rows are nonzero. It is clear that $\mathrm{rowspace}(\boldsymbol{A}) = \mathrm{span}_{\mathsf{F}}(\boldsymbol{r}_1, \dots, \boldsymbol{r}_k)$. Next let $c_1, \dots, c_k \in \mathsf{F}$ satisfy

$$c_1\boldsymbol{r}_1 + \dots + c_k\boldsymbol{r}_k = \mathbf{0}_{\mathsf{F}^n}.$$

This equation constitutes n equations when one asks that it be satisfied componentwise. If the first nonzero entry in $\boldsymbol{r}_1$ appears is j_1st component, then the j_1st of the n equations reads $A(1, j_1)c_1 = 0_{\mathsf{F}}$ where $A(1, j_1) \neq 0_{\mathsf{F}}$. Thus $c_1 = 0_{\mathsf{F}}$. Applying now the same reasoning to the second equation, if the first nonzero component of $\boldsymbol{r}_2$ is the j_2nd component, then the j_2nd equation reads $A(2, j_2)c_2 = 0_{\mathsf{F}}$ where $A(2, j_2) \neq 0_{\mathsf{F}}$. Thus $c_2 = 0_{\mathsf{F}}$. Continuing in this way we conclude that $c_1 = \dots = c_k = 0_{\mathsf{F}}$, showing linear independence of $\{\boldsymbol{r}_1, \dots, \boldsymbol{r}_k\}$. ■

In Theorem 5.1.58 we will provide a general result which indicates the significance of row Hermite form and reduced row echelon form. Specifically, we will see that these forms for matrices, when applied to systems of linear equations, make it easy to explicitly describe the set of solutions for a system of equations. The reader in need of motivation may wish to look ahead to see how this works. Here we will just prove the following result which gives some significance to the notion of reduced row echelon form in terms of row equivalence. Note that the following result applies only to matrices with components in a field.

5.1.47 Theorem (Row equivalence and reduced row echelon form) *Let* F *be field and let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*. Then each equivalence class in* $\mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$ *under the equivalence relation of row equivalence contains exactly one matrix in reduced row echelon form.*

Proof First let us show that every equivalence class contains at least one matrix in reduced row echelon form. Let $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$. If $\boldsymbol{A}$ is the zero matrix, the result holds trivially, so suppose that $\boldsymbol{A}$ is nonzero. Let j_1 be the smallest positive integer for which that j_1st column of $\boldsymbol{A}$ is nonzero, and let i_1 have the property that $\boldsymbol{A}(i_1, j_1) \neq 0_{\mathsf{F}}$. Let $\boldsymbol{A}'_1$ be the matrix obtained from $\boldsymbol{A}$ by swapping the 1st and j_1st rows. Thus $\boldsymbol{A}'_1$ has the form

$$\boldsymbol{A}'_1 = \begin{bmatrix} 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & a^1_{11} & a^1_{12} & \cdots & a^1_{1k_1} \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & b^1_{21} & b^1_{22} & \cdots & b^1_{2k_1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & b^1_{n1} & b^1_{n2} & \cdots & b^1_{nk_1} \end{bmatrix},$$

where $a^1_{11} \neq 0_{\mathsf{F}}$. Now, for $i \in \{2, \dots, n\}$ perform successive elementary row operations of subtracting $(a^1_{11})^{-1} a^1_{i1}$ times the first row from the ith row. The resulting matrix we denote by $\boldsymbol{A}_1$ and we note that this matrix has the form

$$\boldsymbol{A}_1 = \begin{bmatrix} 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & a^1_{11} & a^1_{12} & \cdots & a^1_{1k_1} \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & a^1_{22} & \cdots & a^1_{2k_1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & a^1_{n2} & \cdots & a^1_{nk_1} \end{bmatrix}. \tag{5.6}$$

The construction can then be applied to the matrix

$$\begin{bmatrix} a^1_{22} & \cdots & a^1_{k_1} \\ \vdots & \ddots & \vdots \\ a^1_{n2} & \cdots & a^1_{nk_1} \end{bmatrix} \tag{5.7}$$

to obtain a matrix, row equivalent to this one, and of the form of (5.6). Replacing the block (5.7) in $\boldsymbol{A}_1$ by this new matrix, we obtain a matrix $\boldsymbol{A}_2$, row equivalent to $\boldsymbol{A}_1$, of the form

$$\boldsymbol{A}_2 = \begin{bmatrix} 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & a^1_{11} & a^1_{12} & \cdots & a^1_{1j_2} & a^1_{1(j_2+1)} & a^1_{1(j_2+2)} & \cdots & a^1_{1k_1} \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & a^2_{11} & a^2_{12} & \cdots & a^2_{1k_2} \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & a^2_{22} & \cdots & a^2_{2k_2} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & a^2_{n2} & \cdots & a^2_{nk_2} \end{bmatrix},$$

where $a^2_{11} \neq 0_{\mathsf{F}}$. Proceeding in this way, after a finite number of steps we arrive at a matrix in row Hermite form that is row equivalent to $\boldsymbol{A}$.

Next we indicate how one, by elementary row operations, turns a matrix $\boldsymbol{A}$ in row Hermite form to one in reduced row echelon form. Suppose that the first k rows of $\boldsymbol{A}$ are nonzero, and that the first nonzero entry in the ith row occurs in column $E(i)$ for

$i \in \{1, \dots, k\}$. By multiplying the ith row of $\boldsymbol{A}$ by $(\boldsymbol{A}(i, E(i)))^{-1}$ we can ensure that the $(i, E(i))$-components are equal to 1_{F} for $i \in \{1, \dots, k\}$. Next, for $i' \in \{1, \dots, i-1\}$, we can subtract $\boldsymbol{A}(i', E(i))$ times the ith row from the i'th row to get zeros in the $E(i)$th column, except for the ith row. Thus we arrive at a matrix in reduced row echelon form after a finite number of elementary row operations. This establishes the existence part of the theorem.

For the uniqueness part of the theorem it suffices to show that, if $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are $m \times n$ matrices that are row equivalent *and* in reduced row echelon form, then it must hold that $\boldsymbol{A}_1 = \boldsymbol{A}_2$. We prove this by induction on n. There is only one $m \times 1$ matrix in reduced row echelon form, and this is the matrix

$$\begin{bmatrix} 1_{\mathsf{F}} \\ 0_{\mathsf{F}} \\ \vdots \\ 0_{\mathsf{F}} \end{bmatrix}.$$

Now suppose that any two row equivalent $m \times (n-1)$ matrices that are in reduced row echelon form are equal and let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be row equivalent $m \times n$ matrices in reduced row echelon form. By Proposition 5.1.32 and Theorem 5.1.33 there exists an invertible matrix $\boldsymbol{P}$ such that $\boldsymbol{A}_2 = \boldsymbol{P}\boldsymbol{A}_1$. Now write

$$\boldsymbol{A}_a = \left[\begin{array}{c|c} \boldsymbol{B}_a & \boldsymbol{b}_a \end{array}\right], \qquad a \in \{1, 2\},$$

for $\boldsymbol{B}_a \in \mathrm{Mat}_{m \times (n-1)}(\mathsf{F})$ and $\boldsymbol{b}_a \in \mathrm{Mat}_{m \times 1}(\mathsf{F})$, $a \in \{1, 2\}$. It is easy to see that $\boldsymbol{B}_2 = \boldsymbol{P}\boldsymbol{B}_1$, and so $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are row equivalent by Proposition 5.1.32 and Theorem 5.1.33. One can also easily deduce that $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are in reduced row echelon form since $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are. Therefore, by the induction hypothesis, $\boldsymbol{B}_1 = \boldsymbol{B}_2$. We now consider two cases.

1. The mth row of $\boldsymbol{b}_1$ is equal to 1_{F}: Since $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are row equivalent, $\mathrm{rowspace}(\boldsymbol{A}_1) = \mathrm{rowspace}(\boldsymbol{A}_2)$. By Proposition 5.1.46 it follows that the number of leading ones in $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ must agree. Since the number of leading ones in $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ agree, it follows that the mth row of $\boldsymbol{b}_2$ is 1_{F}. Since $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are in reduced row echelon form, the first $m+1$ components of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are zero. Thus $\boldsymbol{b}_1 = \boldsymbol{b}_2$ and so $\boldsymbol{A}_1 = \boldsymbol{A}_2$.
2. The mth row of $\boldsymbol{b}_1$ is not equal to 1_{F}: As we saw in the previous case, the number of leading ones in $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ must agree, and from this we conclude that the mth row of $\boldsymbol{b}_2$ is also zero. Let us write the r nonzero rows of $\boldsymbol{A}_a$, $a \in \{1, 2\}$, as

$$\left[\begin{array}{c|c} \boldsymbol{r}_j & \rho_j^a \end{array}\right], \qquad j \in \{1, \dots, r\},$$

using the fact that the first $n-1$ columns of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree. Since the rowspaces of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree by Theorem 5.1.50, we know that, for each $j \in \{1, \dots, j\}$, there exists $c_1^j, \dots, c_r^j \in \mathsf{F}$ such that,

$$\left[\begin{array}{c|c} \boldsymbol{r}_j & \rho_j^2 \end{array}\right] = c_1^j \left[\begin{array}{c|c} \boldsymbol{r}_1 & \rho_1^1 \end{array}\right] + \cdots + c_r^j \left[\begin{array}{c|c} \boldsymbol{r}_r & \rho_r^1 \end{array}\right].$$

This, in particular, implies that

$$\boldsymbol{r}_j = c_1^j \boldsymbol{r}_1 + \cdots + c_r^j \boldsymbol{r}_r, \qquad j \in \{1, \dots, r\}.$$

Since the vectors $\{\boldsymbol{r}_1,\dots,\boldsymbol{r}_r\}$ are linearly independent by Proposition 5.1.46, it follows that

$$c_k^j = \begin{cases} 1_{\mathsf{F}}, & j = k, \\ 0_{\mathsf{F}}, & j \neq k. \end{cases}$$

From this it immediately follows that $\rho_j^1 = \rho_j^2$ for each $j \in \{1,\dots,r\}$. Thus $\boldsymbol{A}_1 = \boldsymbol{A}_2$, as desired. ■

Note that the existence part of the proof of the theorem is constructive. Let us illustrate this with an example.

5.1.48 Example (Reduced row echelon form) For a field F, take $\boldsymbol{A} \in \mathrm{Mat}_{3\times4}(\mathsf{F})$ to be

$$\boldsymbol{A} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 2_{\mathsf{F}} \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 2_{\mathsf{F}} & 1_{\mathsf{F}} & 7_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix},$$

where by k_{F}, $k \in \mathbb{Z}$, we mean $k1_{\mathsf{F}}$ (cf. Proposition 4.2.10). We now perform a sequence of elementary row operations.

1. Swap the first and second row:

$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 2_{\mathsf{F}} \\ 2_{\mathsf{F}} & 1_{\mathsf{F}} & 7_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

2. Swap the second and third row:

$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 2_{\mathsf{F}} & 1_{\mathsf{F}} & 7_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 2_{\mathsf{F}} \end{bmatrix}.$$

3. Subtract 2_{F} times the first row from the second row:

$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & -1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 2_{\mathsf{F}} \end{bmatrix}.$$

4. Subtract the second row from the third row:

$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & -1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} \end{bmatrix}.$$

5. Multiply the third row by 3_{F}^{-1} if F does not have characteristic 3 (otherwise, do nothing):

$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & -1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

6. Add the third row to the second row:
$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

7. Subtract the third row from the first row:
$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix},$$
which gives a matrix in reduced row echelon form.

From this example, one can perhaps convince oneself that at least the existence part of Theorem 5.1.47 makes sense. •

5.1.49 Notation (Row reduction) The process by which one starts with some matrix and performs the elementary row operations to put the system in reduced row echelon form is called ***row reduction***. •

Now we can state alternative characterisations of row equivalence.

5.1.50 Theorem (Characterisations of row equivalence) *Let* F *be a field, let* $m, n \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{F})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row equivalent;*

(ii) there exists an invertible matrix $\mathbf{P} \in \mathrm{Mat}_{m\times m}(\mathsf{F})$ *such that* $\mathbf{P}\mathbf{A}_1 = \mathbf{A}_2$*;*

(iii) $\ker(\mathbf{A}_1) = \ker(\mathbf{A}_2)$*;*

(iv) $\mathrm{rowspace}(\mathbf{A}_1) = \mathrm{rowspace}(\mathbf{A}_2)$*.*

Proof (i) $\Longleftrightarrow$ (ii) This follows from Proposition 5.1.32 and Theorem 5.1.33.

(ii) $\Longrightarrow$ (iii) Since $\boldsymbol{P}$ is invertible, $\ker(\boldsymbol{P}) = \{\mathbf{0}_{\mathsf{F}^m}\}$ by Exercise 4.5.23. Thus we have
$$\boldsymbol{A}_1\boldsymbol{x} = \mathbf{0}_{\mathsf{F}^m} \quad\Longleftrightarrow\quad \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{x} = \mathbf{0}_{\mathsf{F}^m} \quad\Longleftrightarrow\quad \boldsymbol{x} \in \ker(\boldsymbol{A}_2).$$

(iii) $\Longrightarrow$ (iv) We use a lemma.

1 Lemma *For* $\mathbf{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$ *we have*
$$\ker(\mathbf{A}) = \left\{\mathbf{x} \in \mathsf{F}^n \;\middle|\; \textstyle\sum_{i=1}^n \mathbf{x}(i)\mathbf{z}(i) = 0_{\mathsf{F}} \textit{ for all } \mathbf{z} \in \mathrm{image}(\mathbf{A}^T)\right\}.$$

Proof This will be proved in greater generality as Lemma 1 in the proof of Proposition 5.2.14. ▼

Now suppose that, for $\boldsymbol{x} \in \mathsf{F}^n$, $\boldsymbol{A}_1\boldsymbol{x} = \mathbf{0}_{\mathsf{F}^m}$ if and only if $\boldsymbol{A}_2\boldsymbol{x} = \mathbf{0}_{\mathsf{F}^m}$. By the lemma this means that
$$\left\{\boldsymbol{x} \in \mathsf{F}^n \;\middle|\; \sum_{i=1}^n \boldsymbol{x}(i)\boldsymbol{z}(i) = 0_{\mathsf{F}} \text{ for all } \boldsymbol{z} \in \mathrm{image}(\boldsymbol{A}_1^T)\right\}$$
$$= \left\{\boldsymbol{x} \in \mathsf{F}^n \;\middle|\; \sum_{i=1}^n \boldsymbol{x}(i)\boldsymbol{z}(i) = 0_{\mathsf{F}} \text{ for all } \boldsymbol{z} \in \mathrm{image}(\boldsymbol{A}_2^T)\right\}.$$

We claim that this implies that $\mathrm{image}(A_1^T) = \mathrm{image}(A_2^T)$. Indeed, if $\mathsf{U} \subseteq \mathsf{F}^n$ is a subspace, let us define

$$\mathsf{U}^\perp = \left\{ x \in \mathsf{F}^n \;\middle|\; \sum_{i=1}^n x(i)z(i) = 0_{\mathsf{F}} \text{ for all } z \in \mathsf{U} \right\}.$$

It is then easy to see that $\mathsf{U}^{\perp\perp} = \mathsf{U}$. In particular, since we have

$$(\mathrm{image}(A_1^T))^\perp = (\mathrm{image}(A_2^T))^\perp,$$

it follows that $\mathrm{image}(A_1^T) = \mathrm{image}(A_2^T)$, as desired. This part of the result then follows from Proposition 5.1.19.

(iv) $\Longrightarrow$ (i) Here we use Theorem 5.1.47 along with the following lemma.

2 Lemma *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*. Then* $\mathrm{rowspace}(\mathbf{A}_1) = \mathrm{rowspace}(\mathbf{A}_2)$ *if* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row equivalent.*

Proof This follows since, if A_2 is obtained from A_1 by an elementary row operation, then clearly the rows of A_2 are linear combinations of the rows of A_1. Moreover, since row operations are invertible by Theorem 5.1.33, it also follows that the rows of A_1 are linear combinations of the rows of A_2. ▼

Since $\mathrm{rowspace}(A_1) = \mathrm{rowspace}(A_2)$, it follows from the lemma and from the existence part of Theorem 5.1.47 that A_1 and A_2 are row equivalent to a matrices in reduced row echelon form whose rowspaces agree. However, if two matrices in reduced row echelon form have equal rowspaces, then these matrices must be equal (why?). By the uniqueness part of Theorem 5.1.47 we know that A_1 and A_2 are row equivalent. ■

Of course, the constructions in this section can be repeated, with appropriate modifications, for column equivalence. For example, one could define "column Hermite form" and "reduced column echelon form." This is not often done, however, and the reason is that reduced row echelon form is useful for solving systems of linear equations as we shall see in Section 5.1.8. Therefore, we content ourselves with stating the column equivalence version of Theorem 5.1.50.

5.1.51 Theorem (Alternative characterisations of column equivalence) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are column equivalent;*

(ii) there exists an invertible matrix $\mathbf{P} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{m}}(\mathsf{F})$ *such that* $\mathbf{A}_1\mathbf{P} = \mathbf{A}_2$*.*

(iii) $\ker(\mathbf{A}_1^T) = \ker(\mathbf{A}_2^T)$*;*

(iv) $\mathrm{colspace}(\mathbf{A}_1) = \mathrm{colspace}(\mathbf{A}_2)$*;*

We close this section by giving an interesting application of reduced row echelon form to the determination of a basis for the columnspace of a matrix.

5.1.52 Theorem (Basis for the columnspace) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*. If the leading ones in the reduced row echelon form for* $\mathbf{A}$ *appear in the columns* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{r}$*, then the columns* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{r}$ *of* $\mathbf{A}$ *form a basis for* $\mathrm{colspace}(\mathbf{A})$*.*

Proof The theorem follows from the following lemma.

1 **Lemma** *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F})$*. For* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{k} \in \{1, \dots, \mathrm{n}\}$*, suppose that columns* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{k}$ *of* $\mathbf{A}$ *are linearly independent. Let* $\mathbf{E} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{m}}(\mathsf{F})$ *be an elementary matrix. Then the columns* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{k}$ *of* $\mathbf{EA}$ *are linearly independent.*

Proof Recall from Definition 5.1.4 that the column vectors of $\boldsymbol{A}$ are denoted by $\boldsymbol{c}(\boldsymbol{A}, 1), \dots, \boldsymbol{c}(\boldsymbol{A}, n)$. Using the definition of matrix multiplication it is easy to show that the column vectors of $\boldsymbol{EA}$ are given by $\boldsymbol{c}(\boldsymbol{EA}, j) = \boldsymbol{Ec}(\boldsymbol{A}, j)$ for $j \in \{1, \dots, n\}$. Thus it suffices to show that if $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\} \subseteq \mathsf{F}^n$ is linearly independent, then so too is $\{\boldsymbol{Ex}_1, \dots, \boldsymbol{Ex}_k\}$. Suppose that

$$c_1\boldsymbol{Ex}_1 + \cdots + c_k\boldsymbol{Ex}_k = \boldsymbol{0}_{\mathsf{F}^m}.$$

Then, by linearity,

$$\boldsymbol{E}(c_1\boldsymbol{x}_1 + \cdots + c_k\boldsymbol{x}_k) = \boldsymbol{0}_{\mathsf{F}^m},$$

giving $c_1\boldsymbol{x}_1 + \cdots + c_k\boldsymbol{x}_k = \boldsymbol{0}_{\mathsf{F}^n}$ by Exercise 4.5.23 and since $\boldsymbol{E}$ is invertible. Thus $c_1 = \cdots = c_k = 0_{\mathsf{F}}$, giving our assertion, and so the lemma. ▼

Now, if $\boldsymbol{B}$ is the reduced row echelon form corresponding to $\boldsymbol{A}$ we have $\boldsymbol{A} = \boldsymbol{E}_1 \cdots \boldsymbol{E}_k\boldsymbol{B}$ for elementary matrices $\boldsymbol{E}_1, \dots, \boldsymbol{E}_k$. By successively applying the lemma, we see that the columns $j_1, \dots, j_r$ of $\boldsymbol{A}$ are linearly independent. Moreover, since $r = \mathrm{rank}(\boldsymbol{A})$ by Proposition 5.1.46, it also holds that the columns $j_1, \dots, j_r$ form a basis for $\mathrm{colspace}(\boldsymbol{A})$ since $\dim_{\mathsf{F}}(\mathrm{colspace}(\boldsymbol{A})) = \mathrm{rank}(\boldsymbol{A})$. ■

Let us illustrate this result with an example.

5.1.53 Example (Basis for the columnspace (Example 5.1.48 cont'd)) For a field F we take

$$\boldsymbol{A} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 2_{\mathsf{F}} \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 1_{\mathsf{F}} \\ 2_{\mathsf{F}} & 1_{\mathsf{F}} & 7_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix},$$

As we saw in Example 5.1.48, the reduced row echelon form for $\boldsymbol{A}$ is given by

$$\begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 3_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix},$$

which gives leading ones in columns 1, 2, and 4. Then the first, second, and fourth columns of $\boldsymbol{A}$, i.e.,

$$\{(0_{\mathsf{F}}, 1_{\mathsf{F}}, 2_{\mathsf{F}}), (1_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}}), (2_{\mathsf{F}}, 1_{\mathsf{F}}, 1_{\mathsf{F}})\},$$

form a basis for $\mathrm{colspace}(\boldsymbol{A})$. This choice of columns as a basis for the column space is not unique, but at least the reduced row echelon form gives *some* selection of columns for a basis. •

5.1.8 Systems of linear equations over fields

Matrices are useful for formulating systems of linear equations. The sorts of equations we are interested in are linear equations of the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$, $\boldsymbol{x} \in \mathsf{F}^n$, and $\boldsymbol{b} \in \mathsf{F}^m$. Given $\boldsymbol{b}$, we are interested in the set of $\boldsymbol{x}$'s that satisfy this equation.

Let us first define precisely what we mean by a system of linear equations.

5.1.54 Definition (System of linear equations over a field) Let F be a field and let I and J be index sets.

(i) A ***system of linear equations*** over F is a pair $(\boldsymbol{A}, \boldsymbol{b}) \in \mathrm{Mat}_{I\times J}(\mathsf{F}) \times \mathsf{F}_0^I$.

(ii) A system of linear equations $(\boldsymbol{A}, \boldsymbol{b})$ is ***homogeneous*** if $\boldsymbol{b}(i) = 0_{\mathsf{F}}$ for every $i \in I$.

(iii) The ***solution set*** for a system of linear equations $(\boldsymbol{A}, \boldsymbol{b})$ is the subset of F_0^J defined by

$$\mathrm{Sol}(\boldsymbol{A}, \boldsymbol{b}) = \{\boldsymbol{x} \in \mathsf{F}_0^J \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

A ***solution*** to the system of linear equations $(\boldsymbol{A}, \boldsymbol{b})$ is an element of the solution set.

(iv) For a system of linear equations $(\boldsymbol{A}, \boldsymbol{b}) \in \mathrm{Mat}_{I\times J}(\mathsf{F}) \times \mathsf{F}_0^I$, the ***augmented matrix*** for the system is the matrix $[\boldsymbol{A}, \boldsymbol{b}]$ over F in $I \times (J \overset{\circ}{\cup} \{j_0\})$ defined by

$$[\boldsymbol{A}, \boldsymbol{b}](i, j) = \begin{cases} \boldsymbol{A}(i, j), & (i, j) \in I \times J, \\ \boldsymbol{b}(i), & (i, j) \in I \times \{j_0\}. \end{cases} \qquad \bullet$$

Intuitively, the augmented matrix for a system of linear equations is formed by adding a column to $\boldsymbol{A}$ consisting of $\boldsymbol{b}$. In the cases when $I = \{1, \dots, n\}$ and $J = \{1, \dots, m\}$, we shall adopt the convention that the augmented matrix have $\boldsymbol{b}$ as its $(n + 1)$st column. That is, we have

$$[\boldsymbol{A}, \boldsymbol{b}] = \begin{bmatrix} \boldsymbol{A}(1, 1) & \cdots & \boldsymbol{A}(1, n), & \boldsymbol{b}(1) \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{A}(m, 1) & \cdots & \boldsymbol{A}(m, n) & \boldsymbol{b}(m) \end{bmatrix}.$$

Much of this section will be devoted to understanding $\mathrm{Sol}(\boldsymbol{A}, \boldsymbol{b})$ when $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{F})$ and $\boldsymbol{b} \in \mathsf{F}^m$. However, before we simplify to this finite-dimensional case, let us make some general observations about the character of $\mathrm{Sol}(\boldsymbol{A}, \boldsymbol{b})$. First we state a result which gives a general, but noncomputational, characterisation of the set of solutions to a system of linear equations.

5.1.55 Proposition (Existence and uniqueness of solutions) *Let F be a field, let I and J be index sets, and let $(\mathbf{A}, \mathbf{b}) \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F}) \times \mathsf{F}_0^{\mathrm{I}}$ be a system of linear equations. Then the following statements hold:*

(i) $\mathrm{Sol}(\mathbf{A}, \mathbf{b})$ *is nonempty if and only if* $\mathbf{b} \in \mathrm{image}(\mathbf{A})$*;*

(ii) *in particular,* $\mathrm{Sol}(\mathbf{A},\mathbf{b})$ *is nonempty for every* $\mathbf{b}\in\mathsf{F}_0^{\mathrm{I}}$ *if and only if* $\mathbf{A}$ *is surjective;*

(iii) $\mathrm{Sol}(\mathbf{A},\mathbf{b})$ *is a singleton if and only if*

- *(a)* $\mathbf{b}\in\mathrm{image}(\mathbf{A})$ *and*
- *(b)* $\mathbf{A}$ *is injective.*

Proof The only nonobvious assertion is the last one, so it is the only one we prove. If $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ is a singleton, then by the first part of the proposition it holds that $\boldsymbol{b}\in\mathrm{image}(\boldsymbol{A})$. If $\boldsymbol{A}$ is not injective then $\ker(\boldsymbol{A})\neq\{\boldsymbol{0}_{\mathsf{F}_0^J}\}$ by Exercise 4.5.23. If $\boldsymbol{x}\in\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ and if $\boldsymbol{x}'\in\ker(\boldsymbol{A})$ then

$$\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{x}')=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{A}\boldsymbol{x}'=\boldsymbol{b}.$$

This shows that $\boldsymbol{A}$ must be injective if $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ is a singleton.

Conversely, suppose that $\boldsymbol{b}\in\mathrm{image}(\boldsymbol{A})$ and that $\boldsymbol{A}$ is injective. Then $\ker(\boldsymbol{A})=\{\boldsymbol{0}_{\mathsf{F}_0^J}\}$ by Exercise 4.5.23. Now let $\boldsymbol{x}_1,\boldsymbol{x}_2\in\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$. Then

$$\boldsymbol{A}(\boldsymbol{x}_1-\boldsymbol{x}_2)=\boldsymbol{A}\boldsymbol{x}_1-\boldsymbol{A}\boldsymbol{x}_2=\boldsymbol{b}-\boldsymbol{b}=\boldsymbol{0}_{\mathsf{F}_0^I}.$$

Thus $\boldsymbol{x}_1-\boldsymbol{x}_2\in\ker(\boldsymbol{A})$, giving $\boldsymbol{x}_1=\boldsymbol{x}_2$. ■

Next we characterise the set of the solutions. The following result says that the set of solutions, when it is nonempty, is an affine subspace, referring to the terminology of Definition 4.5.13.

5.1.56 Proposition (Characterisation of $\mathrm{Sol}(\mathbf{A},\mathbf{b})$) *Let* F *be a field, let* I *and* J *be index sets, and let* $(\mathbf{A},\mathbf{b})\in\mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})\times\mathsf{F}_0^{\mathrm{I}}$ *be a system of linear equations in* F*. Then, for any* $\mathbf{x}_0\in\mathrm{Sol}(\mathbf{A},\mathbf{b})$*,*

$$\mathrm{Sol}(\mathbf{A},\mathbf{b})=\{\mathbf{x}+\mathbf{x}_0\in\mathsf{F}_0^{\mathrm{J}}\mid\ \mathbf{x}\in\mathrm{Sol}(\mathbf{A},\mathbf{0}_{\mathsf{F}_0^{\mathrm{J}}})\}.$$

Proof Let $\boldsymbol{x}_0$ be any solution to $(\boldsymbol{A},\boldsymbol{b})$ as stated in the proposition. If $\boldsymbol{x}\in\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ then

$$\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_0)=\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}_0=\boldsymbol{b}-\boldsymbol{b}=\boldsymbol{0}_{\mathsf{F}_0^I}.$$

Thus $\boldsymbol{x}=\boldsymbol{x}_0+\boldsymbol{x}'$ for $\boldsymbol{x}'\in\mathrm{Sol}(\boldsymbol{A},\boldsymbol{0}_{\mathsf{F}_0^I})$. Conversely, if $\boldsymbol{x}=\boldsymbol{x}_0+\boldsymbol{x}'$ for some $\boldsymbol{x}'\in\mathrm{Sol}(\boldsymbol{A},\boldsymbol{0}_{\mathsf{F}_0^I})$, then

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{A}(\boldsymbol{x}_0+\boldsymbol{x}')=\boldsymbol{A}\boldsymbol{x}_0=\boldsymbol{b},$$

and so $\boldsymbol{x}\in\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$, giving the result. ■

Note that the preceding result does *not* say that $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ is nonempty. It characterises $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ in cases when it *is* nonempty, i.e., when $\boldsymbol{b}\in\mathrm{image}(\boldsymbol{A})$. It is also worth remarking on the general procedure that the result suggests for finding $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$:

1. find *some* element of $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$;
2. find *all* solutions to the homogeneous system $(\boldsymbol{A},\boldsymbol{0}_{\mathsf{F}_0^I})$.

In practice the first step is the most difficult, in some sense. Note that $\mathbf{0}_{\mathsf{F}_0^J} \in \mathrm{Sol}(\boldsymbol{A}, \mathbf{0}_{\mathsf{F}_0^I})$, so the homogeneous system always has solutions. The reader may wish to compare this procedure with, for example, methods for solving inhomogeneous linear differential equations. The idea is the same; one first finds *some* solution (often called a "particular solution"), and then the set of solutions is formed by adding to this particular solution the set of all solutions to the homogeneous system. Moreover, this idea is repeated for many sorts of linear equations, not necessarily algebraic (e.g., ordinary differential equations and partial differential equations).

Let us now proceed to a description of a method for finding solutions in the case where $\boldsymbol{A}$ has finitely many rows and columns. We use row reduction to accomplish this. That this is a feasible thing to do is based on the following result.

5.1.57 Theorem (Elementary row operations do not change the solution set) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $(\mathbf{A}_1, \mathbf{b}_1), (\mathbf{A}_2, \mathbf{b}_2) \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F}) \times \mathsf{F}^{\mathrm{m}}$ *be systems of linear equations. Then* $\mathrm{Sol}(\mathbf{A}_1, \mathbf{b}_1) = \mathrm{Sol}(\mathbf{A}_2, \mathbf{b}_2)$ *if* $[\mathbf{A}_1, \mathbf{b}_1]$ *and* $[\mathbf{A}_2, \mathbf{b}_2]$ *are row equivalent. Conversely, if* $\mathrm{Sol}(\mathbf{A}_1, \mathbf{b}_1) = \mathrm{Sol}(\mathbf{A}_2, \mathbf{b}_2)$ *with both sets of solutions being nonempty, then* $[\mathbf{A}_1, \mathbf{b}_1]$ *and* $[\mathbf{A}_2, \mathbf{b}_2]$ *are row equivalent.*

Proof First suppose that $[\boldsymbol{A}_1, \boldsymbol{b}_1]$ and $[\boldsymbol{A}_2, \boldsymbol{b}_2]$ are row equivalent. It is sufficient to consider the case where $[\boldsymbol{A}_2, \boldsymbol{b}_2]$ is obtained from $[\boldsymbol{A}_1, \boldsymbol{b}_1]$ by an elementary row operation. Moreover, since row and column equivalence are equivalence relations, it suffices to show that, if $[\boldsymbol{A}_1, \boldsymbol{b}_1]$ and $[\boldsymbol{A}_2, \boldsymbol{b}_2]$ are row equivalent, then $\mathrm{Sol}(\boldsymbol{A}_1, \boldsymbol{b}_1) \subseteq \mathrm{Sol}(\boldsymbol{A}_2, \boldsymbol{b}_2)$. We first consider the case where $\mathrm{Sol}(\boldsymbol{A}_1, \boldsymbol{b}_1) \neq \emptyset$. We consider the three types of row operations in succession.

Type (i): Suppose $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}_1, \boldsymbol{b}_1)$. Then

$$\sum_{j=1}^{m} \boldsymbol{A}_1(i, j)\boldsymbol{x}(j) = \boldsymbol{b}_1(j), \qquad i \in \{1, \dots, m\}. \tag{5.8}$$

This represents m equations, and clearly the order in which we write then is inconsequential. Thus it immediately follows that $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}_2, \boldsymbol{b}_2)$.

Type (ii): Suppose that $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}_1, \boldsymbol{b}_1)$ so that (5.8) holds. Let $i_0 \in \{1, \dots, n\}$ have the property that the i_0th row of $[\boldsymbol{A}_2, \boldsymbol{b}_2]$ is equal to u times the i_0th row of $[\boldsymbol{A}_1, \boldsymbol{b}_1]$. It also holds that

$$\begin{aligned} \sum_{j=1}^{m} \boldsymbol{A}_1(i, j)\boldsymbol{x}(j) &= \boldsymbol{b}_1(i), \qquad i \in \{1, \dots, n\} \setminus \{i_0\}, \\ \sum_{j=1}^{m} u\boldsymbol{A}_1(i_0, j)\boldsymbol{x}(j) &= u\boldsymbol{b}_1(i_0). \end{aligned}$$

But this is exactly the assertion that $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}_2, \boldsymbol{b}_2)$.

Type (iii): Again, if $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}_1, \boldsymbol{b}_1)$, then (5.8) holds. Let $i_1, i_2 \in \{1, \dots, n\}$ have the property that the i_1st row of $[\boldsymbol{A}_2, \boldsymbol{b}_2]$ is obtained by adding r times the i_2nd row of

$[A_1, b_1]$ to the i_1st row of $[A_1, b_1]$. Then we have

$$\sum_{j=1}^{m} A_1(i,j)x(j) = b_1(i), \qquad i \in \{1,\dots,n\} \setminus \{i_1\},$$

$$\sum_{j=1}^{m} (A_1(i_1,j) + rA_1(i_2,j))x(j) = b_1(i_1) + rb_1(i_2),$$

which exactly implies that $x \in \mathrm{Sol}(A_2, b_2)$.

If either $\mathrm{Sol}(A_1, b_1)$ or $\mathrm{Sol}(A_2, b_2)$ are empty, it immediately follows that the other is also empty, since our above computations give a means of construction a solution for one system of linear equations given a solution for the other.

Now suppose that $\mathrm{Sol}(A_1, b_1) = \mathrm{Sol}(A_2, b_2)$ and that both sets of solutions are nonempty. By Proposition 5.1.56 it follows that $\ker(A_1) = \ker(A_2)$, and so by Theorem 5.1.50 it follows that A_1 and A_2 are row equivalent. Therefore, by Theorem 5.1.47, A_1 and A_2 have the same reduced row echelon form. Let us then write the reduced row echelon forms for $[A_1, b_1]$ and $[A_2, b_2]$ as

$$\left[\begin{array}{c|c} A_1' & b_1' \end{array}\right], \quad \left[\begin{array}{c|c} A_2' & b_2' \end{array}\right]$$

for $A_1', A_2' \in \mathrm{Mat}_{m\times n}(\mathsf{F})$ and for $b_1', b_2' \in \mathrm{Mat}_{m\times 1}(\mathsf{F})$. Since A_1 and A_2 have the same reduced row echelon form, $A_1' = A_2'$. Moreover, by the first part of the theorem, $\mathrm{Sol}(A_1', b_1') = \mathrm{Sol}(A_1, b_1)$ and $\mathrm{Sol}(A_2', b_2') = \mathrm{Sol}(A_2, b_2)$. In particular, $\mathrm{Sol}(A_1', b_1') = \mathrm{Sol}(A_1', b_2')$. This implies that

$$A_1'(\mathrm{Sol}(A_1', b_1')) = A_1'(\mathrm{Sol}(A_1', b_2')).$$

But this implies that $\{b_1'\} = \{b_2'\}$, meaning that $[A_1, b_1] = [A_2, b_2]$ have the same reduced row echelon form. Thus they are row equivalent by Theorem 5.1.50. ■

Since the set of solutions $\mathrm{Sol}(A, b)$ is only dependent on the equivalence class of the augmented matrix $[A, b]$ under the equivalence relation of row equivalence, one might hope that, by choosing a simple representative of this equivalence class, it is easy to characterise the nature of $\mathrm{Sol}(A, b)$. A convenient choice of this representative is, of course, none other than the reduced row echelon form of Definition 5.1.45. Let us write the form of a system of linear equations when the augmented matrix has been put into reduced row echelon form. Suppose that the leading ones appear in the first r rows and in columns $j_1, \dots, j_r$. Then the equations $Ax = b$ have the form

$$\begin{aligned}
x_{j_1} + a_{1,j_1+1}x_{j_1+1} + \cdots + 0_{\mathsf{F}}x_{j_2} + a_{1,j_2+1}x_{j_2+1} + \cdots + 0_{\mathsf{F}}x_{j_r} + a_{1,j_r+1}x_{j_r+1} + \cdots &= b_1 \\
x_{j_2} + a_{2,j_2+1}x_{j_2+1} + \cdots + 0_{\mathsf{F}}x_{j_r} + a_{2,j_r+1}x_{j_r+1} + \cdots &= b_2 \\
&\vdots \\
x_{j_r} + a_{r,j_r+1}x_{j_r+1} + \cdots &= b_r \\
0_{\mathsf{F}} &= b_{r+1} \\
0_{\mathsf{F}} &= 0_{\mathsf{F}} \\
&\vdots \\
0_{\mathsf{F}} &= 0_{\mathsf{F}}.
\end{aligned} \tag{5.9}$$

Moreover, concerning the values of $b_1, \dots, b_{r+1}$, one of the following two cases holds.

1. $b_{r+1} = 0_{\mathsf{F}}$: In this case the values of $b_1, \dots, b_r$ are unspecified, i.e., they are what they are.
2. $b_{r+1} = 1_{\mathsf{F}}$: In this case $b_1 = \cdots = b_r = 0_{\mathsf{F}}$.

With this form of the equations in reduced row echelon form, we have the following result which characterises the solutions.

5.1.58 Theorem (Reduced row echelon form and solutions of systems of linear equations) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $(\mathbf{A}, \mathbf{b}) \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F}) \times \mathsf{F}^{\mathrm{m}}$ *be a system of linear equations. Let* $\mathrm{r} \in \mathbb{Z}_{>0}$ *and* $\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{r}$ *be as in (5.9). Then*

(i) $\mathrm{Sol}(\mathbf{A}, \mathbf{b})$ *is nonempty if and only if the number of leading ones in the reduced row echelon form for* $[\mathbf{A}, \mathbf{b}]$ *is equal to the number of leading ones in the reduced row echelon form for* $\mathbf{A}$*.*

Now suppose that the number of leading ones in the reduced row echelon form for $[\mathbf{A}, \mathbf{b}]$ *is equal to the number of leading ones in the reduced row echelon form for* $\mathbf{A}$*. Then the following statements hold:*

(ii) we have

$$\begin{aligned}
\mathrm{Sol}(\mathbf{A}, \mathbf{b}) = \{(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}) \in \mathsf{F}^\mathrm{n} \mid\ & \mathrm{x}_\mathrm{j} \in \mathsf{F},\ \mathrm{j} \notin \{\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{r}\}, \\
& \mathrm{x}_\mathrm{j} \textit{ are determined by (5.9)},\ \mathrm{j} \in \{\mathrm{j}_1, \dots, \mathrm{j}_\mathrm{r}\}\};
\end{aligned}$$

(iii) $\mathrm{Sol}(\mathbf{A}, \mathbf{b})$ *is a singleton if and only if the number of leading ones in the reduced row echelon form for* $\mathbf{A}$ *is* n*.*

Proof Since the theorem has only to do with reduced row echelon form, we suppose $[\boldsymbol{A}, \boldsymbol{b}]$ (and, therefore, $\boldsymbol{A}$) to be in reduced row echelon form to save having to repeatedly say "in the reduced row echelon form of."

(i) The fact that the number of leading ones in $[\boldsymbol{A}, \boldsymbol{b}]$ is equal to the number of leading ones in $\boldsymbol{A}$ is exactly the assertion that $\boldsymbol{b}$ lies in the columnspace of $\boldsymbol{A}$. To see this, let r denote the number of leading ones in $\boldsymbol{A}$ and let r' denote the number of leading ones in $[\boldsymbol{A}, \boldsymbol{b}]$. Note that the r columns of $\boldsymbol{A}$ which contain leading ones are the first r standard basis vectors, $\boldsymbol{e}_1, \dots, \boldsymbol{e}_r$, for F^m. If $r' = r$ then clearly $\boldsymbol{b}$ lies in the span of these columns, and so in the columnspace of $\boldsymbol{A}$. Conversely, if $r' > r$, then $\boldsymbol{b} = \boldsymbol{e}_{r+1}$ and so $\boldsymbol{b}$ does not lie in the columnspace of $\boldsymbol{A}$. Thus indeed we see that the assertion in this part of the theorem is equivalent to the assertion that $\boldsymbol{b} \in \operatorname{colspace}(\boldsymbol{A})$. However, by Proposition 5.1.19, this is equivalent to asserting that $\boldsymbol{b} \in \operatorname{image}(\boldsymbol{A})$ which is the obvious necessary and sufficient condition of Proposition 5.1.55 for $\operatorname{Sol}(\boldsymbol{A}, \boldsymbol{b})$ to be nonempty.

(ii) This follows immediately from (5.9).

(iii) From part (ii) we know that if $r = n$ then there are no components that can be freely specified for elements of $\operatorname{Sol}(\boldsymbol{A}, \boldsymbol{b})$. This means that there is only one solution, and it is given by $x_j = b_j$, $j \in \{1, \dots, n\}$. ■

The first part of the theorem has the following reinterpretation.

5.1.59 Corollary (Rank and solutions of systems of linear equations) *Let* F *be a field, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$, *and let* $(\mathbf{A}, \mathbf{b}) \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{F}) \times \mathsf{F}^{\mathrm{m}}$ *be a system of linear equations. Then* $\operatorname{Sol}(\mathbf{A}, \mathbf{b})$ *is nonempty if and only if* $\operatorname{rank}([\mathbf{A}, \mathbf{b}]) = \operatorname{rank}(\mathbf{A})$.

Let us consider this for some examples.

5.1.60 Examples (Reduced row echelon form and solutions of systems of linear equations) We let F be a field. In each case we consider systems for which the augmented matrix is already in reduced row echelon form; doing the row operations is not the point here.

1. Consider the system of linear equations with the augmented matrix

$$\left[\begin{array}{ccc|c} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{array}\right].$$

Since the augmented matrix has three leading ones whereas $\boldsymbol{A}$ has only two, it follows that this system of linear equations, and any system possessing this system as its reduced row echelon form, has no solutions.

2. Next consider the augmented matrix

$$\left[\begin{array}{cccc|c} 0_{\mathsf{F}} & 1_{\mathsf{F}} & a_{13} & 0_{\mathsf{F}} & b_1 \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & b_2 \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{array}\right].$$

Here the system has solutions since the number of leading ones in the augmented matrix agrees with the number of leading ones in $\boldsymbol{A}$. Applying the second part of Theorem 5.1.58 (or simply solving the equations "by hand") one see that

$$\operatorname{Sol}(\boldsymbol{A}, \boldsymbol{b}) = \{(x_1, x_2, x_3, x_4) \mid x_1, x_3 \in \mathsf{F},\ x_2 = b_1 - a_{13}x_3,\ x_4 = b_2\}.$$

In particular, there is not a unique solution.

3. Finally consider the system with the augmented matrix

$$\left[\begin{array}{ccc|c} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & b_1 \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & b_2 \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & b_3 \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{array}\right].$$

This system has the same number of leading ones in $[\boldsymbol{A}, \boldsymbol{b}]$ and $\boldsymbol{A}$, so it possesses solutions. Moreover, the number of leading ones is equal to the number of columns, so there is a unique solution. Indeed, $\mathrm{Sol}(\boldsymbol{A}, \boldsymbol{b}) = \{(b_1, b_2, b_3)\}$. •

5.1.9 Notes

Material in Köthe on solutions of linear equations with arbitrary index sets.

Exercises

5.1.1 Prove Proposition 5.1.6. Are the sufficient conditions given for existence of the product $\boldsymbol{AB}$ also necessary?

5.1.2 Prove Proposition 5.1.7.

5.1.3 Let F be a field, let $m, n \in \mathbb{Z}_{>0}$, and consider $\mathrm{Mat}_{m\times n}(\mathsf{F})$, the $m \times n$ matrices over F in $I \times J$, thought of as an F-vector space via Corollary 5.1.8. Let $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ and $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_m\}$ be the standard bases for F^n and F^m, respectively.

(a) Give a natural basis for $\mathrm{Mat}_{m\times n}(\mathsf{F})$ defined using the standard bases for F^n and F^m.

(b) Conclude that $\dim_{\mathsf{F}}(\mathrm{Mat}_{m\times n}(\mathsf{F})) = mn$.

In the next exercise, you will use the following definition.

5.1.61 Definition (Lie algebra) A ***Lie algebra*** over a field F is a pair $(\mathfrak{g}, [\cdot,\cdot])$ where $\mathfrak{g}$ is an F-vector space that is equipped with a map from $\mathfrak{g}\times\mathfrak{g}$ to $\mathfrak{g}$, denoted by $(u, v) \mapsto [u, v]$, having the following three properties:

(i) for fixed $v \in \mathfrak{g}$ the maps $u \mapsto [u, v]$ and $u \mapsto [v, u]$ are endomorphisms of $\mathfrak{g}$;

(ii) $[v, v] = 0_{\mathfrak{g}}$ for each $v \in \mathfrak{g}$;

(iii) $[u, [v, w]] + [w, [u, v]] + [v, [w, u]] = 0_{\mathfrak{g}}$ for every $u, v, w \in \mathfrak{g}$ (***Jacobi identity***).

The product $[\cdot,\cdot]$ is called the ***Lie bracket*** on $\mathfrak{g}$. •

5.1.4 Let F be a field and let I be an index set. Define a map from $\mathrm{Mat}_{I\times I}(\mathsf{F}) \times \mathrm{Mat}_{I\times I}(\mathsf{F})$ to $\mathrm{Mat}_{I\times I}(\mathsf{F})$ by

$$(\boldsymbol{A}, \boldsymbol{B}) \mapsto [\boldsymbol{A}, \boldsymbol{B}] \triangleq \boldsymbol{AB} - \boldsymbol{BA}.$$

Answer the following questions.

(a) Show that $(\mathrm{Mat}_{I\times I}(\mathsf{F}), [\cdot,\cdot])$ is a Lie algebra.

(b) Show that if F does not have characteristic 2 then $[\boldsymbol{A}, \boldsymbol{B}] = -[\boldsymbol{B}, \boldsymbol{A}]$ for every $\boldsymbol{A}, \boldsymbol{B} \in \mathrm{Mat}_{I \times I}(\mathsf{F})$.

5.1.5 Let F be a field and let I be an index set.

(a) Show that the set of invertible column finite matrices in $\mathrm{Mat}_{I \times I}(\mathsf{F})$ is a group with product given by matrix multiplication.

In the case when $I = \{1, \dots, n\}$ this group of invertible matrices is denoted by $\mathsf{GL}(n; \mathsf{F})$ and is called the ***general linear group*** of order n over F.

(b) Is $\mathsf{GL}(n; \mathsf{F})$ a subalgebra of $\mathrm{Mat}_{n \times n}(\mathsf{F})$?

5.1.6 Let F be a field and let $n \in \mathbb{Z}_{>0}$. We consider $\mathrm{Mat}_{n \times n}(\mathsf{F})$ as a ring by Corollary 5.1.9.

(a) For what values of n is it true that $\mathrm{Mat}_{n \times n}(\mathsf{F})$ is a commutative ring?

(b) For what values of n is it true that $\mathrm{Mat}_{n \times n}(\mathsf{F})$ is an integral domain?

5.1.7 Prove Proposition 5.1.10.

5.1.8 Prove Proposition 5.1.19.

5.1.9 Prove Proposition 5.1.28.

Hint: *Show first that it suffices to consider matrices where one is obtained from the other by a single elementary row operation.*

5.1.10 Prove Proposition 5.1.39.

Section 5.2

Matrices over rings

In this section we generalise the discussion in Section 5.1 of matrices whose entries are elements of a field to matrices whose entries are elements of a ring. The character of this generalisation bears much resemblance to the extension from vector spaces (Section 4.5) to modules (Section 4.8). That is to say, many aspects of the generalisation are simply obtained by replacing the word “field” with the word “ring,” but there are other parts of the generalisation where care must be taken in that some things that hold for matrices over fields do not hold for matrices over rings. Due to the similarities with the presentation in Section 5.1, there will be some unavoidable redundancy.

Do I need to read this section? The material in this section can probably be omitted until it is needed. •

5.2.1 Matrices over rings: definitions and notation

Let us just hop into the definition.

5.2.1 Definition (Matrix over a ring) Let R be a ring and let I and J be index sets. A ***matrix over*** R in $I \times J$ is a map $\boldsymbol{A}\colon I \times J \to \mathsf{R}$. The expression $\boldsymbol{A}(i, j)$, $i \in I$, $j \in J$, is the **(i, j)*th component*** of the matrix $\boldsymbol{A}$, and is said to lie in the **i*th row*** and the **j*th column*** of $\boldsymbol{A}$. If, for each $i_0 \in I$ the set

$$\{j \in J \mid \boldsymbol{A}(i_0, j) \neq 0_{\mathsf{R}}\}$$

is finite, then $\boldsymbol{A}$ is ***row finite***, and, if for each $j_0 \in J$ the set

$$\{i \in I \mid \boldsymbol{A}(i, j_0) \neq 0_{\mathsf{R}}\}$$

is finite, then $\boldsymbol{A}$ is ***column finite***.

The set of matrices over R in $I \times J$ is denoted by $\mathrm{Mat}_{I\times J}(\mathsf{R})$. If $I = \{1, \dots, m\}$ and $J = \{1, \dots, n\}$, then a matrix over R in $I \times J$ is an **m × n *matrix***, and the set of $m \times n$ matrices is denoted by $\mathrm{Mat}_{m\times n}(\mathsf{R})$. •

As with matrices over fields, and indeed even more so for matrices over rings, the case of most interest for matrices over rings will be the case of matrices with finitely many rows and columns. As for fields, these will be represented by an array of the form (5.1), only now the entries are allowed to be elements of a ring.

Of course, all of the examples of matrices over fields given in Example 5.1.2 are also examples of matrices over rings. We thus only consider examples here that extend the existing examples.

5.2.2 Examples (Matrices over rings)

1. As with matrices over fields, the ***zero matrix*** over a ring R in $I \times J$ is the matrix $\boldsymbol{0}_{I\times J}$ defined by $\boldsymbol{0}_{I\times J}(i,j) = 0_{\mathsf{R}}$. The $m \times n$ zero matrix is denoted by $\boldsymbol{0}_{m\times n}$.
2. A ***square*** matrix over a ring R is any matrix in $I \times I$ for an index set I. A square matrix $\boldsymbol{A} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ is ***diagonal*** if $\boldsymbol{A}(i_1, i_2) = 0_{\mathsf{R}}$ whenever $i_1 \neq i_2$. If R is a unit ring, the special diagonal matrix $\boldsymbol{I}_I \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ defined by
$$\boldsymbol{I}_I(i_1, i_2) = \begin{cases} 1_{\mathsf{R}}, & i_1 = i_2, \\ 0_{\mathsf{R}}, & i_1 \neq 1_2 \end{cases}$$
is the ***identity matrix***. If $I = \{1, \dots, n\}$ then we denote $\boldsymbol{I}_n = \boldsymbol{I}_I$.
3. The most common example of matrices over rings we will encounter occurs when $\mathsf{R} = \mathsf{F}[\xi]$ is the polynomial ring over a field F. Note that an $m \times n$ matrix over $\mathsf{F}[\xi]$ is represented as
$$\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{m1} & P_{m2} & \cdots & P_{mn} \end{bmatrix},$$
for $P_{ij} \in \mathsf{F}[\xi]$, $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$. •

One can, of course, define the ***transpose*** of $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ as the matrix $\boldsymbol{A}^T \in \mathrm{Mat}_{J\times I}(\mathsf{R})$ given by $\boldsymbol{A}^T(j,i) = \boldsymbol{A}(i,j)$.

The partitioning of matrices over a ring R can be done just as for matrices over fields. There is also an identical notion of a matrix being block diagonal. One can also define the row and column vectors for a matrix over R as elements of R^J and R^I, respectively. We refer the reader to the discussion for matrices over fields.

5.2.2 The algebra of matrices over rings

Because rings may not have an identity element and may not be commutative, one has to exercise a little more care in considering the algebraic structure of matrices over rings when compared to matrices over fields. Nonetheless, the basic ingredients are the same. In particular, the starting point of the matrix product is the same, except that one now needs to allow for the fact that rings may not be commutative.

5.2.3 Definition (Sum and product of matrices over rings) Let R be a ring and let I, J, and K be index sets.

(i) If $\boldsymbol{A}, \boldsymbol{B} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ then the ***sum*** of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{A} + \boldsymbol{B} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ defined by
$$(\boldsymbol{A} + \boldsymbol{B})(i,j) = \boldsymbol{A}(i,j) + \boldsymbol{B}(i,j).$$

(ii) If $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ and $\boldsymbol{B} \in \mathrm{Mat}_{J\times K}(\mathsf{R})$ then the ***product*** of $\boldsymbol{A}$ and $\boldsymbol{B}$ is the matrix $\boldsymbol{AB} \in \mathrm{Mat}_{I\times K}(\mathsf{R})$ defined by

$$(\boldsymbol{AB})(i,k) = \sum_{j\in J} \boldsymbol{A}(i,j)\boldsymbol{B}(j,k),$$

and is defined whenever the sum is finite.

(iii) If $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ and $r \in \mathsf{R}$ then ***left multiplication*** (resp. ***right multiplication***) of $\boldsymbol{A}$ by a is the matrix $r\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ (resp. $\boldsymbol{A}r \in \mathrm{Mat}_{I\times J}(\mathsf{R})$) defined by $(r\boldsymbol{A})(i,j) = r(\boldsymbol{A}(i,j))$ (resp. $(\boldsymbol{A}r)(i,j) = (\boldsymbol{A}(i,j))r$). •

As with matrices over fields, there are simple sufficient conditions which ensure that the product of matrices makes sense. The proof for matrices over rings follows that for matrices over fields.

5.2.4 Proposition (Definability of the product of matrices over rings) *If* R *is a ring and if* $\mathrm{I}, \mathrm{J},$ *and* K *are index sets, then the following statements hold for* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *and* $\mathbf{B} \in \mathrm{Mat}_{\mathrm{J}\times\mathrm{K}}(\mathsf{R})$*:*

(i) the product $\mathbf{AB}$ *is defined if* $\mathbf{A}$ *is row finite;*

(ii) the product $\mathbf{AB}$ *is defined if* $\mathbf{B}$ *is column finite.*

Moreover, if both $\mathbf{A}$ *and* $\mathbf{B}$ *are column (resp. row) finite, then* $\mathbf{AB}$ *is column (resp. row) finite.*

The sum and product for matrices over rings have the following properties; note that there are some differences from the case of matrices over fields in that a ring may not be a unit ring and in that we have two possibilities for products of matrices with elements of the ring. Nonetheless, the proof of the following result is essentially like that of Proposition 5.1.7 for matrices over fields.

5.2.5 Proposition (Properties of sum and product of matrices over rings) *Let* R *be a ring, let* $\mathrm{I}, \mathrm{J}, \mathrm{K},$ *and* L *be index sets, and let* $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$*,* $\mathbf{B}_1, \mathbf{B}_2 \in \mathrm{Mat}_{\mathrm{J}\times\mathrm{K}}(\mathsf{R})$*,* $\mathbf{C}_1 \in \mathrm{Mat}_{\mathrm{K}\times\mathrm{L}}(\mathsf{R})$*, and* $\mathrm{r}_1, \mathrm{r}_2 \in \mathsf{R}$*. Then the following equalities hold:*

(i) $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{A}_2 + \mathbf{A}_1$*;*

(ii) $(\mathbf{A}_1 + \mathbf{A}_2) + \mathbf{A}_3 = \mathbf{A}_1 + (\mathbf{A}_2 + \mathbf{A}_3)$*;*

(iii) $\mathbf{A}_1 + \mathbf{0}_{\mathrm{I}\times\mathrm{J}} = \mathbf{A}_1$*;*

(iv) if $-\mathbf{A}_1 \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *is defined by* $(-\mathbf{A}_1)(\mathrm{i},\mathrm{j}) = -(\mathbf{A}_1(\mathrm{i},\mathrm{j}))$*,* $\mathrm{i} \in \mathrm{I}$*,* $\mathrm{j} \in \mathrm{J}$*, then* $\mathbf{A}_1 + (-\mathbf{A}_1) = \mathbf{0}_{\mathrm{I}\times\mathrm{J}}$*;*

(v) if $\mathbf{A}_1$ *is row finite, or if* $\mathbf{B}_1$ *and* $\mathbf{B}_2$ *are column finite, then* $\mathbf{A}_1(\mathbf{B}_1 + \mathbf{B}_2) = \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_1\mathbf{B}_2$*;*

(vi) if $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row finite, or if* $\mathbf{B}_1$ *is column finite, then* $(\mathbf{A}_1 + \mathbf{A}_2)\mathbf{B}_1 = \mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_1$*;*

(vii) if $\mathbf{A}_1$ *and* $\mathbf{B}_1$ *are row finite, or if* $\mathbf{B}_1$ *and* $\mathbf{C}_1$ *are column finite, then* $(\mathbf{A}_1\mathbf{B}_1)\mathbf{C}_1 = \mathbf{A}_1(\mathbf{B}_1\mathbf{C}_1)$*;*

(viii) if R *is a unit ring then* $\mathbf{I}_{\mathrm{J}}\mathbf{A}_1 = \mathbf{A}_1\mathbf{I}_{\mathrm{I}} = \mathbf{A}_1$*;*

(ix) $\mathrm{r}_1(\mathrm{r}_2\mathbf{A}_1) = (\mathrm{r}_1\mathrm{r}_2)\mathbf{A}_1$*;*

(x) $(\mathbf{A}_1 r_1)r_2 = \mathbf{A}_1(r_1 r_2)$;

(xi) $(r_1 + r_2)\mathbf{A}_1 = r_1\mathbf{A}_1 + r_2\mathbf{A}_1$;

(xii) $\mathbf{A}_1(r_1 + r_2) = \mathbf{A}_1 r_1 + \mathbf{A}_1 r_2$;

(xiii) $r_1(\mathbf{A}_1 + \mathbf{A}_2) = r_1\mathbf{A}_1 + r_1\mathbf{A}_2$;

(xiv) $(\mathbf{A}_1 + \mathbf{A}_2)r_1 = \mathbf{A}_1 r_1 + \mathbf{A}_2 r_1$;

(xv) *if* R *is a unit ring then* $1_{\mathsf{R}}\mathbf{A}_1 = \mathbf{A}_1 1_{\mathsf{R}} = \mathbf{A}_1$.

Thus, for matrices over rings we have the following structure, echoing Corollaries 5.1.8 and 5.1.9 for matrices over fields.

5.2.6 Corollary (Matrices over rings as elements of a module) *If* R *is a ring and if* I *and* J *are index sets, then*

(i) $\mathrm{Mat}_{I\times J}(\mathsf{R})$ *is a left* R*-module with addition given by the sum of matrices and with multiplication being given by left multiplication of a matrix by a scalar,*

(ii) $\mathrm{Mat}_{I\times J}(\mathsf{R})$ *is a right* R*-module with addition given by the sum of matrices and with multiplication being given by right multiplication of a matrix by a scalar, and*

(iii) *if* R *is additionally a unit ring, then* $\mathrm{Mat}_{I\times J}(\mathsf{R})$ *is a unity left* R*-module and a unity right* R*-module.*

5.2.7 Corollary (Matrices over rings as elements of an algebra) *If* R *is a ring and if* I *is an index set, then*

(i) *the set of column finite matrices in* $\mathrm{Mat}_{I\times I}(\mathsf{R})$ *is a left* R*-algebra with the left* R*-module structure of Corollary 5.2.6 and with the product given by the product of matrices,*

(ii) *the set of column finite matrices in* $\mathrm{Mat}_{I\times I}(\mathsf{R})$ *is a right* R*-algebra with the right* R*-module structure of Corollary 5.2.6 and with the product given by the product of matrices, and*

(iii) *if* R *is additionally a unit ring, then the set of column finite matrices in* $\mathrm{Mat}_{I\times J}(\mathsf{R})$ *is a left unity* R*-algebra and a right unity* R*-algebra whose ring structure has* $\mathbf{I}_I$ *as a unity element.*

We have the following result which characterises the properties of transpose with respect to matrix algebra. Note that one needs to be careful in generalising Proposition 5.1.10 since R may not be commutative.

5.2.8 Proposition (Matrix algebra and matrix transpose) *Let* R *be a field, let* I *and* J *be index sets, and let* $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{I\times J}(\mathsf{R})$*. Then the following statements hold:*

(i) $(\mathbf{A}_1 + \mathbf{A}_2)^T = \mathbf{A}_1^T + \mathbf{A}_2^T$;

(ii) *if* R *is additionally commutative, then the product* $\mathbf{A}_1\mathbf{A}_2$ *is defined if and only if the product* $\mathbf{A}_2^T\mathbf{A}_1^T$ *is defined, and when these are defined we have* $(\mathbf{A}_1\mathbf{A}_2)^T = \mathbf{A}_2^T\mathbf{A}_1^T$.

5.2.3 Matrices as homomorphisms

Just as one can, under suitable hypotheses, associate to a matrix over a field a linear map between certain vector spaces, to a matrix over a ring one can, again under suitable restrictions, associate to a matrix over a ring a homomorphism of certain modules. Due to the fact that rings may not be commutative, we need to make modifications in our definitions of matrix-vector product. We recall from Example 4.8.36 and Notation 4.8.37 the R-modules R^I and R^I_0, defined using an index set I.

5.2.9 Definition (Matrix-vector product) Let R be a ring and let I and J be index sets. For a matrix $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ and $\boldsymbol{x} \in \mathsf{R}^J$,

(i) the ***right product*** of $\boldsymbol{A}$ and $\boldsymbol{x}$ is the element $\boldsymbol{Ax}$ of R^I given by

$$(\boldsymbol{Ax})(i) = \sum_{j\in J} \boldsymbol{A}(i,j)\boldsymbol{x}(j),$$

and

(ii) the ***left product*** of $\boldsymbol{A}$ and $\boldsymbol{x}$ is the element $\boldsymbol{xA}$ of R^I given by

$$(\boldsymbol{xA})(i) = \sum_{j\in J} \boldsymbol{x}(j)\boldsymbol{A}(i,j),$$

and these are defined whenever the sums involved are finite. •

Of course, one needs, as with Proposition 5.1.12 for matrices over fields, conditions which guarantee the definedness of the matrix-vector product.

5.2.10 Proposition (Definability and properties of the matrix-vector product) *Let* R *be a ring, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *and* $\mathbf{x} \in \mathsf{R}^{\mathrm{J}}$*. Then the following statements hold:*

(i) if $\mathbf{x} \in \mathsf{R}^{\mathrm{J}}_0$ *then* $\mathbf{Ax}$ *and* $\mathbf{xA}$ *are defined;*

(ii) if $\mathbf{x} \in \mathsf{R}^{\mathrm{J}}_0$ *and if* $\mathbf{A}$ *is column finite then* $\mathbf{Ax}$ *and* $\mathbf{xA}$ *are defined and are elements of* $\mathsf{R}^{\mathrm{I}}_0$*;*

(iii) if $\mathbf{A}$ *is row finite then* $\mathbf{Ax}$ *and* $\mathbf{xA}$ *are defined.*

Now we state our main result in this section which associates a homomorphism to a matrix over a ring. Because of the two possible ways of defining matrix-vector product, this leads to multiple versions of homomorphisms adapted to either the left or right module structure of R^I (see Example 4.8.2–7). In the case where the ring is commutative, the result boils down to a transcription of Theorem 5.1.13, exchanging “field” with “commutative ring” (see Exercise 5.2.1).

5.2.11 Theorem (Matrices as homomorphisms) *Let* R *be a unit ring, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$*. Then the following statements hold:*

(i) if $\mathbf{A}$ *is column finite, then the map* $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ *is an* R*-homomorphism from the right* R*-module* $\mathsf{R}_0^{\mathrm{J}}$ *to the right* R*-module* $\mathsf{R}_0^{\mathrm{I}}$*;*

(ii) if $\mathbf{A}$ *is column finite, then the map* $\mathbf{x} \mapsto \mathbf{x}\mathbf{A}$ *is an* R*-homomorphism from the left* R*-module* $\mathsf{R}_0^{\mathrm{J}}$ *to the left* R*-module* $\mathsf{R}_0^{\mathrm{I}}$*;*

(iii) if $\mathbf{A}$ *is row finite, then the map* $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ *is an* R*-homomorphism from the right* R*-module* R^{J} *to the right* R*-module* R^{I}*;*

(iv) if $\mathbf{A}$ *is row finite, then the map* $\mathbf{x} \mapsto \mathbf{x}\mathbf{A}$ *is an* R*-homomorphism from the left* R*-module* R^{J} *to the left* R*-module* R^{I}*.*

Moreover, the following statements also hold:

(v) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^{\mathrm{J}}; \mathsf{R}_0^{\mathrm{I}})$ *is a homomorphism of right* R*-modules, then there exists a unique column finite matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ *for all* $\mathbf{x} \in \mathsf{R}_0^{\mathrm{J}}$*;*

(vi) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^{\mathrm{J}}; \mathsf{R}_0^{\mathrm{I}})$ *is a homomorphism of left* R*-modules, then there exists a unique column finite matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{x}\mathbf{A}$ *for all* $\mathbf{x} \in \mathsf{R}_0^{\mathrm{J}}$*;*

(vii) if J *is finite, then, if* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^{\mathrm{J}}; \mathsf{R}^{\mathrm{I}})$ *is a homomorphism of right* R*-modules, then there exists a unique (necessarily row finite) matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ *for all* $\mathbf{x} \in \mathsf{R}^{\mathrm{J}}$*;*

(viii) if J *is finite, then, if* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^{\mathrm{J}}; \mathsf{R}^{\mathrm{I}})$ *is a homomorphism of left* R*-modules, then there exists a unique (necessarily row finite) matrix* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *such that* $\mathsf{L}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ *for all* $\mathbf{x} \in \mathsf{R}^{\mathrm{J}}$*.*

Proof Much of the proof follows along the same lines as the proof of Theorem 5.1.13. Therefore, in the proof here we shall only point out the parts of the proof that differ from Theorem 5.1.13.

(i) We let $\mathsf{L}_{\boldsymbol{A}}$ be the given map from R_0^J to R_0^I. The only place where the proof differs from Theorem 5.1.13(i) is in the linearity with respect to multiplication by elements of the ring. For this we check that for $r \in \mathsf{R}$ and $\boldsymbol{x} \in \mathsf{R}_0^J$ we have

$$\mathsf{L}_{\boldsymbol{A}}(\boldsymbol{x}r)(i) = \sum_{j\in J} \boldsymbol{A}(i,j)(\boldsymbol{x}r)(j) = \left(\sum_{j\in J} \boldsymbol{A}(i,j)\boldsymbol{x}(j)\right) r = (\mathsf{L}_{\boldsymbol{A}}(\boldsymbol{x}))r,$$

as desired.

(ii) The proof here follows, *mutatis mutandis*, along the lines of the preceding part of the proof.

(iii) Again, the only thing to check that differs from the proof of Theorem 5.1.13(ii) is the linearity with respect to multiplication by elements of the ring. But this goes exactly like the corresponding computation for part (i).

(iv) The proof here follows, *mutatis mutandis*, along the lines of the preceding part of the proof.

(v) As in the proof of part (iii) of Theorem 5.1.13, we let $\{\boldsymbol{e}_i\}_{i\in I}$ and $\{\boldsymbol{f}_j\}_{j\in J}$ be the standard bases for R_0^I and R_0^J, respectively. For $j \in J$ we have

$$\mathsf{L}(\boldsymbol{f}_j) = \boldsymbol{e}_{i_1} a_{i_1 j} + \cdots + \boldsymbol{e}_{i_{k_j}} a_{i_{k_j} j}$$

for some unique $k_j \in \mathbb{Z}_{\geq 0}$, some unique basis elements $\{\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_{k_j}}\}$, and some unique nonzero $a_{i_1 j}, \dots, a_{i_{k_j} j} \in \mathsf{R}$ since $\{\boldsymbol{e}_i\}_{i \in I}$ is a basis for R^I_0. We then define $\boldsymbol{A} \in \mathrm{Mat}_{I \times J}(\mathsf{R})$ by

$$\boldsymbol{A}(i, j) = \begin{cases} a_{ij}, & i \in \{i_1, \dots, i_{k_j}\}, \\ 0_{\mathsf{R}}, & \text{otherwise.} \end{cases}$$

To check that $\mathsf{L}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, let $\boldsymbol{x} \in \mathsf{R}^J_0$ and write

$$\boldsymbol{x} = \boldsymbol{f}_{j_1} x_1 + \dots + \boldsymbol{f}_{j_k} x_k$$

for some $x_1, \dots, x_k \in \mathsf{R}$. Now let $\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_m}$ be standard basis elements with the property that

$$\mathsf{L}(\boldsymbol{f}_{j_l}) \in \mathrm{span}_{\mathsf{R}}(\boldsymbol{e}_{i_1}, \dots, \boldsymbol{e}_{i_m})$$

for each $l \in \{1, \dots, k\}$. Then

$$\begin{aligned} \mathsf{L}(\boldsymbol{x}) &= \mathsf{L}(\boldsymbol{f}_{j_1} x_1 + \dots + \boldsymbol{f}_{j_k} x_k) \\ &= (\mathsf{L}(\boldsymbol{f}_{j_1}))x_1 + \dots + (\mathsf{L}(\boldsymbol{f}_{j_k}))x_k \\ &= \boldsymbol{e}_{i_1} a_{i_1 j_1} x_1 + \dots + \boldsymbol{e}_{i_m} a_{i_m j_1} x_1 + \dots + \boldsymbol{e}_{i_1} a_{i_1 j_k} x_k + \dots + \boldsymbol{e}_{i_m} a_{i_m j_k} x_k \\ &= \boldsymbol{A}\boldsymbol{x}, \end{aligned}$$

as desired.

(vi) The proof here follows, *mutatis mutandis*, along the lines of the proof of the preceding part of the theorem.

(vii) Suppose that $J = \{1, \dots, n\}$ so that $\mathsf{R}^J = \mathsf{R}^n$, and let $\{\boldsymbol{f}_1, \dots, \boldsymbol{f}_n\}$ be the standard basis for R^n. Let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^n; \mathsf{R}^I)$ and define $\boldsymbol{A} \in \mathrm{Mat}_{I \times J}(\mathsf{R})$ by asking that, for $(i, j) \in I \times J$, $\mathsf{L}(\boldsymbol{f}_j)(i) = \boldsymbol{A}(i, j)$. We claim that $\mathsf{L}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathsf{R}^n$. Indeed, for $\boldsymbol{x} \in \mathsf{R}^n$ write

$$\boldsymbol{x} = \boldsymbol{f}_1 \boldsymbol{x}(1) + \dots + \boldsymbol{f}_n \boldsymbol{x}(n),$$

and then compute, for $i \in I$,

$$\begin{aligned} \mathsf{L}(\boldsymbol{x})(i) &= \mathsf{L}(\boldsymbol{f}_1 \boldsymbol{x}(1) + \dots + \boldsymbol{f}_n \boldsymbol{x}(n))(i) \\ &= (\mathsf{L}(\boldsymbol{f}_1))(i)\boldsymbol{x}(1) + \dots + (\mathsf{L}(\boldsymbol{f}_n)(i))\boldsymbol{x}(n) \\ &= \boldsymbol{A}(i, 1)\boldsymbol{x}(1) + \dots + \boldsymbol{A}(i, n)\boldsymbol{x}(n) = (\boldsymbol{A}\boldsymbol{x})(i), \end{aligned}$$

as desired.

(viii) The proof here follows like that for the preceding part of the proof. ■

We shall adopt the following notation when thinking of a matrix $\boldsymbol{A}$ over a unit ring as a homomorphism.

1. We denote by $\boldsymbol{A}_r \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^J_0; \mathsf{R}^I_0)$ (if $\boldsymbol{A}$ is column finite) or $\boldsymbol{A}_r \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^J; \mathsf{R}^I)$ (if $\boldsymbol{A}$ is row finite) the homomorphism of right R-modules. We write $\boldsymbol{A}_r(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ in this case.
2. We denote by $\boldsymbol{A}_l \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^J_0; \mathsf{R}^I_0)$ (if $\boldsymbol{A}$ is column finite) or $\boldsymbol{A}_l \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^J; \mathsf{R}^I)$ (if $\boldsymbol{A}$ is row finite) the homomorphism of left R-modules. We write $\boldsymbol{A}_l(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{A}$ in this case.

When R is commutative, then $\boldsymbol{A}_r = \boldsymbol{A}_l$ when one makes the natural correspondence between the right R-modules R^J and R^I and the left R-modules R^J and R^I (see Example 4.8.2–8).

The product of matrices corresponds to the composition of homomorphisms, exactly in analogy to the case of matrices over fields, except that one now needs notions for both the left and right matrix-vector product.

5.2.12 Proposition (Matrix product and composition of homomorphisms) *Let* R *be a unit ring, let* I, J, *and* K *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *and* $\mathbf{B} \in \mathrm{Mat}_{\mathrm{J}\times\mathrm{K}}(\mathsf{R})$. *Then the following statements hold:*

(i) if $\mathbf{A}$ *and* $\mathbf{B}$ *are column finite then the matrix corresponding to the composition of the homomorphisms* $\mathbf{A}_\mathrm{r} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}_0^\mathrm{J};\mathsf{R}_0^\mathrm{I})$ *and* $\mathbf{B}_\mathrm{r} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}_0^\mathrm{K};\mathsf{R}_0^\mathrm{J})$ *is* $\mathbf{AB}$*;*

(ii) if $\mathbf{A}$ *and* $\mathbf{B}$ *are column finite then the matrix corresponding to the composition of the homomorphisms* $\mathbf{A}_\mathrm{l} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}_0^\mathrm{J};\mathsf{R}_0^\mathrm{I})$ *and* $\mathbf{B}_\mathrm{l} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}_0^\mathrm{K};\mathsf{R}_0^\mathrm{J})$ *is* $\mathbf{AB}$*;*

(iii) if $\mathbf{A}$ *and* $\mathbf{B}$ *are row finite then the matrix corresponding to the composition of the homomorphisms* $\mathbf{A}_\mathrm{r} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}^\mathrm{J};\mathsf{R}^\mathrm{I})$ *and* $\mathbf{B}_\mathrm{r} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}^\mathrm{K};\mathsf{R}^\mathrm{J})$ *is* $\mathbf{AB}$*;*

(iv) if $\mathbf{A}$ *and* $\mathbf{B}$ *are row finite then the matrix corresponding to the composition of the homomorphisms* $\mathbf{A}_\mathrm{r} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}^\mathrm{J};\mathsf{R}^\mathrm{I})$ *and* $\mathbf{B}_\mathrm{r} \in \mathrm{Hom}_\mathsf{R}(\mathsf{R}^\mathrm{K};\mathsf{R}^\mathrm{J})$ *is* $\mathbf{AB}$.

Proof (i) Let $\{e_i\}_{i\in I}$, $\{f_j\}_{j\in J}$, and $\{g_k\}_{k\in K}$ be the standard bases for R_0^I, R_0^J, and R_0^K, respectively. We compute

$$\boldsymbol{A}_r \circ \boldsymbol{B}_r(\boldsymbol{g}_k) = \boldsymbol{A}_r\left(\sum_{j\in J} \boldsymbol{f}_j \boldsymbol{B}(j,k)\right) = \sum_{j\in J} \boldsymbol{A}_r(\boldsymbol{f}_j)\boldsymbol{B}(j,k) = \sum_{j\in J}\sum_{i\in I} \boldsymbol{e}_i \boldsymbol{A}(i,j)\boldsymbol{B}(j,k),$$

where all sums are finite since $\boldsymbol{A}$ and $\boldsymbol{B}$ are column finite. This directly gives

$$\boldsymbol{A}_r \circ \boldsymbol{B}_r(\boldsymbol{g}_k) = \sum_{i\in I} \boldsymbol{e}_i(\boldsymbol{AB})(i,k),$$

using the definition of right matrix product. A reference to the proof of Theorem 5.2.11 shows that $\boldsymbol{AB}$ is the matrix associated to the homomorphism $\boldsymbol{A}_r \circ \boldsymbol{B}_r$, as desired.

(ii) The proof here goes exactly as in the previous part of the proof, but using the left module structure.

(iii) Let $\boldsymbol{z} \in \mathsf{F}^K$ and $i \in I$ and compute

$$\begin{aligned}(\boldsymbol{A}_r \circ \boldsymbol{B}_r)(\boldsymbol{z})(i) &= \sum_{j\in J} \boldsymbol{A}(i,j)(\boldsymbol{Bz})(j) = \sum_{j\in J} \boldsymbol{A}(i,j)\left(\sum_{k\in K} \boldsymbol{B}(j,k)\boldsymbol{z}(k)\right)\\ &= \sum_{k\in K} (\boldsymbol{AB})(i,k)\boldsymbol{z}(k) = (\boldsymbol{AB})_r(\boldsymbol{z})(i),\end{aligned}$$

giving $\boldsymbol{A}_r \circ \boldsymbol{B}_r(\boldsymbol{z}) = (\boldsymbol{AB})_r(\boldsymbol{z})$, as desired.

(iv) The proof here goes exactly as in the previous part of the proof, but using the left module structure. ■

As with matrices over rings, the transpose of a matrix can be regarded as a homomorphism. To give a characterisation of the transpose as a homomorphism, note that—analogously to the situation with fields discussed before Theorem 5.1.16—elements of R^I are elements of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^I_0;\mathsf{R})$, this being the case for both the left and right module structure on R^I_0. For $\boldsymbol{y} \in \mathsf{R}^I$ we denote by $\mathsf{L}_{r,\boldsymbol{y}} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^I_0;\mathsf{R})$ the homomorphism of right R-modules defined by

$$\mathsf{L}_{r,\boldsymbol{y}}(\boldsymbol{x}) = \sum_{i\in I} \boldsymbol{y}(i)\boldsymbol{x}(i)$$

and we denote by $\mathsf{L}_{l,\boldsymbol{y}} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^I_0;\mathsf{R})$ the homomorphism of left R-modules defined by

$$\mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{x}) = \sum_{i\in I} \boldsymbol{x}(i)\boldsymbol{y}(i).$$

With this notation, we have the following result.

5.2.13 Theorem (Transpose as a homomorphism) *Let* R *be a unit ring, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *be column finite (so defining homomorphisms* $\mathbf{A}_{\mathrm{r}} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^{\mathrm{J}}_0;\mathsf{R}^{\mathrm{I}}_0)$ *and* $\mathbf{A}_{\mathrm{l}} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^{\mathrm{J}}_0;\mathsf{R}^{\mathrm{I}}_0)$ *of right and left* R*-modules, respectively). Then the following statements hold:*

(i) the map $\mathbf{y} \mapsto \mathbf{A}^{\mathrm{T}}\mathbf{y}$ *is a homomorphism of the right* R*-modules* R^{I} *and* R^{J}*, and furthermore satisfies*

$$\mathsf{L}_{\mathrm{l},\mathbf{A}^{\mathrm{T}}\mathbf{y}}(\mathbf{x}) = \mathsf{L}_{\mathrm{l},\mathbf{y}}(\mathbf{x}\mathbf{A});$$

(ii) the map $\mathbf{y} \mapsto \mathbf{y}\mathbf{A}^{\mathrm{T}}$ *is a homomorphism of the left* R*-modules* R^{I} *and* R^{J}*, and furthermore satisfies*

$$\mathsf{L}_{\mathrm{r},\mathbf{y}\mathbf{A}^{\mathrm{T}}}(\mathbf{x}) = \mathsf{L}_{\mathrm{r},\mathbf{y}}(\mathbf{A}\mathbf{x}).$$

Proof That the maps $\boldsymbol{y} \mapsto \boldsymbol{A}^T\boldsymbol{y}$ and $\boldsymbol{y} \mapsto \boldsymbol{y}\boldsymbol{A}^T$ are homomorphisms of right and left R-modules, respectively, as claimed follows from Theorem 5.2.11 since $\boldsymbol{A}^T$ is row finite. We also directly compute that both $\mathsf{L}_{l,\boldsymbol{A}^T\boldsymbol{y}}(\boldsymbol{x})$ and $\mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{A}\boldsymbol{x})$ are given by

$$\sum_{i\in I}\sum_{j\in J} \boldsymbol{x}(j)\boldsymbol{A}(i,j)\boldsymbol{y}(i)$$

and that both $\mathsf{L}_{r,\boldsymbol{y}\boldsymbol{A}^T}(\boldsymbol{x})$ and $\mathsf{L}_{r,\boldsymbol{y}}(\boldsymbol{x}\boldsymbol{A})$ are given by

$$\sum_{i\in I}\sum_{j\in J} \boldsymbol{y}(j)\boldsymbol{A}(i,j)\boldsymbol{x}(i). \qquad ■$$

In Proposition 5.1.17 we gave some complementary properties of matrices over fields and their transposes. Specifically we showed that a column finite matrix $\boldsymbol{A}$ over a field is injective (resp. surjective) if and only if $\boldsymbol{A}^T$ is surjective (resp. injective). For matrices over rings, one must first take care to properly account for the left and right module structures. But then, after one does this, only one half of Proposition 5.1.17 remains true for matrices over rings as we show in the following result and example.

5.2.14 Proposition (Properties of transpose as a homomorphism) *Let* R *be a unit ring, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ *be column finite. Then*

(i) $\mathbf{A}_r^T$ *is injective if* $\mathbf{A}_l$ *is surjective and*

(ii) $\mathbf{A}_l^T$ *is injective if* $\mathbf{A}_r$ *is surjective.*

Proof First we give a lemma.

1 Lemma *The following statements hold:*

(i) $\ker(\mathbf{A}_r^T) = \{\mathbf{y} \in \mathsf{R}^I \mid \mathsf{L}_{l,\mathbf{y}}(\mathbf{z}) = 0_{\mathsf{R}}$ *for all* $\mathbf{z} \in \mathrm{image}(\mathbf{A}_l)\}$*;*

(ii) $\ker(\mathbf{A}_l^T) = \{\mathbf{y} \in \mathsf{R}^I \mid \mathsf{L}_{r,\mathbf{y}}(\mathbf{z}) = 0_{\mathsf{R}}$ *for all* $\mathbf{z} \in \mathrm{image}(\mathbf{A}_r)\}$.

Proof We shall only prove the first part of the lemma, since the second part is proved in the same manner. We rely on the fact that $\mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{x}) = 0_{\mathsf{R}}$ for all $\boldsymbol{x} \in \mathsf{R}_0^I$ if and only if $\boldsymbol{y} = \mathbf{0}_{\mathsf{R}^I}$. Clearly the "if" part of this statement is true. For the "only if" part, suppose that $\boldsymbol{y} \in \mathsf{R}^I$ is nonzero. Then $\boldsymbol{y}(i_0) \neq 0_{\mathsf{R}}$ for some $i_0 \in I$. It then follows that, if $\{e_i\}_{i\in I}$ is the standard basis for R_0^I, we have $\mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{e}_{i_0}) = \boldsymbol{y}(i_0) \neq 0_{\mathsf{R}}$, giving the claim.

With this fact and Theorem 5.2.13 in mind, we have the following computation:

$$\begin{aligned}\ker(\boldsymbol{A}_r^T) &= \{\boldsymbol{y} \in \mathsf{R}^I \mid \boldsymbol{A}^T\boldsymbol{y} = \mathbf{0}_{\mathsf{R}^J}\}\\ &= \{\boldsymbol{y} \in \mathsf{R}^I \mid \mathsf{L}_{l,\boldsymbol{A}^T\boldsymbol{y}}(\boldsymbol{x}) = 0_{\mathsf{R}} \text{ for all } \boldsymbol{x} \in \mathsf{R}_0^J\}\\ &= \{\boldsymbol{y} \in \mathsf{R}^I \mid \mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{x}\boldsymbol{A}) = 0_{\mathsf{R}} \text{ for all } \boldsymbol{x} \in \mathsf{R}_0^J\}\\ &= \{\boldsymbol{y} \in \mathsf{R}^I \mid \mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{z}) = 0_{\mathsf{R}} \text{ for all } \boldsymbol{z} \in \mathrm{image}(\boldsymbol{A}_l)\},\end{aligned}$$

as desired. ▼

Continuing with the proof, we prove only the first assertion since the second follows *mutatis mutandis.*

Suppose that $\boldsymbol{A}_l$ is surjective and let $\boldsymbol{y} \in \ker(\boldsymbol{A}_r^T)$. Since $\boldsymbol{A}_l$ is surjective the lemma tells us that $\mathsf{L}_{l,\boldsymbol{y}}(\boldsymbol{z}) = 0_{\mathsf{R}}$ for every $\boldsymbol{z} \in \mathsf{R}_0^I$. In particular, taking $\boldsymbol{z} = \boldsymbol{e}_i$ for $i \in I$ shows that $\boldsymbol{y}(i) = 0_{\mathsf{R}}$ for $i \in I$. Thus $\boldsymbol{y} = \mathbf{0}_{\mathsf{R}^I}$, and so that $\boldsymbol{A}_r^T$ is injective by Exercise 4.8.3. ■

The following example shows that the converses in either part of the preceding proposition do not generally hold, even for matrices over rather ordinary rings.

5.2.15 Example (Injective matrix with a non-surjective transpose) We take $\mathsf{R} = \mathbb{Z}$ and consider the matrix $\boldsymbol{A} \in \mathrm{Mat}_{2\times 2}(\mathbb{Z})$ given by

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

Since $\mathbb{Z}$ is commutative, we do not need to concern ourselves with the distinction between the left and right module structures, and so we denote the element of $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}^2;\mathbb{Z}^2)$ associated with $\boldsymbol{A}$ (resp. $\boldsymbol{A}^T$) simply by $\boldsymbol{A}$ (resp. $\boldsymbol{A}^T$). We note that $\boldsymbol{A}^T = \boldsymbol{A}$. We claim that $\boldsymbol{A}$, as a homomorphism from $\mathbb{Z}^2$ to $\mathbb{Z}^2$, is injective but that $\boldsymbol{A}^T$ is not surjective. To see that $\boldsymbol{A}$ is injective compute $\boldsymbol{A}(j_1, j_2) = (j_1, 2j_2)$, and that $(j_1, 2j_2) = (0, 0)$ only if $j_1 = j_2 = 0$. Thus $\ker(\boldsymbol{A}) = \{(0, 0)\}$ and so $\boldsymbol{A}$ is injective by Exercise 4.8.3. Since $(0, 1) \notin \mathrm{image}(\boldsymbol{A})$ it follows that $\boldsymbol{A}$, and so $\boldsymbol{A}^T$, is not surjective.

•

We conclude with definitions and interpretations of columnspace and rowspace for matrices over rings. Here again one must take care about the left and right module structures.

5.2.16 Definition (Columnspace and rowspace) Let R be a ring, let I and J be index sets, and let $A \in \mathrm{Mat}_{I\times J}(\mathsf{R})$.

(i) The ***left columnspace*** of A is the submodule of R^I, thought of as a left R-module, generated by the column vectors of A, and is denoted by $\mathrm{colspace}_l(A)$.

(ii) The ***right columnspace*** of A is the submodule of R^I, thought of as a right R-module, generated by the column vectors of A, and is denoted by $\mathrm{colspace}_r(A)$.

(iii) The ***left rowspace*** of A is the submodule of R^J, thought of as a left R-module, generated by the column vectors of A, and is denoted by $\mathrm{rowspace}_l(A)$.

(iv) The ***right rowspace*** of A is the submodule of R^J, thought of as a right R-module, generated by the column vectors of A, and is denoted by $\mathrm{rowspace}_r(A)$. •

The following result indicates the meaning of columnspace and rowspace in terms of the homomorphisms associated with a matrix.

5.2.17 Proposition (Interpretation of columnspace and rowspace) *Let* R *be a unit ring, let* I *and* J *be index sets, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$*. Then the following statements hold:*

(i) if $\mathbf{A}$ *is column finite then* $\mathrm{colspace}_{\mathrm{l}}(\mathbf{A}) = \mathrm{image}(\mathbf{A}_{\mathrm{l}})$ *and* $\mathrm{colspace}_{\mathrm{r}}(\mathbf{A}) = \mathrm{image}(\mathbf{A}_{\mathrm{r}})$*;*

(ii) if $\mathbf{A}$ *is row finite then* $\mathrm{rowspace}_{\mathrm{l}}(\mathbf{A}) = \mathrm{image}(\mathbf{A}_{\mathrm{l}}^{\mathrm{T}})$ *and* $\mathrm{rowspace}_{\mathrm{r}}(\mathbf{A}) = \mathrm{image}(\mathbf{A}_{\mathrm{r}}^{\mathrm{T}})$*.*

5.2.4 Invertible matrices over rings

The discussion of invertible matrices over rings proceeds much as that for invertible matrices over fields, but with a few modifications to allow for the fact that rings may not be commutative. Indeed, this can be seen immediately in the following result which characterises the nature of the linearity associated with the inverse of a homomorphism associated with a matrix.

5.2.18 Proposition (The inverse of an isomorphism is a homomorphism) *If* R *is a unit ring, if* I *and* J *are index sets, then the following statements hold:*

(i) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^{\mathrm{J}}; \mathsf{R}_0^{\mathrm{I}})$ *is an isomorphism of left* R*-modules (resp. right* R*-modules), then the inverse of* L *is an element of* $\mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^{\mathrm{I}}; \mathsf{R}_0^{\mathrm{J}})$*, and so is a homomorphism of left* R*-modules (resp. right* R*-modules);*

(ii) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^{\mathrm{J}}; \mathsf{R}^{\mathrm{I}})$ *is an isomorphism of left* R*-modules (resp. right* R*-modules), then the inverse of* L *is an element of* $\mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^{\mathrm{I}}; \mathsf{R}^{\mathrm{J}})$*, and so is a homomorphism of left* R*-modules (resp. right* R*-modules).*

Proof We shall prove the result in the case that $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^J; \mathsf{R}_0^I)$ (resp. $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}^J; \mathsf{R}^I)$) is a homomorphism of left R-modules; the case of right R-modules follows in a similar vein.

For $\boldsymbol{y}, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathsf{R}_0^I$ (resp. $\boldsymbol{y}, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathsf{R}^I$) let $\boldsymbol{x} = \mathsf{L}^{-1}(\boldsymbol{x})$, $\boldsymbol{x}_1 = \mathsf{L}^{-1}(\boldsymbol{y}_1)$, and $\boldsymbol{x}_2 = \mathsf{L}^{-1}(\boldsymbol{y}_2)$. Then compute

$$\mathsf{L}^{-1}(\boldsymbol{y}_1 + \boldsymbol{y}_2) = \mathsf{L}^{-1}(\mathsf{L}(\boldsymbol{x}_1) + \mathsf{L}(\boldsymbol{x}_2)) = \mathsf{L}^{-1} \circ \mathsf{L}(\boldsymbol{x}_1 + \boldsymbol{x}_2) = \boldsymbol{x}_1 + \boldsymbol{x}_2 = \mathsf{L}^{-1}(\boldsymbol{y}_1) + \mathsf{L}^{-1}(\boldsymbol{x}_2)$$

and, for $r \in \mathsf{R}$, compute

$$\mathsf{L}^{-1}(r\boldsymbol{y}) = \mathsf{L}^{-1}(r\mathsf{L}(\boldsymbol{x})) = \mathsf{L}^{-1} \circ \mathsf{L}(r\boldsymbol{x}) = r\boldsymbol{x} = r\mathsf{L}^{-1}(\boldsymbol{y}),$$

which gives the result. ■

When defining the inverse of a matrix, there is a potential issue that arises from the fact that, associated to a matrix, there are two homomorphisms, one of left modules and one of right modules. As far as inverses go, however, this difference is of no consequence, as the following result indicates.

5.2.19 Proposition (Inverse and homomorphisms associated to a matrix) *Let* R *be a unit ring, let* I *be an index set, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{R})$. *Then the following statements hold:*

(i) if $\mathbf{A}$ *is column finite, then* $\mathbf{A}_{\mathrm{l}}$ *is an isomorphism if and only if* $\mathbf{A}_{\mathrm{r}}$ *is an isomorphism, and, moreover, the column finite matrix associated to the inverse of* $\mathbf{A}_{\mathrm{r}}$ *is the same as the column finite matrix associated to the inverse of* $\mathbf{A}_{\mathrm{l}}$*;*

(ii) if R *is commutative and if* $\mathbf{A}$ *is row finite, then* $\mathbf{A}_{\mathrm{l}}$ *is an isomorphism if and only if* $\mathbf{A}_{\mathrm{r}}$ *is an isomorphism, and, moreover, the row finite matrix associated to the inverse of* $\mathbf{A}_{\mathrm{r}}$ *is the same as the row finite matrix associated to the inverse of* $\mathbf{A}_{\mathrm{l}}$*.*

Moreover, in each of the two cases, the matrix $\mathbf{A}^{-1}$ *associated to the inverse homomorphism is uniquely determined by the relation*

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_{\mathrm{I}}.$$

Proof (i) First we note that, by Theorem 5.2.11 and Proposition 5.2.18, $\boldsymbol{A}_l$ and $\boldsymbol{A}_r$ are isomorphisms if and only if there are homomorphisms $\boldsymbol{A}_l^{-1} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^I; \mathsf{R}_0^I)$ and $\boldsymbol{A}_r^{-1} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{R}_0^I; \mathsf{R}_0^I)$ of left and right R-modules, respectively, such that

$$\boldsymbol{A}_r \circ \boldsymbol{A}_r^{-1} = \boldsymbol{A}_r^{-1} \circ \boldsymbol{A}_r = \mathrm{id}_{\mathsf{R}_0^I}, \quad \boldsymbol{A}_l \circ \boldsymbol{A}_l^{-1} = \boldsymbol{A}_l^{-1} \circ \boldsymbol{A}_r = \mathrm{id}_{\mathsf{R}_0^I}. \tag{5.10}$$

For convenience, let us denote by $\boldsymbol{A}_r$ and $\boldsymbol{A}_r^{-1}$ the matrices associated by Theorem 5.1.13 to the homomorphisms $\boldsymbol{A}_r$ and $\boldsymbol{A}_r^{-1}$, and similarly for $\boldsymbol{A}_l$ and $\boldsymbol{A}_l^{-1}$. In terms of matrices, the relations (5.10) take the form that

$$(\boldsymbol{A}_r\boldsymbol{A}_r^{-1})\boldsymbol{x} = (\boldsymbol{A}_r^{-1}\boldsymbol{A}_r)\boldsymbol{x} = \boldsymbol{x}, \quad \boldsymbol{x}(\boldsymbol{A}_l^{-1}\boldsymbol{A}_l) = \boldsymbol{x}(\boldsymbol{A}_l\boldsymbol{A}_l^{-1}) = \boldsymbol{x}$$

for every $\boldsymbol{x} \in \mathsf{R}_0^I$. This shows that $\boldsymbol{A}_r$ is an isomorphism if and only if there exists a matrix $\boldsymbol{A}_r^{-1}$ such that

$$\boldsymbol{A}_r\boldsymbol{A}_r^{-1} = \boldsymbol{A}_r^{-1}\boldsymbol{A}_r = \boldsymbol{I}_n$$

and that $\boldsymbol{A}_l$ is an isomorphism if and only if there exists a matrix $\boldsymbol{A}_l^{-1}$ such that

$$\boldsymbol{A}_l\boldsymbol{A}_l^{-1} = \boldsymbol{A}_l^{-1}\boldsymbol{A}_l = \boldsymbol{I}_n.$$

Now suppose that $\boldsymbol{A}_r$ is an isomorphism and let $\boldsymbol{A}_r^{-1}$ be its inverse. Using the fact that, as matrices, we have $\boldsymbol{A}_r = \boldsymbol{A}$, this means that

$$\boldsymbol{A}\boldsymbol{A}_r^{-1} = \boldsymbol{A}_r^{-1}\boldsymbol{A} = \boldsymbol{I}_I.$$

However, since it also holds that, as matrices, we have $\boldsymbol{A}_l = \boldsymbol{A}$, we may conclude that $\boldsymbol{A}_r^{-1}$ is the inverse of $\boldsymbol{A}_l$.

(ii) Let us first show that if $\boldsymbol{A}$ is row finite and if $\boldsymbol{A}_r$ (resp. $\boldsymbol{A}_l$) is an isomorphism, then the inverse of $\boldsymbol{A}_r$ (resp. $\boldsymbol{A}_l$) is represented by a row finite matrix. Since $\boldsymbol{A}_r^T$ is column finite, by Theorem 5.2.11 there exists a column finite matrix $\boldsymbol{B} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ such that

$$\boldsymbol{B}\boldsymbol{A}_r^T = \boldsymbol{A}_r^T\boldsymbol{B} = \boldsymbol{I}_I,$$

or, in terms of components,

$$\sum_{i'\in I} \boldsymbol{B}(i_1,i')\boldsymbol{A}(i_2,i') = \sum_{i'\in I} \boldsymbol{A}(i',i_1)\boldsymbol{B}(i',i_2) = \begin{cases} 1_{\mathsf{R}}, & i_1 = i_2, \\ 0_{\mathsf{R}}, & i_1 \neq i_2. \end{cases}$$

Therefore,

$$\sum_{i'\in I} \boldsymbol{B}(i',i_1)\boldsymbol{A}(i',i_2) = \sum_{i'\in I} \boldsymbol{A}(i_1,i')\boldsymbol{B}(i_2,i') = \begin{cases} 1_{\mathsf{R}}, & i_1 = i_2, \\ 0_{\mathsf{R}}, & i_1 \neq i_2. \end{cases}$$

This gives

$$\boldsymbol{B}^T\boldsymbol{A}_r = \boldsymbol{A}_r\boldsymbol{B}^T = \boldsymbol{I}_I,$$

showing that the row finite matrix $\boldsymbol{B}^T$ represents the inverse of $\boldsymbol{A}_r$ as per part (iii) of Theorem 5.2.11. The verification for $\boldsymbol{A}_l$ follows along the same lines.

Once one has this correspondence between inverses of isomorphisms associated to a row finite matrix and another row finite matrix, the proof proceeds as in part (i).

The final assertion of the proposition was proved during the course of the proof above. ■

Theorem 5.2.11 and Propositions 5.2.18 and 5.2.19 make possible the following definition.

5.2.20 Definition (Invertible matrix over a ring) Let R be a unit ring, let I be an index set, and let $\boldsymbol{A} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$.

(i) If $\boldsymbol{A}$ is column finite, then it is ***invertible*** if it is an isomorphism from (either the left or right R-module) R_0^I to itself.

(ii) If $\boldsymbol{A}$ is row finite, then it is ***invertible*** if it is an isomorphism from (either the left or right R-module) R^I to itself.

The ***inverse*** of an invertible matrix $\boldsymbol{A}$ is the matrix $\boldsymbol{A}^{-1} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ associated to the inverse of the isomorphism from R_0^I to R_0^I (or from R^I to R^I) associated with $\boldsymbol{A}$. •

For readers used to matrices over fields defined with finitely many rows and columns, care should be taken to note that $\boldsymbol{BA} = \boldsymbol{I}_I$ does not necessarily imply that $\boldsymbol{AB} = \boldsymbol{I}_I$. For matrices over fields this can fail if I is not finite (Example 5.1.44), but always holds if I is finite. However, for rings this can fail even for finite matrices, as the following example shows.

5.2.21 Example (A counterexample on invertibility of finite matrices) We recall from Example 4.2.7–7 the details behind this example. We let $\mathsf{R} = \mathrm{Hom}_{\mathbb{R}}(\mathbb{R}[\xi]; \mathbb{R}[\xi])$ the linear maps between the $\mathbb{R}$-vector space of polynomials with real coefficients. By r_d and r_i we denote the linear maps corresponding, respectively, to differentiation and to integration with zero constant coefficient. Then the matrices $[r_d], [r_i] \in \mathrm{Mat}_{1\times 1}(\mathsf{R})$ are 1×1 matrices. Note that $[r_d][r_i] = \boldsymbol{I}_1$ but that $[r_i][r_d] \neq \boldsymbol{I}_1$. •

Let us next consider some of the simpler properties of invertible matrices.

5.2.22 Proposition (Properties of the matrix inverse) *Let* R *be a unit ring, let* I *be an index set, and let* $\mathbf{A}, \mathbf{B} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{R})$*. Then the following statements hold:*

(i) if $\mathbf{A}$ *is invertible, then so is* $\mathbf{A}^{-1}$*, and* $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$*;*

(ii) if R *is commutative, then* $\mathbf{A}$ *is invertible if and only if* $\mathbf{A}^{\mathrm{T}}$ *is invertible and, if* $\mathbf{A}$ *is invertible, then* $(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}$*;*

(iii) if $\mathbf{A}$ *and* $\mathbf{B}$ *are invertible, then* $\mathbf{AB}$ *is invertible and* $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$*.*

Proof (i) We assume that $\boldsymbol{A}$ is column finite, the row finite case following in a similar manner. As a map of sets we know that $\boldsymbol{A}_r \colon \mathsf{R}_0^I \to \mathsf{R}_0^I$ being invertible implies the existence of a unique map $\boldsymbol{A}_r^{-1} \colon \mathsf{R}_0^I \to \mathsf{R}_0^I$ such that $\boldsymbol{A}_r \circ \boldsymbol{A}_r^{-1} = \boldsymbol{A}_r^{-1} \circ \boldsymbol{A}_r = \mathrm{id}_{\mathsf{R}_0^I}$ by Proposition 1.3.9. Therefore, by Proposition 5.2.18 we know that this set map $\boldsymbol{A}_r^{-1}$ must be exactly the linear map associated to the inverse matrix for $\boldsymbol{A}_r$. Note that $\boldsymbol{A}_r \circ \boldsymbol{A}_r^{-1} = \boldsymbol{A}_r^{-1} \circ \boldsymbol{A}_r = \mathrm{id}_{\mathsf{R}_0^I}$ then implies, by Proposition 1.3.9, that $\boldsymbol{A}_r^{-1}$ is invertible with inverse equal to $\boldsymbol{A}_r$. This is equivalent to the matrix $\boldsymbol{A}^{-1}$ being invertible with inverse $\boldsymbol{A}$.

(ii) Suppose that $\boldsymbol{A}$ is invertible and column finite, so that its inverse is also invertible and column finite, and satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_I$$

by Proposition 5.2.19. By Proposition 5.2.8 it makes sense to take the transpose of this equation to get

$$(\boldsymbol{A}^{-1})^T\boldsymbol{A}^T = \boldsymbol{A}^T(\boldsymbol{A}^{-1})^T = \boldsymbol{I}_I.$$

Since $\boldsymbol{I}_I$ corresponds to the identity map on F^I it follows from Proposition 1.3.9 that $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1}$.

(iii) We assume that $\boldsymbol{A}$ is column finite, the row finite case following similarly. We note that, thinking of matrices as linear maps and using Proposition 5.2.12,

$$(\boldsymbol{B}_r^{-1} \circ \boldsymbol{A}_r^{-1}) \circ (\boldsymbol{A}_r \circ \boldsymbol{B}_r) = (\boldsymbol{A}_r \circ \boldsymbol{B}_r) \circ (\boldsymbol{B}_r^{-1} \circ \boldsymbol{A}_r^{-1}) = \mathrm{id}_{\mathsf{F}_0^I}.$$

By Proposition 1.3.9 this implies that $\boldsymbol{A}_r \circ \boldsymbol{B}_r$ is invertible with inverse $\boldsymbol{B}_r^{-1} \circ \boldsymbol{A}_r^{-1}$. This in turn implies that $\boldsymbol{AB}$ is invertible and that its inverse is $\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$. ■

5.2.23 Notation (Inverse of transpose) In cases where the equality $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$ makes sense (e.g., when $\boldsymbol{A}$ is column and row finite) then one often writes

$$\boldsymbol{A}^{-T} = (\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T. \qquad \bullet$$

5.2.5 Elementary and secondary operations and elementary and secondary matrices

As we saw in Sections 5.1.5, 5.1.6, and 5.1.7, the notion of an elementary matrix is an important one for understanding the problem of equivalence for matrices over fields. The problem of equivalence for matrices over rings is far more complicated, and we will not attempt any sort of general treatment. Our objective will be to provide a useful classification of matrices over principal ideal domains, and especially Euclidean rings, since the cases we will be interested in are of this sort. That one ought to be able to do more in these cases is suggested by the results of Section 4.9.

Before we get to specific sorts of rings, we define various sorts of operations on matrices over general rings. The first bunch of these echo those for matrices over fields, beginning with elementary row operations.

5.2.24 Definition (Elementary row operation) Let R be a ring, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{R})$. The matrix $\boldsymbol{A}_2$ is obtained by an ***elementary row operation*** from $\boldsymbol{A}_1$ if one of the following hold:

(i) there exists distinct $i_1, i_2 \in \{1, \dots, n\}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i, j) = \begin{cases} \boldsymbol{A}_1(i, j), & i \notin \{i_1, i_2\}, \\ \boldsymbol{A}_1(i_2, j), & i = i_1, \\ \boldsymbol{A}_1(i_1, j), & i = i_2, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree except that the i_1st and i_2nd rows are interchanged;

(ii) there exists $i_0 \in \{1, \dots, n\}$ and a unit $u \in \mathsf{R}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i, j) = \begin{cases} \boldsymbol{A}_1(i, j), & i \neq i_0, \\ u\boldsymbol{A}_1(i, j), & i = i_0 \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree, except that the i_0th row of $\boldsymbol{A}_2$ is the i_0th row of $\boldsymbol{A}_1$ multiplied by u on the left;

(iii) there exists distinct $i_1, i_2 \in \{1, \dots, n\}$ and $r \in \mathsf{R}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i, j) = \begin{cases} \boldsymbol{A}_1(i, j), & i \neq i_1, \\ \boldsymbol{A}_1(i, j) + r\boldsymbol{A}_1(i_2, j), & i = i_1, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree except that the i_1st row of $\boldsymbol{A}_2$ is obtained by adding r times the i_2nd row of $\boldsymbol{A}_1$ to the i_1st row of $\boldsymbol{A}_1$. •

Of course, one also has elementary column operations.

5.2.25 Definition (Elementary column operation) Let R be a ring, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{R})$. The matrix $\boldsymbol{A}_2$ is obtained by an ***elementary column operation*** from $\boldsymbol{A}_1$ if one of the following hold:

(i) there exists distinct $j_1, j_2 \in \{1, \dots, m\}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & j \notin \{j_1, j_2\}, \\ \boldsymbol{A}_1(i,j_2), & j = j_1, \\ \boldsymbol{A}_1(i,j_1), & j = j_2, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree except that the j_1st and j_2nd columns are interchanged;

(ii) there exists $j_0 \in \{1, \dots, m\}$ and a unit $u \in \mathsf{R}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & j \neq j_0, \\ \boldsymbol{A}_1(i,j)u, & j = j_0 \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree, except that the j_0th column of $\boldsymbol{A}_2$ is the j_0th column of $\boldsymbol{A}_1$ multiplied by u on the right;

(iii) there exists distinct $j_1, j_2 \in \{1, \dots, m\}$ and $r \in \mathsf{R}$ such that, for $(i, j) \in \{1, \dots, n\} \times \{1, \dots, m\}$,

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & j \neq j_1, \\ \boldsymbol{A}_1(i,j) + \boldsymbol{A}_1(i,j_2)r, & j = j_1, \end{cases}$$

i.e., $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ agree except that the j_1st column of $\boldsymbol{A}_2$ is obtained by adding the j_2nd column of $\boldsymbol{A}_1$, multiplied by r, to the j_1st column of $\boldsymbol{A}_1$. •

For reasons that are not at the moment completely clear, for matrices over rings it is useful to allow a second type of row and column operation.

5.2.26 Definition (Secondary row and column operations) Let R be a unit ring, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{R})$.

(i) The matrix $\boldsymbol{A}_2$ is obtained by a ***secondary row operation*** from $\boldsymbol{A}_1$ if

$$\boldsymbol{A}_2(i,j) = \begin{cases} \boldsymbol{A}_1(i,j), & i \geq 2, \\ a\boldsymbol{A}_1(1,j) + b\boldsymbol{A}_1(2,j), & i = 1, \\ c\boldsymbol{A}_1(1,j) + d\boldsymbol{A}_1(2,j), & i = 2, \end{cases}$$

for $a, b, c, d \in \mathsf{R}$ satisfying

$$ad - bc = da - cb = da - bc = ad - cb = 1_{\mathsf{R}} \tag{5.11}$$

and

$$ba - ab = cd - dc = db - bd = ac - ca = 0_{\mathsf{R}}. \tag{5.12}$$

(ii) The matrix A_2 is obtained by a ***secondary column operation*** from A_1 if

$$A_2(i,j) = \begin{cases} A_1(i,j), & j \geq 2, \\ A_1(i,1)a + A_1(i,2)c, & j = 1, \\ A_1(i,1)b + A_1(i,2)d, & j = 2, \end{cases}$$

for $a, b, c, d \in \mathsf{R}$ satisfying (5.11) and (5.12). •

The reader might justifiably ask, "Why are these referred to as row and column operations?" since the answer here is not so evident as for elementary operations. The reader will probably best understand this after understanding the matrix versions of elementary and secondary operations stated as Proposition 5.2.30. Also, the meaning of the conditions on the ring elements a, b, c, and d are also probably best understood after one understands the matrix versions of the operations; see Remark 5.2.32.

In any case, let us define the notions of row equivalence and column equivalence for commutative unit rings.

5.2.27 Definition (Row equivalence, column equivalence) Let R be a commutative unit ring, let $m, n \in \mathbb{Z}_{>0}$, and let $A_1, A_2 \in \mathrm{Mat}_{m\times n}(\mathsf{R})$.

(i) The matrix A_2 is ***row equivalent*** to A_1 if there exists $k \in \mathbb{Z}_{>0}$ and matrices $A'_1, \dots, A'_k \in \mathrm{Mat}_{m\times n}(\mathsf{R})$ such that $A'_1 = A_1$, $A'_k = A_2$, and A'_{j+1} is obtained by either an elementary row operation or a secondary operation from A'_j for each $j \in \{1, \dots, k-1\}$.

(ii) The matrix A_2 is ***column equivalent*** to A_1 if there exists $k \in \mathbb{Z}_{>0}$ and matrices $A'_1, \dots, A'_k \in \mathrm{Mat}_{m\times n}(\mathsf{R})$ such that $A'_1 = A_1$, $A'_k = A_2$, and A'_{j+1} is obtained by either an elementary column operation or a secondary operation from A'_j for each $j \in \{1, \dots, k-1\}$. •

The reader will notice that we now have competing definitions of row and column equivalence for fields. This will not be problematic, however, since we will show in Proposition 5.2.33 that secondary operations can be realised as sequences of elementary operations for Euclidean domains, and so particularly for fields. We shall not have much to say about row and column equivalence for matrices over rings. What we do say is covered in Section 5.2.7.

As for row and column equivalence for matrices over fields, one can show that row equivalence and column equivalence define equivalence relations for matrices over commutative unit rings. The situation here is not quite as trivial as for matrices over fields, as the rôle of secondary operations complicates matters slightly. However, a reference to the proof of Proposition 5.2.30 will set things straight. In like manner, one can prove fairly easily, once one understands the rôle of secondary operations, that A_1 and A_2 are row equivalent if and only if A_1^T and A_2^T are column equivalent. We leave it for the reader to verify this.

For matrices over fields we have seen that it is useful to realise elementary row and column operations as multiplication by elementary row and column matrices. The same is true for matrices over rings, of course. Moreover, it is by casting secondary operations in this way that we can best understand them.

5.2.28 Definition (Elementary and secondary row and column matrices) Let R be a unit ring and let $n \in \mathbb{Z}_{>0}$. A matrix $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$

(i) is an ***elementary row matrix*** if $\boldsymbol{A}$ is obtained by an elementary row operation from $\boldsymbol{I}_n$,

(ii) is an ***elementary column matrix*** if $\boldsymbol{A}$ is obtained by an elementary column operation from $\boldsymbol{I}_n$,

(iii) is a ***secondary row matrix*** if it is obtained by a secondary row operation from $\boldsymbol{I}_n$, and

(iv) is a ***secondary column matrix*** if it is obtained by a secondary column operation from $\boldsymbol{I}_n$. •

The following result now usefully characterises the elementary row and column matrices and the secondary matrices.

5.2.29 Proposition (Elementary and secondary row and column operations and transpose) *If* R *is a unit ring and if* $n \in \mathbb{Z}_{>0}$*, then the following statements for* $\mathbf{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ *are equivalent:*

(i) $\mathbf{A}$ *is an elementary row matrix;*

(ii) $\mathbf{A}$ *is an elementary column matrix;*

(iii) $\mathbf{A}^T$ *is an elementary row matrix;*

(iv) $\mathbf{A}^T$ *is an elementary column matrix.*

Also, the following statements are equivalent:

(v) $\mathbf{A}$ *is a secondary row matrix;*

(vi) $\mathbf{A}$ *is a secondary column matrix;*

(vii) $\mathbf{A}^T$ *is a secondary row matrix;*

(viii) $\mathbf{A}^T$ *is a secondary column matrix.*

Proof The verification of the equivalence of the first four statements goes in exactly the same way as the proof of Proposition 5.1.31, taking care that the ring may not be commutative. Therefore, we only prove the equivalence of the last four statements. To do this, the following lemma is the key observation, and indeed gives the most insightful characterisation of a secondary row or column operation.

1 Lemma *If* $\mathbf{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ *is obtained from* $\mathbf{I}_n$ *by a secondary row or column operation corresponding to elements* $\mathsf{a}, \mathsf{b}, \mathsf{c}, \mathsf{d} \in \mathsf{R}$ *as in Definition 5.2.26, then*

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}' & \mathbf{0}_{2\times(n-2)} \\ \hline \mathbf{0}_{(n-2)\times 2} & \mathbf{I}_{n-2} \end{array}\right],$$

where

$$\mathbf{A}' = \begin{bmatrix} \mathrm{a} & \mathrm{b} \\ \mathrm{c} & \mathrm{d} \end{bmatrix}.$$

Proof Follows directly from the definition. ▼

Now we can easily prove the equivalence of the last four statements in the proposition. Indeed, the condition for a matrix to be a secondary row or column matrix is simply a condition on the ring elements $a, b, c, d \in \mathsf{R}$ that comprise the top left 2×2 block of the matrix, i.e., the matrix A' appearing in the proof of the lemma above. These conditions are (5.11) and (5.12). It is easy to check that these conditions are satisfied for A' if and only if they are satisfied for $(A')^T$, and this gives the result. ■

As for matrices over fields, we shall call elementary row and column matrices simply ***elementary matrices***. We shall also call a matrix that is either a secondary row matrix or a secondary column matrix a ***secondary matrix***.

As with matrices over fields, we can realise elementary row and column operations as multiplication by elementary matrices. The same now holds for secondary matrices, as the following result indicates.

5.2.30 Proposition (Elementary and secondary matrices, and elementary and secondary row and column operations) *Let* R *be a unit ring, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$, *and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$. *Then the following statements hold:*

(i) *if* $\mathbf{A}_2$ *is obtained from* $\mathbf{A}_1$ *by an elementary row operation and if* $\mathbf{E}_{\mathrm{m}} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{m}}(\mathsf{R})$ *is the elementary matrix obtained by applying the same row operation to* $\mathbf{I}_{\mathrm{m}}$, *then* $\mathbf{A}_2 = \mathbf{E}_{\mathrm{m}}\mathbf{A}_1$;

(ii) *if* $\mathbf{A}_2$ *is obtained from* $\mathbf{A}_1$ *by an elementary column operation and if* $\mathbf{E}_{\mathrm{n}} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$ *is the elementary matrix obtained by applying the same column operation to* $\mathbf{I}_{\mathrm{n}}$, *then* $\mathbf{A}_2 = \mathbf{A}_1\mathbf{E}_{\mathrm{n}}$;

(iii) *if* $\mathbf{A}_2$ *is obtained from* $\mathbf{A}_1$ *by a secondary row operation and if* $\mathbf{S}_{\mathrm{m}} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{m}}(\mathsf{F})$ *is the secondary matrix obtained by applying the same row operation to* $\mathbf{I}_{\mathrm{m}}$, *then* $\mathbf{A}_2 = \mathbf{S}_{\mathrm{m}}\mathbf{A}_1$;

(iv) *if* $\mathbf{A}_2$ *is obtained from* $\mathbf{A}_1$ *by a secondary column operation and if* $\mathbf{S}_{\mathrm{n}} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$ *is the secondary matrix obtained by applying the same column operation to* $\mathbf{I}_{\mathrm{n}}$, *then* $\mathbf{A}_2 = \mathbf{A}_1\mathbf{S}_{\mathrm{n}}$.

Proof The same proof as given for Proposition 5.1.32 works here for the first two parts of the result, taking appropriate care with the fact that the ring may not be commutative. Thus we need prove only the second two parts. And here we only prove the third part since the fourth follows by an entirely similar argument. From the lemma in the proof of Proposition 5.2.29 we know that by applying a secondary row operation associated to $a, b, c, d \in \mathsf{R}$ to the identity matrix we get the matrix

$$S = \left[\begin{array}{c|c} S' & \mathbf{0}_{2\times(n-2)} \\ \hline \mathbf{0}_{(n-2)\times 2} & I_{n-2} \end{array}\right],$$

where

$$S' = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

It is then a simple matter of checking matrix multiplication to see that the matrix $\boldsymbol{S}\boldsymbol{A}_1$ is the matrix obtained from $\boldsymbol{A}_1$ by applying the secondary row operation corresponding to $a, b, c, d \in \mathsf{R}$. ■

For matrices over fields we showed in Theorem 5.1.33 that every invertible matrix is a finite product of elementary matrices. An analogous result is too much to hope for for general rings. However, for principal ideal domains the hoped for result is true, and really gives validity to the notion of elementary and secondary operations for matrices over rings. Without such a result, these constructions would be valueless. And, indeed, their value for matrices over rings is already limited by the fact that the important implication in the following result only applies to principal ideal domains.

5.2.31 Theorem (Invertible matrices and products of elementary and secondary matrices) *Let* R *be a unit ring and let* $\mathrm{n} \in \mathbb{Z}_{>0}$. *For* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$ *consider the following statements:*

(i) $\mathbf{A}$ *is invertible;*

(ii) $\mathbf{A}$ *is a product of a finite number of elementary and secondary matrices.*

Then

(iii) (ii) $\Longrightarrow$ *(i) and*

(iv) (i) $\Longrightarrow$ *(ii) if* R *is a principal ideal domain.*

Proof (iii) In Theorem 5.1.33 we proved this implication for elementary matrices over fields. That proof is easily adapted, taking care with the possible noncommutativity of the ring, to general rings to show that every elementary matrix over a unit ring is invertible. By Proposition 5.2.22 it then suffices to show that every secondary matrix is invertible. By the lemma of Proposition 5.2.29 it suffices to show that the matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is invertible provided that $a, b, c, d \in \mathsf{R}$ satisfy (5.11) and (5.12). This, however, follows by the computation

$$\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} 1_{\mathsf{R}} & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & 1_{\mathsf{R}} \end{bmatrix}.$$

(iv) Our proof relies on Theorem 5.2.43 below. By that theorem, for any $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ there exists $\boldsymbol{P}, \boldsymbol{Q} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ such that

$$\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q} = \left[\begin{array}{c|c} \boldsymbol{D}_r & \mathbf{0}_{r\times(n-r)} \\ \hline \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{array}\right],$$

with $\boldsymbol{D}_r$ being the diagonal matrix

$$\boldsymbol{D}_r = \begin{bmatrix} d_1 & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_r \end{bmatrix}$$

and with $d_1|\cdots|d_r$.

We claim that $\boldsymbol{A}$ is invertible if and only if (1) $r = n$ and (2) each of $d_1, \dots, d_n$ is a unit. To see this we first note that the implication part (iii) proved above gives $\boldsymbol{P}$ and $\boldsymbol{Q}$ as invertible. Thus $\boldsymbol{A}$ is invertible if and only if

$$\left[\begin{array}{c|c} \boldsymbol{D}_r & \boldsymbol{0}_{r\times(n-r)} \\ \hline \boldsymbol{0}_{(m-r)\times r} & \boldsymbol{0}_{(m-r)\times(n-r)} \end{array}\right]$$

is invertible. It is clear (why?) that if this matrix is invertible then $r = n$. Now suppose that the matrix $\boldsymbol{D}_n$ is invertible. Then there exists a matrix $\boldsymbol{D}'_n$ such that $\boldsymbol{D}'_n\boldsymbol{D}_n = \boldsymbol{D}_n\boldsymbol{D}'_n = \boldsymbol{I}_n$. One can do the matrix multiplication to see that $\boldsymbol{D}'_n$ must be a diagonal matrix with diagonal entries $d'_1, \dots, d'_n$ which satisfy $d'_j d_j = 1_{\mathsf{R}}$, $j \in \{1, \dots, n\}$. Thus $d_1, \dots, d_n$ are units and $d'_j = d_j^{-1}$, $j \in \{1, \dots, n\}$. This gives our assertion that $\boldsymbol{A}$ is invertible if and only if $n = r$ and $d_1, \dots, d_n$ are units.

Now we simply write

$$\boldsymbol{A} = \boldsymbol{P} \begin{bmatrix} d_1 & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & 1_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & 1_{\mathsf{R}} \end{bmatrix} \begin{bmatrix} 1_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & 1_{\mathsf{R}} \end{bmatrix} \cdots \begin{bmatrix} 1_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & 1_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_n \end{bmatrix} \boldsymbol{Q}.$$

Since the n matrices in the middle on the right-hand side are elementary matrices, this part of the result follows. ■

5.2.32 Remark (On secondary operations) It is fairly easy to understand the character of elementary operations. These manipulate rows and columns in a way that can be "undone" (meaning elementary matrices are invertible). Proposition 5.2.30 also makes clear the way one should think of secondary operations. These manipulate the first two rows or columns of a matrix in the most general way that can be "undone" (meaning that secondary matrices are invertible). While secondary operations are only applied to the first two rows or columns, by performing elementary operations of row or column swapping, these can be made on any two distinct rows or columns. Thus secondary operations are the most general things that one can do with two rows or columns. •

As a final result in this section, let us show that the possible definitions of row equivalence and column equivalence for matrices over fields agree. We do this by first considering secondary matrices over Euclidean domains.

5.2.33 Proposition (Equivalence of secondary and elementary operations for Euclidean domains) *If* R *is a Euclidean domain, if* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{S} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$ *is a secondary matrix, then there exists elementary matrices* $\mathbf{E}_1, \dots, \mathbf{E}_\mathrm{k} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$ *such that* $\mathbf{S} = \mathbf{E}_1 \cdots \mathbf{E}_\mathrm{k}$*.*

Proof We denote by $\delta\colon \mathsf{R} \to \mathbb{Z}_{\geq 0}$ the degree function.

By the lemma of Proposition 5.2.29 and using the fact that R is commutative, it suffices to show that any matrix

$$\boldsymbol{S} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{Mat}_{2\times 2}(\mathsf{R})$$

for which $ad - bc = 1_{\mathsf{R}}$ can be written as a product of elementary matrices. We first prove a lemma.

1 Lemma *If* R *is a Euclidean domain and if* $\mathbf{A} \in \mathrm{Mat}_{2\times 2}(\mathsf{R})$ *then* $\mathbf{A}$ *can be transformed into a diagonal matrix by a finite number of elementary row and column operations.*

Proof Let us write

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Let us first consider the case when $a = b = 0_{\mathsf{R}}$. If $c = 0_{\mathsf{R}}$ then the result is already true. If $c \neq 0_{\mathsf{R}}$ then there exists $q_1, d_1 \in \mathsf{R}$ such that $d = q_1 c + d_1$ with $\delta(d_1) < \delta(c)$. By an elementary column operation we then transform $\boldsymbol{A}$ into

$$\begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ c & d \end{bmatrix} \longrightarrow \begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ c & d - q_1 c \end{bmatrix} = \begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ c & d_1 \end{bmatrix}$$

with $\delta(d_1) < \delta(c)$. If $d_1 = 0_{\mathsf{R}}$ then by a row or column swap, the lemma is proved. If $d_1 \neq 0_{\mathsf{R}}$, we write $c = q_2 d_1 + c_1$ with $\delta(c_1) < \delta(d_1)$. We then perform elementary column operations to get

$$\begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ c & d_1 \end{bmatrix} \longrightarrow \begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ c - q_2 d_1 & d_1 \end{bmatrix} = \begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ c_1 & d_1 \end{bmatrix}$$

with $\delta(c_1) < \delta(d_1) < \delta(c)$. This produces a sequence of ring elements with strictly decreasing degree. This sequence of degrees must terminate with zero, and will do so when we have made elementary column operations to transform $\boldsymbol{A}$ into a matrix of the form

$$\begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & r \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0_{\mathsf{R}} & 0_{\mathsf{R}} \\ r & 0_{\mathsf{R}} \end{bmatrix}.$$

In the first case we are done, and in the second case we are done after a column swap.

Now suppose that a and b are both nonzero. We shall first use elementary column operations to transform $\boldsymbol{A}$ into a matrix whose top right entry is zero. First write $b = q_1 a + b_1$ with $\delta(b_1) < \delta(a)$. We then perform the following elementary column operations:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \longrightarrow \begin{bmatrix} a & b - q_1 a \\ c & d - q_1 c \end{bmatrix} = \begin{bmatrix} a & b_1 \\ c & d - q_1 c \end{bmatrix}$$

with $\delta(b_1) < \delta(a)$. If $b_1 = 0_{\mathsf{R}}$ then we have arrived at our desired form with a zero as the top right entry. Otherwise, write $a = q_2 b_1 + a_1$ with $\delta(a_1) < \delta(b_1)$ and perform the elementary column operations

$$\begin{bmatrix} a & b_1 \\ c & d - q_1 c \end{bmatrix} \longrightarrow \begin{bmatrix} a - q_2 b_1 & b_1 \\ c - q_2(d - q_1 c) & d - q_1 c \end{bmatrix} = \begin{bmatrix} a_1 & b_1 \\ c - q_2(d - q_1 c) & d - q_1 c \end{bmatrix}$$

with $\delta(a_1) < \delta(b_1) < \delta(a)$. We again arrive at a strictly decreasing sequence of degree, and so must eventually arrive at a matrix of the form

$$\begin{bmatrix} a' & 0_{\mathsf{R}} \\ c' & d' \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0_{\mathsf{R}} & b' \\ c' & d' \end{bmatrix},$$

giving, possibly after an additional column swap, our desired matrix

$$\begin{bmatrix} a' & 0_{\mathsf{R}} \\ c' & d' \end{bmatrix}$$

with a zero as the top right entry and with $\delta(a') < \delta(a)$.

Now we perform elementary row operations to additionally ensure that the bottom left entry can be transformed to zero. If $c' = 0_{\mathsf{R}}$ then this is already done. Otherwise write $c' = q_1 a' + c_1$ with $\delta(c'_1) < \delta(a')$ and perform the elementary row operations

$$\begin{bmatrix} a' & 0_{\mathsf{R}} \\ c' & d' \end{bmatrix} \longrightarrow \begin{bmatrix} a' & 0_{\mathsf{R}} \\ c' - q_1 a' & d' \end{bmatrix} = \begin{bmatrix} a' & 0_{\mathsf{R}} \\ c'_1 & d' \end{bmatrix}$$

with $\delta(c'_1) < \delta(a')$. If $c'_1 = 0_{\mathsf{R}}$ then the proof is complete. Otherwise write $a' = q_2 c'_1 + a'_1$ with $\delta(a'_1) < \delta(c'_1)$ and perform the elementary row operations

$$\begin{bmatrix} a' & 0_{\mathsf{R}} \\ c'_1 & d' \end{bmatrix} \longrightarrow \begin{bmatrix} a' - q_2 c'_1 & -q_2 d' \\ c'_1 & d' \end{bmatrix} = \begin{bmatrix} a'_1 & -q_2 d' \\ c'_1 & d' \end{bmatrix}$$

with $\delta(a'_1) < \delta(c'_1) < \delta(a')$. By our (by now) usual argument involving strictly decreasing sequences of degrees we arrive at a matrix of the form

$$\begin{bmatrix} 0_{\mathsf{R}} & b'' \\ c'' & d'' \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a'' & b'' \\ 0_{\mathsf{R}} & d'' \end{bmatrix}$$

with $\delta(c'') < \delta(a')$ in the first case, or $\delta(a'') < \delta(a')$ in the second case. A row swap then gives a matrix of the form

$$\begin{bmatrix} a'' & b'' \\ 0_{\mathsf{R}} & d'' \end{bmatrix}$$

with $\delta(a'') < \delta(a') < \delta(a)$.

The above arguments can then be repeated, giving rise to a sequence of matrices, obtained by sequences of elementary row and column operations, of the form

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \longrightarrow \begin{bmatrix} a' & 0_{\mathsf{R}} \\ c' & d' \end{bmatrix} \longrightarrow \begin{bmatrix} a'' & b'' \\ 0_{\mathsf{R}} & d'' \end{bmatrix} \longrightarrow \begin{bmatrix} a''' & 0_{\mathsf{R}} \\ c''' & d''' \end{bmatrix} \longrightarrow \cdots$$

with $\delta(a) > \delta(a') > \delta(a'') > \delta(a''') > \cdots$. Moreover, the starting point for moving from the second step to the third is that $c' \neq 0_{\mathsf{R}}$, the starting point for moving from the third step to the fourth is that $b'' \neq 0_{\mathsf{R}}$, and so on. Thus since the sequence of degree of the top left elements is strictly decreasing, this construction must terminate with a diagonal matrix after a finite number of steps. ▼

Let us now proceed with the proof. We first note that by part (iii) of Theorem 5.2.31 the secondary matrix

$$\boldsymbol{S} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is invertible. By the lemma above and by Proposition 5.2.30 there exists matrices $\boldsymbol{P}, \boldsymbol{Q} \in \mathrm{Mat}_{2\times 2}(\mathsf{R})$, each a finite product of elementary matrices, such that

$$\boldsymbol{PSQ} = \boldsymbol{D} \triangleq \begin{bmatrix} d_1 & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 \end{bmatrix}$$

for some $d_1, d_2 \in \mathsf{R}$. By Theorem 5.2.31 the matrices $\boldsymbol{P}$ and $\boldsymbol{Q}$ are invertible and so $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{S}\boldsymbol{Q}^{-1}$, whence $\boldsymbol{D}$ is also invertible by Proposition 5.2.22. Note that $\boldsymbol{D}$ being invertible precludes either of d_1 or d_2 from being zero (why?). Thus there exists $p, q, r, s \in \mathsf{R}$ such that

$$\begin{bmatrix} p & q \\ r & s \end{bmatrix}\begin{bmatrix} d_1 & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 \end{bmatrix} = \begin{bmatrix} d_1 & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 \end{bmatrix}\begin{bmatrix} p & q \\ r & s \end{bmatrix} = \begin{bmatrix} 1_{\mathsf{R}} & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & 1_{\mathsf{R}} \end{bmatrix}.$$

Doing the matrix multiplication gives

$$pd_1 = sd_2 = 1_{\mathsf{R}}, \quad qd_2 = rd_1 = qd_1 = sd_2 = 0_{\mathsf{R}}.$$

Since neither of d_1 and d_2 are zero and since R is an integral domain this implies that $q = r = 0_{\mathsf{R}}$ and that d_1 and d_2 are units with $p = d_1^{-1}$ and $s = d_2^{-1}$. Now write

$$\boldsymbol{S} = \boldsymbol{P}\begin{bmatrix} d_1 & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & 1_{\mathsf{R}} \end{bmatrix}\begin{bmatrix} 1_{\mathsf{R}} & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 \end{bmatrix}\boldsymbol{Q},$$

which renders $\boldsymbol{S}$ as a finite product of elementary matrices, as desired, since the middle two matrices are elementary matrices. ■

Combining the previous result with Proposition 4.3.2 gives the following result. In the statement, the definition being used for row and column equivalence is that given by Definition 5.2.27 for matrices over rings.

5.2.34 Corollary (Row and column equivalence for matrices over Euclidean domains) *Let R be a Euclidean domain, let $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$, and let $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$. Then the following two statements are equivalent:*

(i) $\mathbf{A}_1$ and $\mathbf{A}_2$ are row equivalent;

(ii) there exists $\mathrm{k} \in \mathbb{Z}_{>0}$ and matrices $\mathbf{A}'_1, \dots, \mathbf{A}'_{\mathrm{k}} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$ such that $\mathbf{A}'_1 = \mathbf{A}_1$, $\mathbf{A}'_{\mathrm{k}} = \mathbf{A}_2$, and $\mathbf{A}'_{\mathrm{j}+1}$ is obtained by an elementary row operation from $\mathbf{A}'_{\mathrm{j}}$ for each $\mathrm{j} \in \{1, \dots, \mathrm{k}-1\}$.

Also, the following two statements are equivalent:

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are column equivalent;*

(ii) *there exists* $\mathrm{k} \in \mathbb{Z}_{>0}$ *and matrices* $\mathbf{A}'_1, \dots, \mathbf{A}'_\mathrm{k} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$ *such that* $\mathbf{A}'_1 = \mathbf{A}_1$, $\mathbf{A}'_\mathrm{k} = \mathbf{A}_2$*, and* $\mathbf{A}'_{\mathrm{j}+1}$ *is obtained by an elementary column operation from* $\mathbf{A}'_\mathrm{j}$ *for each* $\mathrm{j} \in \{1, \dots, \mathrm{k}-1\}$.

In particular, the result holds if R *is a field.*

5.2.6 Rank and equivalence for matrices over rings

We have not really seen sharply up to this point the reason why matrices over rings should be so much different than matrices over fields. In this section we shall see that even defining the rank of a matrix over a general ring is not something one can do in all cases. There are a few reasons for this.

1. Although the rowspace and columnspace of a matrix over a ring are submodules of a free module, this does not suffice for them to be themselves free, cf. Example 4.8.20. Thus the notion of rank may not be definable for the rowspace and columnspace.
2. Even if the rowspace and columnspace are free, they may not have a well-defined rank, cf. Example 4.8.22. This would again make the definition of row rank and column rank problematic.

Neither of these difficulties arise when the ring is a principal ideal domain (the first difficulty evaporates by virtue of Theorem 4.9.1 and the second by virtue of Theorem 4.8.25). Thus we shall see that a distinguished rôle is played by matrices over principal ideal domains. Fortunately, these are the rings that are of most interest to us, particularly the case of the principal ideal domain given by the polynomial ring over a field.

As we did in Section 5.1.6 we restrict to matrices with finite numbers of rows and columns.

Our definitions of rank for this section are the following.

5.2.35 Definition (Row rank and column rank) Let R be a commutative unit ring, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{R})$.

(i) If the rowspace of $\boldsymbol{A}$ is free, the ***row rank*** of $\boldsymbol{A}$ is the rank of the rowspace.

(ii) If the columnspace of $\boldsymbol{A}$ is free, the ***column rank*** of $\boldsymbol{A}$ is the rank of the columnspace. •

As mentioned above, by Theorem 4.9.1 we have the following result.

5.2.36 Proposition (Definability of row rank and column rank of matrices over principal ideal domains) *If* R *is a principal ideal domain, if* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$*, then the row rank and column rank of* $\mathbf{A}$ *can be defined.*

Proof This follows from Theorem 4.9.1 since the rowspace and columnspace are submodules of the free modules R^m and R^n, respectively. ■

One of the fundamental questions that arises is the relationship between the row rank and the column rank, even when they can be defined. The relationship of equality does not generally hold, as the following example indicates.

5.2.37 Example (Row rank and column rank do not generally agree) We let $\mathsf{R} = \mathbb{Z}_{30} = \mathbb{Z}/30\mathbb{Z}$ and consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 + 30\mathbb{Z} & 1 + 30\mathbb{Z} & -1 + 30\mathbb{Z} \\ 0 + 30\mathbb{Z} & 2 + 30\mathbb{Z} & 3 + 30\mathbb{Z} \end{bmatrix}$$

over R. We claim that $\boldsymbol{A}$ has row rank 2 and column rank 1.

To get our claim about the row rank, it suffices to show that the rows of $\boldsymbol{A}$ are linearly independent. So suppose that

$$(j_1 + 30\mathbb{Z})(1 + 30\mathbb{Z}, 1 + 30\mathbb{Z}, -1 + 30\mathbb{Z}) + (j_2 + 30\mathbb{Z})(0 + 30\mathbb{Z}, 2 + 30\mathbb{Z}, 3 + 30\mathbb{Z}) \\ = (0 + 30\mathbb{Z}, 0 + 30\mathbb{Z}, 0 + 30\mathbb{Z}),$$

for $j_1, j_2 \in \mathbb{Z}$. Thus

$$j_1 + 30\mathbb{Z} = j_1 + 2j_2 + 30\mathbb{Z} = -j_1 + 3j_2 + 30\mathbb{Z} = 0 + 30\mathbb{Z}.$$

Therefore, $j_1 = 30k_1$ for some $k_1 \in \mathbb{Z}$ immediately, and thence $2j_2 = 30k_2$ for some $k_2 \in \mathbb{Z}$ and $3j_2 = 30l_2$ for some $l_2 \in \mathbb{Z}$. Thus $j_2 = 15k_2' = 10l_2'$ for some $k_2', l_2' \in \mathbb{Z}$. Thus $15|j_2$ and $10|j_2$ from which we deduce that j_2 must be a multiple of the least common multiple of $\{10, 15\}$, i.e., a multiple of 30. This shows that the rows of $\boldsymbol{A}$ are linearly independent and so the row rank of $\boldsymbol{A}$ is 2.

To show that the column rank of $\boldsymbol{A}$ is well-defined and equal to 1, we show that any two columns of $\boldsymbol{A}$ are linearly independent. To see this note that

$$(15 + 30\mathbb{Z})(1 + 30\mathbb{Z}, 0 + 30\mathbb{Z}) + (15 + 30\mathbb{Z})(1 + 30\mathbb{Z}, 2 + 30\mathbb{Z}) = \\ (0 + 30\mathbb{Z}, 0 + 30\mathbb{Z}),$$

$$(10 + 30\mathbb{Z})(1 + 30\mathbb{Z}, 0 + 30\mathbb{Z}) + (10 + 30\mathbb{Z})(-1 + 30\mathbb{Z}, 3 + 30\mathbb{Z}) = \\ (0 + 30\mathbb{Z}, 0 + 30\mathbb{Z}),$$

$$(24 + 30\mathbb{Z})(1 + 30\mathbb{Z}, 2 + 30\mathbb{Z}) + (-6 + 30\mathbb{Z})(-1 + 30\mathbb{Z}, 3 + 30\mathbb{Z}) = \\ (0 + 30\mathbb{Z}, 0 + 30\mathbb{Z}),$$

showing that the first and second, the first and third, and the second and third columns of $\boldsymbol{A}$, respectively, are linearly dependent. •

The example shows that we cannot just go ahead and define the rank of a matrix to be either the column rank or the row rank, since these may not agree for general rings, even commutative unit rings. However, for principal ideal domains, the row and column ranks do, in fact, agree. We state this result here, although our proof relies on Theorem 5.2.43 below. Note that the proof of Theorem 5.1.35 *does not* carry over to this case; for example, since it is not generally possible to extend a basis for a submodule to a basis for the module, even for rings over principal ideal domains (cf. Example 4.9.2).

5.2.38 Theorem (Row rank and column rank agree for matrices over principal ideal domains) *If* R *is a principal ideal domain, if* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$*, then* $\dim_{\mathsf{R}}(\mathrm{image}(\mathbf{A})) = \dim_{\mathsf{R}}(\mathrm{image}(\mathbf{A}^{\mathrm{T}}))$*. In particular, the row rank and the column rank of* $\mathbf{A}$ *are the same.*

Proof By Lemma 1 of Theorem 5.2.43 $\boldsymbol{A}$ is equivalent to a matrix of the form

$$\boldsymbol{B} = \left[\begin{array}{c|c} \boldsymbol{D}_r & \boldsymbol{0}_{r\times(n-r)} \\ \hline \boldsymbol{0}_{(m-r)\times r} & \boldsymbol{0}_{(m-r)\times(n-r)} \end{array}\right],$$

with $\boldsymbol{D}_r$ being the diagonal matrix

$$\boldsymbol{D}_r = \begin{bmatrix} d_1 & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_r \end{bmatrix},$$

for nonzero $d_1, \dots, d_r \in \mathsf{R}$ (we do not need here the additional divisibility properties of these diagonal elements). That is, $\boldsymbol{A} = \boldsymbol{PBQ}$ for invertible $\boldsymbol{P} \in \mathrm{Mat}_{m\times m}(\mathsf{R})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$. Thus $\boldsymbol{A}^T = \boldsymbol{Q}^T\boldsymbol{B}^T\boldsymbol{P}^T$, whence $\boldsymbol{A}^T$ is equivalent to $\boldsymbol{B}^T$ by Proposition 5.2.22. Now by Lemma 2 of Theorem 5.2.43 it follows that the row and column ranks of both $\boldsymbol{A}$ and $\boldsymbol{A}^T$ are equal to r. ■

We can thus define rank for matrices over principal ideal domains.

5.2.39 Definition (Rank of matrices over principal ideal domains) If R is a principal ideal domain, if $m, n \in \mathbb{Z}_{>0}$, and if $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{R})$, then the ***rank*** of $\boldsymbol{A}$ is equal to the column or row rank of $\boldsymbol{A}$, and is denoted by $\mathrm{rank}(\boldsymbol{A})$. •

From Theorem 5.2.38 the following result holds.

5.2.40 Corollary (Rank and rank of transpose agree for matrices over principal ideal domains) *If* R *is a principal ideal domain, if* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$*, then* $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}_2^{\mathrm{T}})$*.*

Now we discuss equivalence of matrices over rings. We give the definition in the general setting, although we will not be able to say anything useful except in the case when the matrices are over a principal ideal domain and have finitely many rows and columns. The equivalence of arbitrary matrices over general rings is very complicated.

5.2.41 Definition (Equivalence of matrices over rings) Let R be a ring, let I and J be index sets, and let $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{I\times J}(\mathsf{R})$ be column finite. The matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are ***equivalent*** if there exist column finite invertible matrices $\boldsymbol{P} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{J\times J}(\mathsf{R})$ such that $\boldsymbol{A}_2 = \boldsymbol{PA}_1\boldsymbol{Q}$. •

Just as for matrices over fields (Proposition 5.1.39), the notion of two matrices being equivalent as in the definition defines an equivalence relation in $\mathrm{Mat}_{I\times J}(\mathsf{R})$. The problem for equivalence of matrices over general rings being as complicated as it is, we simply consider the case of matrices over principal ideal domains. Fortunately, all cases of matrices over rings that will be of interest to us are covered by the following theorem.

Before we state this result, let us state a useful result concerning elementary and secondary operations and rank.

5.2.42 Proposition (Elementary and secondary operations and rank) *Let* R *be a commutative unit ring, let* $m, n \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{R})$*, supposing that the rowspace and columnspace of* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are free. Then the following statements hold:*

(i) if $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row equivalent then the row ranks of* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *agree;*

(ii) if $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are column equivalent then the column ranks of* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *agree.*

Proof The idea here is exactly like the corresponding proof for matrices over fields given in Proposition 5.1.40. ■

5.2.43 Theorem (Characterisation of equivalence for matrices over principal ideal domains) *Let* R *be a principal ideal domain, let* $m, n \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{m\times n}(\mathsf{R})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are equivalent;*

(ii) there exists $r \in \mathbb{Z}_{\geq 0}$ *and nonzero ideals* $\mathsf{I}_1, \dots, \mathsf{I}_r \subseteq \mathsf{R}$ *uniquely determined by the conditions*

(a) $\mathsf{I}_1 \subseteq \cdots \subseteq \mathsf{I}_r$ *and*

(b) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are equivalent to a matrix of the form*

$$\left[\begin{array}{c|c} \mathbf{D}_r & \mathbf{0}_{r\times(n-r)} \\ \hline \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{array}\right],$$

with $\mathbf{D}_r$ *being the diagonal matrix*

$$\mathbf{D}_r = \begin{bmatrix} d_1 & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_r \end{bmatrix},$$

and where $\mathsf{I}_j = (d_j)$ *for* $j \in \{1, \dots, r\}$*.*

Proof We first prove the following lemma, from which the rest of the proof is fairly easily deduced.

1 Lemma *Let* R *be a principal ideal domain, let* $m, n \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{m\times n}(\mathsf{R})$*. Then there exists* $r \in \mathbb{Z}_{\geq 0}$*, and* $d_1, \dots, d_r \in \mathsf{R}$ *such that* $\mathbf{A}$ *can be transformed by a finite sequence of elementary and secondary row and column operations into a matrix of the form*

$$\left[\begin{array}{c|c} \mathbf{D}_r & \mathbf{0}_{r\times(n-r)} \\ \hline \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{array}\right],$$

with $\mathbf{D}_\mathrm{r}$ *being the diagonal matrix*

$$\mathbf{D}_\mathrm{r} = \begin{bmatrix} \mathrm{d}_1 & 0_\mathsf{R} & \cdots & 0_\mathsf{R} \\ 0_\mathsf{R} & \mathrm{d}_2 & \cdots & 0_\mathsf{R} \\ \vdots & \vdots & \ddots & \vdots \\ 0_\mathsf{R} & 0_\mathsf{R} & \cdots & \mathrm{d}_\mathrm{r} \end{bmatrix},$$

and where $\mathrm{d}_1|\cdots|\mathrm{d}_\mathrm{r}$.

Proof The proof of the lemma is rather tedious. Let us first describe a few of the basic procedures that we will use throughout the proof. In these descriptions, $\boldsymbol{M}$ is a general $m \times n$ matrix.

1. *Zeroing elements in the first row I:* We suppose that $\boldsymbol{M}(1,1) \neq 0_\mathsf{R}$. The objective is to perform an elementary or secondary operation that transforms $\boldsymbol{M}$ into a matrix with a zero as the $(1,j)$th component for some $j \in \{2,\ldots,n\}$. In this procedure, we suppose that $\boldsymbol{M}(1,1)|\boldsymbol{M}(1,j)$ so that there exists $r \in \mathsf{R}$ such that $\boldsymbol{M}(1,j) = r\boldsymbol{M}(1,1)$. Then subtracting r times the first column from the jth column does the job. Let us record some features of the transformed matrix.

   (a) As desired, the transformed matrix has a zero in the $(1,j)$th position.

   (b) The first column of the transformed matrix is equal to the first column of $\boldsymbol{M}$.

   (c) If $a \in \mathsf{R}$ is a common divisor for all elements of $\boldsymbol{M}$, a is also a common divisor for all elements of the transformed matrix. This is true because the components of transformed matrix are comprised of linear combinations of components of $\boldsymbol{M}$.

2. *Zeroing elements in the first row II:* The objective here is as in the last procedure: use elementary and secondary operations to transform $\boldsymbol{M}$ into a matrix with zero as the $(1,j)$th component, for some $j \in \{2,\ldots,n\}$, supposing that $\boldsymbol{M}(1,1) \neq 0_\mathsf{R}$. Here we suppose that $\boldsymbol{M}(1,1) \nmid \boldsymbol{M}(1,j)$. We let d be a greatest common divisor for $\boldsymbol{M}(1,1)$ and $\boldsymbol{M}(1,j)$ so that, by Proposition 4.2.77(ii) there exists $r_1, r_2 \in \mathsf{R}$ such that $r_1\boldsymbol{M}(1,1) + r_2\boldsymbol{M}(1,j) = d$. Also let $d_1, d_2 \in \mathsf{R}$ satisfy $dd_1 = \boldsymbol{M}(1,1)$ and $dd_2 = \boldsymbol{M}(1,j)$. With this notation one has

   $$d_1\boldsymbol{M}(1,j) - d_2\boldsymbol{M}(1,1) = d_1dd_2 - d_2dd_1 = 0_\mathsf{R}.$$

   Define a secondary matrix $\boldsymbol{S}$ whose top left 2×2 block is

   $$\begin{bmatrix} r_1 & -d_2 \\ r_2 & d_1 \end{bmatrix}.$$

   This is indeed a secondary matrix by virtue of the identity

   $$r_1dd_1 + r_2dd_2 = d(r_1d_1 + r_2d_2) = d \quad \Longrightarrow \quad r_1d_1 + r_2d_2 = 1_\mathsf{R}$$

   which follows since $d \neq 0_\mathsf{R}$.

   Now (1) swap the jth and second columns, (2) then apply secondary column transformation corresponding to $\boldsymbol{S}$, and (3) then again swap the jth and second columns. One directly sees that the $(1,j)$th component of the resulting matrix is zero. Let us record some properties of the transformed matrix.

(a) As desired, the transformed matrix has a zero in the $(1, j)$th position.

(b) The first column of the transformed matrix may not be equal to the first column of $\boldsymbol{M}$.

(c) The $(1, 1)$ component of the matrix is directly shown to be d which is a nonzero divisor of $\boldsymbol{M}(1, 1)$. Moreover, $\boldsymbol{M}(1, 1) \nmid d$.

(d) If $a \in \mathsf{R}$ is a common divisor for all elements of $\boldsymbol{M}$, a is also a common divisor for all elements of the transformed matrix. This is true because the components of transformed matrix are comprised of linear combinations of components of $\boldsymbol{M}$.

3. *Zeroing elements in the first column I:* We suppose that $\boldsymbol{M}(1, 1) \neq 0_{\mathsf{R}}$ and that the objective is to transform $\boldsymbol{M}$ into a matrix whose $(i, 1)$st entry is zero for some $i \in \{2, \dots, m\}$. We suppose in this procedure that $\boldsymbol{M}(1, 1)|\boldsymbol{M}(i, 1)$ so that $\boldsymbol{M}(i, 1) = r\boldsymbol{M}(1, 1)$ for some $r \in \mathsf{R}$. By subtracting r times the first row from the ith row we achieve the desired result. Let us record some properties of the transformed matrix.

   (a) As desired, the transformed matrix has a zero in the $(i, 1)$st position.

   (b) The first row of the transformed matrix is equal to the first row of $\boldsymbol{M}$.

   (c) If $a \in \mathsf{R}$ is a common divisor for all elements of $\boldsymbol{M}$, a is also a common divisor for all elements of the transformed matrix. This is true because the components of transformed matrix are comprised of linear combinations of components of $\boldsymbol{M}$.

4. *Zeroing elements in the first column II:* Again we suppose that $\boldsymbol{M}(1, 1) \neq 0_{\mathsf{R}}$ and that the objective is to transform $\boldsymbol{M}$ into a matrix whose $(i, 1)$st entry is zero for some $i \in \{2, \dots, m\}$. In this case we suppose that $\boldsymbol{M}(1, 1) \nmid \boldsymbol{M}(i, 1)$. We let d be a greatest common divisor for $\boldsymbol{M}(1, 1)$ and $\boldsymbol{M}(i, 1)$ so that, by Proposition 4.2.77(ii), there exists $r_1, r_2 \in \mathsf{R}$ for which $r_1\boldsymbol{M}(1, 1) + r_2\boldsymbol{M}(i, 1) = d$. We also let $d_1, d_2 \in \mathsf{R}$ satisfy $dd_1 = \boldsymbol{M}(1, 1)$ and $dd_2 = \boldsymbol{M}(i, 1)$. Then we have

   $$d_1\boldsymbol{M}(i, 1) - d_2\boldsymbol{M}(1, 1) = d_1dd_2 - d_2dd_1 = 0_{\mathsf{R}}.$$

   Now define a secondary matrix $\boldsymbol{S}$ whose top left 2×2 block is

   $$\begin{bmatrix} r_1 & r_2 \\ -d_2 & d_1 \end{bmatrix}.$$

   This is a secondary matrix since $r_1d_2 + r_2d_2 = 1_{\mathsf{R}}$, just as we argued in procedure 2. Now (1) swap the ith and second rows, (2) then apply the secondary row transformation corresponding to $\boldsymbol{S}$, and (3) then again swap the ith and second rows. The resulting matrix can be directly checked to have a zero as its $(i, 1)$st entry. Let us record some properties of the transformed matrix.

   (a) As desired, the transformed matrix has a zero in the $(i, 1)$st position.

   (b) The first row of the transformed matrix may not be equal to the first row of $\boldsymbol{M}$.

   (c) The $(1, 1)$ component of the matrix is directly shown to be d which is a nonzero divisor of $\boldsymbol{M}(1, 1)$. Moreover, $\boldsymbol{M}(1, 1) \nmid d$.

(d) If $a \in \mathsf{R}$ is a common divisor for all elements of $\boldsymbol{M}$, a is also a common divisor for all elements of the transformed matrix. This is true because the components of transformed matrix are comprised of linear combinations of components of $\boldsymbol{M}$.

5. *Making the* $(1,1)$ *component a divisor for another component:* Here we suppose that the matrix $\boldsymbol{M}$ has the form
$$\begin{bmatrix} \boldsymbol{M}(1,1) & \boldsymbol{0}_{1\times(n-1)} \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{M}' \end{bmatrix}.$$
The objective is to transform $\boldsymbol{M}$, using elementary and secondary operations, into a matrix whose (i,j)th component has its $(1,1)$ component as a divisor for some $i \in \{2,\dots,m\}$ and $j \in \{2,\dots,n\}$. The procedure is this. If $\boldsymbol{M}(1,1)|\boldsymbol{M}(i,j)$ already, then nothing needs to be done. Otherwise, add row i to row 1 to get a matrix $\boldsymbol{N}$. The first row of $\boldsymbol{N}$ is the ith row of $\boldsymbol{M}$, except in the first entry where one adds $\boldsymbol{M}(1,1)$. If the $\boldsymbol{N}(1,1)|\boldsymbol{N}(1,j)$ then we are done since $\boldsymbol{N}(1,j) = \boldsymbol{N}(i,j)$. Otherwise, apply Procedure 2 to $\boldsymbol{N}$ and note that the $(1,1)$ component in the resulting matrix is a divisor of $\boldsymbol{N}(1,j) = \boldsymbol{N}(i,j)$. This is what we wanted. Let us make some observations about the transformed matrix.

(a) As desired the $(1,1)$ component divides the (i,j) component.

(b) The first row of the transformed matrix may have nonzero $(1,j)$ components for $j \in \{2,\dots,n\}$. Thus the transformed matrix may not have the same form as $\boldsymbol{M}$.

In what follows, we shall refer to the procedures above as Procedures 1–5.

If $\boldsymbol{A} = \boldsymbol{0}_{m\times n}$ then there is nothing to prove. So we suppose that $\boldsymbol{A}$ is nonzero. Now we carry out a sequence of operations as follows.

1. Repeatedly apply Procedures 1 and/or 2 to transform $\boldsymbol{A}$ into a matrix $\boldsymbol{B}_1$ whose first row is zero except for the first entry which we denote by a_1.
2. If a_1 divides all elements in the first column of $\boldsymbol{B}_2$, repeatedly apply Procedure 3 to arrive at a matrix $\boldsymbol{B}_2$ whose first row and column is zero except for the first element. Stop.
3. If a_1 does not divide all elements in the first column of $\boldsymbol{B}_1$ then repeatedly apply Procedures 3 and/or 4 to transform $\boldsymbol{B}_1$ into a matrix $\boldsymbol{B}_2$ whose first column is zero except for the first entry. The first entry we denote by a_2 and we note that $a_2|a_1$. Also note that the first row of $\boldsymbol{B}_2$ may have nonzero entries in positions 2 through n.
4. If a_2 divides all elements in the first row of $\boldsymbol{B}_2$ then repeatedly apply Procedure 1 to arrive at a matrix $\boldsymbol{B}_3$ whose first row and column is zero except for the first element. Stop.
5. If a_2 does not divide all elements in the first column of $\boldsymbol{B}_2$ then repeatedly apply Procedures 1 and/or 2 to transform $\boldsymbol{B}_2$ into a matrix $\boldsymbol{B}_3$ whose first row is zero except for the first entry. The first entry we denote by a_3 and we note that $a_3|a_2$. Also note that the first column of $\boldsymbol{B}_3$ may have nonzero entries in positions 2 through m.
6. Repeat steps 2 through 5.

We claim that this process must terminate in a finite number of steps. Suppose otherwise. Then it must be the case that we successively apply Procedures 2 and 4 (possibly along with Procedures 1 and 3 as well) to arrive at a sequence $\boldsymbol{B}_1, \boldsymbol{B}_2, \dots$ of matrices whose $(1,1)$ components we denote by $b_1, b_2, \dots$ and which satisfy $b_1 | b_2 | \cdots$. Moreover, $b_{j+1} \nmid b_j$ for $j \in \mathbb{Z}_{>0}$. However, this is not possible since it would imply that b_1 has an infinite number of prime factors, violating the condition that R is a principal ideal domain.

The above argument shows that $\boldsymbol{A}$ can be transformed by a finite number of elementary and secondary operations into a matrix of the form

$$\boldsymbol{C}_1 = \left[\begin{array}{c|c} c_1 & \boldsymbol{0}_{1\times(n-1)} \\ \hline \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{C}_1' \end{array}\right].$$

Now consider a sequence of operations as follows.

7. If c_1 divides every component of $\boldsymbol{C}_1'$ then stop.
8. If c_1 does not divide every component of $\boldsymbol{C}_1'$ then $c_1 \nmid \boldsymbol{C}_1'(i-1, j-1)$ for some $(i,j) \in \{2, \dots, m\} \times \{2, \dots, n\}$. Apply Procedure 5 to obtain a matrix $\boldsymbol{C}_2'$ for which $\boldsymbol{C}_2'(1,1) | c_1$, but for which the first row may have nonzero entries in positions $2, \dots, n$.
9. Apply steps 1 through 6 above to $\boldsymbol{C}_2'$ to obtain a matrix of the form
$$\boldsymbol{C}_2 = \left[\begin{array}{c|c} c_2 & \boldsymbol{0}_{1\times(n-1)} \\ \hline \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{C}_2' \end{array}\right]$$
where $c_2 | c_1$ but $c_1 \nmid c_2$.
10. Repeat steps 7 through 9 on $\boldsymbol{C}_2$.

We claim that this sequence of operations must terminate after a finite number of steps. This is so because otherwise one would produce an infinite number of nonassociate divisors for c_1, which contradicts R being a principal ideal domain. Thus we arrive, by a finite sequence of elementary and secondary operations applied to $\boldsymbol{C}_1$, at a matrix of the form

$$\left[\begin{array}{c|c} d_1 & \boldsymbol{0}_{(n-1)\times 1} \\ \hline \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{A}_1 \end{array}\right],$$

where d_1 is nonzero and divides all entries of the matrix $\boldsymbol{A}_1$. Now one applies the procedure above (i.e., steps 1 through 10) to the matrix $\boldsymbol{A}_1$. If $\boldsymbol{A}_1 = \boldsymbol{0}_{(m-1)\times(n-1)}$ then the proof is complete. Otherwise one proceeds to transform $\boldsymbol{A}_1$ into a matrix of the form

$$\left[\begin{array}{c|c} d_2 & \boldsymbol{0}_{(n-2)\times 1} \\ \hline \boldsymbol{0}_{(m-2)\times 1} & \boldsymbol{A}_2 \end{array}\right],$$

where d_2 divides all components of $\boldsymbol{A}_2$. We also claim that $d_1 | d_2$. This follows since d_1 divides all entries of $\boldsymbol{A}_1$ and since all of the operations performed to transform $\boldsymbol{A}_1$ preserve common divisors, just as stated above for Procedures 1–5.

Now the lemma follows by induction. ▼

Ugh. . . now we proceed with the proof of the theorem. By the lemma there exists $r_1, r_2 \in \mathbb{Z}_{\geq 0}$ and $d_{11}, \dots, d_{1r_1} \in \mathsf{R}$ and $d_{21}, \dots, d_{2r_2} \in \mathsf{R}$ such that $\boldsymbol{A}_1$ is equivalent to

$$\boldsymbol{B}_1 = \left[\begin{array}{c|c} \boldsymbol{D}_1 & \boldsymbol{0}_{r_1\times(n-r_1)} \\ \hline \boldsymbol{0}_{(m-r_1)\times r_1} & \boldsymbol{0}_{(m-r_1)\times(n-r_1)} \end{array}\right],$$

with $\boldsymbol{D}_1$ being the diagonal matrix

$$\boldsymbol{D}_1 = \begin{bmatrix} d_{11} & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_{12} & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_{1r_1} \end{bmatrix}$$

and $\boldsymbol{A}_2$ is equivalent to

$$\boldsymbol{B}_2 = \left[\begin{array}{c|c} \boldsymbol{D}_2 & \boldsymbol{0}_{r_2\times(n-r_2)} \\ \hline \boldsymbol{0}_{(m-r_2)\times r_2} & \boldsymbol{0}_{(m-r_2)\times(n-r_2)} \end{array}\right],$$

with $\boldsymbol{D}_2$ being the diagonal matrix

$$\boldsymbol{D}_2 = \begin{bmatrix} d_{21} & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_{22} & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_{2r_2} \end{bmatrix}.$$

We also have $d_{11}|\cdots|d_{1r_1}$ and $d_{21}|\cdots|d_{2r_2}$.

(i) $\Longrightarrow$ (ii) Equivalence of matrices being an equivalence relation, the matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are equivalent if and only if the matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are equivalent. Thus $r \in \mathbb{Z}_{>0}$ and ideals $\mathsf{I}_1, \dots, \mathsf{I}_r$ exist as stated in the theorem by taking, for example, $r = r_1$ and $\mathsf{I}_j = (d_{1j})$ for $r \in \{1, \dots, r\}$. That $\mathsf{I}_1 \subseteq \cdots \subseteq \mathsf{I}_r$ follows from Proposition 4.2.61 since $d_{11}|\cdots|d_{1r}$. Now we show that r and the ideals are uniquely determined by the stated conditions in part (ii). The following lemma constitutes the first step in this.

2 Lemma *Let* R *be an integral domain, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$*. If* $\mathbf{A}$ *can be transformed by a finite sequence of elementary and secondary row and column operations into a matrix of the form*

$$\left[\begin{array}{c|c} \mathbf{D}_{\mathrm{k}} & \mathbf{0}_{\mathrm{k}\times(\mathrm{n}-\mathrm{k})} \\ \hline \mathbf{0}_{(\mathrm{m}-\mathrm{k})\times\mathrm{k}} & \mathbf{0}_{(\mathrm{m}-\mathrm{k})\times(\mathrm{n}-\mathrm{k})} \end{array}\right]$$

where $\mathbf{D}_{\mathrm{k}}$ *is a diagonal matrix, all of whose diagonal entries are nonzero, then the row rank and column rank of* $\mathbf{A}$ *can be defined and are both equal to* k*.*

Proof Let us denote by $\boldsymbol{B}$ the matrix in the statement of the lemma with $\boldsymbol{D}_k$ in the top left corner. By Proposition 5.2.30 and Theorem 5.2.31(iii) we have

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{B}\boldsymbol{Q}$$

for invertible $\boldsymbol{P} \in \mathrm{Mat}_{m\times m}(\mathsf{R})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$. Thus $\boldsymbol{A}$ is column equivalent to $\boldsymbol{P}\boldsymbol{B}$, and so, by Proposition 5.2.42, $\boldsymbol{A}$ has column rank k if $\boldsymbol{P}\boldsymbol{B}$ has column rank k. To show this we write $\boldsymbol{B}$ in terms of its columns:

$$\boldsymbol{B} = \left[\begin{array}{c|c|c|c|c|c} r_1\boldsymbol{e}_1 & \dots & r_k\boldsymbol{e}_k & \boldsymbol{0}_{\mathsf{R}^m} & \cdots & \boldsymbol{0}_{\mathsf{R}^m} \end{array}\right]$$

so that

$$\boldsymbol{P}\boldsymbol{B} = \left[\begin{array}{c|c|c|c|c|c} r_1\boldsymbol{P}\boldsymbol{e}_1 & \dots & r_k\boldsymbol{P}\boldsymbol{e}_k & \boldsymbol{0}_{\mathsf{R}^m} & \cdots & \boldsymbol{0}_{\mathsf{R}^m} \end{array}\right],$$

where $r_1, \dots, r_k$ are the diagonal elements of $\boldsymbol{D}_k$. From this expression it is clear that the column rank of $\boldsymbol{PB}$ is at most k. To show that it is equal to k we show that the first k columns of $\boldsymbol{PB}$ are linearly independent. Suppose that

$$a_1 r_1 \boldsymbol{P}\boldsymbol{e}_1 + \dots + a_k r_k \boldsymbol{P}\boldsymbol{e}_k = \boldsymbol{0}_{\mathsf{R}^m}$$

for $a_1, \dots, a_k \in \mathsf{R}$. Then since $\boldsymbol{P}$ is invertible we can multiply by $\boldsymbol{P}^{-1}$ on the left to obtain

$$a_1 r_1 \boldsymbol{e}_1 + \dots + a_k r_k \boldsymbol{e}_k = \boldsymbol{0}_{\mathsf{R}^m},$$

which gives $a_j r_j = 0_{\mathsf{R}}$, $j \in \{1, \dots, k\}$, since $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_k\}$ is linearly independent. Since R is an integral domain and since $r_1, \dots, r_k$ are nonzero, it follows that $a_j = 0_{\mathsf{R}}$, $j \in \{1, \dots, k\}$. Thus the first k columns of $\boldsymbol{PB}$ are linearly independent, and so the column rank of $\boldsymbol{PB}$, and so of $\boldsymbol{A}$, is k.

An entirely analogous computation can be used to show that the row rank of $\boldsymbol{A}$ is k. ▼

Returning now to our matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, the preceding lemma implies that $r_1 = r_2$ if these matrices are equivalent. Now note that

$$\operatorname{image}(\boldsymbol{B}_1) = \operatorname{span}_{\mathsf{R}}(d_{11}\boldsymbol{e}_1) \oplus \dots \oplus \operatorname{span}_{\mathsf{R}}(d_{1r}\boldsymbol{e}_r) \tag{5.13}$$

and so $\mathsf{R}^m / \operatorname{image}(\boldsymbol{B}_1)$ is isomorphic to

$$\mathsf{R}/(d_{11}) \oplus \dots \oplus \mathsf{R}/(d_{1r}) \oplus \mathsf{R} \oplus \dots \oplus \mathsf{R}$$

by Exercise 4.8.4 and by Proposition 4.9.7. In like manner, $\mathsf{R}^m / \operatorname{image}(\boldsymbol{B}_2)$ is isomorphic to

$$\mathsf{R}/(d_{21}) \oplus \dots \oplus \mathsf{R}/(d_{2r}) \oplus \mathsf{R} \oplus \dots \oplus \mathsf{R}. \tag{5.14}$$

Now $\operatorname{image}(\boldsymbol{B}_1)$ and $\operatorname{image}(\boldsymbol{B}_2)$ are isomorphic if $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are equivalent (why?). Therefore, the R-modules in (5.13) and (5.14) are isomorphic. Since $d_{11}|\cdots|d_{1r_1}$ and $d_{21}|\cdots|d_{2r_2}$ we can apply the uniqueness part of Theorem 4.9.21 to deduce that $(d_{1j}) = (d_{2j})$ for $j \in \{1, \dots, r\}$. Thus means that $d_{2j} = u_j d_{2j}$ for some unit $u_j \in \mathsf{R}$ for each $j \in \{1, \dots, r\}$. This gives the uniqueness of the ideals generated by the nonzero diagonals of the matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$.

(ii) $\Longrightarrow$ (i) This is a tautology. ■

The preceding proof is extremely tedious, although it has the advantage of providing an explicit construction of the matrix in part (ii) of the theorem by using elementary and secondary row and column operations.

The form of the matrix in part (ii) of the theorem has a name.

5.2.44 Definition (Smith normal form) Let R be a principal ideal domain, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{R})$ if $r \in \mathbb{Z}_{\geq 0}$ and $d_1, \dots, d_r \in \mathsf{R}$ satisfy the conditions

(i) $d_1 | \dots | d_r$ and

(ii) $\boldsymbol{A}$ is equivalent to the matrix

$$\boldsymbol{B} = \left[\begin{array}{c|c} \boldsymbol{D}_r & \boldsymbol{0}_{r\times(n-r)} \\ \hline \boldsymbol{0}_{(m-r)\times r} & \boldsymbol{0}_{(m-r)\times(n-r)} \end{array}\right],$$

with $\boldsymbol{D}_r$ being the diagonal matrix

$$\boldsymbol{D}_r = \begin{bmatrix} d_1 & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & d_2 & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & d_r \end{bmatrix},$$

then the matrix $\boldsymbol{B}$ is the ***Smith normal form*** for $\boldsymbol{A}$. •

Let us give an example of determining the Smith normal form of a matrix over a principal ideal domain.

5.2.45 Example (Smith normal form) We work with the ring $\mathbb{Z}$ and consider the matrix

$$\boldsymbol{A} = \begin{bmatrix} 8 & 4 & 8 \\ 4 & 8 & 4 \end{bmatrix}.$$

We now carry out row and column operations.

1. Swap rows:
$$\begin{bmatrix} 4 & 8 & 4 \\ 8 & 4 & 8 \end{bmatrix}.$$
2. Subtract 2 times row 1 from row 2:
$$\begin{bmatrix} 4 & 8 & 4 \\ 0 & -12 & 0 \end{bmatrix}.$$
3. Subtract 2 times column 1 from column 2:
$$\begin{bmatrix} 4 & 0 & 4 \\ 0 & -12 & 0 \end{bmatrix}.$$
4. Subtract column 1 from column 3:
$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & -12 & 0 \end{bmatrix}.$$
5. Multiply row 2 by −1:
$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 12 & 0 \end{bmatrix}.$$

This gives the Smith normal form in this case. •

Theorem 5.2.43 leads to the following theorem concerning equivalence of matrices over principal ideal domains. The reader should compare the following theorem to the analogous Theorem 5.1.42 for fields.

5.2.46 Theorem (Equivalence for invertible matrices over principal ideal domains) *Let* R *be a principal ideal domain, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}$ *is invertible;*

(ii) $\mathbf{A}$ *is equivalent to* $\mathbf{I}_{\mathrm{n}}$*;*

(iii) $\mathbf{A}$ *is row equivalent to* $\mathbf{I}_{\mathrm{n}}$*;*

(iv) there exists a unique invertible matrix $\mathbf{P}$ *such that* $\mathbf{PA} = \mathbf{I}_{\mathrm{n}}$*;*

(v) $\mathbf{A}$ *is column equivalent to* $\mathbf{I}_{\mathrm{n}}$*;*

(vi) there exists a unique invertible matrix $\mathbf{Q}$ *such that* $\mathbf{AQ} = \mathbf{I}_{\mathrm{n}}$*.*

Moreover, any of the six equivalent statements above implies that $\mathrm{rank}(\mathbf{A}) = \mathrm{n}$.

Proof (i) $\Longrightarrow$ (ii) This is part (iv) of Theorem 5.2.31, along with part (iii) of the same theorem and Proposition 5.2.30. In fact, Theorem 5.2.31 gives the equivalence of parts (i) and (ii).

(ii) $\Longrightarrow$ (iii) Note that part (ii) and the existence part of (iii) are equivalent by Theorem 5.2.31 and Proposition 5.2.30. We shall use this equivalence to more easily denote row equivalence. By Theorem 5.2.49 there exists an invertible matrix $\boldsymbol{P}'$ such that $\boldsymbol{P}'\boldsymbol{A}$ is in row Hermite form. Since $\boldsymbol{A}$ is invertible, the row Hermite matrix $\boldsymbol{P}'\boldsymbol{A}$ must also be invertible. Therefore, this row Hermite matrix cannot have any zero rows (why?) and so must be an upper triangular matrix whose diagonal elements are nonzero. Moreover, since $\boldsymbol{P}'\boldsymbol{A}$ is invertible, by Theorem 5.3.10 it follows that the product of its diagonal elements must be a unit. Thus each element of the diagonal must itself be a unit. Now we refer to the proof of Theorem 5.1.47 where we showed that for a matrix over a field one could, by elementary row operations, make all elements above the leading ones zero, and that these row operations leave the elements below the leading ones untouched. The same argument can be applied here to each column of $\boldsymbol{P}'\boldsymbol{A}$ to arrive at a matrix $\boldsymbol{P}''\boldsymbol{P}'\boldsymbol{A}$ that is diagonal and whose diagonal entries are units. Such a matrix can obviously be transformed into the identity matrix by elementary row operations. This gives an invertible matrix $\boldsymbol{P}$ such that $\boldsymbol{PA} = \boldsymbol{I}_n$. Thus $\boldsymbol{A}$ is row equivalent to $\boldsymbol{I}_n$.

(iii) $\Longrightarrow$ (iv) As mentioned in the preceding part of the proof, the existence part of this implication follows from Theorem 5.2.31 and Proposition 5.2.30. Thus there exists an invertible matrix $\boldsymbol{P}$ such that $\boldsymbol{PA} = \boldsymbol{I}_n$. Thus $\boldsymbol{A} = \boldsymbol{P}^{-1}$ which shows that $\boldsymbol{A}$ is invertible, and so $\boldsymbol{P}$ is uniquely determined by the requirement that $\boldsymbol{P} = \boldsymbol{A}^{-1}$.

(iv) $\Longrightarrow$ (v) If $\boldsymbol{PA} = \boldsymbol{I}_n$ for $\boldsymbol{P}$ invertible then $\boldsymbol{A} = \boldsymbol{P}^{-1}$ which shows that $\boldsymbol{A}$ is invertible and so $\boldsymbol{P} = \boldsymbol{A}^{-1}$. Thus $\boldsymbol{AP} = \boldsymbol{I}_n$, giving this part of the result by Theorem 5.2.31 and Proposition 5.2.30.

(v) $\Longrightarrow$ (vi) The existence part of this implication follows from Theorem 5.2.31 and Proposition 5.2.30. The uniqueness follows since, if $\boldsymbol{AQ} = \boldsymbol{I}_n$ for $\boldsymbol{Q}$ invertible, then $\boldsymbol{A} = \boldsymbol{Q}^{-1}$ which implies that $\boldsymbol{A}$ is invertible, whence $\boldsymbol{Q} = \boldsymbol{A}^{-1}$.

(vi) $\Longrightarrow$ (i) It is trivial that (vi) implies (ii), and the desired implication follows since, as mentioned above, (i) and (ii) are equivalent.

The final assertion of the theorem follows since, for example, if $\boldsymbol{A}$ is equivalent to $\boldsymbol{I}_n$, then $\mathrm{rank}(\boldsymbol{A}) = n$ since $\boldsymbol{I}_n$ is a Smith normal form for $\boldsymbol{A}$. ■

Conspicuously missing from the list of equivalent statements in the theorem is the condition that $\operatorname{rank}(\boldsymbol{A}) = n$. Indeed, it does not belong in the list, and we refer to Exercise 5.2.4 for an exploration of some of the consequences of this.

5.2.7 Characterisations of row and column equivalence for matrices over rings

In this section we say a few words about row and column equivalence for matrices over rings. Given all of our caveats thus far concerning matrices defined over general rings, it should not be surprising that we will not have much to say about such matrices, restricting instead to the case of matrices over principal ideal domains.

For matrices over rings, reduced row echelon form is too restrictive to be very useful. However, we can make use of the generalisation of row Hermite form.

5.2.47 Definition (Row Hermite form) Let R be a ring, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \operatorname{Mat}_{m\times n}(\mathsf{R})$. For $i \in \{1, \dots, m\}$ denote

$$E(i) = \begin{cases} \min\{j \in \{1,\dots,n\} \mid \boldsymbol{A}(i,j) \neq 0_{\mathsf{R}}\}, & \text{the } i\text{th row of } \boldsymbol{A} \text{ is nonzero}, \\ \infty, & \text{the } i\text{th row of } \boldsymbol{A} \text{ is zero}. \end{cases}$$

Then $\boldsymbol{A}$ is in ***row Hermite form*** if there exists $k \in \{1, \dots, m\}$ such that

(i) the first k rows of $\boldsymbol{A}$ are nonzero and the last $n - k$ rows of $\boldsymbol{A}$ are zero and such that

(ii) $i_1 < i_2$, $i_1, i_2 \in \{1, \dots, k\}$, implies that $E(i_1) < E(i_2)$. •

We refer to the equation following Definition 5.1.45 to remind the reader of the character of a matrix in row Hermite form.

One can say a few useful things about matrices in row Hermite form over fairly general rings.

5.2.48 Proposition (Rowspace of a matrix in row Hermite form) *If* R *is an integral domain, if* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and if* $\mathbf{A} \in \operatorname{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$ *is in row Hermite form, then the nonzero rows of* $\mathbf{A}$ *form a basis for* $\operatorname{rowspace}(\mathbf{A})$.

Proof The argument here exactly follows that for Proposition 5.1.46, making use of the fact that R is an integral domain to establish linear independence of the rows. ■

Now let us show that a matrix over a principal ideal domain can always be put into row Hermite form by a finite sequence of elementary and secondary row operations. Note that the result we get here is nowhere near as strong as that for matrices over fields. In particular, we use the uniqueness that we have for the reduced row echelon form.

5.2.49 Theorem (Row equivalence and row Hermite form) *Let* R *be a principal ideal domain and let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$. *Then each equivalence class in* $\mathsf{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$ *under the equivalence relation of row equivalence contains at least one matrix in row Hermite form.*

Proof The proof is a blending of the techniques used to prove Theorems 5.1.47 and 5.2.43. Thus we shall be a little sketchy about the details since the reader understanding the other two proofs will quickly understand this one.

Let $\boldsymbol{A} \in \mathsf{Mat}_{m\times n}(\mathsf{R})$. If $\boldsymbol{A}$ is the zero matrix, then we are done since this matrix is in row Hermite form. Let j_1 be the smallest positive integer for which the j_1st column of $\boldsymbol{A}$ is nonzero, and let i_1 have the property that $\boldsymbol{A}(i_1, j_1) \neq 0_{\mathsf{R}}$. Let $\boldsymbol{A}'_1$ be the matrix obtained from $\boldsymbol{A}$ by swapping the 1st and j_1st rows. Thus $\boldsymbol{A}'_1$ has the form

$$\boldsymbol{A}'_1 = \begin{bmatrix} 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & a^1_{11} & a^1_{12} & \cdots & a^1_{1k_1} \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & b^1_{21} & b^1_{22} & \cdots & b^1_{2k_1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & b^1_{n1} & b^1_{n2} & \cdots & b^1_{nk_1} \end{bmatrix},$$

where $a^1_{11} \neq 0_{\mathsf{F}}$. Now repeatedly apply Procedures 3 and/or 4 from the proof of Lemma 1 in Theorem 5.2.43 to the j_1st column of $\boldsymbol{A}'_1$ to make all entries in this column zero except the first. The result is a matrix of the form

$$\boldsymbol{A}_1 = \begin{bmatrix} 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & a^1_{11} & a^1_{12} & \cdots & a^1_{1k_1} \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & a^1_{22} & \cdots & a^1_{2k_1} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} & a^1_{n2} & \cdots & a^1_{nk_1} \end{bmatrix}.$$

Now repeat the procedure on the matrix

$$\begin{bmatrix} a^1_{22} & \cdots & a^1_{k_1} \\ \vdots & \ddots & \vdots \\ a^1_{n2} & \cdots & a^1_{nk_1} \end{bmatrix},$$

and then continue inductively to give the desired row Hermite form. ■

Let us give a simple example of this construction. Generally, the process of putting a matrix into row Hermite form can be involved, since it requires computing greatest common divisors, cf. the proof of Lemma 1 of Theorem 5.2.43.

5.2.50 Example (Row Hermite form) For the ring $\mathbb{Z}$, take $\boldsymbol{A} \in \mathsf{Mat}_{3\times 4}(\mathbb{Z})$ to be

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 3 & 1 \\ 2 & 1 & 7 & 1 \end{bmatrix}.$$

We now perform a sequence of elementary row operations.

1. Swap the first and second row:

$$\begin{bmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & 2 \\ 2 & 1 & 7 & 1 \end{bmatrix}.$$

2. Swap the second and third row:

$$\begin{bmatrix} 1 & 0 & 3 & 1 \\ 2 & 1 & 7 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}.$$

3. Subtract 2 times the first row from the second row:

$$\begin{bmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 1 & 1 & 2 \end{bmatrix}.$$

4. Subtract the second row from the third row:

$$\begin{bmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

which gives a matrix in row Hermite form. •

Analogously to Theorem 5.1.50 for fields, we have the following result for matrices over principal ideal domains.

5.2.51 Theorem (Characterisations of row equivalence) *Let* R *be a principal ideal domain, let* $\mathrm{m}, \mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{n}}(\mathsf{R})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are row equivalent;*

(ii) there exists an invertible matrix $\mathbf{P} \in \mathrm{Mat}_{\mathrm{m}\times\mathrm{m}}(\mathsf{R})$ *such that* $\mathbf{P}\mathbf{A}_1 = \mathbf{A}_2$*;*

(iii) $\mathrm{rowspace}(\mathbf{A}_1) = \mathrm{rowspace}(\mathbf{A}_2)$.

Moreover, any of the preceding equivalent statements implies that $\ker(\mathbf{A}_1) = \ker(\mathbf{A}_2)$.

Proof (i) $\Longrightarrow$ (ii) This follows from Theorem 5.2.31 and Proposition 5.2.30.

(ii) $\Longrightarrow$ (iii) One can easily see that if $\boldsymbol{P}\boldsymbol{A}_1 = \boldsymbol{A}_2$ then the rows of $\boldsymbol{A}_2$ are linear combinations of the rows of $\boldsymbol{A}_1$. Moreover, the rows of $\boldsymbol{A}_1$ are, by similar reasoning, linear combinations of the rows of $\boldsymbol{A}_2$ since $\boldsymbol{A}_1 = \boldsymbol{P}^{-1}\boldsymbol{A}_2$.

(iii) $\Longrightarrow$ (i) By Theorem 5.2.49 the matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ possess row Hermite forms $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, respectively. Moreover,

$$\mathrm{rowspace}(\boldsymbol{A}_1) = \mathrm{rowspace}(\boldsymbol{B}_1), \quad \mathrm{rowspace}(\boldsymbol{A}_2) = \mathrm{rowspace}(\boldsymbol{B}_2)$$

by the previous part of the proof. Therefore, the rowspaces of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ agree. By Proposition 5.2.48 the nonzero rows of the matrices $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ form a basis for their rowspace. Let us denote the rows of $\boldsymbol{B}_a$ by the vectors $\{\boldsymbol{r}_{a1},\dots,\boldsymbol{r}_{am}\}$, $a \in \{1,2\}$. Therefore, by Theorem 5.5.17, there exists an invertible matrix $\boldsymbol{P}$ such that

$$\boldsymbol{r}_{2j} = \sum_{i=1}^{m} \boldsymbol{P}(j,i)\boldsymbol{r}_{1i}.$$

But this means that $\boldsymbol{B}_2 = \boldsymbol{P}\boldsymbol{B}_1$. The row equivalence of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, and in consequence of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, follows from Theorem 5.2.31 and Proposition 5.2.30.

The last assertion is proved by noting that part (ii) implies that

$$\boldsymbol{A}_1\boldsymbol{x} = \boldsymbol{0}_{\mathsf{R}^m} \quad\iff\quad \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{x} = \boldsymbol{0}_{\mathsf{R}^m} \quad\iff\quad \boldsymbol{A}_2\boldsymbol{x} = \boldsymbol{0}_{\mathsf{R}^m},$$

using the fact that $\ker(\boldsymbol{P}) = \{\boldsymbol{0}_{\mathsf{R}^m}\}$. ■

Note that the assertion that $\ker(\boldsymbol{A}_1) = \ker(\boldsymbol{A}_2)$ is *not* generally equivalent to the other three statements in the theorem. We ask the reader to consider this in Exercise 5.2.7.

Of course, one also has the analogous theorem for column equivalence. We do not state this, but refer the reader to Theorem 5.1.51 to make the trivial transcriptions required.

5.2.8 Systems of linear equations over rings

Systems of linear equations over rings are more complicated than their brethren over fields. In this section we give the definitions and basic results, and indicate where some of the differences lie with the situation when compared to the results of Section 5.1.8.

First we give the definitions. For simplicity we deal with commutative rings. The reader can easily generalise to the multiple definitions needed for left and right products for noncommutative rings.

5.2.52 Definition (System of linear equations over a ring) Let R be a commutative ring and let I and J be index sets.

(i) A ***system of linear equations*** over R is a pair $(\boldsymbol{A},\boldsymbol{b}) \in \mathrm{Mat}_{I\times J}(\mathsf{R}) \times \mathsf{R}_0^I$.

(ii) A system of linear equations $(\boldsymbol{A},\boldsymbol{b})$ is ***homogeneous*** if $\boldsymbol{b}(i) = 0_{\mathsf{R}}$ for every $i \in I$.

(iii) The ***solution set*** for a system of linear equations $(\boldsymbol{A},\boldsymbol{b})$ is the subset of R_0^J defined by

$$\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b}) = \{\boldsymbol{x} \in \mathsf{R}_0^J \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

A ***solution*** to the system of linear equations $(\boldsymbol{A},\boldsymbol{b})$ is an element of the solution set. •

We have the following simple characterisation of solutions of systems of equations.

5.2.53 Proposition (Existence and uniqueness of solutions) *Let* R *be a commutative ring, let* I *and* J *be index sets, and let* $(\mathbf{A},\mathbf{b}) \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R}) \times \mathsf{R}_0^{\mathrm{I}}$ *be a system of linear equations. Then the following statements hold:*

(i) $\mathrm{Sol}(\mathbf{A},\mathbf{b})$ *is nonempty if and only if* $\mathbf{b} \in \mathrm{image}(\mathbf{A})$*;*

(ii) in particular, $\mathrm{Sol}(\mathbf{A},\mathbf{b})$ *is nonempty for every* $\mathbf{b} \in \mathsf{R}_0^{\mathrm{I}}$ *if and only if* $\mathbf{A}$ *is surjective;*

(iii) $\mathrm{Sol}(\mathbf{A},\mathbf{b})$ *is a singleton if and only if*

(a) $\mathbf{b} \in \mathrm{image}(\mathbf{A})$ *and*

(b) $\mathbf{A}$ *is injective.*

Proof The proof follows that for Proposition 5.1.55, except that one appeals to Exercise 4.8.3 rather than Exercise 4.5.23. ■

Now we can give a "geometric" interpretation of the set of solutions of a system of linear equations.

5.2.54 Proposition (Characterisation of Sol(A, b)) *Let* R *be a commutative ring, let* I *and* J *be index sets, and let* $(\mathbf{A},\mathbf{b}) \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R}) \times \mathsf{R}_0^{\mathrm{I}}$ *be a system of linear equations in* R*. Then, for any* $\mathbf{x}_0 \in \mathrm{Sol}(\mathbf{A},\mathbf{b})$*,*

$$\mathrm{Sol}(\mathbf{A},\mathbf{b}) = \{\mathbf{x}+\mathbf{x}_0 \in \mathsf{R}_0^{\mathrm{J}} \mid \mathbf{x} \in \mathrm{Sol}(\mathbf{A},\mathbf{0}_{\mathsf{R}_0^{\mathrm{J}}})\}.$$

Proof The proof exactly mirrors that of Proposition 5.1.56. ■

Thus far we see that the generalities concerning systems of linear equations over a ring differ little from those for systems of linear equations over a field. To illustrate the differences we consider a couple of examples which should not be too surprising given the considerations of Section 5.2.6.

5.2.55 Examples (System of linear equations over a ring)

1. Let $\mathsf{R} = \mathbb{Z}$ and take

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}.$$

Note that $\mathrm{rank}(\boldsymbol{A}) = 2$. Also note that

$$\mathrm{image}(\boldsymbol{A}) = \{(j,2k) \in \mathbb{Z}^2 \mid j,k \in \mathbb{Z}\}.$$

Thus, while $\boldsymbol{A}$ has maximal rank, it is not the case that $\mathrm{Sol}(\boldsymbol{A},\boldsymbol{b})$ is nonempty for every $\boldsymbol{b}$. Indeed, b_2 must be even for solutions to exist. But when solutions to exist, they are unique.

2. Let $\mathsf{R} = \mathbb{Z}_4 = \mathbb{Z}/4\mathbb{Z}$ and take

$$\boldsymbol{A} = \begin{bmatrix} 1+4\mathbb{Z} & 0+4\mathbb{Z} \\ 0+4\mathbb{Z} & 2+4\mathbb{Z} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} b_1+4\mathbb{Z} \\ b_2+4\mathbb{Z} \end{bmatrix}.$$

Here we see that

$$\mathrm{image}(\boldsymbol{A}) = \{(j+4\mathbb{Z}, 2k+4\mathbb{Z}) \mid j,k \in \mathbb{Z}\}.$$

Thus, while $\operatorname{rank}(\boldsymbol{A}) = 2$ is maximal, we again only have $\operatorname{Sol}(\boldsymbol{A}, \boldsymbol{b}) \neq \emptyset$ when b_2 is even. In this case, there is an additional feature as well. While, $\operatorname{rank}(\boldsymbol{A})$ is maximal, we have

$$\operatorname{Sol}(\boldsymbol{A}, \boldsymbol{0}_{\mathsf{R}^2}) = \{(0 + 4\mathbb{Z}, 2k + 4\mathbb{Z}) \mid k \in \mathbb{Z}\},$$

implying that, when solutions exist, they are not unique. •

Exercises

5.2.1 State and prove the simplified version of Theorem 5.2.11 that results when the ring R is commutative.

5.2.2 Show that if R is a ring without a unity element and if I is an index set, then $\mathrm{id}_{\mathsf{R}_0^I} \in \operatorname{Hom}_{\mathsf{R}}(\mathsf{R}_0^I; \mathsf{R}_0^I)$, but that there is no matrix $\boldsymbol{A} \in \operatorname{Mat}_{I\times I}(\mathsf{R})$ such that $\boldsymbol{A}\boldsymbol{x} = \mathrm{id}_{\mathsf{R}_0^I}(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathsf{R}_0^I$.

5.2.3 Let R be a ring with unity element and let I be an index set.

(a) Show that the set of invertible column finite matrices in $\operatorname{Mat}_{I\times I}(\mathsf{R})$ is a group with product given by matrix multiplication.

In the case when $I = \{1, \dots, n\}$ this group of invertible matrices is denoted by $\mathsf{GL}(n; \mathsf{R})$ and is called the ***general linear group*** of order n over R.

(b) Is $\mathsf{GL}(n; \mathsf{R})$ a subalgebra of $\operatorname{Mat}_{n\times n}(\mathsf{R})$?

5.2.4 Answer the following questions:

(a) Find $\boldsymbol{A} \in \operatorname{Mat}_{n\times n}(\mathbb{Z})$ that has rank n but that is not invertible.

(b) Find $\boldsymbol{A} \in \operatorname{Mat}_{n\times n}(\mathbb{Z})$ that is injective, but does not possess a left-inverse that is a $\mathbb{Z}$-module homomorphism from $\mathbb{Z}^n$ to $\mathbb{Z}^n$. Find a left-inverse for $\boldsymbol{A}$ that is just a map from $\mathbb{Z}^n$ to $\mathbb{Z}^n$.

(c) Can you find $\boldsymbol{A} \in \operatorname{Mat}_{n\times n}(\mathbb{Z})$ that is surjective, but does not possess a right-inverse that is a $\mathbb{Z}$-module homomorphism?

5.2.5 Let R be a commutative unit ring and let $m, n \in \mathbb{Z}_{>0}$. Show that row equivalence and column equivalence define equivalence relations in $\operatorname{Mat}_{m\times n}(\mathsf{R})$.

5.2.6 Let R be a commutative unit ring and let $m, n \in \mathbb{Z}_{>0}$. Show that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are row equivalent if and only if $\boldsymbol{A}_1^T$ and $\boldsymbol{A}_2^T$ are column equivalent.

5.2.7 Find $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \operatorname{Mat}_{m\times n}(\mathbb{Z})$ such that $\ker(\boldsymbol{A}_1) = \ker(\boldsymbol{A}_2)$ but that $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are not row equivalent.

Section 5.3

Determinant and trace

The determinant is a useful computational tool that is also sometimes helpful in proving basic facts about matrices. In this section we discuss the determinant and some of its properties. We also discuss the trace of a matrix, as this will come up in Proposition 5.8.18 and Theorem V-5.2.6.

Throughout this section we will consider matrices over commutative unit rings since there is almost nothing that is easier were we to restrict to matrices over fields. And we will definitely be interested in matrices whose entries are elements from rings and not just from fields. However, for readers on the "fields only" program, little is lost by replacing "commutative unit ring" with "field" (or "$\mathbb{R}$" or "$\mathbb{C}$" for that matter), at least for getting started with the determinant.

Do I need to read this section? If you are familiar with the determinant, its computation, and its properties, then you can forgo this section. What is perhaps true, even for readers meeting the preceding conditions, is that they may not be aware of how useful determinants are for dealing with matrices whose entries are elements of a ring, and not elements of a field. This will come up for us in Section 5.8.4, and the reader might want to come back to determinants when they get to that material. •

5.3.1 Definition and basic properties of determinant

The determinant assigns to every square matrix with finitely many rows and columns a scalar. Thus the determinant is a function on the square matrices taking values in the ring from which the matrix takes its entries. The determinant, as we shall see, can be thought of as being a function of the rows or columns of a matrix. Specifically, the determinant is a multilinear function of the entries in a matrix, and we refer the reader ahead to Section 5.6 for a discussion of multilinearity.

We use the following result which gives a natural class of multilinear maps. In the statement of the result $\{e_1, \dots, e_n\}$ denotes the standard basis for R^n.

5.3.1 Theorem (Multilinear maps on $(\mathsf{R}^n)^n$) *If* R *is a commutative unit ring then, for* $\mathrm{r} \in \mathsf{R}$*, there exists a unique multilinear map* $\phi_{\mathrm{r}}\colon (\mathsf{R}^n)^n \to \mathsf{R}$ *such that*

$$\phi(\mathrm{e}_1, \dots, \mathrm{e}_n) = \mathrm{r}.$$

Proof We show the existence of ϕ_r by giving an explicit definition. Let $x_1, \dots, x_n \in \mathsf{R}^n$ and define

$$\phi_r(x_1, \dots, x_n) = r \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma) x_1(\sigma(1)) \cdots x_n(\sigma(n)).$$

Let us first show that this is an alternating multilinear map. First we show that ϕ_r is multilinear. Let $j_0 \in \{1, \dots, n\}$, let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n, \boldsymbol{y}_{j_0} \in \mathsf{R}^n$, and let $a \in \mathsf{R}$, and compute

$$\begin{aligned}
\phi_r(\boldsymbol{x}_1, \dots, &a\boldsymbol{x}_{j_0} + \boldsymbol{y}_0, \dots, \boldsymbol{x}_n) \\
&= r \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma)\boldsymbol{x}_1(\sigma(1)) \cdots (a\boldsymbol{x}_{j_0}(\sigma(j_0)) + \boldsymbol{y}_{j_0}(\sigma(j_0))) \cdots \boldsymbol{x}_n(\sigma(n)) \\
&= ra \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma)\boldsymbol{x}_1(\sigma(1)) \cdots \boldsymbol{x}_{j_0}(\sigma(j_0)) \cdots \boldsymbol{x}_n(\sigma(n)) \\
&\quad + \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma)\boldsymbol{x}_1(\sigma(1)) \cdots \boldsymbol{y}_{j_0}(\sigma(j_0)) \cdots \boldsymbol{x}_n(\sigma(n)) \\
&= a\phi_r(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{j_0}, \dots, \boldsymbol{x}_n) + \phi_r(\boldsymbol{x}_1, \dots, \boldsymbol{y}_{j_0}, \dots, \boldsymbol{x}_n),
\end{aligned}$$

giving multilinearity of ϕ_r.

To show that ϕ_r is alternating, let $i, j \in \{1, \dots, n\}$ be distinct. Denote by $\sigma_{ij} \in \mathfrak{S}_n$ the permutation that swaps i and j. Let $\mathfrak{E}_n$ denote the set of even permutations. We claim that the map

$$f_{ij} \colon \sigma \mapsto \sigma \circ \sigma_{ij}$$

is a bijection from $\mathfrak{E}_n$ to the set of odd permutations. Certainly $f_{ij}(\sigma)$ is odd if σ is even. That f_{ij} is injective follows since

$$f_{ij}(\sigma_1) = f_{ij}(\sigma_2) \quad \implies \quad \sigma_1 \circ \sigma_{ji} = \sigma_2 \circ \sigma_{ij} \quad \implies \quad \sigma_1 = \sigma_2.$$

Moreover, if σ is an odd permutation then $\sigma = (\sigma \circ \sigma_{ji}) \circ \sigma_{ij}$ and so f_{ij} is also surjective. This means that

$$\mathfrak{S}_n = \mathfrak{E}_n \cup \{f_{ij}(\sigma) \mid \sigma \in \mathfrak{E}_n\}.$$

Thus we have

$$\begin{aligned}
\phi_r(\boldsymbol{x}_1, &\dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_n) \\
&= r \sum_{\sigma \in \mathfrak{E}_n} \boldsymbol{x}_1(\sigma(1)) \cdots \boldsymbol{x}_i(\sigma(i)) \cdots \boldsymbol{x}_j(\sigma(j)) \cdots \boldsymbol{x}_n(\sigma(n)) \\
&\quad - r \sum_{\substack{f_{ij}(\sigma) \\ \sigma \in \mathfrak{E}_n}} \boldsymbol{x}_1(\sigma(1)) \cdots \boldsymbol{x}_i(\sigma(j)) \cdots \boldsymbol{x}_j(\sigma(i)) \cdots \boldsymbol{x}_n(\sigma(n)).
\end{aligned}$$

If $\boldsymbol{x}_i = \boldsymbol{x}_j$ it follows that

$$\phi_r(\boldsymbol{x}_1, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_n) = 0_{\mathsf{R}}.$$

Thus ϕ_r is alternating.

Now, since

$$\boldsymbol{e}_j(i) = \begin{cases} 1_{\mathsf{R}}, & i = j, \\ 0_{\mathsf{R}}, & \text{otherwise}, \end{cases}$$

it follows that the only term in the sum

$$\phi_r(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = r \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma)\boldsymbol{e}_1(\sigma(1)) \cdots \boldsymbol{e}_n(\sigma(n))$$

is that when σ is the identity map, and this then gives

$$\phi_r(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = r.$$

This gives the existence of ϕ_r.

Now we show the uniqueness of ϕ_r. Suppose that ψ_r is an alternating multilinear map such that

$$\psi_r(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = r.$$

Then

$$\begin{aligned}\psi_r(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) &= \psi_r\left(\sum_{j_1=1}^{n} \boldsymbol{x}_1(j_1)\boldsymbol{e}_{j_1}, \dots, \sum_{j_n=1}^{n} \boldsymbol{x}_n(j_n)\boldsymbol{e}_{j_n}\right) \\ &= \sum_{j_1=1}^{n} \cdots \sum_{j_n=1}^{n} \boldsymbol{x}_1(j_1) \cdots \boldsymbol{x}_n(j_n)\psi_r(\boldsymbol{e}_{j_1}, \dots, \boldsymbol{e}_{j_n}),\end{aligned}$$

using multilinearity. Since ψ_r is alternating only terms in the sum for which $j_1, \dots, j_n$ are distinct will remain. That is to say,

$$\psi_r(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \sum_{\sigma \in \mathfrak{S}_n} \boldsymbol{x}_1(\sigma(1)) \cdots \boldsymbol{x}_n(\sigma(n))\psi_r(\boldsymbol{e}_{\sigma(1)}, \dots, \boldsymbol{e}_{\sigma(n)}).$$

Since ψ_r is alternating it is skew-symmetric by Proposition 5.6.5 and so

$$\psi_r(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma)\boldsymbol{x}_1(\sigma(1)) \cdots \boldsymbol{x}_n(\sigma(n))\psi_r(\boldsymbol{e}_1, \dots, \boldsymbol{e}_n) = \phi_r(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n),$$

as desired. ■

It is now relatively straightforward to give the definition of the determinant of a square matrix. We recall from Definition 5.1.4 the notion of the row vectors of a matrix.

5.3.2 Definition (Determinant) Let R be a commutative unit ring and let $n \in \mathbb{Z}_{>0}$. The ***determinant*** is the map $\det\colon \mathrm{Mat}_{n\times n}(\mathsf{R}) \to \mathsf{R}$ defined by

$$\det \boldsymbol{A} = \phi_{1_{\mathsf{R}}}(\boldsymbol{r}(\boldsymbol{A}, 1), \dots, \boldsymbol{r}(\boldsymbol{A}, n)). \qquad \bullet$$

Since the definition of ϕ_r in Theorem 5.3.1 is constructive, we can go ahead and give a formula for the determinant:

$$\det \boldsymbol{A} = \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sign}(\sigma)\boldsymbol{A}(1, \sigma(1)) \cdots \boldsymbol{A}(n, \sigma(n)). \tag{5.15}$$

Let us give this formula for small n.

1. $n = 1$: $\det \boldsymbol{A} = \boldsymbol{A}(1, 1)$.
2. $n = 2$: $\det \boldsymbol{A} = \boldsymbol{A}(1, 1)\boldsymbol{A}(2, 2) - \boldsymbol{A}(1, 2)\boldsymbol{A}(2, 1)$.

3. $n = 3$: $A(1,1)A(2,2)A(3,3) + A(1,2)A(2,3)A(3,1) + A(1,3)A(2,1)A(3,2) - A(1,3)A(2,2)A(1,3) - A(1,1)A(2,3)A(3,2) - A(1,2)A(2,1)A(3,3)$.

It is common to see graphical tricks for computing the determinant of 2 × 2 and 3 × 3 matrices. These tricks do not carry over to larger matrices, so one should exercise care when using them.

Let us give some of the basic properties of the determinant that follow relatively easily from the definition.

5.3.3 Proposition (Elementary properties of determinant) *Let* R *be a commutative unit ring, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A}, \mathbf{B} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$*. Then the following statements hold:*

(i) $\det \mathbf{I}_{\mathrm{n}} = 1_{\mathsf{R}}$*;*

(ii) $\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B})$*;*

(iii) $\det \mathbf{A}^{\mathrm{T}} = \det \mathbf{A}$*;*

(iv) if $\mathbf{A}$ *is upper or lower triangular then*

$$\det \mathbf{A} = \mathbf{A}(1,1) \cdots \mathbf{A}(\mathrm{n},\mathrm{n}).$$

Proof We shall only prove those assertions that are not obvious.

(ii) Let us fix $\boldsymbol{B} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ and define $\phi_{\boldsymbol{B}} \colon \mathrm{Mat}_{n\times n}(\mathsf{R}) \to \mathsf{R}$ by $\phi_{\boldsymbol{B}}(\boldsymbol{A}) = \det(\boldsymbol{AB})$. We claim that this is an alternating multilinear function of the rows of $\boldsymbol{A}$. We have

$$\begin{aligned}\phi_{\boldsymbol{B}}(\boldsymbol{A}) &= \sum_{\sigma\in\mathfrak{S}_n} \mathrm{sign}(\sigma)\boldsymbol{AB}(1,\sigma(1)) \cdots \boldsymbol{AB}(n,\sigma(n)) \\ &= \sum_{\sigma\in\mathfrak{S}_n} \mathrm{sign}(\sigma)\left(\sum_{j_1=1}^{n} \boldsymbol{A}(1,j_1)\boldsymbol{B}(j_1,\sigma(1)\right) \cdots \left(\sum_{j_n=1}^{n} \boldsymbol{A}(n,j_n)\boldsymbol{B}(j_n,\sigma(n)\right).\end{aligned}$$

With this expression at hand, it is then a matter of direct computation to check that $\phi_{\boldsymbol{B}}$ is an alternating multilinear function of $\boldsymbol{A}$. By the definition and characterisation of ϕ_r in Theorem 5.3.1 and from the definition of det it follows that $\phi_{\boldsymbol{B}} = r \det$ for some r. Thus $\det(\boldsymbol{AB}) = r \det A$ for some $r \in \mathsf{R}$. But

$$\phi_{\boldsymbol{B}}(\boldsymbol{I}_n) = \det(\boldsymbol{I}_n\boldsymbol{B}) = \det \boldsymbol{B} = r \det \boldsymbol{I}_n = r.$$

Thus $r = \det \boldsymbol{B}$ and so $\phi_{\boldsymbol{B}} = \det A \det B$ as claimed.

(iii) Let $\sigma \in \mathfrak{S}_n$ and note that, by rearranging the terms in the product, we have

$$\boldsymbol{A}(1,\sigma(1)) \cdots \boldsymbol{A}(n,\sigma(n)) = \boldsymbol{A}(1,\sigma^{-1}(1)) \cdots \boldsymbol{A}(n,\sigma^{-1}(n)).$$

Thus

$$\begin{aligned}\det \boldsymbol{A}^T &= \sum_{\sigma\in\mathfrak{S}_n} \boldsymbol{A}(\sigma(1),1) \cdots \boldsymbol{A}(\sigma(n),n) \\ &= \sum_{\sigma\in\mathfrak{S}_n} \boldsymbol{A}(1,\sigma^{-1}(1)) \cdots \boldsymbol{A}(n,\sigma^{-1}(n)) = \det \boldsymbol{A}.\end{aligned}$$

(iv) This is part of Exercise 5.3.7. ■

The definition we give of the determinant, in terms of the linear map ϕ_{1_R} defined in the proof of Theorem 5.3.1, may seem a little roundabout, given that one can simply directly give the explicit formula (5.15). However, the definition we give is useful for investigating various properties of the determinant.

It is not uncommon in elementary treatments of the determinant to give a computationally motivated definition involving expansions about a row or column. Let us show how these computational formulae follow from our definition. First let us define some useful terminology.

5.3.4 Definition (Minor, cofactor) Let R be a commutative unit ring, let $n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$. For $i, j \in \{1, \dots, n\}$ let $\hat{\boldsymbol{A}}(i, j)$ be the $(n-1)\times(n-1)$-matrix obtained by deleting the ith row and jth column of $\boldsymbol{A}$.

(i) The **(i, j)*th minor*** is $\det \hat{\boldsymbol{A}}(i, j)$.

(ii) The **(i, j)*th cofactor*** is $(-1)^{i+j} \det \hat{\boldsymbol{A}}(i, j)$.

The ***cofactor matrix*** of $\boldsymbol{A}$ is the matrix $\mathrm{Cof}(\boldsymbol{A}) \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ whose (i, j)th entry of the (i, j)th cofactor. •

A simple example illustrates these concepts.

5.3.5 Example (Minor, cofactor) Consider a general 2×2 matrix

$$\boldsymbol{A} = \begin{bmatrix} A(1,1) & A(1,2) \\ A(2,1) & A(2,2) \end{bmatrix}.$$

We then have

$$\hat{\boldsymbol{A}}(1,1) = \begin{bmatrix} A(2,2) \end{bmatrix}, \quad \hat{\boldsymbol{A}}(1,2) = \begin{bmatrix} A(2,1) \end{bmatrix}, \\ \hat{\boldsymbol{A}}(2,1) = \begin{bmatrix} A(1,2) \end{bmatrix}, \quad \hat{\boldsymbol{A}}(2,2) = \begin{bmatrix} A(1,1) \end{bmatrix}.$$

The (i, j)th minors are then the determinants of these matrices, which are $A(2,2)$, $A(2,1)$, $A(1,2)$, and $A(1,1)$, respectively. The cofactor matrix is

$$\mathrm{Cof}(\boldsymbol{A}) = \begin{bmatrix} A(2,2) & -A(2,1) \\ -A(1,2) & A(2,2) \end{bmatrix}.$$

Larger matrices will merely be a larger test of your computing ability. •

The following result then gives what is sometimes given as the definition of the determinant.

5.3.6 Proposition (Computation of determinant using cofactors) *Let* R *be a commutative unit ring, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, let* $\mathrm{i_0}, \mathrm{j_0} \in \{1, \dots, \mathrm{n}\}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$*. Then*

$$\det \mathbf{A} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{A}(\mathrm{i}, \mathrm{j}_0)\mathrm{Cof}(\mathbf{A})(\mathrm{i}, \mathrm{j}_0) = \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathbf{A}(\mathrm{i}_0, \mathrm{j})\mathrm{Cof}(\mathbf{A})(\mathrm{i}_0, \mathrm{j}).$$

Proof We shall prove that

$$\det \boldsymbol{A} = \sum_{i=1}^{n} \boldsymbol{A}(i, j_0)\mathrm{Cof}(\boldsymbol{A})(i, j_0)$$

by showing that the map

$$\phi\colon \boldsymbol{A} \mapsto \sum_{i=1}^{n} (-1)^{i+j_0}\boldsymbol{A}(i, j_0) \det \hat{\boldsymbol{A}}(i, j_0)$$

is an alternating multilinear function on the rows of $\boldsymbol{A}$ with the property that $\phi(\boldsymbol{I}_n) = 1_{\mathsf{R}}$. This will give the first equality of the proposition by the definition of the determinant and by Theorem 5.3.1. The second equality can be proved using the fact, proved in Proposition 5.3.3, that $\det \boldsymbol{A}^T = \det \boldsymbol{A}$.

First we show that ϕ is a multilinear function of its rows. Let $i' \in \{1, \dots, n\}$. We take the ith row of $\boldsymbol{A}$ for $i \neq i'$ to be $\boldsymbol{r}_i \in \mathsf{R}^n$ and we take the i'th row to be $a\boldsymbol{r}_{i'} + \boldsymbol{s}_{i'}$ for $\boldsymbol{r}_{i'}, \boldsymbol{s}_{i'} \in \mathsf{R}^n$ and for $a \in \mathsf{R}$. We let $\boldsymbol{B}$ be the matrix with rows $\boldsymbol{r}_i$, $i \in \{1, \dots, n\}$, and let $\boldsymbol{C}$ be the same matrix, but with the i'th row being $\boldsymbol{s}_{i'}$. We will show that

$$\phi(\boldsymbol{A}) = a\phi(\boldsymbol{B}) + \phi(\boldsymbol{C}), \tag{5.16}$$

which will show that ϕ is multilinear. If $i = i'$ then

$$\det \hat{\boldsymbol{A}}(i, j_0) = \det \hat{\boldsymbol{B}}(i, j_0) = \det \hat{\boldsymbol{C}}(i, j_0)$$

and so

$$\boldsymbol{A}(i, j_0) \det \hat{\boldsymbol{A}}(i, j_0) = a\boldsymbol{B}(i, j_0) \det \hat{\boldsymbol{B}}(i, j_0) + \boldsymbol{C}(i, j_0) \det \hat{\boldsymbol{C}}(i, j_0).$$

For $i \neq i'$ we have

$$\det \hat{\boldsymbol{A}}(i, j_0) = a \det \hat{\boldsymbol{B}}(i, j_0) + \det \hat{\boldsymbol{A}}(i, j_0)$$

since the determinant is multilinear. Thus

$$\boldsymbol{A}(i, j_0) \det \hat{\boldsymbol{A}}(i, j_0) = a\boldsymbol{B}(i, j_0) \det \hat{\boldsymbol{B}}(i, j_0) + \boldsymbol{C}(i, j_0) \det \hat{\boldsymbol{C}}(i, j_0)$$

since $\boldsymbol{A}(i, j_0) = \boldsymbol{B}(i, j_0) = \boldsymbol{C}(i, j_0)$ in this case. Thus (5.16) does indeed hold.

Now let us show that if distinct rows i_1 and i_2 of $\boldsymbol{A}$ are equal then $\phi(\boldsymbol{A}) = 0_{\mathsf{R}}$. Without loss of generality suppose that $i_1 < i_2$. As long as $i \notin \{i_1, i_2\}$ then $\det \hat{\boldsymbol{A}}(i, j_0) = 0_{\mathsf{R}}$ since the matrix $\hat{\boldsymbol{A}}(i, j_0)$ will have two equal rows in this case. Now note that $\hat{\boldsymbol{A}}(i_1, j_0)$ is obtained from $\hat{\boldsymbol{A}}(i_2, j_0)$ by successively swapping row i_2 with rows $i_2 - 1, \dots, i_1 + 1$. Thus

$$\det \hat{\boldsymbol{A}}(i_1, j_0) = (-1)^{i_2 - i_1 - 1} \det \hat{\boldsymbol{A}}(i_2, j_0)$$

since the determinant is a skew-symmetric function of its rows. Therefore, we have

$$\begin{aligned}\phi(\boldsymbol{A}) &= (-1)^{i_1 + j_0} \det \hat{\boldsymbol{A}}(i_1, j_0) + (-1)^{i_2 + j_0} \det \hat{\boldsymbol{A}}(i_2, j_0)\\ &= (-1)^{i_1 + j_0 + i_2 - i_1 - 1} \det \hat{\boldsymbol{A}}(i_2, j_0) + (-1)^{i_2 + j_0} \det \hat{\boldsymbol{A}}(i_2, j_0) = 0_{\mathsf{R}}.\end{aligned}$$

Thus ϕ is an alternating function of its rows.

Finally we note that a direct computation shows that

$$\det \hat{\boldsymbol{I}}_n(i, j_0) = \begin{cases} 1_{\mathsf{R}}, & i = j_0, \\ 0_{\mathsf{R}}, & \text{otherwise.} \end{cases}$$

This gives $\phi(\boldsymbol{I}_n) = 1_{\mathsf{R}}$, which gives $\phi = \det$. ■

The expression

$$\sum_{i=1}^{n} \boldsymbol{A}(i, j_0)\mathrm{Cof}(\boldsymbol{A})(i, j_0)$$

is the ***expansion of* det A *along column* $\mathbf{j_0}$** and the expression

$$\sum_{j=1}^{n} \boldsymbol{A}(i_0, j)\mathrm{Cof}(\boldsymbol{A})(i_0, j)$$

is the ***expansion of* det A *along row* $\mathbf{i_0}$**.

Let us illustrate the application of Proposition 5.3.6 in a simple case.

5.3.7 Example (Computing the determinant using cofactors) Let $\boldsymbol{A}$ be a 3×3 matrix. Expanding $\det \boldsymbol{A}$ along the first row gives

$$\det \boldsymbol{A} = \boldsymbol{A}(1,1)\det\begin{bmatrix} \boldsymbol{A}(2,2) & \boldsymbol{A}(2,3) \\ \boldsymbol{A}(3,2) & \boldsymbol{A}(3,3) \end{bmatrix} \\ - \boldsymbol{A}(1,2)\det\begin{bmatrix} \boldsymbol{A}(2,1) & \boldsymbol{A}(2,3) \\ \boldsymbol{A}(3,1) & \boldsymbol{A}(3,3) \end{bmatrix} + \boldsymbol{A}(1,3)\det\begin{bmatrix} \boldsymbol{A}(2,1) & \boldsymbol{A}(2,2) \\ \boldsymbol{A}(3,1) & \boldsymbol{A}(3,2) \end{bmatrix}$$

and expanding $\det \boldsymbol{A}$ along the second column gives

$$\det \boldsymbol{A} = -\boldsymbol{A}(1,2)\det\begin{bmatrix} \boldsymbol{A}(2,1) & \boldsymbol{A}(2,3) \\ \boldsymbol{A}(3,1) & \boldsymbol{A}(3,3) \end{bmatrix} \\ + \boldsymbol{A}(2,2)\det\begin{bmatrix} \boldsymbol{A}(1,1) & \boldsymbol{A}(1,3) \\ \boldsymbol{A}(3,1) & \boldsymbol{A}(3,3) \end{bmatrix} - \boldsymbol{A}(3,2)\det\begin{bmatrix} \boldsymbol{A}(1,1) & \boldsymbol{A}(1,3) \\ \boldsymbol{A}(2,1) & \boldsymbol{A}(2,3) \end{bmatrix}.$$

Thus we see that the expansions produce determinants of matrices that have one fewer row and column than $\boldsymbol{A}$. If one is forced to compute a determinant by hand, this is the method of choice. •

5.3.2 Determinant and invertibility

To this point the determinant seems like an entirely pointless construction. We shall now begin to explore the value of the determinant. First let us explore the relationship between the determinant and invertibility of matrices. In order to do so we introduce the following idea.

5.3.8 Definition (Adjugate) Let R be a commutative unit ring, let $n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$. The ***adjugate*** of $\boldsymbol{A}$ is the matrix $\mathrm{Adj}(\boldsymbol{A}) \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ defined by

$$\mathrm{Adj}(\boldsymbol{A})(i,j) = (-1)^{i+j}\det\hat{\boldsymbol{A}}(j,i). \qquad \bullet$$

Let us illustrate the adjugate.

5.3.9 Example (Adjugate (Example 5.3.5 cont'd)) For the general 2×2 matrix the adjugate is

$$\mathrm{Adj}(\boldsymbol{A}) = \begin{bmatrix} A(2,2) & -A(1,2) \\ -A(2,1) & A(2,2) \end{bmatrix}. \qquad \bullet$$

Thus the adjugate is the transpose of the cofactor matrix. Sometimes what we call the adjugate is called the "adjoint." However, adjoint has another very different usage in functional analysis, and so we choose to use terminology that discriminates the two concepts.

5.3.10 Theorem (Determinant and invertibility) *Let* R *be a commutative unit ring, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$*. Then*

(i) $\mathrm{Adj}(\mathbf{A})\mathbf{A} = \mathbf{A}\mathrm{Adj}(\mathbf{A}) = (\det\mathbf{A})\mathbf{I}_{\mathrm{n}}$,

(ii) $\mathbf{A}$ *is invertible if and only if* $\det\mathbf{A}$ *is a unit, and*

(iii) if $\mathbf{A}$ *is invertible then*

$$\det\mathbf{A}^{-1} = (\det\mathbf{A})^{-1}, \qquad \mathbf{A}^{-1} = (\det\mathbf{A})^{-1}\mathrm{Adj}(\mathbf{A}).$$

Proof (i) Note that

$$\boldsymbol{A}\mathrm{Adj}(\boldsymbol{A})(i,i) = \sum_{l=1}^{n}(-1)^{i+l}\boldsymbol{A}(i,l)\det\hat{\boldsymbol{A}}(i,l) = \det\boldsymbol{A}$$

for each $i \in \{1,\dots,n\}$, using Proposition 5.3.6. For $i \neq j$ let $\boldsymbol{B}$ be the matrix which agree with $\boldsymbol{A}$ except that the jth row of $\boldsymbol{B}$ is the ith row of $\boldsymbol{A}$. Thus $\boldsymbol{B}$ has two equal rows and so $\det B = 0_{\mathsf{R}}$. Moreover,

$$\boldsymbol{B}(i,l) = \boldsymbol{A}(i,l) = \boldsymbol{B}(j,l), \quad \det\hat{\boldsymbol{A}}(j,l) = \det\hat{\boldsymbol{B}}(j,l)$$

for all $l \in \{1,\dots,n\}$. Therefore,

$$\begin{aligned}\boldsymbol{A}\mathrm{Adj}(\boldsymbol{A})(i,j) &= \sum_{l=1}^{n}(-1)^{j+l}\boldsymbol{A}(i,l)\det\hat{\boldsymbol{A}}(j,l)\\ &= \sum_{l=1}^{n}(-1)^{j+l}\boldsymbol{B}(j,l)\det\hat{\boldsymbol{B}}(j,l) = \det\boldsymbol{B} = 0_{\mathsf{R}},\end{aligned}$$

using Proposition 5.3.6. Therefore, $\boldsymbol{A}\mathrm{Adj}(\boldsymbol{A}) = \det(\boldsymbol{A})\boldsymbol{I}_n$. Since $\mathrm{Adj}(\boldsymbol{A}^T) = \mathrm{Adj}(\boldsymbol{A})^T$ (why?) it then holds that

$$(\det\boldsymbol{A})\boldsymbol{I}_n = (\det\boldsymbol{A}^T)\boldsymbol{I}_n = \boldsymbol{A}^T\mathrm{Adj}(\boldsymbol{A}^T) = \boldsymbol{A}^T\mathrm{Adj}(\boldsymbol{A})^T = \mathrm{Adj}(\boldsymbol{A}\boldsymbol{A})^T.$$

Thus

$$\mathrm{Adj}(\boldsymbol{A})\boldsymbol{A} = (\det \boldsymbol{A})\boldsymbol{I}_n^T = (\det \boldsymbol{A})\boldsymbol{I}_n,$$

and so $\boldsymbol{A}\mathrm{Adj}(\boldsymbol{A}) = \mathrm{Adj}(\boldsymbol{A})\boldsymbol{A} = (\det \boldsymbol{A})\boldsymbol{I}_n$, as desired.

(ii) Suppose that $\boldsymbol{A}$ is invertible. Then, using Proposition 5.3.3,

$$\det \boldsymbol{I}_n = (\det \boldsymbol{A})(\det \boldsymbol{A}^{-1}) = (\det \boldsymbol{A}^{-1})(\det \boldsymbol{A}) = 1_{\mathsf{R}}.$$

Thus $\det \boldsymbol{A}$ is a unit and its inverse is $\det \boldsymbol{A}^{-1}$.

Now suppose that $\det \boldsymbol{A}$ is a unit. Then, by part (i).

$$\boldsymbol{A}(\det \boldsymbol{A})^{-1}\mathrm{Adj}(\boldsymbol{A}) = (\det \boldsymbol{A})^{-1}\mathrm{Adj}(\boldsymbol{A})\boldsymbol{A} = \boldsymbol{I}_n,$$

giving $\boldsymbol{A}$ as invertible with inverse $(\det \boldsymbol{A})^{-1}\mathrm{Adj}(\boldsymbol{A})$.

(iii) This was proved as part of our proving part (ii). ■

The formula for $\boldsymbol{A}^{-1}$ given as part (iii) in the preceding theorem is interesting in that it gives an explicit formula for the inverse in terms of the components of $\boldsymbol{A}$. However, in practice this is usually an inefficient means of computing the inverse. Nonetheless, we shall use it in Sections V-?? and V-?? to say some things about inverses of matrices with polynomial entries.

Let us compute an inverse using the adjugate.

5.3.11 Example (Inverse using adjugate (Example 5.3.5 cont'd)) For a general invertible 2 × 2 matrix we have

$$\begin{aligned}\boldsymbol{A}^{-1} &= (\det \boldsymbol{A})^{-1}\mathrm{Adj}(\boldsymbol{A})\\ &= (\boldsymbol{A}(1,1)\boldsymbol{A}(2,2) - \boldsymbol{A}(1,2)\boldsymbol{A}(2,1))^{-1}\begin{bmatrix}\boldsymbol{A}(2,2) & -\boldsymbol{A}(1,2)\\ -\boldsymbol{A}(2,1) & \boldsymbol{A}(2,2)\end{bmatrix},\end{aligned}$$

which may be a familiar formula. •

5.3.3 Determinant, systems of equations, and linear independence

In this section we explore another use for determinants, namely in solving linear equations and for determining linear independence. We work only with finite matrices. In this case we recall from Definitions 5.1.54 and 5.2.52 that a system of linear equations over a commutative ring R is a pair $(\boldsymbol{A}, \boldsymbol{b}) \in \mathrm{Mat}_{m\times n}(\mathsf{R}) \times \mathsf{R}^m$. The set of solutions is then

$$\mathrm{Sol}(\boldsymbol{A}, \boldsymbol{b}) = \{\boldsymbol{x} \in \mathsf{R}^n \mid \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\}.$$

In the case when $m = n$ there is a rule for computing solutions using determinants if $\boldsymbol{A}$ is invertible.

5.3.12 Proposition (Cramer's Rule) *Let* R *be a commutative unit ring, let* $\mathrm{n} \in \mathbb{Z}_{>0}$*, and let* $(\mathbf{A}, \mathbf{b}) \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R}) \times \mathsf{R}^{\mathrm{n}}$ *be a system of linear equations over* R*. If* $\det \mathbf{A}$ *is a unit then* $\mathrm{Sol}(\mathbf{A}, \mathbf{b}) = \{\mathbf{x}\}$ *where*

$$\mathbf{x}(\mathrm{j}) = (\det \mathbf{A})^{-1} \sum_{\mathrm{i}=1}^{\mathrm{n}} (-1)^{\mathrm{i}+\mathrm{j}} \mathbf{b}(\mathrm{i}) \det \hat{\mathbf{A}}(\mathrm{i}, \mathrm{j}), \qquad \mathrm{j} \in \{1, \dots, \mathrm{n}\}.$$

Proof First of all, if $\det \boldsymbol{A}$ is a unit then $\boldsymbol{A}$ is invertible by Theorem 5.3.10 and so $\mathrm{Sol}(\boldsymbol{A}, \boldsymbol{b})$ is a singleton. Moreover, using the formula for $\boldsymbol{A}^{-1}$ from Theorem 5.3.10, th unique solution is

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} = (\det \boldsymbol{A})^{-1} \mathrm{Adj}(\boldsymbol{A})\boldsymbol{b},$$

which is exactly the expression given. ■

The expression

$$\sum_{i=1}^{n} (-1)^{i+j} \boldsymbol{b}(i) \det \hat{\boldsymbol{A}}(i, j)$$

bears a little thinking about. Indeed, a moments reflection shows that if one takes $\boldsymbol{B}_j$ to be the matrix equal to $\boldsymbol{A}$, but with the jth column replaced by $\boldsymbol{b}$, then we have

$$\det \boldsymbol{B}_j = \sum_{i=1}^{n} (-1)^{i+j} \boldsymbol{b}(i) \det \hat{\boldsymbol{A}}(i, j),$$

as the expression on the right is the expansion of $\det \boldsymbol{B}_j$ along the jth column. Let us illustrate this with an example.

5.3.13 Example (Cramer's Rule) We take a general system of linear equations $(\boldsymbol{A}, \boldsymbol{b})$ in two variables:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}(1,1) & \boldsymbol{A}(1,2) \\ \boldsymbol{A}(2,1) & \boldsymbol{A}(2,2) \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}(1) \\ \boldsymbol{b}(2) \end{bmatrix}.$$

We then have

$$\boldsymbol{B}_1 = \begin{bmatrix} \boldsymbol{b}(1) & \boldsymbol{A}(1,2) \\ \boldsymbol{b}(2) & \boldsymbol{A}(2,2) \end{bmatrix}, \quad \boldsymbol{B}_2 = \begin{bmatrix} \boldsymbol{A}(1,1) & \boldsymbol{b}(1) \\ \boldsymbol{A}(2,1) & \boldsymbol{b}(2) \end{bmatrix}.$$

Thus, using Cramer's Rule, the unique solution is the vector $\boldsymbol{x} \in \mathsf{R}^2$ given by

$$\begin{aligned} \boldsymbol{x}(1) &= (\det \boldsymbol{A})^{-1} \det \boldsymbol{B}_1 \\ &= (\boldsymbol{A}(1,1)\boldsymbol{A}(2,2) - \boldsymbol{A}(1,2)\boldsymbol{A}(2,1))^{-1} \boldsymbol{A}(2,2)\boldsymbol{b}(1) - \boldsymbol{A}(1,2)\boldsymbol{b}(2), \\ \boldsymbol{x}(2) &= (\det \boldsymbol{A})^{-1} \det \boldsymbol{B}_2 \\ &= (\boldsymbol{A}(1,1)\boldsymbol{A}(2,2) - \boldsymbol{A}(1,2)\boldsymbol{A}(2,1))^{-1} \boldsymbol{A}(1,1)\boldsymbol{b}(2) - \boldsymbol{A}(2,1)\boldsymbol{b}(1). \end{aligned}$$

This, of course, agrees with $\boldsymbol{A}^{-1}\boldsymbol{b}$. •

Now let us consider an important special case of systems of linear equations: those that are homogeneous. The following result characterises the solution of these in terms of the determinant of the coefficient matrix for the system.

5.3.14 Proposition (Solutions to homogeneous systems of linear equations) *Let* R *be a commutative unit ring, let* $\mathrm{n} \in \mathbb{Z}_{>0}$, *and let* $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{R})$. *If* $\mathbf{x} \in \mathrm{Sol}(\mathbf{A}, \mathbf{0}_{\mathsf{R}^{\mathrm{n}}})$ *then* $\det \mathbf{A}\mathbf{x}(\mathrm{j}) = 0_{\mathsf{R}}$ *for each* $\mathrm{j} \in \{1, \dots, \mathrm{n}\}$. *In particular, if* $\det \mathbf{A} \neq 0_{\mathsf{R}}$ *and* R *is an integral domain, then* $\mathrm{Sol}(\mathbf{A}, \mathbf{0}_{\mathsf{R}^{\mathrm{n}}}) = \{\mathbf{0}_{\mathsf{R}^{\mathrm{n}}}\}$.

Proof Let $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}, \mathbf{0}_{\mathsf{R}^n})$. For $j \in \{1, \dots, n\}$ denote by $\boldsymbol{B}_j$ the diagonal matrix whose diagonal entries are 1_{R} except the jth diagonal, which is $\boldsymbol{x}(j)$. Since $\det \boldsymbol{B}_j = \boldsymbol{x}(j)$ by Proposition 5.3.3(iv) we have

$$\det \boldsymbol{A}\boldsymbol{x}(j) = \det \boldsymbol{A} \det \boldsymbol{B}_j = \det(\boldsymbol{A}\boldsymbol{B}_j), \qquad j \in \{1, \dots, n\}.$$

Let us show that $\det(\boldsymbol{A}\boldsymbol{B}_j) = 0_{\mathsf{R}}$. First, a direct computation shows that the matrix $\boldsymbol{A}\boldsymbol{B}_j$ agrees with $\boldsymbol{A}$ except that the entries in the jth column are multiplied by $\boldsymbol{x}(j)$. Now fix $j \in \{1, \dots, n\}$ and let $\boldsymbol{C}_j$ be the matrix obtained by adding $\boldsymbol{x}(i)$ times the ith column of $\boldsymbol{A}\boldsymbol{B}_j$ to the jth column of $\boldsymbol{A}\boldsymbol{B}_j$ for all $i \in \{1, \dots, n\} \setminus \{j\}$. Since $\boldsymbol{C}_j$ is obtained by elementary column operations of the second type from $\boldsymbol{A}\boldsymbol{B}_j$, we have $\det \boldsymbol{C}_j = \det(\boldsymbol{A}\boldsymbol{B}_j)$ (cf. Exercise 5.3.1). Moreover, a direct calculation shows that the jth column of $\boldsymbol{C}_j$ is equal to the jth column of $\boldsymbol{A}\boldsymbol{x}$ which is zero since $\boldsymbol{x} \in \mathrm{Sol}(\boldsymbol{A}, \mathbf{0}_{\mathsf{R}^n})$. Thus $\det \boldsymbol{C}_j = 0_{\mathsf{R}}$ (by, say, expanding the determinant about the jth column) and so $\det \boldsymbol{A}\boldsymbol{x}(j) = \det(\boldsymbol{A}\boldsymbol{B}_j) = 0_{\mathsf{R}}$, as desired.

The final assertion of the proposition is clear. ■

This result has useful applications for determining when a collection of vectors is linearly independent. To do so we must give some constructions using determinants for matrices that are not necessarily square. The following definition is a generalisation of Definition 5.3.4, although the word usage is not in exact correspondence.

5.3.15 Definition (Minors for a nonsquare matrix) Let R be a commutative unit ring, let $m, n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{m\times n}(\mathsf{R})$. Let $k \in \{1, \dots, \min\{m, n\}\}$ and let $I = \{i_1, \dots, i_k\} \subseteq \{1, \dots, m\}$ and $J = \{j_1, \dots, j_k\} \subseteq \{1, \dots, n\}$ satisfy

$$i_1 < \cdots < i_k, \quad j_1 < \cdots < j_k.$$

Denote $\boldsymbol{A}(I, J) \in \mathrm{Mat}_{k\times k}(\mathsf{R})$ the matrix defined by

$$\boldsymbol{A}(I, J)(a, b) = \boldsymbol{A}(i_a, j_b), \qquad a, b \in \{1, \dots, k\}.$$

The ***(I, J)th minor*** of $\boldsymbol{A}$ is $\det \boldsymbol{A}(I, J)$. A $\mathbf{k} \times \mathbf{k}$***-minor*** of $\boldsymbol{A}$ is an (I, J)th minor for some I, J having cardinality k. •

Some authors call the matrices $\boldsymbol{A}(I, J)$ the minors of $\boldsymbol{A}$.

Let us now state how one may determine linear independence of a collection of vectors by using minors.

5.3.16 Proposition (Determining linear independence using determinants) *Let* R *be an integral domain, let* $k, n \in \mathbb{Z}_{>0}$ *satisfy* $k \leq n$*, and let* $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathsf{R}^n$*. Denote by* $\mathbf{X} \in \mathrm{Mat}_{n\times k}(\mathsf{R})$ *the matrix whose* j*th column is* $\mathbf{x}_j$*,* $j \in \{1, \dots, k\}$*. Then the set* $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ *is linearly independent if and only if some* $k \times k$*-minor of* $\mathbf{X}$ *is nonzero.*

Proof First of all, note that $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$ is linearly dependent if and only if there exists a nonzero vector $(c_1, \dots, c_k) \in \mathsf{R}^k$ such that $\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}_{\mathsf{R}^n}$. That is, $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$ is linearly dependent if and only if $\mathrm{Sol}(\boldsymbol{X}, \boldsymbol{0}_{\mathsf{R}^n})$ contains vectors other than the zero vector.

Suppose that $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$ is linear dependent so that $\mathrm{Sol}(\boldsymbol{X}, \boldsymbol{0}_{\mathsf{R}^n})$ contains a nonzero vector $\boldsymbol{c}$. Let $I = \{i_1, \dots, i_k\}$ satisfy $i_1 < \cdots < i_k$ and let $J = \{1, \dots, j\}$. Since $\boldsymbol{X}\boldsymbol{c} = \boldsymbol{0}_{\mathsf{R}^n}$ it holds that $\boldsymbol{X}(I, J)\boldsymbol{c} = \boldsymbol{0}_{\mathsf{R}^n}$. Therefore, by Proposition 5.3.14, $\det \boldsymbol{X}(I, J)\boldsymbol{c}(j) = 0_{\mathsf{R}}$ for all $j \in \{1, \dots, k\}$. Since $\boldsymbol{c}(j) \neq 0_{\mathsf{R}}$ for some $j \in \{1, \dots, k\}$ and since R is an integral domain, it follows that $\det \boldsymbol{X}(I, J) = 0_{\mathsf{R}}$. Therefore, all $k \times k$-minors are zero.

Now suppose that some $k \times k$-minor is nonzero. Then there exists $I = \{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$ such that $i_1 < \cdots < i_k$ for which $\det \boldsymbol{X}(I, J) \neq 0_{\mathsf{R}}$ for $J = \{1, \dots, k\}$. Now let $\boldsymbol{c} \in \mathrm{Sol}(\boldsymbol{X}, \boldsymbol{0}_{\mathsf{R}^n})$, and note that this implies that $\boldsymbol{X}(I, J)\boldsymbol{c} = \boldsymbol{0}_{\mathsf{R}^n}$. By Proposition 5.3.14 we then have $\det \boldsymbol{X}(I, J)\boldsymbol{c}(j) = 0_{\mathsf{R}}$ for all $j \in \{1, \dots, k\}$. Since R is an integral domain, $\boldsymbol{c} = \boldsymbol{0}_{\mathsf{R}^k}$ and so $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$ is linearly independent. ■

When $k > n$ the situation is more complicated, at least for general rings. However, for principal ideal domains (and so for fields in particular), it holds that if $k > n$ then any collection of vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathsf{R}^n$ is linearly dependent; this follows from Theorem 4.9.1.

5.3.4 Trace and its properties

We shall not attempt to motivate the uses for trace as we did for determinant. Indeed, the trace is simply not as useful for us as is the determinant. Nevertheless, it does come up in a few places, so we give the definition and basic properties here.

5.3.17 Definition (Trace) Let R be a commutative unit ring, let $n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$. The ***trace*** of $\boldsymbol{A}$ is

$$\mathrm{tr}\,\boldsymbol{A} = \boldsymbol{A}(1,1) + \cdots + \boldsymbol{A}(n,n). \qquad \bullet$$

The basic properties of trace are as follows.

5.3.18 Proposition (Properties of trace) *Let* R *be a commutative unit ring, let* $n \in \mathbb{Z}_{>0}$*, let* $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$*, and let* $r \in \mathsf{R}$*. Then the following statements hold:*

(i) $\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}\,\mathbf{A} + \mathrm{tr}\,\mathbf{B}$;

(ii) $\mathrm{tr}(r\mathbf{A}) = r\,\mathrm{tr}\,\mathbf{A}$;

(iii) $\mathrm{tr}\,\mathbf{A}^T = \mathrm{tr}\,\mathbf{A}$;

(iv) $\mathrm{tr}(\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{A})$;

(v) $\mathrm{tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \mathrm{tr}(\mathbf{C}\mathbf{A}\mathbf{B}) = \mathrm{tr}(\mathbf{B}\mathbf{C}\mathbf{A})$.

Proof We shall only prove the not completely trivial parts of the result.

(iv) We have

$$\begin{aligned}\operatorname{tr}(\boldsymbol{AB}) &= \sum_{i=1}^{n}(\boldsymbol{AB})(i,i) = \sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{A}(i,j)\boldsymbol{B}(j,i)\\ &= \sum_{j=1}^{n}\sum_{i=1}^{n}\boldsymbol{B}(j,i)\boldsymbol{A}(i,j) = \sum_{j=1}^{n}(\boldsymbol{BA})(j,j) = \operatorname{tr}(\boldsymbol{BA}).\end{aligned}$$

(v) Using part (iv) we have

$$\operatorname{tr}(\boldsymbol{A}(\boldsymbol{BC})) = \operatorname{tr}((\boldsymbol{BC})\boldsymbol{A}) = \operatorname{tr}((\boldsymbol{AB})\boldsymbol{C}) = \operatorname{tr}(\boldsymbol{C}(\boldsymbol{AB})). \qquad \blacksquare$$

Exercises

5.3.1 Recall the three flavours of elementary row (resp. column) operations that lead to elementary matrices: (1) swapping rows (resp. columns), (2) adding a multiple of a row (resp. column) to another row (resp. column), (3) multiplying a row (resp. column) by a unit u.

(a) What is the determinant of an elementary matrix of the first sort?

(b) What is the determinant of an elementary matrix of the second sort?

(c) What is the determinant of an elementary matrix of the third sort?

(d) For matrices over principal ideal domains, what is the determinant of a secondary matrix?

5.3.2 Let R be a commutative unit ring and let $\boldsymbol{A}, \boldsymbol{B} \in \operatorname{Mat}_{n\times n}(\mathsf{R})$ be invertible. Show that $\boldsymbol{A} + r\boldsymbol{B}$ is invertible for all but a finite number of $r \in \mathsf{R}$.

5.3.3 Let R be a commutative unit ring and let $\boldsymbol{A} \in \operatorname{Mat}_{n\times n}(\mathsf{R})$. Show that $\det \boldsymbol{A} = 0_{\mathsf{R}}$ if $\boldsymbol{A}$ has a zero row or column.

In the next exercise you will need to recall Definition 5.1.61. You will also need the following definition.

Definition (Lie subalgebra) Let F be a field and let $(\mathfrak{g}, [\cdot,\cdot])$ be a Lie algebra over F. A ***Lie subalgebra*** of $\mathfrak{g}$ is a subspace $\mathfrak{h}$ of $\mathfrak{g}$ such that $[u, v] \in \mathfrak{h}$ for every $u, v \in \mathfrak{h}$. •

Of course, a Lie subalgebra of a Lie algebra is itself a Lie algebra.

5.3.4 Let F be a field, let $n \in \mathbb{Z}_{>0}$, and consider the Lie algebra $(\operatorname{Mat}_{n\times n}(\mathsf{F}), [\cdot,\cdot])$ of Exercise 5.1.4.

(a) Show that the set of matrices with trace 0_{F} is a Lie subalgebra of $\operatorname{Mat}_{n\times n}(\mathsf{F})$.

(b) Suppose that the map $m_n \colon \boldsymbol{x} \mapsto n\boldsymbol{x}$ (by $n\boldsymbol{x}$ we mean the n-fold sum of $\boldsymbol{x}$ with itself) is an isomorphism of F^n. Define $\operatorname{Tr}_0 \colon \operatorname{Mat}_{n\times n}(\mathsf{F}) \to \operatorname{Mat}_{n\times n}(\mathsf{F})$ by

$$\operatorname{Tr}_0(\boldsymbol{A}) = \boldsymbol{A} - \operatorname{tr}(\boldsymbol{A}) m_n^{-1} \circ \boldsymbol{I}_n.$$

Show that $\operatorname{tr}(\operatorname{Tr}_0(\boldsymbol{A})) = 0_{\mathsf{F}}$ for all $\boldsymbol{A} \in \operatorname{Mat}_{n\times n}(\mathsf{F})$.

5.3.5 Let F be a field, let $n \in \mathbb{Z}_{>0}$, and recall from Exercise 5.1.5 that $\mathsf{GL}(n;\mathsf{F})$ denotes the group of invertible $n \times n$ matrices over F.

(a) Show that the subset of matrices with determinant 1_{F} is a subgroup of $\mathsf{GL}(n;\mathsf{F})$.

This subgroup of invertible matrices with determinant 1_{F} is denoted by $\mathsf{SL}(n;\mathsf{F})$ and is called the ***special linear group*** of order n over F.

(b) Is $\mathsf{SL}(n;\mathsf{F})$ a subalgebra of $\mathrm{Mat}_{n\times n}(\mathsf{F})$?

5.3.6 Let R be a commutative unit ring, let $n \in \mathbb{Z}_{>0}$, and recall from Exercise 5.2.3 that $\mathsf{GL}(n;\mathsf{R})$ denotes the group of invertible $n \times n$ matrices over R.

(a) Show that the subset of matrices with determinant 1_{R} is a subgroup of $\mathsf{GL}(n;\mathsf{R})$.

This subgroup of invertible matrices with determinant 1_{R} is denoted by $\mathsf{SL}(n;\mathsf{R})$ and is called the ***special linear group*** of order n over R.

(b) Is $\mathsf{SL}(n;\mathsf{R})$ a subalgebra of $\mathrm{Mat}_{n\times n}(\mathsf{R})$?

(c) Describe the difference between $\mathsf{GL}(n;\mathbb{Z})$ and $\mathsf{SL}(n;\mathbb{Z})$. Contrast this with the difference between $\mathsf{SL}(n;\mathbb{R})$ and $\mathsf{GL}(n;\mathbb{R})$.

5.3.7 Let R be a commutative unit ring, let $n \in \mathbb{Z}_{>0}$, and let $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ be upper triangular.

(a) Show that $\det \boldsymbol{A}$ is equal to the product of the diagonal elements of R.

(b) If $\boldsymbol{A}$ is invertible, show that $\boldsymbol{A}^{-1}$ is upper triangular.

5.3.8 Let R be a commutative unit ring, let $\boldsymbol{A}_j \in \mathrm{Mat}_{n_j\times n_j}(\mathsf{R})$ for $j \in \{1,\dots,k\}$, and consider the partitioned matrix

$$\boldsymbol{A} = \left[\begin{array}{c|c|c|c} \boldsymbol{A}_j & * & \cdots & * \\ \hline \boldsymbol{0}_{n_2\times n_1} & \boldsymbol{A}_2 & \cdots & * \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{0}_{n_k\times n_1} & \boldsymbol{0}_{n_k\times n_2} & \cdots & \boldsymbol{A}_k \end{array}\right],$$

where an entry $*$ means an arbitrary matrix of the appropriate size. Show that

$$\det \boldsymbol{A} = \det \boldsymbol{A}_1 \cdots \det \boldsymbol{A}_k.$$

Section 5.4

Linear algebra over fields

In this section we carefully study linear maps between vector spaces. It is worth making a few comments on the relationship between this section and Section 5.1. First note that by Theorem 5.1.13 there is an exact correspondence between matrices in $\mathrm{Mat}_{I\times J}(\mathsf{F})$ and linear maps between F_0^J and F_0^I. It is also true that, given vector spaces U and V, there exist sets (for example, bases for U and V) J and I such that U is isomorphic to F_0^J and V is isomorphic to F_0^I (this is the content of Theorem 4.5.38). Therefore, it is reasonable to expect that there be a correspondence between linear maps from U to V and matrices in $\mathrm{Mat}_{I\times J}(\mathsf{F})$. This is in fact the case (see Proposition 5.4.25). For this reason, certain topics in this section will have exact analogues in Section 5.1. However, for the most part the development will proceed rather independently of Section 5.1. Indeed, it is usually the case that topics in common between this section and Section 5.1 are most naturally developed in the more abstract setup used in this section. In particular, the abstract development is more revealing of the "geometry" of linear algebra, as opposed to the more computational flavour of a treatment based in matrix manipulation.

Now a word of warning. In Section 5.1 quite a lot of attention was paid to the rôle of the transpose of a matrix. Since much of what is discussed in Section 5.1 extends in a simple way to this section, it is worth pointing out that this extension is not so simple for the concept of transpose. This can even be seen in Section 5.1 where a column finite matrix $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ gives rise to a linear map from F_0^J to F_0^I, but its transpose does not necessarily give rise to a linear map from F_0^I to F_0^J, but *does* give rise to a linear map from F^I to F^J. For readers who are only familiar with matrices with finite numbers of rows and columns, there is a natural tendency to think that if U and V are finite-dimensional vector spaces and if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$, then there is a natural map, called the transpose of L, from V to U. *This is not the case.* As we shall see in Theorem 5.7.22, the transpose is naturally a linear map between the between the algebraic duals V' and U'. The transpose will be completely absent from the present section. The bottom line is this: Unless the statement, "The transpose of a linear map is a linear map between dual spaces," is known to you, erase any preconceptions you have about the transpose.

Do I need to read this section? It is important to understand abstract linear algebra in order to understand many of the topics in these volumes. Therefore, any time spent understanding the material in this section will be time well spent. •

5.4.1 Subspaces and vector spaces associated to a linear map

We let F be a field, and in this section U and V will denote F-vector spaces, and L will denote an element of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$.

Associated with every linear map $\mathsf{L}\colon \mathsf{U} \to \mathsf{V}$ are two subspaces, one each of U and V, and two associated quotient spaces which are useful for understanding the character of L. The subspaces we have seen before, but we redefine them here in order to preserve the coherence of the presentation.

5.4.1 Definition (Image, kernel, cokernel, coimage) Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$.

(i) The ***image*** of L is the subspace of V given by $\mathrm{image}(\mathsf{L}) = \{\mathsf{L}(u) \mid u \in \mathsf{U}\}$.

(ii) The ***kernel*** of L is the subspace of U given by $\ker(\mathsf{L}) = \{u \in \mathsf{U} \mid \mathsf{L}(u) = 0_{\mathsf{V}}\}$.

(iii) The ***cokernel*** of L is the F-vector space $\mathrm{coker}(\mathsf{L}) = \mathsf{V}/\,\mathrm{image}(\mathsf{L})$.

(iv) The ***coimage*** of L is the F-vector space $\mathrm{coimage}(\mathsf{L}) = \mathsf{U}/\ker(\mathsf{L})$.

(v) The ***rank*** of L is $\mathrm{rank}(\mathsf{L}) = \dim_{\mathsf{F}}(\mathrm{image}(\mathsf{L}))$.

(vi) The ***nullity*** of L is $\mathrm{nullity}(\mathsf{L}) = \dim_{\mathsf{F}}(\ker(\mathsf{L}))$.

(vii) The ***defect*** of L is $\mathrm{defect}(\mathsf{L}) = \dim_{\mathsf{F}}(\mathrm{coker}(\mathsf{L}))$. •

Associated to the linear map L are then three other linear maps between various of the above subspaces. The idea is that, if L is not injective (resp. surjective, bijective), there is still a "part" of L that is injective (resp. surjective, bijective).

5.4.2 Theorem (Linear maps derived from a linear map) *Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$. Then the following statements hold:*

(i) there exists a unique linear map $\mathsf{L}_{\mathrm{inj}} \in \mathrm{Hom}_{\mathsf{F}}(\mathrm{coimage}(\mathsf{L});\mathsf{V})$ such that the diagram

$$\begin{tikzcd}
\mathsf{U} \arrow[rr, "\mathsf{L}"] \arrow[dr, "\pi_{\mathrm{coimage}(\mathsf{L})}"'] & & \mathsf{V} \\
& \mathrm{coimage}(\mathsf{L}) \arrow[ur, "\mathsf{L}_{\mathrm{inj}}"'] &
\end{tikzcd}$$

commutes, where $\pi_{\mathrm{coimage}(\mathsf{L})}\colon \mathsf{U} \to \mathrm{coimage}(\mathsf{L})$ is the canonical projection;

(ii) there exists a unique linear map $\mathsf{L}_{\mathrm{srj}} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathrm{image}(\mathsf{L}))$ such that the diagram

$$\begin{tikzcd}
\mathsf{U} \arrow[rr, "\mathsf{L}"] \arrow[dr, "\mathsf{L}_{\mathrm{srj}}"'] & & \mathsf{V} \\
& \mathrm{image}(\mathsf{L}) \arrow[ur, "\mathrm{i}_{\mathrm{image}(\mathsf{L})}"'] &
\end{tikzcd}$$

commutes;

(iii) *there exists a unique linear map* $\mathsf{L}_{\mathrm{bij}} \in \mathrm{Hom}_{\mathsf{F}}(\mathrm{coimage}(\mathsf{L}); \mathrm{image}(\mathsf{L}))$ *such that the diagram*

$$\begin{tikzcd}
\mathsf{U} \arrow[r, "\mathsf{L}"] \arrow[d, "\pi_{\mathrm{coimage}(\mathsf{L})}"'] & \mathsf{V} \\
\mathrm{coimage}(\mathsf{L}) \arrow[r, "\mathsf{L}_{\mathrm{bij}}"'] & \mathrm{image}(\mathsf{L}) \arrow[u, "\mathrm{i}_{\mathrm{image}(\mathsf{L})}"']
\end{tikzcd}$$

commutes.

Moreover, $\mathsf{L}_{\mathrm{inj}}$ *is injective,* $\mathsf{L}_{\mathrm{srj}}$ *is surjective, and* $\mathsf{L}_{\mathrm{bij}}$ *is bijective.*

Proof (i) Define $\mathsf{L}_{\mathrm{inj}}$ by $\mathsf{L}_{\mathrm{inj}}(u + \ker(\mathsf{L})) = \mathsf{L}(u)$. We claim first that $\mathsf{L}_{\mathrm{inj}}$ is well-defined. Indeed, suppose that $u_1 + \ker(\mathsf{L}) = u_2 + \ker(\mathsf{L})$ for $u_1, u_2 \in \mathsf{U}$. Then there exists $u \in \ker(\mathsf{L})$ such that $u_2 = u_1 + u$. Then

$$\mathsf{L}(u_2) = \mathsf{L}(u_1 + u) = \mathsf{L}(u_1) + \mathsf{L}(u) = \mathsf{L}(u_1),$$

thus showing that $\mathsf{L}_{\mathrm{inj}}$ is well-defined. By definition, the diagram in this part of the theorem commutes. Moreover, it is clear that any linear map from coimage(L) which makes the diagram commute must be the same as $\mathsf{L}_{\mathrm{inj}}$. To see that $\mathsf{L}_{\mathrm{inj}}$ is injective, suppose that $\mathsf{L}_{\mathrm{inj}}(u_1 + \ker(\mathsf{L})) = \mathsf{L}_{\mathrm{inj}}(u_2 + \ker(\mathsf{L}))$. Then $\mathsf{L}(u_1 - u_2) = 0_{\mathsf{V}}$, whence $u_1 - u_2 \in \ker(\mathsf{L})$. Thus $u_1 + \ker(\mathsf{L}) = u_2 + \ker(\mathsf{L})$ and so $\mathsf{L}_{\mathrm{inj}}$ is injective.

(ii) We simply define $\mathsf{L}_{\mathrm{srj}}$ by $\mathsf{L}_{\mathrm{srj}}(u) = \mathsf{L}(u)$. The commutativity of the diagram given, and the uniqueness of the map that forces the commutativity of this diagram, are immediate. By definition of image(L) it follows that $\mathsf{L}_{\mathrm{srj}}$ is surjective.

(iii) We define $\mathsf{L}_{\mathrm{bij}}$ by $\mathsf{L}_{\mathrm{bij}}(u + \ker(\mathsf{L})) = \mathsf{L}_{\mathrm{srj}}(u)$. The well-definedness of $\mathsf{L}_{\mathrm{bij}}$ follows as above where we showed the well-definedness of $\mathsf{L}_{\mathrm{inj}}$. The commutativity of the given diagram, along with the uniqueness of the map that makes the map commutative, follow directly. That $\mathsf{L}_{\mathrm{bij}}$ is injective follows in the same manner as the injectivity of $\mathsf{L}_{\mathrm{inj}}$, and the surjectivity of $\mathsf{L}_{\mathrm{bij}}$ follows from the definition of image(L). ■

The result has the following corollary that shows that coimage(L) and coker(L) can be thought of as measuring the extent to which L is not injective or surjective, respectively.

5.4.3 Corollary (Interpretation of coimage(L) and coker(L)) *Let* F *be a field, let* U *and* V *be* F*-vector spaces, and let* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$*. Then*

(i) L *is injective if and only if* $\dim_{\mathsf{F}}(\mathrm{coimage}(\mathsf{L})) = 0$ *and*

(ii) L *is surjective if and only if* $\dim_{\mathsf{F}}(\mathrm{coker}(\mathsf{L})) = 0$.

Proof This follows from the fact that if U′ is a subspace of an F-vector space V′, then $\dim_{\mathsf{F}}(\mathsf{V}'/\mathsf{U}') = 0$ if and only if $\mathsf{U}' = \mathsf{V}'$ (cf. Theorem 4.5.56). ■

Theorem 5.4.2 also has the following well-known corollary which relates the dimensions of various of the subspaces associated to L.

5.4.4 Corollary (Rank–Nullity Formula) *If F is a field, if U and V are F-vector spaces, and if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$, then*

$$\dim_{\mathsf{F}}(\mathsf{U}) = \mathrm{rank}(\mathsf{L}) + \mathrm{nullity}(\mathsf{L}).$$

Proof This follows from Proposition 4.5.50 and Theorem 4.5.56. ■

5.4.2 Invariant subspaces

In the special case where $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V})$, i.e., L is a linear map from V to itself, an important sort of subspace associated with L can arise. These will be rather important in our understanding of the structure of endomorphisms, so we devote some time to these subspaces.

5.4.5 Definition (Invariant subspace) Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V})$. A subspace U of V is ***L-invariant***, or an ***invariant subspace*** for L, if $\mathsf{L}(u) \in \mathsf{U}$ for every $u \in \mathsf{U}$. •

Let us first consider the situation arising when one has multiple invariant subspaces.

5.4.6 Proposition (Sums and intersections of invariant subspaces are invariant subspaces) *Let F be a field, let V be an F-vector space, and let $(\mathsf{U}_j)_{j\in J}$ be a family of L-invariant subspaces. Then*

(i) $\sum_{j\in J} \mathsf{U}_j$ *is an L-invariant subspace and*

(ii) $\cap_{j\in J}\mathsf{U}_j$ *is an L-invariant subspace.*

Proof If $v \in \sum_{j\in J} \mathsf{U}_j$ then $v = u_1 + \cdots + u_k$ where $u_l \in \mathsf{U}_{j_l}$ for some $k \in \mathbb{Z}_{>0}$ and for $j_1,\dots,j_k \in J$. Then $\mathsf{L}(u_l) \in \mathsf{U}_{j_l}$ for each $l \in \{1,\dots,k\}$, and so

$$\mathsf{L}(v) = \mathsf{L}(u_1) + \cdots + \mathsf{L}(u_k) \in \sum_{j\in J} \mathsf{U}_j,$$

showing that $\sum_{j\in J} \mathsf{U}_j$ is L-invariant.

Next suppose that $v \in \cap_{j\in J}\mathsf{U}_j$. Then, for each $j \in \mathbb{Z}_{>0}$, $\mathsf{L}(v) \in \mathsf{U}_j$ since $v \in \mathsf{U}_j$. Thus $\mathsf{L}(v) \in \cap_{j\in J}\mathsf{U}_j$, and so $\cap_{j\in J}\mathsf{U}_j$ is L-invariant. ■

The importance of invariant subspaces is characterised by the following simple result.

5.4.7 Proposition (Properties of invariant subspaces) *Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V})$. For an L-invariant subspace U, the following statements hold:*

(i) $\mathsf{L}|\mathsf{U} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{U})$;

(ii) the map $\mathrm{v} + \mathsf{U} \mapsto \mathsf{L}(\mathrm{v}) + \mathsf{U}$ *is a well-defined endomorphism of* V/U.

Proof The first statement is obvious from the definition of L-invariance. For the second statement, suppose that $v_1 + \mathsf{U} = v_2 + \mathsf{U}$. Then there exists $u \in \mathsf{U}$ such that $v_2 = v_1 + u$. Then

$$\mathsf{L}(v_2) + \mathsf{U} = \mathsf{L}(v_1 + u) + \mathsf{U} = (\mathsf{L}(v_1) + \mathsf{L}(u)) + \mathsf{U} = \mathsf{L}(v_1) + \mathsf{U},$$

using the first part of the proposition. Thus the map is well-defined. That it is also linear is clear from the definition of the vector space structure on V/U. ■

We shall see in Section 5.8 that invariant subspaces form an important and surprisingly complicated rôle in the characterisation of endomorphisms of finite-dimensional vector spaces. In this development the notion of invariant subspaces possessing invariant an complement. It is convenient to have notation to represent this.

5.4.8 Definition (Direct sum of endomorphisms) Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Suppose that $\mathsf{U}_1, \dots, \mathsf{U}_k \subseteq \mathsf{V}$ are subspaces such that

(i) $\mathsf{V} = \mathsf{U}_1 \oplus \cdots \oplus \mathsf{U}_k$ and

(ii) U_j is L-invariant for each $j \in \{1, \dots, k\}$.

Denote $\mathsf{L}_j = \mathsf{L}|\mathsf{U}_j \in \mathrm{End}_{\mathsf{F}}(\mathsf{U}_j)$. Then L is the ***direct sum*** of $\mathsf{L}_1, \dots, \mathsf{L}_k$, and we write $\mathsf{L} = \mathsf{L}_1 \oplus \cdots \oplus \mathsf{L}_k$. •

The idea is very simple, and the essence is expressed by the following result.

5.4.9 Proposition (Characterisation of direct sum of endomorphisms) *Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Suppose that $\mathsf{U}_1, \dots, \mathsf{U}_{\mathrm{k}} \subseteq \mathsf{V}$ are subspaces such that*

(i) $\mathsf{V} = \mathsf{U}_1 \oplus \cdots \oplus \mathsf{U}_{\mathrm{k}}$ *and*

(ii) U_{j} *is* L*-invariant for each* $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$.

Denote $\mathsf{L}_{\mathrm{j}} = \mathsf{L}|\mathsf{U}_{\mathrm{j}}$ *and for* $\mathrm{v} \in \mathsf{V}$ *let*

$$\mathrm{v} = \mathrm{v}_1 + \cdots + \mathrm{v}_{\mathrm{k}}, \qquad \mathrm{v}_{\mathrm{j}} \in \mathsf{V}_{\mathrm{j}},\ \mathrm{j} \in \{1, \dots, \mathrm{k}\},$$

be the decomposition corresponding to the direct sum. Then

$$\mathsf{L}(\mathrm{v}) = \mathsf{L}_1(\mathrm{v}_1) + \cdots + \mathsf{L}_{\mathrm{k}}(\mathrm{v}_{\mathrm{k}}),$$

and the expression on the right is the decomposition of the expression on the left corresponding to the direct sum.

Proof The reader is asked to give the proof as Exercise 5.4.3. ■

Another useful characterisation of direct sums of endomorphisms involves their matrix representatives. We refer to (5.2) for the notation used in the following result for a block diagonal matrix.

5.4.10 Proposition (The matrix representative of a direct sum) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Suppose that* $\mathsf{U}_1, \dots, \mathsf{U}_k \subseteq \mathsf{V}$ *are subspaces such that*

(i) $\mathsf{V} = \mathsf{U}_1 \oplus \cdots \oplus \mathsf{U}_k$ *and*

(ii) U_j *is* L*-invariant for each* $j \in \{1, \dots, k\}$*.*

Denote $\mathsf{L}_j = \mathsf{L}|\mathsf{U}_j$ *so that* $\mathsf{L} = \mathsf{L}_1 \oplus \cdots \oplus \mathsf{L}_k$*. Let* $\mathscr{B}$ *be a basis for* V *and suppose that* $\mathscr{B} = \mathscr{B}_1 \cup \cdots \cup \mathscr{B}_k$ *with* $\mathscr{B}_j$ *a basis for* U_j*,* $j \in \{1, \dots, k\}$*. Then*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \mathrm{diag}([\mathsf{L}_1]_{\mathscr{B}_1}^{\mathscr{B}_1}, \dots, [\mathsf{L}_k]_{\mathscr{B}_k}^{\mathscr{B}_k})$$

Proof This is Exercise 5.4.4. ■

Sometimes we will be interested in cases when $\mathsf{U} \subseteq \mathsf{V}$ is a subspace that is *not* invariant under $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$, but still can be used to define an invariant subspace as in the following definition.

5.4.11 Definition (Smallest invariant subspace containing a subspace) Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The smallest L-invariant subspace of V containing U is denoted by $\langle \mathsf{L}, \mathsf{U} \rangle$. •

Note that the definition makes sense since

1. V is an L-invariant subspace containing U (i.e., such subspaces exist) and
2. if W_1 and W_2 are two L-invariant subspaces containing U, then $\mathsf{W}_1 \cap \mathsf{W}_2$ is itself an L-invariant subspace containing U (as the reader may verify in Exercise 5.4.1).

Therefore, we have

$$\langle \mathsf{L}, \mathsf{U} \rangle = \cap \{\mathsf{W} \mid \ \mathsf{W} \text{ is an } \mathsf{L}\text{-invariant subspace containing } \mathsf{U}\}.$$

The next result gives an explicit characterisation of $\langle \mathsf{L}, \mathsf{U} \rangle$.

5.4.12 Theorem (Characterisation of ⟨L, U⟩) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then*

$$\langle \mathsf{L}, \mathsf{U} \rangle = \mathrm{span}_{\mathsf{F}}(\mathsf{L}^j(u) \mid \ u \in \mathsf{U},\ j \in \mathbb{Z}_{\geq 0}).$$

Proof Let $\mathsf{W} = \mathrm{span}_{\mathsf{F}}(\mathsf{L}^j(u) \mid \ u \in \mathsf{U},\ j \in \mathbb{Z}_{\geq 0})$. Clearly W contains U. If $w \in \mathsf{W}$ then we can write

$$w = c_1 \mathsf{L}^{j_1}(u_1) + \cdots + c_k \mathsf{L}^{j_k}(u_k)$$

for $c_1, \dots, c_k \in \mathsf{F}$, $j_1, \dots, j_k \in \mathbb{Z}_{\geq 0}$, and $u_1, \dots, u_k \in \mathsf{U}$. Then

$$\mathsf{L}(w) = c_1 \mathsf{L}^{j_1+1}(u_1) + \cdots + c_k \mathsf{L}^{j_k+1}(u_k),$$

which is clearly again an element of W. Thus W is L-invariant. Therefore $\langle \mathsf{L}, \mathsf{U} \rangle \subseteq \mathsf{W}$.

Now let W' be any L-invariant subspace containing U. Then $u \in \mathsf{W}$ for all $u \in \mathsf{U}$. Since W' is L-invariant, $\mathsf{L}(u) \in \mathsf{W}'$ for all $u \in \mathsf{U}$. Continuing to use the L-invariance of W' gives $\mathsf{L}^j(u) \in \mathsf{W}'$ for all $j \in \mathbb{Z}_{\geq 0}$ and $u \in \mathsf{U}$. Thus $\mathsf{W} \subseteq \mathsf{W}'$. In particular, $\mathsf{W} \subseteq \langle \mathsf{L}, \mathsf{U} \rangle$, giving the result. ■

Now we turn to a characterisation of a particular invariant subspace associated to an endomorphism. The construction is a little involved, and seems a little pointless at present. However, it will form the essential part of the definition of algebraic multiplicity in Definition 5.4.57. We consider some constructions involving the kernels of powers of an endomorphism. Thus we let F be a field, V be an F-vector space, and $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. We consider the sequence $(\ker(\mathsf{L}^j))_{j\in\mathbb{Z}_{>0}}$ of subspaces of V, and we note that $\ker(\mathsf{L}^j) \subseteq \ker(\mathsf{L}^{j+1})$ for each $j \in \mathbb{Z}_{>0}$, i.e., the sequence is increasing with respect to the partial order of inclusion of subsets. We denote by U_{L} the subspace generated by $\cup_{j\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^j)$. Since the sequence

$$\ker(\mathsf{L}) \subseteq \ker(\mathsf{L}^2) \subseteq \cdots \subseteq \ker(\mathsf{L}^j) \subseteq \cdots$$

is partially ordered by inclusion, in fact we simply have $\mathsf{U}_{\mathsf{L}} = \cup_{j\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^j)$ (cf. Exercise 4.5.17). Since the subspace U_{L} will be essential in our definition of algebraic multiplicity, let us make a few comments on its properties and its computation in practice.

5.4.13 Theorem (Characterisation of U_{L}) *Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The subspace $\mathsf{U}_{\mathsf{L}} = \cup_{\mathrm{j}\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^{\mathrm{j}})$ is the smallest subspace of V with the properties that*

(i) U_{L} is L-invariant and

(ii) the map induced by L on the quotient $\mathsf{V}/\mathsf{U}_{\mathsf{L}}$ and defined by $\mathrm{v} + \mathsf{U}_{\mathsf{L}} \mapsto \mathsf{L}(\mathrm{v}) + \mathsf{U}_{\mathsf{L}}$ (cf. Proposition 5.4.7) is injective.

Proof Let us first prove that U_{L} has the two properties stated in the proposition. Let $v \in \mathsf{U}_{\mathsf{L}}$ so that $v \in \ker(\mathsf{L}^k)$ for some $k \in \mathbb{Z}_{>0}$. Then $\mathsf{L} \circ \mathsf{L}^k(v) = \mathsf{L}^k(\mathsf{L}(v)) = 0_{\mathsf{V}}$, and so $\mathsf{L}(v) \in \ker(\mathsf{L}^k) \subseteq \mathsf{U}_{\mathsf{L}}$, showing that U_{L} is L-invariant.

Denote by L_0 the endomorphism of $\mathsf{V}/\mathsf{U}_{\mathsf{L}}$ as stated. Let $v \in \mathsf{V}$ be such that $\mathsf{L}_0(v + \mathsf{U}_{\mathsf{L}}) = 0_{\mathsf{V}} + \mathsf{U}_{\mathsf{L}}$. By definition of L_0 this implies that $\mathsf{L}(v) \in \mathsf{U}_{\mathsf{L}}$. Therefore $\mathsf{L}(v) \in \ker(\mathsf{L}^k)$ for some $k \in \mathbb{Z}_{>0}$. That is, $\mathsf{L}^k \circ \mathsf{L}(v) = 0_{\mathsf{V}}$, and so $v \in \ker(\mathsf{L}^{k+1}) \subseteq \mathsf{U}_{\mathsf{L}}$. Thus $v + \mathsf{U}_{\mathsf{L}} = 0_{\mathsf{V}} + \mathsf{U}_{\mathsf{L}}$, so showing that L_0 is injective.

Now we show that U_{L} is the smallest subspace with the two stated properties. Thus we let U'_{L} be a subspace with the two properties, and we denote by $\mathsf{L}'_0 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V}/\mathsf{U}'_{\mathsf{L}})$ the induced endomorphism. We first claim that $\ker(\mathsf{L}) \subseteq \mathsf{U}'_{\mathsf{L}}$. To see this, suppose that $\ker(\mathsf{L})$ is not contained in U'_{L}. Then there exists $v \in \ker(\mathsf{L}) - \mathsf{U}'_{\mathsf{L}}$. Then

$$\mathsf{L}'_0(v + \mathsf{U}'_{\mathsf{L}}) = \mathsf{L}(v) + \mathsf{U}'_{\mathsf{L}} = 0_{\mathsf{V}} + \mathsf{U}'_{\mathsf{L}},$$

so that $\ker(\mathsf{L}'_0) \neq \{0_{\mathsf{V}} + \mathsf{U}'_{\mathsf{L}}\}$. Thus L'_0 is not injective by Exercise 4.5.23.

We next claim that $\ker(\mathsf{L}^j) \subseteq \mathsf{U}'_{\mathsf{L}}$ for $j \in \mathbb{Z}_{>0}$. We have already proved this when $j = 1$, so suppose it true for $j \in \{1, \dots, k\}$ and let $v \in \ker(\mathsf{L}^{k+1})$. Thus $\mathsf{L}^{k+1}(v) = \mathsf{L}^k(\mathsf{L}(v)) = 0_{\mathsf{V}}$. Thus $\mathsf{L}(v) \in \ker(\mathsf{L}^k) \subseteq \mathsf{U}'_{\mathsf{L}}$ by the induction hypothesis. Thus $\mathsf{L}'_0(v + \mathsf{U}'_{\mathsf{L}}) = 0_{\mathsf{V}} + \mathsf{U}'_{\mathsf{L}}$, and so $v \in \mathsf{U}'_{\mathsf{L}}$ since L_0 is injective. This shows that $\ker(\mathsf{L}^j) \subseteq \mathsf{U}'_{\mathsf{L}}$ for each $j \in \mathbb{Z}_{>0}$, as desired. Therefore, by definition of U_{L} and since U'_{L} is a subspace, we have $\mathsf{U}_{\mathsf{L}} \subseteq \mathsf{U}'_{\mathsf{L}}$, which completes the proof. ■

This result has the following corollary which is useful in limiting the computations one must do in practice when computing the algebraic multiplicity. The result says, roughly, that if the sequence

$$\ker(\mathsf{L}) \subseteq \ker(\mathsf{L}^2) \subseteq \cdots \subseteq \ker(\mathsf{L}^j) \subseteq \cdots$$

has two neighbouring terms which are equal, then all remaining terms in the sequence are also equal. This makes the computation of U_L simpler in these cases.

5.4.14 Corollary (Computation of U_L) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_\mathsf{F}(\mathsf{V})$*. If, for some* $\mathrm{k} \in \mathbb{Z}_{>0}$*,* $\ker(\mathsf{L}^\mathrm{k}) = \ker(\mathsf{L}^{\mathrm{k}+1})$*, then* $\ker(\mathsf{L}^\mathrm{j}) = \ker(\mathsf{L}^\mathrm{j})$ *for all* $\mathrm{j} \geq \mathrm{k}$*, and, moreover,* $\mathsf{U}_\mathsf{L} = \ker(\mathsf{L}^\mathrm{k})$*.*

Proof The result will follow from the definition of U_L if we can show that $\mathsf{U}_\mathsf{L} = \ker(\mathsf{L}^k)$. Since $\ker(\mathsf{L}^k) \subseteq \mathsf{U}_\mathsf{L}$, this will follow if we can show that $\ker(\mathsf{L}^k)$ is L-invariant and that the map induced by L on the quotient $\mathsf{V}/\ker(\mathsf{L}^k)$ is injective. First let $v \in \ker(\mathsf{L}^k)$. Then, since $\ker(\mathsf{L}^{k+1}) = \ker(\mathsf{L}^k)$, $\mathsf{L}^{k+1}(v) = \mathsf{L}^k(\mathsf{L}(v)) = 0_\mathsf{V}$, showing that $\mathsf{L}(v) \in \ker(\mathsf{L}^k)$. Thus $\ker(\mathsf{L}^k)$ is L-invariant. Now let L_0 be the induced endomorphism on $\mathsf{V}/\ker(\mathsf{L}^k)$ and suppose that $\mathsf{L}_0(v + \ker(\mathsf{L}^k)) = 0_\mathsf{V} + \ker(\mathsf{L}^k)$. Then, by definition of L_0, $\mathsf{L}(v) \in \ker(\mathsf{L}^k)$. Therefore, $\mathsf{L}^k(\mathsf{L}(v)) = \mathsf{L}^{k+1}(v) = 0_\mathsf{V}$, and so $v \in \ker(\mathsf{L}^{k+1}) = \ker(\mathsf{L}^k)$. This shows that $\ker(\mathsf{L}_0) = \{0_\mathsf{V} + \ker(\mathsf{L}^k)\}$, or that L_0 is injective by Exercise 4.5.23. ■

5.4.3 The algebra of linear maps

Sets of linear maps, like sets of matrices, have defined on them certain algebraic operations which endow them with familiar algebraic structures.

5.4.15 Definition (Sum and scalar multiplication for linear maps) Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L}, \mathsf{K} \in \mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})$ and $a \in \mathsf{F}$.

(i) The ***sum*** of L and K is the element $\mathsf{L} + \mathsf{K}$ of $\mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})$ defined by

$$(\mathsf{L} + \mathsf{K})(u) = \mathsf{L}(u) + \mathsf{K}(u).$$

(ii) ***Multiplication*** of L with a is the element $a\mathsf{L}$ of $\mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})$ defined by $(a\mathsf{L})(u) = a(\mathsf{L}(u))$. •

The following result records the manner in which the two preceding operations interact with the composition of linear maps. The result mirrors Proposition 5.1.7 for matrices, and this will be made precise in Theorem 5.4.22. In the statement of the result, we denote by $0_{\mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})}, -\mathsf{L} \in \mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})$ the linear maps defined by

$$0_{\mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})}(u) = 0_\mathsf{V}, \quad -\mathsf{L}(u) = -(\mathsf{L}(u)), \qquad u \in \mathsf{U}.$$

Thus $0_{\mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})}$ is the zero linear map, and $-\mathsf{L}$ is negative L. We shall also adopt the notation that, for $k \in \mathbb{Z}_{\geq 0}$, L^k is the k-fold composition of L with itself. If $k = 0$ then we adopt the convention that $\mathsf{L}^0 = \mathrm{id}_\mathsf{V}$.

5.4.16 Proposition (Properties of sum and composition of linear maps) *Let* F *be a field, let* U, V, W, *and* X *be* F*-vector spaces, and let* $\mathsf{K}_1, \mathsf{K}_2, \mathsf{K}_3 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$, $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}; \mathsf{W})$, $\mathsf{M}_1 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{W}; \mathsf{X})$, *and* $a_1, a_2 \in \mathsf{F}$*. Then the following equalities hold:*

(i) $\mathsf{K}_1 + \mathsf{K}_2 = \mathsf{K}_2 + \mathsf{K}_1$;

(ii) $(\mathsf{K}_1 + \mathsf{K}_2) + \mathsf{K}_3 = \mathsf{K}_1 + (\mathsf{K}_2 + \mathsf{K}_3)$;

(iii) $\mathsf{K}_1 + 0_{\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})} = \mathsf{K}_1$;

(iv) $\mathsf{K}_1 + (-\mathsf{K}_1) = 0_{\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})}$;

(v) $\mathsf{L}_1 \circ (\mathsf{K}_1 + \mathsf{K}_2) = \mathsf{L}_1 \circ \mathsf{K}_1 + \mathsf{L}_1 \circ \mathsf{K}_2$;

(vi) $(\mathsf{L}_1 + \mathsf{L}_2) \circ \mathsf{K}_1 = \mathsf{L}_1 \circ \mathsf{K}_1 + \mathsf{L}_2 \circ \mathsf{K}_1$;

(vii) $(\mathsf{M}_1 \circ \mathsf{K}_1) \circ \mathsf{L}_1 = \mathsf{M}_1 \circ (\mathsf{K}_1 \circ \mathsf{L}_1)$;

(viii) $a_1(a_2\mathsf{K}_1) = (a_1a_2)\mathsf{K}_1$;

(ix) $(a_1 + a_2)\mathsf{K}_1 = a_1\mathsf{K}_1 + a_2\mathsf{K}_1$;

(x) $a_1(\mathsf{K}_1 + \mathsf{K}_2) = a_1\mathsf{K}_1 + a_1\mathsf{K}_2$.

Proof This is Exercise 5.4.7. ■

With these properties of linear maps, we can endow various sets of linear maps with familiar algebraic structures. The next two results mirror Corollaries 5.1.8 and 5.1.9.

5.4.17 Corollary (Linear maps as elements of a vector space) *If* F *is a field and if* U *and* V *are* F*-vector spaces, then* $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ *is an* F*-vector space with addition given by the sum of linear maps and with multiplication being given by multiplication of a matrix by a scalar.*

5.4.18 Corollary (Linear maps as elements of an algebra) *If* F *is a field and if* V *is an* F*-vector space, then* $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *is an* F*-algebra with the vector space structure of Corollary 5.4.17 and with the product given by the composition of linear maps.*

Note that for endomorphisms it makes sense to ask that they commute.

5.4.19 Definition (Commuting endomorphisms) Let F be a field, let V be an F-vector space. Endomorphisms $\mathsf{L}, \mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ ***commute*** if $\mathsf{L} \circ \mathsf{K} = \mathsf{K} \circ \mathsf{L}$. •

Not all endomorphisms commute (Exercise 5.4.9). In Exercises 5.4.10 and 5.4.13 we ask the reader to examine some properties of commuting endomorphisms.

5.4.4 Linear maps and matrices

In this section we make precise the connection between linear maps and matrices. As we shall see, the two concepts are linked by specific choices of basis. This then gives rise to the question, "What is the effect of choosing different bases," and so to the notion of change of basis formulae.

Let us get things started by making precise the connection between linear maps and matrices upon the choice of a basis. We let F be a field, and U and V be F-vector spaces. We let $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ be bases for U and V, respectively. In order to make the connection to matrices notationally convenient, it is useful to introduce index sets I and J that index the elements of the bases $\mathscr{B}_{\mathsf{V}}$ and $\mathscr{B}_{\mathsf{U}}$, respectively. Precisely, we let I and J be sets for which there exist bijections $\phi_{\mathsf{V}}\colon I \to \mathscr{B}_{\mathsf{V}}$ and $\phi_{\mathsf{U}}\colon J \to \mathscr{B}_{\mathsf{U}}$. Now let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$. For $u \in \mathscr{B}_{\mathsf{U}}$ denote $j = \phi_{\mathsf{U}}^{-1}(u) \in J$. Since $\mathscr{B}_{\mathsf{V}}$ is a basis for V, for each $i \in I$ there exists $a_{ij} \in \mathsf{F}$, with a_{ij} being nonzero for only finitely many i, such that

$$\mathsf{L}(u) = \sum_{i \in I} a_{ij}\phi_{\mathsf{V}}(i),$$

i.e., $\mathsf{L}(u)$ is a finite linear combination of basis vectors from $\mathscr{B}_{\mathsf{V}}$. The matrix $[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}\colon I \times J \to \mathsf{F}$ defined by $[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}(i,j) = a_{ij}$. Summarising this, we have the following definition.

5.4.20 Definition (Matrix representative of linear map relative to bases) Let F be a field, let U and V be F-vector spaces, let $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ be bases for U and V, respectively, and let I and J be sets for which there exist bijections $\phi_{\mathsf{U}}\colon J \to \mathscr{B}_{\mathsf{U}}$ and $\phi_{\mathsf{V}}\colon I \to \mathscr{B}_{\mathsf{V}}$. If $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ then the ***matrix representative*** of L with respect to the bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ and to the bijections ϕ_{U} and ϕ_{V} is the map $[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}\colon I \times J \to \mathsf{F}$ as defined above. •

Let us represent this a little more explicitly in the case where U and V are finite-dimensional, having bases $\mathscr{B}_{\mathsf{U}} = \{u_1, \dots, u_n\}$ and $\mathscr{B}_{\mathsf{V}} = \{v_1, \dots, v_m\}$, respectively. In this case we can write

$$\begin{aligned}
\mathsf{L}(u_1) &= a_{11}v_1 + a_{21}v_2 + \dots + a_{m1}v_m,\\
\mathsf{L}(u_2) &= a_{12}v_1 + a_{22}v_2 + \dots + a_{m2}v_m,\\
&\;\;\vdots\\
\mathsf{L}(u_n) &= a_{1n}v_1 + a_{2n}v_2 + \dots + a_{mn}v_m,
\end{aligned} \tag{5.17}$$

for uniquely defined a_{ij}, $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$. In this case the matrix representative for L with respect to the bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ is

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \tag{5.18}$$

Note carefully the way one assembles the coefficients in (5.17) into the matrix in (5.18); what appear as "rows" in (5.17) are the columns in the matrix representative.

Note that a matrix associated to a linear map $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ is defined only once one defines bases for both U and V. As we shall see below, different bases generally give rise to different matrices for the same linear map. A helpful way to think about the correspondence between a linear map and its matrix representative is as follows. Let us first recall from Section 4.5.5 some constructions associated with a vector space V in the presence of a basis $\mathscr{B}$. In Theorem 4.5.45 we showed that $\mathscr{B}$ gives rise to an F-isomorphism from V to $\bigoplus_{u\in\mathscr{B}} \mathsf{F}$. In order to match this with our matrix constructions, let I be an index set such that there exists a bijection $\phi_{\mathscr{B}} \colon I \to \mathscr{B}$. Then the conclusion of Theorem 4.5.45 is that, associated to $\mathscr{B}$ and $\phi_{\mathscr{B}}$, there is an isomorphism $\iota_{\mathscr{B}} \colon \mathsf{V} \to \mathsf{F}_0^I$ defined as follows. If $v \in \mathsf{V}$ then we can write

$$v = c_1\phi_{\mathscr{B}}(i_1) + \cdots + c_k\phi_{\mathscr{B}}(i_k)$$

for some unique $c_1, \dots, c_k \in \mathsf{F}^*$ and $i_1, \dots, i_k \in I$. We then define

$$\iota_{\mathscr{B}}(v)(i) = \begin{cases} \phi_{\mathscr{B}}(i), & i \in \{1, \dots, i_k\}, \\ 0_{\mathsf{F}}, & i \notin \{i_1, \dots, i_k\}. \end{cases}$$

With this construction in place, the following result tells us that $[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}$ is what L "looks like" when we use the isomorphisms $\iota_{\mathscr{B}_{\mathsf{U}}} \colon \mathsf{U} \to \mathsf{F}_0^J$ and $\iota_{\mathscr{B}_{\mathsf{V}}} \colon \mathsf{V} \to \mathsf{F}_0^I$.

5.4.21 Theorem (Interpretation of matrix representative) *Let F be a field, let U and V be F-vector spaces, let $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ be bases for U and V, respectively, and let I and J be sets for which there exist bijections $\phi_{\mathsf{U}} \colon \mathrm{J} \to \mathscr{B}_{\mathsf{U}}$ and $\phi_{\mathsf{V}} \colon \mathrm{I} \to \mathscr{B}_{\mathsf{V}}$. If $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ then, with the isomorphisms $\iota_{\mathscr{B}_{\mathsf{U}}} \colon \mathsf{U} \to \mathsf{F}_0^{\mathrm{J}}$ and $\iota_{\mathscr{B}_{\mathsf{V}}} \colon \mathsf{V} \to \mathsf{F}_0^{\mathrm{I}}$ as defined above, the following diagram commutes:*

$$\begin{tikzcd}
\mathsf{U} \arrow[r, "\mathsf{L}"] \arrow[d, "\iota_{\mathscr{B}_{\mathsf{U}}}"'] & \mathsf{V} \arrow[d, "\iota_{\mathscr{B}_{\mathsf{V}}}"] \\
\mathsf{F}_0^{\mathrm{J}} \arrow[r, "{[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}}"'] & \mathsf{F}_0^{\mathrm{I}}
\end{tikzcd}$$

In particular, the map $\mathsf{L} \mapsto \iota_{\mathscr{B}_{\mathsf{V}}} \circ \mathsf{L} \circ \iota_{\mathscr{B}_{\mathsf{U}}}^{-1}$ is an isomorphism of the F-vector spaces $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U},\mathsf{V})$ and $\mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$.

Proof Let $u \in \mathsf{U}$ and write

$$u = c_1\phi_{\mathsf{U}}(j_1) + \cdots + c_k\phi_{\mathsf{U}}(j_k)$$

for unique $c_1, \dots, c_k \in \mathsf{F}^*$ and $j_1, \dots, j_k \in J$. Then, using the definition of $\iota_{\mathscr{B}_{\mathsf{U}}}$, we compute, for $i \in I$,

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}} \circ \iota_{\mathscr{B}_{\mathsf{U}}}(u)(i) = \sum_{l=1}^{k} [\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}(i, j_l)c_l. \tag{5.19}$$

On the other hand, using the definition of the matrix representative,

$$\mathsf{L}(u) = \sum_{l=1}^{k} c_l\mathsf{L}(\phi_{\mathsf{U}}(j_l)) = \sum_{i\in I}\sum_{l=1}^{k} c_l[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}(i, j_l)\phi_{\mathsf{V}}(i). \tag{5.20}$$

The result now follows by comparing (5.19) and (5.20) and from the definition of $\iota_{\mathscr{B}_{\mathsf{V}}}$.

The final assertion of the theorem follows by noting that the given map is a bijection, and by directly checking that it is linear. ■

The next result tells us that the matrix representative of the composition of linear maps is the product of the matrix representatives.

5.4.22 Theorem (Matrix representative of a composition) *Let F be a field, let U, V, and W be F-vector spaces, let $\mathscr{B}_{\mathsf{U}}$, $\mathscr{B}_{\mathsf{V}}$, and $\mathscr{B}_{\mathsf{W}}$ be bases for U, V, and W, respectively, and let I, J, and K be index sets for which there exists bijections $\phi_{\mathsf{U}}\colon I \to \mathscr{B}_{\mathsf{U}}$, $\phi_{\mathsf{V}}\colon J \to \mathscr{B}_{\mathsf{V}}$, and $\phi_{\mathsf{W}}\colon K \to \mathscr{B}_{\mathsf{W}}$. If $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ and $\mathsf{M} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{W})$, then*

$$[\mathsf{M}\circ\mathsf{L}]^{\mathscr{B}_{\mathsf{W}}}_{\mathscr{B}_{\mathsf{U}}} = [\mathsf{M}]^{\mathscr{B}_{\mathsf{W}}}_{\mathscr{B}_{\mathsf{V}}}[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}.$$

Proof For $i \in I$ compute, using the definition of matrix representative and using linearity,

$$\begin{aligned}(\mathsf{M}\circ\mathsf{L})(\phi_{\mathsf{U}}(i)) &= \mathsf{M}\Bigg(\sum_{j\in J}[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(j,i)\phi_{\mathsf{V}}(j)\Bigg)\\ &= \sum_{j\in J}[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(j,i)\mathsf{M}(\phi_{\mathsf{V}}(j))\\ &= \sum_{j\in J}\sum_{k\in K}[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(j,i)[\mathsf{M}]^{\mathscr{B}_{\mathsf{W}}}_{\mathscr{B}_{\mathsf{V}}}(k,j)\phi_{\mathsf{W}}(k)\\ &= \sum_{k\in K}[\mathsf{M}\circ\mathsf{L}]^{\mathscr{B}_{\mathsf{W}}}_{\mathscr{B}_{\mathsf{U}}}(k,i)\phi_{\mathsf{W}}(k),\end{aligned}$$

which gives the result. ■

For endomorphisms of a vector space, the preceding result, combined with Theorem 5.4.21 has the following corollary.

5.4.23 Corollary (Matrix representatives of endomorphisms) *Let F be a field, let V be an F-vector space, let $\mathscr{B}$ be a basis for V, let I be an index set for which there exists a bijection $\phi\colon I \to \mathscr{B}$, and let $\iota_{\mathscr{B}}\colon \mathsf{V} \to \mathsf{F}^I_0$ be the isomorphism defined preceding the statement of Theorem 5.4.21. Then the map $\mathsf{L} \mapsto \iota_{\mathscr{B}}\circ\mathsf{L}\circ\iota^{-1}_{\mathscr{B}}$ is an isomorphism of the F algebras $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ and $\mathrm{Mat}_{I\times I}(\mathsf{F})$.*

Let us give an example of how to construct the matrix representative of a linear map in a simple case.

5.4.24 Example (Matrix representative of a linear map) Let $\mathsf{U} = \mathsf{F}^2$ and $\mathsf{V} = \mathsf{F}^3$. Define an element L of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ by

$$\mathsf{L}(x_1,x_2) = (x_1 - 2_{\mathsf{F}}x_2, x_2, 3_{\mathsf{F}}x_1 - x_2),$$

where $k_{\mathsf{F}} = k1_{\mathsf{F}}$, recalling the notation of Proposition 4.2.10. As bases for F^2 and F^3, let us use the standard bases which we denote by $\{\boldsymbol{e}_1,\boldsymbol{e}_2\}$ and $\{\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3\}$. We take

$I = \{1,2,3\}$ and $J = \{1,2\}$, and use the obvious bijections $\phi_{\mathsf{U}}(j) = \boldsymbol{e}_j$, $j \in \{1,2\}$, and $\phi_{\mathsf{V}}(i) = \boldsymbol{f}_i$, $i \in \{1,2,3\}$. It is then a simple computation to see that

$$\begin{aligned}\mathsf{L}(\boldsymbol{e}_1) &= (1_{\mathsf{F}}, 0_{\mathsf{F}}, 3_{\mathsf{F}}) = 1_{\mathsf{F}}\boldsymbol{f}_1 + 0_{\mathsf{F}}\boldsymbol{f}_2 + 3_{\mathsf{F}}\boldsymbol{f}_3,\\ \mathsf{L}(\boldsymbol{e}_2) &= (-2_{\mathsf{F}}, 1_{\mathsf{F}}, -1_{\mathsf{F}}) = -2_{\mathsf{F}}\boldsymbol{f}_1 + 1_{\mathsf{F}}\boldsymbol{f}_2 - 1_{\mathsf{F}}\boldsymbol{f}_3.\end{aligned}$$

Thus the matrix representative for L associated to these bases and bijections is given by

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}} = \begin{bmatrix} 1_{\mathsf{F}} & -2_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 3_{\mathsf{F}} & -1_{\mathsf{F}} \end{bmatrix}. \qquad \bullet$$

Note that in the preceding example, since the linear map is between F^2 and F^3, by Theorem 5.1.13 we already know that there is an exact correspondence between such linear maps and matrices. The following result records how this previous correspondence between linear maps and matrices meshes with the one of Definition 5.4.20. The result says, roughly, that the linear map defined by a matrix has the matrix itself as its matrix representative with respect to the standard bases.

5.4.25 Proposition (Matrix representative associated to standard bases) *Let* F *be a field, let* I *and* J *be index sets, let* $\mathsf{U} = \mathsf{F}_0^{\mathrm{J}}$ *and* $\mathsf{V} = \mathsf{F}_0^{\mathrm{I}}$*, and let* $\mathscr{B}_{\mathsf{U}} = \{\mathbf{e}_{\mathrm{j}}\}_{\mathrm{j}\in\mathrm{J}}$ *and* $\mathscr{B}_{\mathsf{V}} = \{\mathbf{f}_{\mathrm{i}}\}_{\mathrm{i}\in\mathrm{I}}$ *be the standard bases for* $\mathsf{F}_0^{\mathrm{J}}$ *and* $\mathsf{F}_0^{\mathrm{I}}$*, respectively. Let* $\phi_{\mathsf{U}}\colon \mathrm{J} \to \mathscr{B}_{\mathsf{U}}$ *and* $\phi_{\mathsf{V}}\colon \mathrm{I} \to \mathscr{B}_{\mathsf{V}}$ *be defined by*

$$\phi_{\mathsf{U}}(\mathrm{j}) = \mathbf{e}_{\mathrm{j}},\ \mathrm{j} \in \mathrm{J}, \quad \phi_{\mathsf{V}}(\mathrm{i}) = \mathbf{f}_{\mathrm{i}},\ \mathrm{i} \in \mathrm{I}.$$

If $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *is column finite and so defines an element of* $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ *by Theorem 5.1.13, then* $[\mathbf{A}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}} = \mathbf{A}$.

Proof This follows directly from the computations in the proof of Theorem 5.1.13 and the constructions preceding Definition 5.4.20. ■

Note that if one chooses bases for F_0^J and F_0^I that are not the standard bases, then the preceding result is generally not true. Thus the result is as much about the standard bases as it is about the linear map defined by a matrix $\boldsymbol{A}$. We shall explore this further below when we discuss changing bases.

5.4.5 Changing of bases

Now let us address the rather important matter of changes of basis. We first study how one can construct an invertible matrix, called the change of basis matrix, associated with any two bases for the same vector space. Then we examine how the change of basis matrix is used to relate matrix representatives relative to different bases.

The following result provides the existence of a matrix that relates two different bases. The statement of the result relies on the fact that two bases have the same cardinality (Theorem 4.5.25).

5.4.26 Proposition (Existence of change of basis matrix) *Let* F *be a field, let* V *be an* F*-vector space, let* $\mathscr{B}$ *and* $\mathscr{B}'$ *be bases for* V*, and let* I *be an index set for which there exist bijections* $\phi\colon \mathrm{I} \to \mathscr{B}$ *and* $\phi'\colon \mathrm{I} \to \mathscr{B}'$*. Then there exists a unique invertible column finite matrix* $\mathbf{P} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{F})$ *such that*

$$\phi(\mathrm{i}_0) = \sum_{\mathrm{i}\in\mathrm{I}} \mathbf{P}(\mathrm{i}, \mathrm{i}_0)\phi'(\mathrm{i})$$

for each $\mathrm{i}_0 \in \mathrm{I}$.

Proof Let $i_0 \in I$. Since $\mathscr{B}'$ is a basis there exists unique $i_1, \dots, i_k \in I$ and $c_1, \dots, c_k \in \mathsf{F}^*$ such that

$$\phi(i_0) = c_1\phi'(i_1) + \dots + c_k\phi'(i_k).$$

One then defines $\boldsymbol{P}$ by asking that $\boldsymbol{P}(i, i_0) = c_j$ if $i = i_j$ for some $j \in \{1, \dots, k\}$, and that $\boldsymbol{P}(i, i_0) = 0_{\mathsf{F}}$ for $i \in I\setminus\{i_1, \dots, i_k\}$. Note that $\boldsymbol{P}$ thus defined is column finite. We next show that it is invertible. To do this, we construct its inverse. By swapping the rôles of $\mathscr{B}$ and $\mathscr{B}'$, our above argument gives the existence a column finite matrix $\boldsymbol{Q} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ such that

$$\phi'(i_0) = \sum_{i\in I} \boldsymbol{Q}(i, i_0)\phi(i)$$

for each $i_0 \in I$. We claim that $\boldsymbol{QP} = \boldsymbol{I}_I$. Let $i_0 \in I$ and note that

$$\phi(i_0) = \sum_{i\in I} \boldsymbol{P}(i, i_0)\phi'(i) = \sum_{i\in I}\sum_{i'\in I} \boldsymbol{P}(i, i_0)\boldsymbol{Q}(i', i)\phi(i') = \sum_{i'\in I}(\boldsymbol{QP})(i', i_0)\phi(i').$$

Since $\{\phi(i)\}_{i\in I}$ is a basis, and in particular is linearly independent, this implies that

$$(\boldsymbol{QP})(i', i_0) = \begin{cases} 1_{\mathsf{F}}, & i' = i_0, \\ 0_{\mathsf{F}}, & i' \neq i_0. \end{cases}$$

In other words, $\boldsymbol{QP} = \boldsymbol{I}_I$. In like manner we compute

$$\phi'(i_0) = \sum_{i\in I} \boldsymbol{Q}(i, i_0)\phi(i) = \sum_{i\in I}\sum_{i'\in I} \boldsymbol{Q}(i, i_0)\boldsymbol{P}(i', i)\phi'(i') = \sum_{i'\in I}(\boldsymbol{PQ})(i', i_0)\phi'(i'),$$

which leads to the conclusion, using the fact that $\{\phi'(i)\}_{i\in I}$ is linearly independent, that $\boldsymbol{PQ} = \boldsymbol{I}_I$. Thus $\boldsymbol{Q}$ is the inverse of $\boldsymbol{P}$. ■

We introduce some notation and terminology associated with the matrix $\boldsymbol{P}$ of the preceding result.

5.4.27 Definition (Change of basis matrix) Let F be a field, let V be an F-vector space, let $\mathscr{B}$ and $\mathscr{B}'$ be bases for V, and let I be an index set for which there exist bijections $\phi\colon I \to \mathscr{B}$ and $\phi'\colon I \to \mathscr{B}'$. The matrix $\boldsymbol{P}$ of Proposition 5.4.26 is called the ***change of basis matrix*** from the basis $\mathscr{B}$ to the basis $\mathscr{B}'$, and is denoted by $\boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}$. •

As a corollary to Proposition 5.4.26 we have the following result.

5.4.28 Corollary (Inverse of change of basis matrix) *Let* F *be a field, let* V *be an* F*-vector space, let* $\mathscr{B}$ *and* $\mathscr{B}'$ *be bases for* V*, and let* I *be an index set for which there exist bijections* $\phi\colon \mathrm{I} \to \mathscr{B}$ *and* $\phi'\colon \mathrm{I} \to \mathscr{B}'$*. Then* $\mathbf{P}_{\mathscr{B}'}^{\mathscr{B}} = (\mathbf{P}_{\mathscr{B}}^{\mathscr{B}'})^{-1}$*.*

Proof This was shown during the course of the proof of Proposition 5.4.26. ■

If one has more than two bases, the change of basis matrices can be related in a simple way.

5.4.29 Proposition (Product of change of basis matrices is a change of basis matrix) *Let* F *be a field, let* V *be an* F*-vector space, let* $\mathscr{B}$*,* $\mathscr{B}'$*, and* $\mathscr{B}''$ *be bases for* V*, and let* I *be an index set for which there exist bijections* $\phi\colon \mathrm{I} \to \mathscr{B}$*,* $\phi'\colon \mathrm{I} \to \mathscr{B}'$*, and* $\phi''\colon \mathrm{I} \to \mathscr{B}''$*. Then*

$$\mathbf{P}_{\mathscr{B}}^{\mathscr{B}''} = \mathbf{P}_{\mathscr{B}'}^{\mathscr{B}''}\mathbf{P}_{\mathscr{B}}^{\mathscr{B}'}.$$

Proof Let $i_0 \in I$ and compute

$$\begin{aligned}\phi(i_0) &= \sum_{i\in I} \boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}(i,i_0)\phi'(i)\\ &= \sum_{i\in I}\sum_{i'\in I} \boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}(i,i_0)\boldsymbol{P}_{\mathscr{B}'}^{\mathscr{B}''}(i',i)\phi''(i')\\ &= \sum_{i'\in I}(\boldsymbol{P}_{\mathscr{B}'}^{\mathscr{B}''}\boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'})(i',i_0)\phi''(i'),\end{aligned}$$

giving the result by definition of $\boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}''}$. ■

As our final property of change of basis matrices, let us show that the definition of the change of basis matrix can, in some sense, be inverted.

5.4.30 Proposition (Invertible matrices give rise to changes of basis) *Let* F *be a field, let* V *be an* F*-vector space, let* $\mathscr{B}$ *be a basis for* V*, and let* I *be an index set such that there exists a bijection* $\phi\colon \mathrm{I} \to \mathscr{B}$*. Then, given an invertible column matrix* $\mathbf{P} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{F})$*, there exists a basis* $\mathscr{B}'$ *for* V *such that* $\mathbf{P} = \mathbf{P}_{\mathscr{B}}^{\mathscr{B}'}$*.*

Proof We define $\mathscr{B}'$ by defining an injective map $\phi'\colon I \to \mathsf{V}$ and taking $\mathscr{B}' = \mathrm{image}(\phi)$. We define ϕ' by

$$\phi'(i_0) = \sum_{i\in I} \boldsymbol{P}^{-1}(i,i_0)\phi(i).$$

Let us show that $\{\phi'(i)\}_{i\in I}$ is a basis for V. First we prove linear independence. Suppose that $c_1,\dots,c_k \in \mathsf{F}$ and $i_1,\dots,i_k \in I$ satisfy

$$c_1\phi'(i_1) + \dots + c_k\phi'(i_k) = 0_{\mathsf{V}}.$$

Then

$$\sum_{j=1}^{k}\sum_{i\in I} \boldsymbol{P}^{-1}(i,i_j)c_j\phi(i) = 0_{\mathsf{V}}.$$

Since $\{\phi(i)\}_{i\in I}$ is a basis we have

$$\sum_{j=1}^{k} \boldsymbol{P}^{-1}(i,i_j)c_j = 0_{\mathsf{F}}, \qquad i \in I. \tag{5.21}$$

Now, if we define $\boldsymbol{c} \in \mathsf{F}_0^I$ by

$$\boldsymbol{c}(i) = \begin{cases} c_j, & i = i_j,\ j \in \{1, \dots, k\}, \\ 0_{\mathsf{F}}, & i \notin \{i_1, \dots, i_k\}, \end{cases}$$

then (5.21) is simply $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{0}_{\mathsf{F}_0^I}$, and so $\boldsymbol{c} = \boldsymbol{0}_{\mathsf{F}_0^I}$ by Exercise 4.5.23. Thus $c_1 = \cdots = c_k = 0_{\mathsf{F}}$, giving linear independence. Now we show that $\{\phi'(i)\}_{i \in I}$ generates V. Note that

$$\sum_{i' \in I} \boldsymbol{P}(i', i_0)\phi(i') = \sum_{i \in I} \sum_{i' \in I} \boldsymbol{P}(i', i_0)\boldsymbol{P}^{-1}(i, i')\phi(i) = \phi(i_0). \tag{5.22}$$

Therefore, every element in the basis $\{\phi(i)\}_{i \in I}$ is a finite linear combination of vectors from $\{\phi'(i)\}_{i \in I}$. Since every vector in V is a finite linear combination of vectors from $\{\phi(i)\}_{i \in I}$, it then immediately follows that every vector in V is a finite linear combination of vectors from $\{\phi'(i)\}_{i \in I}$. Thus $\mathscr{B}' = \{\phi'(i)\}_{i \in I}$ is a basis.

It follows (5.22) that $\boldsymbol{P} = \boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}$. ■

The result than says that there is a 1–1 correspondence between invertible matrices and changes of bases, provided one is given an initial basis.

Let us determine the change of basis matrix in an example.

5.4.31 Example (Change of basis matrix) Let $\mathsf{V} = \mathsf{F}^3$ and consider two bases

$$\mathscr{B} = \{\boldsymbol{f}_1 = (1_{\mathsf{F}}, 0_{\mathsf{F}}, 0_{\mathsf{F}}), \boldsymbol{f}_2 = (0_{\mathsf{F}}, 1_{\mathsf{F}}, 0_{\mathsf{F}}), \boldsymbol{f}_3 = (0_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}})\},$$
$$\mathscr{B}' = \{\boldsymbol{f}_1' = (1_{\mathsf{F}}, 1_{\mathsf{F}}, 1_{\mathsf{F}}), \boldsymbol{f}_2' = (0_{\mathsf{F}}, 1_{\mathsf{F}}, 1_{\mathsf{F}}), \boldsymbol{f}_3' = (0_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}})\}.$$

Note that $\mathscr{B}$ is the standard basis, and we leave to the reader the verification that $\mathscr{B}'$ is also a basis. To compute the change of basis matrices we note that

$$\begin{aligned} \boldsymbol{f}_1' &= 1_{\mathsf{F}}\boldsymbol{f}_1 + 1_{\mathsf{F}}\boldsymbol{f}_2 + 1_{\mathsf{F}}\boldsymbol{f}_3, \\ \boldsymbol{f}_2' &= 1_{\mathsf{F}}\boldsymbol{f}_2 + 1_{\mathsf{F}}\boldsymbol{f}_3, \\ \boldsymbol{f}_3' &= 1_{\mathsf{F}}\boldsymbol{f}_3. \end{aligned}$$

Using the definition of change of basis matrix we then immediately have

$$\boldsymbol{P}_{\mathscr{B}'}^{\mathscr{B}} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 1_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

By Corollary 5.4.28, to compute $\boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}$ we need only compute the inverse of $\boldsymbol{P}_{\mathscr{B}'}^{\mathscr{B}}$. One can directly verify that

$$\boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'} = (\boldsymbol{P}_{\mathscr{B}'}^{\mathscr{B}})^{-1} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ -1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & -1_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

Note that it is sometimes the case that it is easier to compute the inverse of the change of basis matrix, and then compute the change of basis matrix by performing the matrix inverse. •

We may now address the matter of how matrix representations change when one changes bases.

5.4.32 Theorem (Change of basis formula) *Let* F *be a field, let* U *and* V *be* F*-vector spaces, let* $\mathscr{B}_{\mathsf{U}}$ *and* $\mathscr{B}'_{\mathsf{U}}$ *be bases for* U*, let* $\mathscr{B}_{\mathsf{V}}$ *and* $\mathscr{B}'_{\mathsf{V}}$ *be bases for* V*, and let* I *and* J *be sets for which there exist bijections* $\phi_{\mathsf{U}}\colon \mathrm{J} \to \mathscr{B}_{\mathsf{U}}$*,* $\phi'_{\mathsf{U}}\colon \mathrm{J} \to \mathscr{B}'_{\mathsf{U}}$*,* $\phi_{\mathsf{V}}\colon \mathrm{I} \to \mathscr{B}_{\mathsf{V}}$*, and* $\phi'_{\mathsf{V}}\colon \mathrm{I} \to \mathscr{B}'_{\mathsf{V}}$*. If* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ *then*

$$[\mathsf{L}]^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{U}}} = \mathbf{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(\mathbf{P}^{\mathscr{B}'_{\mathsf{U}}}_{\mathscr{B}_{\mathsf{U}}})^{-1}.$$

This relation is called the ***change of basis formula****.*

Proof For $j_0 \in J$ we compute

$$\begin{aligned}
\mathsf{L}(\phi'_{\mathsf{U}}(j_0)) &= \mathsf{L}\Bigg(\sum_{j\in J} \boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}}(j,j_0)\phi_{\mathsf{U}}(j)\Bigg)\\
&= \sum_{j\in J} \boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}}(j,j_0)\mathsf{L}(\phi_{\mathsf{U}}(j))\\
&= \sum_{j\in J}\sum_{i\in I} \boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}}(j,j_0)[L]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(i,j)\phi_{\mathsf{V}}(i)\\
&= \sum_{j\in J}\sum_{i\in I}\sum_{i'\in I} \boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}}(j,j_0)[L]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(i,j)\boldsymbol{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}(i',i)\phi'_{\mathsf{V}}(i')\\
&= \sum_{i'\in I} (\boldsymbol{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}[L]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}\boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}})(i',j_0)\phi'_{\mathsf{V}}(i').
\end{aligned}$$

The result now follows from the definition of the matrix representative and from Corollary 5.4.28. ■

Let us apply the change of basis formula in an example.

5.4.33 Example (Change of basis formula) Let take $\mathsf{U} = \mathsf{F}^2$, $\mathsf{V} = \mathsf{F}^3$, and, as in Example 5.4.24, take the linear map

$$\mathsf{L}(x_1, x_2) = (x_1 - 2_{\mathsf{F}}x_2, x_2, 3_{\mathsf{F}}x_1 - x_2).$$

As bases for U we take

$$\begin{aligned}
\mathscr{B}_{\mathsf{U}} &= \{\boldsymbol{e}_1 = (1_{\mathsf{F}}, 0_{\mathsf{F}}), \boldsymbol{e}_2 = (0_{\mathsf{F}}, 1_{\mathsf{F}})\},\\
\mathscr{B}'_{\mathsf{U}} &= \{\boldsymbol{e}'_1 = (1_{\mathsf{F}}, 1_{\mathsf{F}}), \boldsymbol{e}'_2 = (0_{\mathsf{F}}, 1_{\mathsf{F}})\},
\end{aligned}$$

and as bases for V we take, as in Example 5.4.31,

$$\begin{aligned}
\mathscr{B}_{\mathsf{V}} &= \{\boldsymbol{f}_1 = (1_{\mathsf{F}}, 0_{\mathsf{F}}, 0_{\mathsf{F}}), \boldsymbol{f}_2 = (0_{\mathsf{F}}, 1_{\mathsf{F}}, 0_{\mathsf{F}}), \boldsymbol{f}_3 = (0_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}})\},\\
\mathscr{B}'_{\mathsf{V}} &= \{\boldsymbol{f}'_1 = (1_{\mathsf{F}}, 1_{\mathsf{F}}, 1_{\mathsf{F}}), \boldsymbol{f}'_2 = (0_{\mathsf{F}}, 1_{\mathsf{F}}, 1_{\mathsf{F}}), \boldsymbol{f}'_3 = (0_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}})\}.
\end{aligned}$$

In Example 5.4.31 we determined that

$$\boldsymbol{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ -1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & -1_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

To determine $\boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}}$ we note that

$$\boldsymbol{e}'_1 = 1_{\mathsf{F}}\boldsymbol{e}_1 + 1_{\mathsf{F}}\boldsymbol{e}_2, \quad \boldsymbol{e}'_2 = 1_{\mathsf{F}}\boldsymbol{e}_2.$$

Therefore, using the definition of the change of basis matrix,

$$\boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}'_{\mathsf{U}}} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 1_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

As we saw in Example 5.4.24, the matrix representative of L with respect to the bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ is

$$[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}} = \begin{bmatrix} 1_{\mathsf{F}} & -2_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 3_{\mathsf{F}} & -1_{\mathsf{F}} \end{bmatrix}.$$

Therefore, we use Theorem 5.4.32 to determine that the matrix representative of L relative to the bases $\mathscr{B}'_{\mathsf{U}}$ and $\mathscr{B}'_{\mathsf{V}}$ is

$$[\mathsf{L}]^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{U}}} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ -1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & -1_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix} \begin{bmatrix} 1_{\mathsf{F}} & -2_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 3_{\mathsf{F}} & -1_{\mathsf{F}} \end{bmatrix} \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 1_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix} = \begin{bmatrix} -1_{\mathsf{F}} & -2_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 1_{\mathsf{F}} & -2_{\mathsf{F}} \end{bmatrix}.$$

One could also have computed $[\mathsf{L}]^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{U}}}$ by using the definition of matrix representative applied to the bases $\mathscr{B}'_{\mathsf{U}}$ and $\mathscr{B}'_{\mathsf{V}}$. Often, when computing matrix representatives, relative to nonstandard bases, of elements of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^n; \mathsf{F}^m)$, it is easier to determine the matrix representative relative to the standard bases, and then use the change of basis formula. •

5.4.6 Determinant and trace of an endomorphism

In this section we indicate how matrix representatives can be used to define determinants of endomorphisms of finite-dimensional vector spaces. To motivate this we recall from Proposition 5.3.3(ii) that if $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ and if $\boldsymbol{P} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ is invertible, then

$$\det(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}) = \det\boldsymbol{P}\det\boldsymbol{A}\det\boldsymbol{P}^{-1} = \det\boldsymbol{A}.$$

Therefore, by Theorem 5.4.32, if V is a finite-dimensional F-vector space, if $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$, and if $\mathscr{B}$ and $\mathscr{B}'$ are bases for V, we have

$$\det[\mathsf{L}]^{\mathscr{B}'}_{\mathscr{B}'} = \det(\boldsymbol{P}^{\mathscr{B}'}_{\mathscr{B}}[\mathsf{L}]^{\mathscr{B}}_{\mathscr{B}}(\boldsymbol{P}^{\mathscr{B}'}_{\mathscr{B}})^{-1}) = \det[\mathsf{L}]^{\mathscr{B}}_{\mathscr{B}}.$$

That is to say, the determinant of the matrix representative of an endomorphism is independent of choice of basis. With this in mind, we make the following definition.

5.4.34 Definition (Determinant of an endomorphism) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The ***determinant*** of L is given by $\det \mathsf{L} = \det[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$, where $\mathscr{B}$ is any basis for V. •

Determinants of endomorphisms inherit many of the properties of determinants of square matrices.

5.4.35 Theorem (Properties of determinant) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L}, \mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then the following statements hold:*

(i) $\det \mathrm{id}_{\mathsf{V}} = 1_{\mathsf{F}}$*;*

(ii) $\det(\mathsf{L} \circ \mathsf{K}) = \det \mathsf{L} \det \mathsf{K}$*;*

(iii) L *is invertible if and only if* $\det \mathsf{L} \neq 0_{\mathsf{F}}$*;*

(iv) if L *is invertible then* $\det(\mathsf{L}^{-1}) = (\det \mathsf{L})^{-1}$.

The manner in which one defines the trace is similar. Here one uses the properties of trace in Proposition 5.3.18 to see that if $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ and if $\boldsymbol{P} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ is invertible then we have

$$\mathrm{tr}(\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{P}) = \mathrm{tr}\,\boldsymbol{A}.$$

Thus the following definition makes sense.

5.4.36 Definition (Trace of an endomorphism) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The ***trace*** of L is given by $\mathrm{tr}\,\mathsf{L} = \mathrm{tr}[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$, where $\mathscr{B}$ is any basis for V. •

The trace has the following properties.

5.4.37 Proposition (Properties of trace) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, let* $\mathsf{L}, \mathsf{K}, \mathsf{M} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*, and let* $\mathsf{a} \in \mathsf{F}$*. Then the following statements hold:*

(i) $\mathrm{tr}(\mathsf{L} + \mathsf{K}) = \mathrm{tr}\,\mathsf{L} + \mathrm{tr}\,\mathsf{K}$*;*

(ii) $\mathrm{tr}(\mathsf{a}\mathsf{L}) = \mathsf{a}\,\mathrm{tr}\,\mathsf{L}$*;*

(iii) $\mathrm{tr}(\mathsf{L} \circ \mathsf{K}) = \mathrm{tr}(\mathsf{K} \circ \mathsf{L})$*;*

(iv) $\mathrm{tr}(\mathsf{L} \circ \mathsf{K} \circ \mathsf{M}) = \mathrm{tr}(\mathsf{M} \circ \mathsf{L} \circ \mathsf{K}) = \mathrm{tr}(\mathsf{K} \circ \mathsf{M} \circ \mathsf{L})$.

5.4.7 Equivalence of linear maps

In Definition 5.1.38 we introduced the idea of equivalence of matrices, saying that matrices $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ are equivalent when there exists invertible matrices $\boldsymbol{P} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{J\times J}(\mathsf{F})$ such that $\boldsymbol{A}_2 = \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{Q}$. Note that, by Theorem 5.4.32, this is exactly the sort of relationship that relates matrix representatives of the same linear map relative to different choices of basis. In this section we flesh out the implications of this observation in a manner analogous to that of Section 5.1.6.

To begin this process, we begin with a definition of equivalence for linear maps.

5.4.38 Definition (Equivalence of linear maps) Let F be a field and let U and V be F-vector spaces. Maps $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ are ***equivalent*** if there exists bases $\mathscr{B}_{\mathsf{U},1}$ and $\mathscr{B}_{\mathsf{U},2}$ for U, and bases $\mathscr{B}_{\mathsf{V},1}$ and $\mathscr{B}_{\mathsf{V},2}$ for V such that $[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V},1}}_{\mathscr{B}_{\mathsf{U},1}} = [\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V},2}}_{\mathscr{B}_{\mathsf{U},2}}$. •

We note that for *endomorphisms* there is a more natural notion of equivalence, called "similarity," which we discuss in Section 5.8.1.

The following result now relates equivalence of linear maps and equivalence of matrices.

5.4.39 Proposition (Equivalent linear maps and equivalent matrices) *Let F be a field, let U and V be F-vector spaces, and let $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ be bases for U and V, respectively. Then the following statements are equivalent:*

(i) the linear maps $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ are equivalent;

(ii) the matrices $[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}$ and $[\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}$ are equivalent;

(iii) there exists invertible endomorphisms $\mathsf{P} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ and $\mathsf{Q} \in \mathrm{End}_{\mathsf{F}}(\mathsf{U})$ such that $\mathsf{L}_2 = \mathsf{P} \circ \mathsf{L}_1 \circ \mathsf{Q}$.

Proof The equivalence of part (iii) with the other parts is Exercise 5.5.4. Thus we shall only prove the equivalence of (i) and (ii).

First suppose that L_1 and L_2 are equivalent and let $\mathscr{B}'_{\mathsf{U}}$ and $\mathscr{B}''_{\mathsf{U}}$ be bases for U, and let $\mathscr{B}'_{\mathsf{V}}$ and $\mathscr{B}''_{\mathsf{V}}$ be bases for V such that $[\mathsf{L}_1]^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{U}}} = [\mathsf{L}_2]^{\mathscr{B}''_{\mathsf{V}}}_{\mathscr{B}''_{\mathsf{U}}}$. By Theorem 5.4.32 we have

$$[\mathsf{L}_1]^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{U}}} = \boldsymbol{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(\boldsymbol{P}^{\mathscr{B}'_{\mathsf{U}}}_{\mathscr{B}_{\mathsf{U}}})^{-1}, \quad [\mathsf{L}_2]^{\mathscr{B}''_{\mathsf{V}}}_{\mathscr{B}''_{\mathsf{U}}} = \boldsymbol{P}^{\mathscr{B}''_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}[\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(\boldsymbol{P}^{\mathscr{B}''_{\mathsf{U}}}_{\mathscr{B}_{\mathsf{U}}})^{-1}.$$

This immediately gives

$$[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}} = \left(\boldsymbol{P}^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{V}}}\boldsymbol{P}^{\mathscr{B}''_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}\right)[\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}\left(\boldsymbol{P}^{\mathscr{B}_{\mathsf{U}}}_{\mathscr{B}''_{\mathsf{U}}}\boldsymbol{P}^{\mathscr{B}'_{\mathsf{U}}}_{\mathscr{B}_{\mathsf{U}}}\right)$$

which gives the equivalence of $[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}$ and $[\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}$ since the product of invertible matrices is invertible by Proposition 5.1.24.

Now suppose that $[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}$ and $[\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}$ are equivalent. Let I and J be the sets for which there exist bijections $\phi_{\mathsf{U}}\colon J \to \mathscr{B}_{\mathsf{U}}$ and $\phi_{\mathsf{V}}\colon I \to \mathscr{B}_{\mathsf{V}}$, and let $\boldsymbol{P} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{J\times J}(\mathsf{F})$ be such that $[\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}} = \boldsymbol{P}[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}\boldsymbol{Q}$. By Proposition 5.4.30 let $\mathscr{B}'_{\mathsf{U}}$ and $\mathscr{B}'_{\mathsf{V}}$ be bases for U and V, respectively, such that

$$\boldsymbol{P} = \boldsymbol{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}, \quad \boldsymbol{Q}^{-1} = \boldsymbol{P}^{\mathscr{B}'_{\mathsf{U}}}_{\mathscr{B}_{\mathsf{U}}}.$$

Then

$$[\mathsf{L}_1]^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}'_{\mathsf{U}}} = \boldsymbol{P}^{\mathscr{B}'_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{V}}}[\mathsf{L}_1]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(\boldsymbol{P}^{\mathscr{B}'_{\mathsf{U}}}_{\mathscr{B}_{\mathsf{U}}})^{-1} = [\mathsf{L}_2]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}},$$

meaning that L_1 and L_2 are equivalent. ■

There is a simple way to represent equivalence as per part (iii) of the proposition:

$$\begin{tikzcd} \mathsf{U} \arrow[r, "\mathsf{L}_1"] \arrow[d, "\mathsf{Q}"'] & \mathsf{V} \arrow[d, "\mathsf{P}"] \\ \mathsf{U} \arrow[r, "\mathsf{L}_2"'] & \mathsf{V} \end{tikzcd}$$

The intuition is that L_2 is the same as L_2 if one looks at things in the right way.

Another manner of interpreting the relationship between equivalent linear maps and equivalent matrices is the following.

5.4.40 Corollary (Equivalent linear maps and equivalent matrices) *Let F be a field, let U and V be F-vector spaces, let $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ be bases for U and V, respectively, and let I and J be index sets such that there exist bijections $\phi_{\mathsf{U}}\colon \mathrm{J} \to \mathscr{B}_{\mathsf{U}}$ and $\phi_{\mathsf{V}}\colon \mathrm{I} \to \mathscr{B}_{\mathsf{V}}$. Then the F-isomorphism of $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ and $\mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ given in Theorem 5.4.21 maps equivalence classes in $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ (with equivalence being as in Definition 5.4.38) to equivalence classes in $\mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ (with equivalence being as in Definition 5.1.38).*

In Theorem 5.1.41(iii) we gave a simple representative for each equivalence class of matrices in the case where there were finitely many rows and columns. To prove this equivalence we use row and column operations. Let us prove a result for equivalence of general linear maps that uses a more geometrically insightful proof.

5.4.41 Theorem (A simple representative for equivalence classes of linear maps) *Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$. Then there exists*

(i) bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ for U and V, respectively, and

(ii) sets I, J, and K with bijections $\phi_{\mathsf{U}}\colon \mathrm{J} \to \mathscr{B}_{\mathsf{U}}$ and $\phi_{\mathsf{V}}\colon \mathrm{I} \to \mathscr{B}_{\mathsf{V}}$

such that the following properties hold:

(iii) $\mathrm{J} = \mathrm{K} \cup \mathrm{J}'$ with $\mathrm{K} \cap \mathrm{J}' = \varnothing$;

(iv) $\mathrm{I} = \mathrm{K} \cup \mathrm{I}'$ with $\mathrm{K} \cap \mathrm{I}' = \varnothing$;

(v) the matrix representative of L relative to the bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ satisfies

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}}(\mathrm{i},\mathrm{j}) = \begin{cases} 1_{\mathsf{F}}, & \mathrm{i},\mathrm{j} \in \mathrm{K},\ \mathrm{i} = \mathrm{j}, \\ 0_{\mathsf{F}}, & \textit{otherwise}. \end{cases}$$

Proof Let $\mathscr{B}''_{\mathsf{U}}$ be a basis for $\ker(\mathsf{L})$ and, by Theorem 4.5.26, let $\mathscr{B}'_{\mathsf{U}}$ be a linearly independent set such that $\mathscr{B}_{\mathsf{U}} = \mathscr{B}'_{\mathsf{U}} \cup \mathscr{B}''_{\mathsf{U}}$ is a basis for U. Let K and J' be sets such that there exist bijections $\phi'_{\mathsf{U}}\colon K \to \mathscr{B}'_{\mathsf{U}}$ and $\phi''_{\mathsf{U}}\colon J' \to \mathscr{B}''_{\mathsf{U}}$. Define $J = K \overset{\circ}{\cup} J'$ and define $\phi_{\mathsf{U}}\colon J \to \mathscr{B}_{\mathsf{U}}$ by

$$\phi_{\mathsf{U}}(j) = \begin{cases} \phi'_{\mathsf{U}}(j), & j \in K, \\ \phi''_{\mathsf{U}}(j), & j \in J'. \end{cases}$$

Let $\mathsf{U}' = \mathrm{span}_{\mathsf{F}}(\mathscr{B}'_{\mathsf{U}})$ so that $\mathsf{U} = \mathsf{U}' \oplus \ker(\mathsf{L})$.

Define $\mathscr{B}'_{\mathsf{V}} = \{\mathsf{L}(u) \mid u \in \mathscr{B}'_{\mathsf{U}}\}$. We claim that $\mathscr{B}'_{\mathsf{V}}$ is a basis for image(L). To see this, let $v \in \text{image}(\mathsf{L})$ and write $v = \mathsf{L}(u)$ for some $u \in \mathsf{U}$. We can then write $u = u' + u''$ for unique $u' \in \mathsf{U}'$ and $u'' \in \ker(\mathsf{L})$. We then have $\mathsf{L}(u) = \mathsf{L}(u')$. Since u' is a finite linear combination of elements of $\mathscr{B}'_{\mathsf{U}}$, $v = \mathsf{L}(u')$ is a finite linear combination of elements of $\mathscr{B}'_{\mathsf{V}}$ by linearity of L. Thus $\text{image}(\mathsf{L}) = \text{span}_{\mathsf{F}}(\mathscr{B}'_{\mathsf{V}})$. Now let $v_1, \dots, v_k \in \mathscr{B}'_{\mathsf{V}}$ and write $v_j = \mathsf{L}(u_j)$ for $u_j \in \mathscr{B}'_{\mathsf{U}}$, $j \in \{1, \dots, k\}$. Then, if $c_1 v_1 + \dots + c_k v_k = 0_{\mathsf{V}}$, it follows by linearity of L that

$$c_1 \mathsf{L}(u_1) + \dots + c_k \mathsf{L}(u_k) = \mathsf{L}(c_1 u_1 + \dots + c_k u_k) = 0_{\mathsf{V}}.$$

Thus $c_1 u_1 + \dots + c_k u_k \in \ker(\mathsf{L})$, implying that $c_1 u_1 + \dots + c_k u_k = 0_{\mathsf{U}}$ by Proposition 4.5.37. Therefore, $c_1 = \dots = c_k = 0_{\mathsf{F}}$, giving linear independence of $\mathscr{B}'_{\mathsf{V}}$. Thus $\mathscr{B}'_{\mathsf{V}}$ is a basis for image(L) as claimed.

Using Theorem 4.5.26 let $\mathscr{B}''_{\mathsf{V}} \subseteq \mathsf{V}$ be a linearly set such that $\mathscr{B}_{\mathsf{V}} = \mathscr{B}'_{\mathsf{V}} \cup \mathscr{B}''_{\mathsf{V}}$ is a basis for V. Since L is a bijection from $\mathscr{B}'_{\mathsf{U}}$ to $\mathscr{B}'_{\mathsf{V}}$, there exists a bijection $\phi'_{\mathsf{U}} \colon K \to \mathscr{B}'_{\mathsf{U}}$. Let I' be a set such that there exists a bijection $\phi''_{\mathsf{V}} \colon I' \to \mathscr{B}''_{\mathsf{V}}$ and take $I = K \overset{\circ}{\cup} I'$. Then define a bijection $\phi_{\mathsf{V}} \colon I \to \mathscr{B}_{\mathsf{V}}$ by

$$\phi_{\mathsf{V}}(i) = \begin{cases} \phi'_{\mathsf{V}}(i), & i \in K, \\ \phi''_{\mathsf{V}}(i), & i \in I'. \end{cases}$$

We now claim that, with respect to the bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$, the matrix representative of L is as claimed. First, if $j \in K$ and if $i \in I$ then we have

$$\mathsf{L}(\phi_{\mathsf{U}}(j)) = \phi_{\mathsf{V}}(j),$$

from which we immediately deduce that

$$[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(i, j) = \begin{cases} 1_{\mathsf{F}}, & i, j \in K,\ i = j, \\ 0_{\mathsf{F}}, & \text{otherwise} \end{cases}$$

in this case. If $j \in J'$, since $\phi_{\mathsf{U}}(j) \in \ker(\mathsf{L})$, it follows that $[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}}(i, j) = 0_{\mathsf{F}}$ for any $i \in I$. This gives the desired matrix representative. ■

After understanding the words and symbols, the theorem says that one can find bases for U and V that are each partitioned into two subsets, and such that the resulting matrix representative for L is partitioned as follows:

$$[\mathsf{L}]^{\mathscr{B}_{\mathsf{V}}}_{\mathscr{B}_{\mathsf{U}}} = \begin{bmatrix} \boldsymbol{I}_K & \boldsymbol{0}_{K \times J'} \\ \boldsymbol{0}_{I' \times K} & \boldsymbol{0}_{I' \times J'} \end{bmatrix}. \tag{5.23}$$

We may now indicate a simple characterisation of the equivalence classes in $\text{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$.

5.4.42 Corollary (Characterisation of equivalent linear maps) *Let* F *be a field, and let* U *and* V *be* F*-vector space. For linear maps* $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$*, the following statements hold:*

(i) L_1 *and* L_2 *are equivalent if and only if* $\mathrm{rank}(\mathsf{L}_1) = \mathrm{rank}(\mathsf{L}_2)$*,* $\mathrm{nullity}(\mathsf{L}_1) = \mathrm{nullity}(\mathsf{L}_2)$*, and* $\mathrm{defect}(\mathsf{L}_1) = \mathrm{defect}(\mathsf{L}_2)$*;*

(ii) if U *is finite-dimensional, then* L_1 *and* L_2 *are equivalent if and only if* $\mathrm{rank}(\mathsf{L}_1) = \mathrm{rank}(\mathsf{L}_2)$*;*

(iii) if V *is finite-dimensional, then* L_1 *and* L_2 *are equivalent if and only if* $\mathrm{rank}(\mathsf{L}_1) = \mathrm{rank}(\mathsf{L}_2)$ *and* $\mathrm{nullity}(\mathsf{L}_1) = \mathrm{nullity}(\mathsf{L}_2)$.

Proof (i) Suppose that L_1 and L_2 are equivalent. By the proof of Theorem 5.4.41 it is possible to find bases $\mathscr{B}_{\mathsf{U},1}$ and $\mathscr{B}_{\mathsf{U},2}$ for U and bases $\mathscr{B}_{\mathsf{V},1}$ and $\mathscr{B}_{\mathsf{V},2}$ for V such that

$$[\mathsf{L}_1]_{\mathscr{B}_{\mathsf{U},1}}^{\mathscr{B}_{\mathsf{V},1}} = \begin{bmatrix} \boldsymbol{I}_{K_1} & \boldsymbol{0}_{K_1\times J_1'} \\ \boldsymbol{0}_{I_1'\times K_1} & \boldsymbol{0}_{I_1'\times J_1'} \end{bmatrix}, \qquad [\mathsf{L}_2]_{\mathscr{B}_{\mathsf{U},2}}^{\mathscr{B}_{\mathsf{V},2}} = \begin{bmatrix} \boldsymbol{I}_{K_2} & \boldsymbol{0}_{K_2\times J_2'} \\ \boldsymbol{0}_{I_2'\times K_2} & \boldsymbol{0}_{I_2'\times J_2'} \end{bmatrix},$$

where $\mathrm{card}(K_a) = \mathrm{rank}(\mathsf{L}_a)$, $\mathrm{card}(J_a') = \mathrm{nullity}(\mathsf{L}_a)$, and $\mathrm{card}(I_a') = \mathrm{defect}(\mathsf{L}_a)$ for $a \in \{1, 2\}$. Moreover, by Proposition 5.4.39, the matrices $[\mathsf{L}_1]_{\mathscr{B}_{\mathsf{U},1}}^{\mathscr{B}_{\mathsf{V},1}}$ and $[\mathsf{L}_2]_{\mathscr{B}_{\mathsf{U},2}}^{\mathscr{B}_{\mathsf{V},2}}$ are equivalent. The following lemma is then useful.

1 Lemma *Let* F *be a field and let* I *and* J *be index sets. If* $\mathbf{A}_1, \mathbf{A}_2 \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$ *are equivalent then* $\mathrm{rank}(\mathbf{A}_1) = \mathrm{rank}(\mathbf{A}_2)$*,* $\mathrm{nullity}(\mathbf{A}_1) = \mathrm{nullity}(\mathbf{A}_2)$*, and* $\mathrm{defect}(\mathbf{A}_1) = \mathrm{defect}(\mathbf{A}_2)$.

Proof Let $\boldsymbol{P} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{J\times J}(\mathsf{F})$ be invertible matrices such that $\boldsymbol{A}_2 = \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{Q}$.

We claim that the map

$$\mathrm{image}(\boldsymbol{A}_1) \ni \boldsymbol{y} \mapsto \boldsymbol{P}\boldsymbol{y} \in \mathsf{F}_0^I$$

is a bijection onto $\mathrm{image}(\boldsymbol{A}_2)$. Indeed, let $\boldsymbol{y} \in \mathrm{image}(\boldsymbol{A}_1)$ and suppose that $\boldsymbol{y} = \boldsymbol{A}_1\boldsymbol{x}$. Then

$$\boldsymbol{y} = \boldsymbol{P}^{-1}\boldsymbol{A}_2\boldsymbol{Q}\boldsymbol{x} \quad \Longrightarrow \quad \boldsymbol{P}\boldsymbol{y} = \boldsymbol{A}_2\boldsymbol{Q}\boldsymbol{x},$$

showing that $\boldsymbol{P}\boldsymbol{y} \in \mathrm{image}(\boldsymbol{A}_2)$. Swapping the rôles of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ shows that if $\boldsymbol{y} \in \mathrm{image}(\boldsymbol{A}_2)$ then $\boldsymbol{P}^{-1}\boldsymbol{y} \in \mathrm{image}(\boldsymbol{A}_1)$. Then $\boldsymbol{P}$ maps $\mathrm{image}(\boldsymbol{A}_1)$ bijectively onto $\mathrm{image}(\boldsymbol{A}_2)$, and so $\mathrm{rank}(\boldsymbol{A}_1) = \mathrm{rank}(\boldsymbol{A}_2)$.

We next claim that the map

$$\ker(\boldsymbol{A}_2) \ni \boldsymbol{x} \mapsto \boldsymbol{Q}\boldsymbol{x} \in \mathsf{F}_0^I$$

is a bijection onto $\ker(\boldsymbol{A}_1)$. Indeed, if $\boldsymbol{x} \in \ker(\boldsymbol{A}_2$ then

$$\boldsymbol{A}_2\boldsymbol{x} = \boldsymbol{0}_{\mathsf{F}_0^I} \quad \Longrightarrow \quad \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{0}_{\mathsf{F}_0^I}.$$

Since $\boldsymbol{P}$ is invertible, by Exercise 4.5.23 it follows that $\boldsymbol{A}_1\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{0}_{\mathsf{F}_0^I}$, or that $\boldsymbol{Q}\boldsymbol{x} \in \ker(\boldsymbol{A}_1)$. An entirely similar computation, but reversing the rôles of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$, shows that if $\boldsymbol{x} \in \ker(\boldsymbol{A}_1)$ then $\boldsymbol{Q}^{-1}\boldsymbol{x} \in \ker(\boldsymbol{A}_2)$. Thus $\boldsymbol{Q}$ does indeed map $\ker(\boldsymbol{A}_2)$ bijectively onto $\ker(\boldsymbol{A}_1)$. Thus $\mathrm{nullity}(\boldsymbol{A}_1) = \mathrm{nullity}(\boldsymbol{A}_2)$.

Finally, we claim that the linear map

$$\mathsf{F}_0^I/\operatorname{image}(\boldsymbol{A}_1) \ni \boldsymbol{y} + \operatorname{image}(\boldsymbol{A}_1) \mapsto \boldsymbol{P}\boldsymbol{y} + \operatorname{image}(\boldsymbol{A}_2) \in \mathsf{F}_0^I/\operatorname{image}(\boldsymbol{A}_2) \tag{5.24}$$

is well-defined and a bijection. To check that the map is well-defined, suppose that $\boldsymbol{y}_1 + \operatorname{image}(\boldsymbol{A}_1) = \boldsymbol{y}_2 + \operatorname{image}(\boldsymbol{A}_1)$. Then there exists $x \in \operatorname{image}(\boldsymbol{A}_1)$ such that $\boldsymbol{y}_2 = \boldsymbol{y}_1 + \boldsymbol{z}$. By the first part of the proof, $\boldsymbol{P}\boldsymbol{z} \in \operatorname{image}(\boldsymbol{A}_2)$, and so $\boldsymbol{P}\boldsymbol{y}_1 + \operatorname{image}(\boldsymbol{A}_2) = \boldsymbol{P}\boldsymbol{y}_2 + \operatorname{image}(\boldsymbol{A}_2)$. To see that the map in (5.24) is injective, suppose that $\boldsymbol{P}\boldsymbol{y} + \operatorname{image}(\boldsymbol{A}_2) = \boldsymbol{0}_{\mathsf{F}_0^I} + \operatorname{image}(\boldsymbol{A}_2)$. Thus $\boldsymbol{P}\boldsymbol{y} \in \operatorname{image}(\boldsymbol{A}_2)$ and so, by the first part of the proof, $\boldsymbol{y} \in \operatorname{image}(\boldsymbol{A}_1)$. Thus $\boldsymbol{y} + \operatorname{image}(\boldsymbol{A}_1) = \boldsymbol{0}_{\mathsf{F}_0^I}$, giving injectivity of the map in (5.24). To show surjectivity, let $\boldsymbol{y} + \operatorname{image}(\boldsymbol{A}_2) \in \mathsf{F}_0^I/\operatorname{image}(\boldsymbol{A}_2)$. One can then see that clearly this vector is the image of $\boldsymbol{P}^{-1}\boldsymbol{y} + \operatorname{image}(\boldsymbol{A}_1)$ under the map (5.24). Thus the map (5.24) is indeed a bijection, and so it follows that $\operatorname{defect}(\boldsymbol{A}_1) = \operatorname{defect}(\boldsymbol{A}_2)$. ▼

From the lemma and the fact that the rank, nullity, and defect of a linear map agree with the rank, nullity, and defect, respectively, of its matrix representative (why?), this part of part (i) follows.

Conversely, if the ranks, nullities, and defects of L_1 and L_2 agree, then, by the proof of Theorem 5.4.41, both possess matrix representatives of the form (5.23) with the sets K, I', and J' having equal cardinalities for both L_1 and L_2. Thus the matrix representatives are equivalent, and then so too are the linear maps by Proposition 5.4.39.

(ii) By the Rank–Nullity Formula, $\operatorname{card}(J'_a) = \dim_{\mathsf{F}}(\mathsf{U}) - \operatorname{card}(K_a)$ for $a \in \{1,2\}$. Thus $\operatorname{nullity}(\mathsf{L}_1) = \operatorname{nullity}(\mathsf{L}_2)$ if and only if $\operatorname{rank}(\mathsf{L}_1) = \operatorname{rank}(\mathsf{L}_2)$. Moreover, by Theorem 4.5.56, $\operatorname{card}(I'_a) = \dim_{\mathsf{F}}(\mathsf{V}) - \operatorname{card}(K_a)$. Since K_a is finite, it follows from Theorem 1.7.17 that $\operatorname{card}(I'_a) = \dim_{\mathsf{F}}(\mathsf{V})$ in the case when $\dim_{\mathsf{F}}(\mathsf{V})$ is infinite. Therefore, in either of the cases when $\dim_{\mathsf{F}}(\mathsf{V})$ is finite or infinite we have $\operatorname{defect}(\mathsf{L}_1) = \operatorname{defect}(\mathsf{L}_2)$. Thus equality of the ranks in this case implies equality of the nullities and the defects.

(iii) By Theorem 4.5.56, $\operatorname{card}(I'_a) = \dim_{\mathsf{F}}(\mathsf{V}) - \operatorname{card}(K_a)$ for $a \in \{1,1\}$. Thus $\operatorname{defect}(\mathsf{L}_1) = \operatorname{defect}(\mathsf{L}_2)$ if and only if $\operatorname{rank}(\mathsf{L}_1) = \operatorname{rank}(\mathsf{L}_2)$. ■

The lemma in the preceding proof immediately gives the following generalisation of Theorem 5.1.41.

5.4.43 Corollary (Equivalence of general matrices over fields) *Let* F *be a field, let* I *and* J *be index sets, and let* $\mathbf{A}_1, \mathbf{A}_2 \in \operatorname{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{F})$*. Then the following statements are equivalent:*

(i) $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are equivalent;*

(ii) $\operatorname{rank}(\mathbf{A}_1) = \operatorname{rank}(\mathbf{A}_2)$, $\operatorname{nullity}(\mathbf{A}_1) = \operatorname{nullity}(\mathbf{A}_2)$, *and* $\operatorname{defect}(\mathbf{A}_1) = \operatorname{defect}(\mathbf{A}_2)$*;*

(iii) there exists sets K, I', *and* J' *such that*

(a) $\mathrm{I} = \mathrm{K} \cup \mathrm{I}'$ *and* $\mathrm{K} \cap \mathrm{I}' = \emptyset$,

(b) $\mathrm{J} = \mathrm{K} \cup \mathrm{J}'$ *and* $\mathrm{K} \cap \mathrm{J}' = \emptyset$, *and*

(c) the matrices $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are equivalent to a matrix of the form (5.23).*

For endomorphisms of finite-dimensional vector spaces, our development of equivalence, along with results from Section 5.1.6, yields the following characterisation of invertible maps.

5.4.44 Corollary (Characterisation of invertible endomorphisms of finite-dimensional vector spaces) *Let* F *be a field and let* V *be a finite-dimensional* F*-vector space. For* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*, the following statements are equivalent:*

(i) L *is invertible;*

(ii) $\mathrm{rank}(\mathsf{L}) = \dim_{\mathsf{F}}(\mathsf{V})$*;*

(iii) $\det \mathsf{L} \neq 0_{\mathsf{F}}$*;*

(iv) L *is equivalent to* id_{V}*;*

(v) L *possesses a left inverse;*

(vi) L *possesses a unique left inverse;*

(vii) L *possesses a right inverse;*

(viii) L *possesses a unique right inverse;*

(ix) L *is injective;*

(x) L *in surjective.*

Proof This follows from Proposition 5.4.39, along with Theorem 5.1.42, Corollary 5.1.43 and Theorem 5.3.10. ■

As is shown in Example 5.1.44, the preceding corollary is not generally true when V is infinite-dimensional.

It is worth taking a moment to understand why, in infinite dimensions, one needs equality of not just rank, but of nullity and defect to characterise equivalence of linear maps. The reason has to do with the arithmetic of infinite cardinal numbers as evidenced by Theorem 1.7.17, for example. Particularly, it is possible to add two nonzero cardinal numbers and not arrive at a larger cardinal number. This may seem strange, but it arises in a straightforward way as the following examples show.

5.4.45 Examples (Equivalent and inequivalent linear maps)

1. Let us take $I = \mathbb{Z}_{>0}$ so that, in the notation of Example 4.5.2–4, $\mathsf{F}_0^I = \mathsf{F}_0^\infty$. We take $\mathsf{U} = \mathsf{V} = \mathsf{F}_0^\infty$ and recall from Theorem 5.1.13 that linear maps from U to V can be regarded as column finite matrices with a countably infinite number of rows and columns. Indeed, we shall represent our linear maps exactly in this way. With this in mind we define $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ by

$$\mathsf{L}_1 = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathsf{L}_2 = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Let us denote by $\{e_j\}_{j\in\mathbb{Z}_{>0}}$ the standard basis for F_0^∞, and make the following observations about the linear maps L_1 and L_2:

(a) $\ker(\mathsf{L}_1) = \operatorname{span}_{\mathsf{F}}(\boldsymbol{e}_j|\ j \text{ even})$;

(b) $\ker(\mathsf{L}_1) = \operatorname{span}_{\mathsf{F}}(\boldsymbol{e}_j|\ j \text{ odd})$;

(c) $\operatorname{image}(\mathsf{L}_1) = \operatorname{span}_{\mathsf{F}}(\boldsymbol{e}_j|\ j \text{ odd})$;

(d) $\operatorname{image}(\mathsf{L}_2) = \operatorname{span}_{\mathsf{F}}(\boldsymbol{e}_j|\ j \text{ even})$.

Therefore,

$$\begin{aligned}\operatorname{rank}(\mathsf{L}_1) &= \operatorname{rank}(\mathsf{L}_2) = \operatorname{card}(\mathbb{Z}_{>0}),\\ \operatorname{defect}(\mathsf{L}_1) &= \operatorname{defect}(\mathsf{L}_2) = \operatorname{card}(\mathbb{Z}_{>0}),\\ \operatorname{nullity}(\mathsf{L}_1) &= \operatorname{nullity}(\mathsf{L}_2) = \operatorname{card}(\mathbb{Z}_{>0}).\end{aligned}$$

Therefore, L_1 and L_2 are equivalent. Indeed, one can check that if we define

$$\boldsymbol{P} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

then $\boldsymbol{A}_2 = \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{P}^{-1}$.

2. We take $\mathsf{U} = \mathsf{V} = \mathsf{F}_0^\infty$. Again, let us simply define a linear map by writing the matrix associated to it. Thus we define $\mathsf{L}_1, \mathsf{L}_2 \in \operatorname{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ by

$$\mathsf{L}_1 = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad \mathsf{L}_2 = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Denoting by $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}}$ the standard basis for F_0^∞, we make the following observations about the linear maps L_1 and L_2:

(a) $\ker(\mathsf{L}_1) = \ker(\mathsf{L}_2) = \{\boldsymbol{0}_{\mathsf{F}_0^\infty}\}$;

(b) $\operatorname{image}(\mathsf{L}_1) = \mathsf{V}$;

(c) $\operatorname{image}(\mathsf{L}_2) = \operatorname{span}_{\mathsf{F}}(\boldsymbol{e}_j|\ j \in \mathbb{Z}_{>0} \setminus \{1\})$.

In particular this shows that

$$\begin{aligned}\operatorname{rank}(\mathsf{L}_1) = \operatorname{rank}(\mathsf{L}_2) &= \operatorname{card}(\mathbb{Z}_{>0}),\\ 0 = \operatorname{defect}(\mathsf{L}_1) < \operatorname{defect}(\mathsf{L}_2) &= 1,\\ \operatorname{nullity}(\mathsf{L}_1) = \operatorname{nullity}(\mathsf{L}_2) &= 0.\end{aligned}$$

Thus while the ranks and nullities of these two linear maps agree, their defects do not, and so they are not equivalent.

3. We take $\mathsf{U} = \mathsf{F}_0^\infty$ and $\mathsf{V} = \mathsf{F}^2$. Thus a linear map from U to V is represented by a matrix with two rows and a countably infinite number of columns. We define $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ by

$$\mathsf{L}_1 = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \end{bmatrix}, \quad \mathsf{L}_2 = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \end{bmatrix}.$$

Let us denote by $\{\boldsymbol{e}_j\}_{j\in\mathbb{Z}_{>0}}$ and $\{\boldsymbol{f}_1, \boldsymbol{f}_2\}$ the standard bases for U and V. About these two linear maps we make the following observations:

(a) $\mathrm{image}(\mathsf{L}_1) = \mathrm{image}(\mathsf{L}_2) = \mathsf{V}$;

(b) $\ker(\mathsf{L}_1) = \{\boldsymbol{0}_{\mathsf{F}_0^\infty}\}$;

(c) $\ker(\mathsf{L}_2) = \{\boldsymbol{e}_1\}$.

In particular this shows that

$$\begin{aligned} \mathrm{rank}(\mathsf{L}_1) &= \mathrm{rank}(\mathsf{L}_2) = 2, \\ \mathrm{defect}(\mathsf{L}_1) &= \mathrm{defect}(\mathsf{L}_2) = 0, \\ 0 = \mathrm{nullity}(\mathsf{L}_1) &< \mathrm{nullity}(\mathsf{L}_2) = 1. \end{aligned}$$

Thus while the ranks and defects of these two linear maps agree, their nullities do not, and so they are not equivalent. •

We close this section by using our characterisations of equivalence to assert the existence of linear left- and right-inverses for injective and surjective linear maps. While the existence of left- and right-inverse, not necessarily linear, follows from the general Proposition 1.3.9, the following result further asserts that these can be chosen to be linear.

5.4.46 Proposition (Linear left and right inverses) *Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$. Then the following statements hold:*

(i) L is injective if and only if there exists $\mathsf{K}_{\mathrm{L}} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V})$ such that $\mathsf{K}_{\mathrm{L}} \circ \mathsf{L} = \mathrm{id}_{\mathsf{U}}$;

(ii) L is surjective if and only if there exists $\mathsf{K}_{\mathrm{R}} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V})$ such that $\mathsf{L} \circ \mathsf{K}_{\mathrm{R}} = \mathrm{id}_{\mathsf{V}}$.

Proof (i) By choosing the bases $\mathscr{B}_{\mathsf{U}}$ and $\mathscr{B}_{\mathsf{V}}$ of Theorem 5.4.41, the injectivity of L ensures that the corresponding matrix representative of L will have the form

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}} = \begin{bmatrix} \boldsymbol{I}_K \\ \boldsymbol{0}_{I'\times K} \end{bmatrix}.$$

This matrix possesses a left-inverse

$$\begin{bmatrix} \boldsymbol{I}_K & \boldsymbol{0}_{K\times I'} \end{bmatrix},$$

and this part of the result follows by taking K_L such that the preceding matrix is equal to $[\mathsf{K}_L]_{\mathscr{B}_{\mathsf{V}}}^{\mathscr{B}_{\mathsf{U}}}$.

(ii) We again write the matrix representative of L using the bases from Theorem 5.4.41, and in this case the surjectivity of L ensures that

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{U}}}^{\mathscr{B}_{\mathsf{V}}} = \begin{bmatrix} \boldsymbol{I}_K & \mathbf{0}_{K\times J'} \end{bmatrix}.$$

This matrix has a right-inverse

$$\begin{bmatrix} \boldsymbol{I}_K \\ \mathbf{0}_{J'\times K} \end{bmatrix},$$

and this part of the result follows by taking K_R such that the preceding matrix is equal to $[\mathsf{K}_R]_{\mathscr{B}_{\mathsf{V}}}^{\mathscr{B}_{\mathsf{U}}}$. ■

5.4.8 Linear equations in vector spaces

Next we turn to the topic of linear equations in vector spaces, these being a generalisation of systems of linear equations as discussed in Section 5.1.8. Much of what we say here is a direct consequence of the results in Section 5.1.8, but we offer complete proofs in any case, to emphasise the geometric flavour of linear maps as opposed to the computational flavour of matrices.

We begin with a definition of the object of interest.

5.4.47 Definition (Linear equation) Let F be a field and let U and V be F-vector spaces.

(i) A ***linear equation*** is a pair $(\mathsf{L}, b) \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V}) \times \mathsf{V}$.

(ii) A linear equation (L, v_0) is ***homogeneous*** if $v_0 = 0_{\mathsf{V}}$.

(iii) The ***solution set*** for a linear equation (L, v_0) is the subset of U defined by

$$\mathrm{Sol}(\mathsf{L}, v_0) = \{u \in \mathsf{U} \mid \mathsf{L}(u) = v_0\}.$$

A ***solution*** of the linear equation (L, v_0) is an element of the solution set.

(iv) For a linear equation (L, v_0), the ***augmented linear map*** for the equation is the element $[\mathsf{L}, v_0] \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U} \oplus \mathsf{F};\mathsf{V})$ defined by

$$[\mathsf{L}, v_0](u, a) = \mathsf{L}(u) + a v_0.$$ •

Let us record the basic result addressing the matter of existence and uniqueness of solutions to linear equations.

5.4.48 Proposition (Existence and uniqueness of solutions) *Let F be a field, let U and V be F-vector spaces, and let $(\mathsf{L}, \mathrm{v}_0) \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V}) \times \mathsf{V}$ be a linear equation. Then the following statements hold:*

(i) $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is nonempty if and only if* $\mathrm{v}_0 \in \mathrm{image}(\mathsf{L})$*;*

(ii) in particular, $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is nonempty for every* $\mathrm{v}_0 \in \mathsf{V}$ *if and only if* L *is surjective;*

(iii) $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is a singleton if and only if*

(a) $\mathrm{v}_0 \in \mathrm{image}(\mathsf{L})$ *and*

(b) L *is injective.*

Proof We only prove the final assertion since the first two are obvious. Suppose that $\mathrm{Sol}(\mathsf{L}, v_0)$ is a singleton. By the first part of the result, $v_0 \in \mathrm{image}(\mathsf{L})$. If L is not injective then, by Exercise 4.5.23, there exists $u \in \mathsf{U} \setminus \{0_{\mathsf{U}}\}$ such that $\mathsf{L}(u) = 0_{\mathsf{V}}$. Now, if $u_0 \in \mathrm{Sol}(\mathsf{L}, v_0)$ then

$$\mathsf{L}(u + u_0) = \mathsf{L}(u) + \mathsf{L}(u_0) = 0_{\mathsf{V}} + v_0 = v_0,$$

showing that $u+u_0 \in \mathrm{Sol}(\mathsf{L}, v_0)$. Thus if L is not injective, $\mathrm{Sol}(\mathsf{L}, v_0)$ cannot be a singleton.

Conversely, suppose that $v_0 \in \mathrm{image}(\mathsf{L})$ and that L is injective. Then there exists $u_0 \in \mathsf{U}$ such that $\mathsf{L}(u_0) = v_0$. Moreover, if $u \in \mathrm{Sol}(\mathsf{L}, v_0)$ then

$$\mathsf{L}(u - u_0) = \mathsf{L}(u) - \mathsf{L}(u_0) = v_0 - v_0 = 0_{\mathsf{V}}.$$

Since L is injective, by Exercise 4.5.23 we have $u = u_0$, and so $\mathrm{Sol}(\mathsf{L}, v_0) = \{u_0\}$. ■

We may also characterise the set of solutions of a linear equation as an affine subspace, following Definition 4.5.13.

5.4.49 Proposition (Characterisation of $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$) *Let* F *be a field, let* U *and* V *be* F*-vector spaces, and let* $(\mathsf{L}, \mathrm{v}_0) \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V}) \times \mathsf{V}$ *be a linear equation. Then* $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$*, if it is nonempty, is an affine subspace of* U *whose linear part is* $\ker(\mathsf{L})$*.*

Proof We assume that $\mathrm{Sol}(\mathsf{L}, v_0)$ is nonempty and so let $u_0 \in \mathrm{Sol}(\mathsf{L}, v_0)$. Let us define

$$\mathsf{A}_{\mathsf{L},u_0} = \{u + u_0 \mid u \in \ker(\mathsf{L})\},$$

which is an affine subspace of U with linear part $\ker(\mathsf{L})$. If $u \in \mathsf{A}_{\mathsf{L},u_0}$ then $u = u_0 + u'$ for $u' \in \ker(\mathsf{L})$ and so

$$\mathsf{L}(u) = \mathsf{L}(u_0 + u') = \mathsf{L}(u_0) + \mathsf{L}(u') = v_0 + 0_{\mathsf{V}} = v_0,$$

so showing that $\mathsf{A}_{\mathsf{L},u_0} \subseteq \mathrm{Sol}(\mathsf{L}, v_0)$. Now suppose that $u \in \mathrm{Sol}(\mathsf{L}, v_0)$ and compute

$$\mathsf{L}(u - u_0) = \mathsf{L}(u) - \mathsf{L}(u_0) = v_0 - v_0 = 0_{\mathsf{V}},$$

so that $u = u_0 + u'$ for $u' \in \ker(\mathsf{L})$. Thus $\mathrm{Sol}(\mathsf{L}, v_0) \subseteq \mathsf{A}_{\mathsf{L},u_0}$. ■

The same comments concerning the determination of $\mathrm{Sol}(\mathsf{L}, v_0)$ in practice can be made as were made for systems of linear equations (see the paragraph following Proposition 5.1.56).

It is also possible to characterise the existence of solutions in terms of the augmented linear map. For matrices this was done in Corollary 5.1.59. Our result here for linear maps has a slightly different form, and we link the two results in a corollary following our next result.

5.4.50 Theorem (Existence and uniqueness of solutions, and the augmented linear map) *Let* F *be a field, let* U *and* V *be* F*-vector spaces, and let* $(\mathsf{L}, \mathrm{v}_0) \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V}) \times \mathsf{V}$ *be a linear equation. Then*

(i) $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is nonempty if and only if* $\mathrm{image}([\mathsf{L}, \mathrm{v}_0]) = \mathrm{image}(\mathsf{L})$*, and*

(ii) $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is a singleton if and only if*

  (a) $\mathrm{image}([\mathsf{L}, \mathrm{v}_0]) = \mathrm{image}(\mathsf{L})$ *and*

  (b) *there exists* $\mathrm{u}_0 \in \mathsf{U}$ *such that*

$$\ker([\mathsf{L}, \mathrm{v}_0]) = \{(\mathrm{u}, \mathrm{a}) \in \mathsf{U} \oplus \mathsf{F} \mid \mathrm{u} = \mathrm{a}\mathrm{u}_0\}.$$

Proof (i) Clearly we always have $\mathrm{image}(\mathsf{L}) \subseteq \mathrm{image}([\mathsf{L}, v_0])$. Thus we shall show that $\mathrm{Sol}(\mathsf{L}, v_0)$ is nonempty if and only if $\mathrm{image}([\mathsf{L}, v_0]) \subseteq \mathrm{image}(\mathsf{L})$. Suppose that $\mathrm{Sol}(\mathsf{L}, v_0)$ is nonempty, say $u_0 \in \mathrm{Sol}(\mathsf{L}, v_0)$, and let $v \in \mathrm{image}([\mathsf{L}, v_0])$. Thus there exists $u \in \mathsf{U}$ and $a \in \mathsf{F}$ such that

$$v = \mathsf{L}(u) + av_0 = \mathsf{L}(u) + a\mathsf{L}(u_0) = \mathsf{L}(u + au_0).$$

Thus $v \in \mathrm{image}(\mathsf{L})$. Conversely suppose that $\mathrm{image}([\mathsf{L}, v_0]) \subseteq \mathrm{image}(\mathsf{L})$. Since $v_0 \in \mathrm{image}(\mathsf{L})$ (we have $v_0 = \mathsf{L}(0_{\mathsf{U}}) + 1_{\mathsf{F}}v_0$), it immediately follows that $v_0 \in \mathrm{Sol}(\mathsf{L}, v_0)$, as desired.

(ii) By the first part of the theorem, if suffices to suppose that $\mathrm{Sol}(\mathsf{L}, v_0)$ is nonempty, and then prove that it is a singleton if and only if there exists $u_0 \in \mathsf{U}$ such that

$$\ker([\mathsf{L}, v_0]) = \{(u, a) \in \mathsf{U} \oplus \mathsf{F} \mid u = au_0\}.$$

First suppose that $\mathrm{Sol}(\mathsf{L}, v_0) = \{u_0\}$ for some $u_0 \in \mathsf{U}$. By Proposition 5.4.48 this implies that L is injective. Thus $(u, a) \in \ker([\mathsf{L}, v_0])$ if and only if

$$0_{\mathsf{V}} = [\mathsf{L}, v_0](u, a) = \mathsf{L}(u) + av_0 = \mathsf{L}(u) + a\mathsf{L}(u_0) = \mathsf{L}(u + au_0).$$

That is to say, by Exercise 4.5.23, $(u, a) \in \ker([\mathsf{L}, v_0])$ if and only if $u \in \mathrm{span}_{\mathsf{F}}(u_0)$.

Conversely, suppose that $u_0 \in \mathsf{U}$ has the property that

$$\ker([\mathsf{L}, v_0]) = \{(u, a) \in \mathsf{U} \oplus \mathsf{F} \mid u = au_0\}.$$

Then let $u \in \mathrm{Sol}(\mathsf{L}, v_0)$ and compute

$$\mathsf{L}(u) = v_0 \quad \Longrightarrow \quad (u, -1_{\mathsf{F}}) \in \ker([\mathsf{L}, v_0]) \quad \Longrightarrow \quad u = -u_0.$$

That is to say, $\mathrm{Sol}(\mathsf{L}, v_0) = \{-u_0\}$. ■

The next corollary follows mirrors Corollary 5.1.59, but we give here a proof that does not rely on row reduction.

5.4.51 Corollary (Rank, and existence and uniqueness of solutions) *Let* F *be a field, let* U *and* V *be* F*-vector spaces, and let* $(\mathsf{L}, \mathrm{v}_0) \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V}) \times \mathsf{V}$ *be a linear equation with* $\mathrm{rank}(\mathsf{L})$ *finite. Then*

(i) $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is nonempty if and only if* $\mathrm{rank}([\mathsf{L}, \mathrm{v}_0]) = \mathrm{rank}(\mathsf{L})$*, and*

(ii) $\mathrm{Sol}(\mathsf{L}, \mathrm{v}_0)$ *is a singleton if and only if*

(a) $\mathrm{rank}([\mathsf{L}, \mathrm{v}_0]) = \mathrm{rank}(\mathsf{L})$*,*

(b) $\dim_{\mathsf{F}}(\ker([\mathsf{L}, \mathrm{v}_0])) = 1$*, and*

(c) $\ker([\mathsf{L}, \mathrm{v}_0]) \cap (\mathsf{U} \oplus \{0_{\mathsf{F}}\}) = \{(0_{\mathsf{U}}, 0_{\mathsf{F}})\}$*.*

Proof (i) Given the form of $[\mathsf{L}, v_0]$ (and referring to the proof of Theorem 5.4.50) we have

$$\mathrm{rank}([\mathsf{L}, v_0]) = \begin{cases} \mathrm{rank}(\mathsf{L}), & v_0 \in \mathrm{image}(\mathsf{L}), \\ \mathrm{rank}(\mathsf{L}) + 1, & v_0 \notin \mathrm{image}(\mathsf{L}). \end{cases}$$

Since $\mathrm{rank}(\mathsf{L})$ is finite the result follows.

(ii) From the first part of the result it suffices to show that $\mathrm{Sol}(\mathsf{L}, v_0)$ is a singleton if and only if $\dim_{\mathsf{F}}(\ker([\mathsf{L}, v_0])) = 1$. First suppose that $\mathrm{Sol}(\mathsf{L}, v_0)$ is a singleton. By Theorem 5.4.50 we have

$$\ker([\mathsf{L}, v_0]) = \{(u, a) \in \mathsf{U} \oplus \mathsf{F} \mid u = a u_0\}$$

for some $u_0 \in \mathsf{U}$. One can easily see that $\{(u_0, 1)\}$ is a basis for $\ker([\mathsf{L}, v_0])$, and so $\dim_{\mathsf{F}}(\ker([\mathsf{L}, v_0])) = 1$. Also, if $(u, a) \in \ker([\mathsf{L}, v_0]) \cap (\mathsf{U} \oplus \{0_{\mathsf{F}}\})$, then $(u, a) = a(u_0, 1_{\mathsf{F}})$ and $a = 0_{\mathsf{F}}$. Thus $(u, a) = (0_{\mathsf{U}}, 0_{\mathsf{F}})$, as claimed.

Conversely, suppose that $\dim_{\mathsf{F}}([\mathsf{L}, v_0]) = 1$ and that $\ker([\mathsf{L}, v_0]) \cap (\mathsf{U} \oplus \{0_{\mathsf{F}}\}) = \{(0_{\mathsf{U}}, 0_{\mathsf{F}})\}$, and let (u_0, a_0) be a basis for $\ker([\mathsf{L}, v_0])$. We claim that $a_0 \neq 0_{\mathsf{F}}$. Indeed, if $a_0 = 0_{\mathsf{F}}$ then we have $(u_0, 0_{\mathsf{F}}) \in \ker([\mathsf{L}, v_0]) \cap (\mathsf{U} \oplus \{0_{\mathsf{F}}\})$. This contradicts the assumption that $\ker([\mathsf{L}, v_0]) \cap (\mathsf{U} \oplus \{0_{\mathsf{F}}\}) = \{(0_{\mathsf{U}}, 0_{\mathsf{F}})\}$ since $u_0 \neq 0_{\mathsf{U}}$. Thus we have

$$\begin{aligned} \ker([\mathsf{L}, v_0]) &= \{a(u_0, a_0) \mid a \in \mathsf{F}\} = \{a a_0 (a_0^{-1} u_0, 1_{\mathsf{F}}) \mid a \in \mathsf{F}\} \\ &= \{a(a_0^{-1} u_0, 1_{\mathsf{F}}) \mid a \in \mathsf{F}\} = \{(u, a) \mid u = a a_0^{-1} u_0\}. \end{aligned}$$

The result now follows from Theorem 5.4.50. ■

Note that the corollary is false if $\mathrm{rank}(\mathsf{L})$ is infinite since, in this case we have, using Theorem 1.7.17,

$$\mathrm{rank}([\mathsf{L}, v_0]) \leq \mathrm{rank}(\mathsf{L}) + 1 = \mathrm{rank}(\mathsf{L})$$

and

$$\mathrm{rank}([\mathsf{L}, v_0]) \geq \mathrm{rank}(\mathsf{L}),$$

i.e., $\mathrm{rank}([\mathsf{L}, v_0]) = \mathrm{rank}(\mathsf{L})$ for *any* $v_0 \in \mathsf{V}$, even when $v_0 \notin \mathrm{image}(\mathsf{L})$. This is entirely analogous to the difference between the finite- and infinite-dimensional parts of Corollary 5.4.42.

5.4.9 Eigenvalues, eigenvectors, and spectral values

In this section we introduce an important concept that can be associated with an endomorphism: the notion of an eigenvalue, or more generally a spectral value. As we shall see in Section 5.8, for finite-dimensional vector spaces the structure of eigenvalues can say a great deal about the character of an endomorphism. For infinite-dimensional vector spaces, matters are more complex, and, while we hint at a few of these complexities here, we shall not have much to say in terms of a more complete discussion of the infinite-dimensional case, which involves the addition of analysis, not just algebra, to the mix.

We begin with the definitions.

5.4.52 Definition (Eigenvalue, eigenvector, spectral value) Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V}$. If $\lambda \in \mathsf{F}$ define $\mathsf{L}_\lambda \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by $\mathsf{L}_\lambda = \mathrm{id}_{\mathsf{V}} - \lambda\mathsf{L}$.

(i) A ***spectral value*** for L is an element $\lambda \in \mathsf{F}$ for which L_λ is not invertible.

(ii) An ***eigenvalue*** for L is a spectral value $\lambda \in \mathsf{F}$ for which L_λ is not injective.

(iii) If $\lambda \in \mathsf{F}$ is an eigenvalue for L, an ***eigenvector*** for λ is a nonzero vector $v \in \ker(\mathsf{L}_\lambda)$. •

Let us first establish the equivalence of the concepts of spectral value and eigenvalue for endomorphisms of finite-dimensional vector spaces.

5.4.53 Proposition (Spectral values and eigenvalues in finite dimensions) *If F be a field, if V be a finite-dimensional F-vector space, and if $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$, then $\lambda \in \mathsf{F}$ is an eigenvalue for L if and only if it is a spectral value for L. Moreover, $\lambda \in \mathsf{F}$ is an eigenvalue for L if and only if $\det \mathsf{L}_\lambda = 0_{\mathsf{F}}$.*

Proof Since eigenvalues are spectral values, we need only show the converse. Thus suppose that L_λ is not invertible. By Corollary 5.4.44 it follows that L_λ is not injective, and so by Exercise 4.5.23 it follows that λ is an eigenvalue for L. That eigenvalues are exactly characterised by the equation $\det \mathsf{L}_\lambda = 0_{\mathsf{F}}$ follows from Theorem 5.4.35(iii). ■

The preceding result introduces an important object, namely the expression $\det \mathsf{L}_\lambda$. The following result characterises this expression in a way that will be important to us in Section 5.8, and which recalls from Proposition 4.4.9 the evaluation homomorphism.

5.4.54 Proposition (Existence of characteristic polynomial for a linear map) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Then there exists a polynomial $\mathsf{C}_{\mathsf{L}} \in \mathsf{F}[\xi]$ with the following properties:*

(i) C_{L} *is monic;*

(ii) $\deg(\mathsf{C}_{\mathsf{L}}) = \dim_{\mathsf{F}}(\mathsf{V})$*;*

(iii) $\mathrm{Ev}_{\mathsf{F}}(\mathsf{C}_{\mathsf{L}})(\lambda) = \det \mathsf{L}_\lambda$ *for all* $\lambda \in \mathsf{F}$.

Moreover, if $\mathrm{card}(\mathsf{F}) \geq \dim_{\mathsf{F}}(\mathsf{V})$ then the preceding three properties uniquely determine C_{L}.

Proof Let $\mathscr{B} = \{v_1, \dots, v_n\}$ be a basis for V and denote $\boldsymbol{A}_{\mathsf{L}} = [\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$. Now define $\hat{\boldsymbol{A}}_{\mathsf{L}} \in \mathrm{Mat}_{n\times n}(\mathsf{F}[\xi])$ by $\hat{\boldsymbol{A}}_{\mathsf{L}} = \xi \boldsymbol{I}_n - \boldsymbol{A}_{\mathsf{L}}$, and then define $C_{\mathsf{L}} \in \mathsf{F}[\xi]$ by $C_{\mathsf{L}} = \det \hat{\boldsymbol{A}}_{\mathsf{L}}$. We claim that C_{L} is monic and has degree n. By Proposition 5.3.6, C_{L} is a sum of terms, each of which is an n-fold product of polynomials of degree at most 1. Moreover, the only term in the sum that is a product of polynomials of degree exactly 1 is the term which is the product of the diagonal elements of $\hat{\boldsymbol{A}}_{\mathsf{L}}$. These diagonal elements are themselves monic, so their product will also be monic, and the degree is obviously equal to the number of terms in the product, which is exactly n. Thus C_{L} is indeed monic and of degree n. If we define $\boldsymbol{A}_{\mathsf{L},\lambda} = \lambda \boldsymbol{I}_n - \boldsymbol{A}_{\mathsf{L}}$ for $\lambda \in \mathsf{F}$, we clearly have $\det \boldsymbol{A}_{\mathsf{L},\lambda} = \mathrm{Ev}_{\mathsf{F}}(C_{\mathsf{L}})(\lambda)$.

For the final assertion of the proposition, suppose that P_1 and P_2 are two distinct polynomials satisfying the three properties given. Then $\mathrm{Ev}_{\mathsf{F}}(P_1 - P_2)(a) = 0_{\mathsf{F}}$ for every $a \in \mathsf{F}$. The nonzero polynomial $P_1 - P_2$ has degree at most $\dim_{\mathsf{F}}(\mathsf{V}) - 1$ and so has at most $n - 1$ roots by Proposition 4.4.26. This is a contradiction if $\mathrm{card}(\mathsf{F}) \geq \dim_{\mathsf{F}}(\mathsf{V})$, and so we must have $P_1 = P_2$. ■

Note that it is possible that there be more than one polynomial having the three properties given in the proposition (see Example 5.4.56–1 below). However, since our interest in these volumes will be primarily with vector spaces over $\mathbb{R}$ and $\mathbb{C}$, the three conditions *do* in fact uniquely prescribe the characteristic polynomial, since $\mathrm{card}(\mathbb{R})$ and $\mathrm{card}(\mathbb{C})$ are both infinite. Moreover, the proof of the proposition gives a natural way of defining one of the possible polynomials satisfying the three conditions of the proposition, so let us indicate, outside the confines of the proof, how this is done:

1. Choose a basis (any basis) $\mathscr{B} = \{v_1, \dots, v_n\}$ for V.
2. Define the matrix $\boldsymbol{A}_{\mathsf{L}} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ by $\boldsymbol{A}_{\mathsf{L}} = [\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$.
3. Define the matrix $\hat{\boldsymbol{A}}_{\mathsf{L}} \in \mathrm{Mat}_{n\times n}(\mathsf{F}[\xi])$ by $\hat{\boldsymbol{A}}_{\mathsf{L}} = \xi \boldsymbol{I}_n - \boldsymbol{A}_{\mathsf{L}}$.
4. Define $C_{\mathsf{L}} \in \mathsf{F}[\xi]$ by $C_{\mathsf{L}} = \det \hat{\boldsymbol{A}}_{\mathsf{L}}$.

The polynomial C_{L} has a name.

5.4.55 Definition (Characteristic polynomial) Let F be a field, let V be a finite-dimensional vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The polynomial $C_{\mathsf{L}} \in \mathsf{F}[\xi]$ defined above is the ***characteristic polynomial*** of L. •

This construction is somewhat unsatisfying in the way that any computation involving bases is unsatisfying. We shall give a more sophisticated view of the characteristic polynomial in Section 5.8.4.

The upshot of the preceding development is that, for endomorphisms of finite-dimensional vector spaces, the determining of eigenvalues, or equivalently spectral values, amounts to computing roots of a monic polynomial whose degree is equal to the dimension of V. We refer the reader to the discussion at the end of Section 4.7.3 as a reminder of the difficulties this imposes on computation of eigenvalues.

Now let us give some examples that illustrate the notions of eigenvalue and spectral value.

5.4.56 Examples (Eigenvalues and spectral values)

1. We let $\mathsf{F} = \mathbb{Z}_2$, $\mathsf{V} = \mathsf{F}^3$, and consider the linear map $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$ given by $\mathsf{V}(k_1 + 2\mathbb{Z}, k_2 + 2\mathbb{Z}, k_3 + 2\mathbb{Z}) = (k_1 + 2\mathbb{Z}, k_2 + 2\mathbb{Z}, k_3 + 2\mathbb{Z})$, i.e., $\mathsf{L} = \mathrm{id}_{\mathsf{V}}$. Let us apply the algorithm above for computing the characteristic polynomial. Using the standard basis

   $$\mathscr{B} = \{(1 + 2\mathbb{Z}, 0 + 2\mathbb{Z}, 0 + 2\mathbb{Z}), (0 + 2\mathbb{Z}, 1 + 2\mathbb{Z}, 0 + 2\mathbb{Z}), (0 + 2\mathbb{Z}, 0 + 2\mathbb{Z}, 1 + 2\mathbb{Z})\}$$

   for F^3 we compute the matrix representative of L to be

   $$\boldsymbol{A}_{\mathsf{L}} = \begin{bmatrix} 1 + 2\mathbb{Z} & 0 + 2\mathbb{Z} & 0 + 2\mathbb{Z} \\ 0 + 2\mathbb{Z} & 1 + 2\mathbb{Z} & 0 + 2\mathbb{Z} \\ 0 + 2\mathbb{Z} & 0 + 2\mathbb{Z} & 1 + 2\mathbb{Z} \end{bmatrix}.$$

   Therefore,

   $$\hat{\boldsymbol{A}}_{\mathsf{L}} = \begin{bmatrix} \xi - (1 + 2\mathbb{Z}) & 0 + 2\mathbb{Z} & 0 + 2\mathbb{Z} \\ 0 + 2\mathbb{Z} & \xi - (1 + 2\mathbb{Z}) & 0 + 2\mathbb{Z} \\ 0 + 2\mathbb{Z} & 0 + 2\mathbb{Z} & \xi - (1 + 2\mathbb{Z}) \end{bmatrix},$$

   and using the standard rules for computing determinants we have

   $$\mathrm{C}_{\mathsf{L}} = \det \hat{\boldsymbol{A}}_{\mathsf{L}} = (\xi - (1 + 2\mathbb{Z}))^3.$$

   We also claim that the polynomial

   $$P'_{\mathsf{L}} = (\xi - (1 + 2\mathbb{Z}))^3 + \xi^2 - \xi$$

   satisfies the three conditions of Proposition 5.4.54. Clearly P'_{L} is monic and has the required degree. But, as we saw in Example 4.4.10, $\mathrm{Ev}_{\mathsf{F}}(\xi^2 - \xi)(a) = 0_{\mathsf{F}}$ for every $a \in \mathsf{F}$. Thus $\mathrm{Ev}_{\mathsf{F}}(\mathrm{C}_{\mathsf{L}}) = \mathrm{Ev}_{\mathsf{F}}(P'_{\mathsf{L}})$. Note that, despite the fact that here are two candidate characteristic polynomials (if candidacy is defined as satisfying the conditions of Proposition 5.4.54), the one we use is the one of Definition 5.4.55, i.e., $\mathrm{C}_{\mathsf{L}} = (\xi - (1 + 2\mathbb{Z}))^3$.

2. We let $\mathsf{F} = \mathbb{R}$, $\mathsf{V} = \mathbb{R}^2$, and consider the two linear maps $\mathsf{L}_1, \mathsf{L}_2 \in \operatorname{End}_{\mathbb{R}}(\mathbb{R}^2)$ defined (as 2×2 matrices) by

   $$\mathsf{L}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathsf{L}_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

   To compute the eigenvalues, we compute the characteristic polynomials to be

   $$\mathrm{C}_{\mathsf{L}_1} = \xi^2 - 1, \quad \mathrm{C}_{\mathsf{L}_2} = \xi^2 + 1.$$

   Note that $\mathrm{C}_{\mathsf{L}_1}$ has roots -1 and 1, while $\mathrm{C}_{\mathsf{L}_2}$ has no roots. Thus L_1 has eigenvalues -1 and 1, while L_2 has no eigenvalues.

3. We let $\mathsf{F} = \mathbb{C}$, $\mathsf{V} = \mathbb{C}^2$, and consider the two linear maps $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathbb{C}}(\mathbb{C}^2)$ defined (as 2×2 matrices) by

$$\mathsf{L}_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \mathsf{L}_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

These are "the same" two linear maps as in the previous example, and so the characteristic polynomials are "the same:"

$$C_{\mathsf{L}_1} = \xi^2 - 1, \quad C_{\mathsf{L}_2} = \xi^2 + 1.$$

But now both have roots since $\mathbb{C}$ is algebraically complete. Specifically, L_1 has eigenvalues -1 and 1, and L_2 has eigenvalues $-\mathrm{i}$ and i. The punchline is that eigenvalues depend on the field over which the vector space is being defined. This is explored further in Section 5.4.10.

4. For a field F we take $\mathsf{V} = \mathsf{F}_0^\infty$ (see Example 4.5.2–4). An endomorphism of V is then, by Theorem 5.1.13, represented by a matrix with an enumerable number of rows and columns. With this representation in mind, let us define $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by

$$\mathsf{L} = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

For $\lambda \in \mathsf{F}$ we then have

$$\mathsf{L}_\lambda = \begin{bmatrix} \lambda & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ -1_{\mathsf{F}} & \lambda & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & -1_{\mathsf{F}} & \lambda & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & -1_{\mathsf{F}} & \lambda & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & -1_{\mathsf{F}} & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

About L_λ we make the following observations.

1 Lemma *The following statements hold:*

(i) L_λ *is injective for all* λ*;*

(ii) L_λ *is surjective if and only if* $\lambda \neq 0_{\mathsf{F}}$*.*

Proof (i) Let $\{e_j\}_{j \in \mathbb{Z}_{>0}}$ be the standard basis for F_0^∞. For $\lambda = 0_{\mathsf{F}}$ one can directly compute that $\mathsf{L}_\lambda(\boldsymbol{x}) = \boldsymbol{0}_{\mathsf{F}_0^\infty}$ implies that $\boldsymbol{x}(j) = 0_{\mathsf{F}}$ for every $j \in \mathbb{Z}_{>0}$. Thus L_λ is injective if $\lambda = 0_{\mathsf{F}}$ by Exercise 4.5.23. If $\lambda \neq 0_{\mathsf{F}}$ and if $\mathsf{L}_\lambda(\boldsymbol{x}) = \boldsymbol{0}_{\mathsf{F}_0^\infty}$, then a direct computation shows that

$$\lambda \boldsymbol{x}(1) = 0_{\mathsf{F}}, \; -\boldsymbol{x}(j) + \lambda \boldsymbol{x}(j+1) = 0_{\mathsf{F}}, \qquad j \in \mathbb{Z}_{>0}.$$

This implies that $x_j = 0_{\mathsf{F}}$ for every $j \in \mathbb{Z}_{>0}$, and so implies that L_λ is injective for $\lambda \neq 0_{\mathsf{F}}$ by Exercise 4.5.23.

(ii) First note that when $\lambda = 0_{\mathsf{F}}$ then L_λ is not surjective since $e_1 \notin \operatorname{image}(\mathsf{L}_\lambda)$. Now suppose that $\lambda \neq 0_{\mathsf{F}}$. Let $y \in \mathsf{F}_0^\infty$. Then, defining $x \in \mathsf{F}_0^\infty$ by

$$x(1) = \lambda^{-1}y(1),\ x(j+1) = \lambda^{-1}(y(j+1) - x(j)), \qquad j \in \mathbb{Z}_{>0},$$

one can directly check that $\mathsf{L}_\lambda(x) = y$. Thus L_λ is surjective when $\lambda \neq 0_{\mathsf{F}}$. ▼

The lemma immediately allows us to conclude that 0_{F} is a spectral value for L, but that L has no eigenvalues. •

Eigenvalues can occur "more than once," i.e., with multiplicity. Complicating matters is the fact that there are two ways of measuring the multiplicity of an eigenvalue. The distinctions between these two sorts of multiplicity will become most clear, at least in the finite-dimensional case, in Section 5.8. Here we primarily consider the definitions and some examples.

Recalling the constructions concerning kernels of powers of endomorphisms described in Theorem 5.4.13, we may define algebraic and geometric multiplicities.

5.4.57 Definition (Eigenspaces, algebraic and geometric multiplicity) Let F be a field, let V be a F-vector space, let $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$, and let $\lambda \in \mathsf{F}$ be an eigenvalue for L.

(i) The ***eigenspace*** for λ is the subspace $\mathsf{W}(\lambda, \mathsf{L}) = \ker(\mathsf{L}_\lambda)$.

(ii) The ***generalised eigenspace*** for λ is the subspace $\overline{\mathsf{W}}(\lambda, \mathsf{L}) = \cup_{j \in \mathbb{Z}_{>0}} \ker(\mathsf{L}_\lambda^j)$.

(iii) The ***geometric multiplicity*** of λ is $m_{\mathrm{g}}(\lambda, \mathsf{L}) = \dim_{\mathsf{F}}(\mathsf{W}(\lambda, \mathsf{L}))$.

(iv) The ***algebraic multiplicity*** of λ is $m_{\mathrm{a}}(\lambda, \mathsf{L}) = \dim_{\mathsf{F}}(\overline{\mathsf{W}}(\lambda, \mathsf{L}))$. •

5.4.58 Remarks (Properties of geometric and algebraic multiplicity)

1. Note that both the geometric and algebraic multiplicity are nonzero.
2. The algebraic and geometric multiplicities can be any nonzero cardinal number (see Example 5.4.61–2).
3. It always holds that $m_{\mathrm{a}}(\lambda, \mathsf{L}) \geq m_{\mathrm{g}}(\lambda, \mathsf{L})$.
4. In Proposition 5.8.31 we will show that, if V is finite-dimensional, then the algebraic multiplicity of an eigenvalue of $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$ is equal to the multiplicity of the corresponding root of the characteristic polynomial. •

It will be useful to know that eigenspaces and generalised eigenspaces are invariant.

5.4.59 Proposition (Invariance of eigenspaces and generalised eigenspaces) *Let* F *be a field, let* V *be an* F*-vector space, let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*, and let* λ *be an eigenvalue for* L*. Then, for any* $\mathrm{j} \in \mathbb{Z}_{>0}$*,* $\ker(\mathsf{L}^{\mathrm{j}}_{\lambda})$ *is an* L*-invariant subspace. As a consequence,* $\mathsf{W}(\lambda, \mathsf{L})$ *and* $\overline{\mathsf{W}}(\lambda, \mathsf{L})$ *are* L*-invariant subspaces.*

Proof We first claim that $\mathsf{L} \circ \mathsf{L}^{j}_{\lambda} = \mathsf{L}^{j}_{\lambda} \circ \mathsf{L}$. We prove this by induction. For $j = 1$ we simply have

$$\mathsf{L} \circ (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) = \lambda \mathsf{L} \circ \mathrm{id}_{\mathsf{V}} - \mathsf{L} \circ \mathsf{L} = \lambda \,\mathrm{id}_{\mathsf{V}} \circ \mathsf{L} - \mathsf{L} \circ \mathsf{L} = (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) \circ \mathsf{L}.$$

Now suppose the claim true for $j \in \{1, \dots, k\}$ and compute

$$\begin{aligned}\mathsf{L} \circ (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})^{k+1} = \mathsf{L} \circ (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) \circ (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})^{k} &= (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) \circ \mathsf{L} \circ (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})^{k} \\ &= (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) \circ (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})^{k} \circ \mathsf{L} = (\lambda \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})^{k+1} \circ \mathsf{L},\end{aligned}$$

giving our claim.

The first assertion of the proposition now follows easily. If $v \in \ker(\mathsf{L}^{j}_{\lambda})$ then we have

$$\mathsf{L}^{j}_{\lambda}(v) = 0_{\mathsf{V}} \quad \Longrightarrow \quad \mathsf{L} \circ \mathsf{L}^{j}_{\lambda}(v) = \mathsf{L}^{j}_{\lambda}(\mathsf{L}(v)) = 0_{\mathsf{V}}$$

so that $\mathsf{L} \in \ker(\mathsf{L}^{j}_{\lambda})$. For the second assertion, it immediately follows that $\mathsf{W}(\lambda, \mathsf{L})$ is L-invariant. The L-invariant of $\overline{\mathsf{W}}(\lambda, \mathsf{L})$ follows since, if $v \in \overline{\mathsf{W}}(\lambda, \mathsf{L})$, then $v \in \ker(\mathsf{L}^{j}_{\lambda})$ for some $j \in \mathbb{Z}_{>0}$. ■

It is fairly clear that, if λ_1 and λ_2 are distinct eigenvalues for $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$, then $\mathsf{W}(\lambda_1, \mathsf{L}) \cap \mathsf{W}(\lambda_2, \mathsf{L}) = \{0_{\mathsf{V}}\}$. It is less clear, although still true, that the corresponding statement for the generalised eigenspaces also holds.

5.4.60 Proposition (Intersections of generalised eigenspaces are zero) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. If* λ_1 *and* λ_2 *are distinct eigenvalues for* L *then* $\overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \cap \overline{\mathsf{W}}(\lambda_2, \mathsf{L}) = \{0_{\mathsf{V}}\}$*.*

Proof We first prove a lemma characterising the intersections of generalised eigenspaces.

1 Lemma $\overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \cap \overline{\mathsf{W}}(\lambda_2, \mathsf{L}) = \cup_{\mathrm{j} \in \mathbb{Z}_{>0}}(\ker(\mathsf{L}^{\mathrm{j}}_{\lambda_1}) \cap \ker(\mathsf{L}^{\mathrm{j}}_{\lambda_2}))$.

Proof By definition we have

$$\overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \cap \overline{\mathsf{W}}(\lambda_2, \mathsf{L}) = \left(\cup_{j \in \mathbb{Z}_{>0}} \ker(\mathsf{L}^{j}_{\lambda_1})\right) \cap \left(\cup_{k \in \mathbb{Z}_{>0}} \ker(\mathsf{L}^{k}_{\lambda_2})\right).$$

By Proposition 1.1.7 we have

$$\begin{aligned}\overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \cap \overline{\mathsf{W}}(\lambda_2, \mathsf{L}) &= \cup_{k \in \mathbb{Z}_{>0}} \left(\left(\cup_{j \in \mathbb{Z}_{>0}} \ker(\mathsf{L}^{j}_{\lambda_1})\right) \cap \ker(\mathsf{L}^{k}_{\lambda_2})\right) \\ &= \cup_{k \in \mathbb{Z}_{>0}} \left(\cup_{j \in \mathbb{Z}_{>0}} \left(\ker(\mathsf{L}^{j}_{\lambda_1}) \cap \ker(\mathsf{L}^{k}_{\lambda_2})\right)\right)\end{aligned}$$

It is clear that the inclusion

$$\cup_{j \in \mathbb{Z}_{>0}}(\ker(\mathsf{L}^{j}_{\lambda_1}) \cap \ker(\mathsf{L}^{j}_{\lambda_2})) \subseteq \cup_{k \in \mathbb{Z}_{>0}} \left(\cup_{j \in \mathbb{Z}_{>0}} \left(\ker(\mathsf{L}^{j}_{\lambda_1}) \cap \ker(\mathsf{L}^{k}_{\lambda_2})\right)\right)$$

holds. If

$$v \in \cup_{k\in\mathbb{Z}_{>0}} \left(\cup_{j\in\mathbb{Z}_{>0}} \left(\ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^k)\right)\right)$$

then there exists $j, k \in \mathbb{Z}_{>0}$ such that $v \in \ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^k)$. If $j = k$ then we immediately have

$$v \in \cup_{j\in\mathbb{Z}_{>0}}(\ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^j)).$$

So suppose, without loss of generality, that $j > k$. Then

$$\ker(\mathsf{L}_{\lambda_2}^k) \subseteq \ker(\mathsf{L}_{\lambda_2}^j),$$

and so we again arrive at

$$v \in \cup_{j\in\mathbb{Z}_{>0}}(\ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^j)),$$

so giving our claim. ▼

We next claim that $\ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^j) = \{0_{\mathsf{V}}\}$ for each $j \in \mathbb{Z}_{>0}$. We prove this by induction on j. For $j = 1$, let $v \in \ker(\mathsf{L}_{\lambda_1}) \cap \ker(\mathsf{L}_{\lambda_2})$. Then

$$\mathsf{L}(v) = \lambda_1 v = \lambda_2 v \quad \Longrightarrow \quad (\lambda_1 - \lambda_2)v = 0_{\mathsf{V}}.$$

It follows from Proposition 4.5.3(vi) that $v = 0_{\mathsf{V}}$. Now suppose that $\ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^j) = \{0_{\mathsf{V}}\}$ for $j \in \{1, \dots, k\}$ and let $v \in \ker(\mathsf{L}_{\lambda_1}^{k+1}) \cap \ker(\mathsf{L}_{\lambda_2}^{k+1})$. Then

$$\mathsf{L}_{\lambda_1}^{k+1}(v) = \mathsf{L}_{\lambda_2}^{k+1}(v) = 0_{\mathsf{V}}.$$

This means that $\mathsf{L}_{\lambda_1}(v) \in \ker(\mathsf{L}_{\lambda_1}^k)$ and $\mathsf{L}_{\lambda_2}(v) \in \ker(\mathsf{L}_{\lambda_2}^k)$. Now we note from Exercises 5.4.10 and 5.4.13 that $\ker(\mathsf{L}_{\lambda_2}^k)$ and $\ker(\mathsf{L}_{\lambda_1}^k)$ are invariant under L_{λ_1} and L_{λ_2}, respectively. Thus we have

$$\mathsf{L}_{\lambda_1}(\mathsf{L}_{\lambda_2}(v)) \in \ker(\mathsf{L}_{\lambda_2}^k), \quad \mathsf{L}_{\lambda_2}(\mathsf{L}_{\lambda_1}(v)) \in \ker(\mathsf{L}_{\lambda_1}^k).$$

Therefore, by the induction hypothesis,

$$\mathsf{L}_{\lambda_1}(\mathsf{L}_{\lambda_2}(v)) = \mathsf{L}_{\lambda_2}(\mathsf{L}_{\lambda_1}(v)) = 0_{\mathsf{V}},$$

since L_{λ_1} and L_{λ_2} commute. Therefore,

$$\mathsf{L}_{\lambda_2}(v) \in \ker(\mathsf{L}_{\lambda_1}) \subseteq \ker(\mathsf{L}_{\lambda_1}^k), \quad \mathsf{L}_{\lambda_1}(v) \in \ker(\mathsf{L}_{\lambda_2}) \subseteq \ker(\mathsf{L}_{\lambda_2}^k).$$

That is, $\mathsf{L}_{\lambda_1}(v), \mathsf{L}_{\lambda_2}(v) \in \ker(\mathsf{L}_{\lambda_1}^k) \cap \ker(\mathsf{L}_{\lambda_2}^k)$. Again by the induction hypothesis, this gives $\mathsf{L}_{\lambda_1}(v) = 0_{\mathsf{V}}$ and $\mathsf{L}_{\lambda_2}(v) = 0_{\mathsf{V}}$. Thus $v \in \ker(\mathsf{L}_{\lambda_1}) \cap \ker(\mathsf{L}_{\lambda_2}) = \{0_{\mathsf{V}}\}$, so giving our claim that $\ker(\mathsf{L}_{\lambda_1}^j) \cap \ker(\mathsf{L}_{\lambda_2}^j) = \{0_{\mathsf{V}}\}$ for each $j \in \mathbb{Z}_{>0}$.

The result now easily follows from this and the above lemma. ■

Let us give some examples that exhibit the character of and relationship between algebraic and geometric multiplicity.

5.4.61 Examples (Algebraic and geometric multiplicity)

1. For a field F take $\mathsf{V} = \mathsf{F}^3$ and define $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by the two 3×3 matrices

$$\mathsf{L}_1 = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & -1_{\mathsf{F}} \end{bmatrix}, \quad \mathsf{L}_2 = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & -1_{\mathsf{F}} \end{bmatrix}.$$

These linear maps both have eigenvalues 0_{F} and -1_{F}. We can readily see that

$$\begin{aligned} \ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1) &= \mathrm{span}_{\mathsf{F}}((1_{\mathsf{F}}, 0_{\mathsf{F}}, 0_{\mathsf{F}}), (0_{\mathsf{F}}, 1_{\mathsf{F}}, 0_{\mathsf{F}})), \\ \ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1) &= \mathrm{span}_{\mathsf{F}}((1_{\mathsf{F}}, 0_{\mathsf{F}}, 0_{\mathsf{F}})), \\ \ker(-1_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1) &= \mathrm{span}_{\mathsf{F}}((0_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}})), \\ \ker(-1_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_2) &= \mathrm{span}_{\mathsf{F}}((0_{\mathsf{F}}, 0_{\mathsf{F}}, 1_{\mathsf{F}})). \end{aligned}$$

From this we deduce that for L_1, $m_{\mathrm{g}}(0_{\mathsf{F}}, \mathsf{L}_1) = 2$ and $m_{\mathrm{g}}(-1_{\mathsf{F}}, \mathsf{L}_1) = 1$, and that for L_2, $m_{\mathrm{g}}(0_{\mathsf{F}}, \mathsf{L}_2) = 1$ and $m_{\mathrm{g}}(-1_{\mathsf{F}}, \mathsf{L}_2) = 1$. To compute the algebraic multiplicities, we must compute the powers of the matrices $\lambda\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}$ where λ runs over the eigenvalues, and L is either L_1 or L_2. For this purpose it is sufficient to compute

$$\begin{aligned} \dim_{\mathsf{F}}(\ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1)) &= 2, & \dim_{\mathsf{F}}(\ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_2)) &= 1, \\ \dim_{\mathsf{F}}(\ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1)^2) &= 2, & \dim_{\mathsf{F}}(\ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_2)^2) &= 2, \\ \dim_{\mathsf{F}}(\ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1)^3) &= 2, & \dim_{\mathsf{F}}(\ker(0_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_2)^3) &= 2, \\ \dim_{\mathsf{F}}(\ker(-1_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1)) &= 1, & \dim_{\mathsf{F}}(\ker(-1_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_2)) &= 1, \\ \dim_{\mathsf{F}}(\ker(-1_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_1)^2) &= 1, & \dim_{\mathsf{F}}(\ker(-1_{\mathsf{F}}\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}_2)^2) &= 1. \end{aligned}$$

We then conclude that $m_{\mathrm{a}}(0_{\mathsf{F}}, \mathsf{L}_1) = m_{\mathrm{a}}(0_{\mathsf{F}}, \mathsf{L}_2) = 2$ and $m_{\mathrm{a}}(-1_{\mathsf{F}}, \mathsf{L}_1) = m_{\mathrm{a}}(-1_{\mathsf{F}}, \mathsf{L}) = 1$.

2. For any set I, take $\mathsf{V} = \mathsf{F}_0^I$ and $\mathsf{L} = 0_{\mathsf{V},\mathsf{V}}$, i.e., L is the zero linear map. Then L has 0_{F} as its only eigenvalue, and it has a geometric multiplicity of $\mathrm{card}(I)$ since $\ker(\mathsf{L}) = \mathsf{V}$. Note that the algebraic multiplicity of the eigenvalue 0_{F} is also equal to $\mathrm{card}(I)$.

3. For a field F we take $\mathsf{V} = \mathsf{F}_0^\infty$ (see Example 4.5.2–4). We let $\{\boldsymbol{e}_j\}_{j \in \mathbb{Z}_{>0}}$ be the standard basis and define $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by

$$\mathsf{L}(\boldsymbol{e}_1 = \boldsymbol{0}_{\mathsf{F}_0^\infty}, \ \mathsf{L}(\boldsymbol{e}_{j-1}) = \boldsymbol{e}_j, \qquad j \in \mathbb{Z}_{>0}.$$

If we represent L as a column finite matrix with countably many rows and columns, then we have

$$\mathsf{L} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

We can directly compute $\ker(\mathsf{L}^k) = \operatorname{span}_{\mathsf{F}}(e_1, \dots, e_k)$. From this we conclude that $m_{\mathrm{g}}(\lambda, \mathsf{L}) = 1$ and $m_{\mathrm{a}}(\lambda, \mathsf{L}) = \operatorname{card}(\mathbb{Z}_{>0})$. Moreover, $\overline{\mathsf{W}}(\lambda, \mathsf{L}) = \mathsf{V}$, even though L^k is nonzero for every $k \in \mathbb{Z}_{>0}$ (in fact, L^k is surjective for every $k \in \mathbb{Z}_{>0}$). •

5.4.10 Complexification of $\mathbb{R}$-linear maps

When dealing with linear maps between $\mathbb{R}$-vector spaces, it is sometimes useful to consider instead the linear map between the complexifications of these vector spaces. This is particularly the case when dealing with eigenvalues of endomorphisms. In this section we indicate the manner in which a linear map between $\mathbb{R}$-vector spaces induces a linear map between the respective complexifications, and the resulting constructions that arise concerning eigenvalues.

We recall from Section 4.5.7 that if V is a $\mathbb{R}$-vector space, then the complexification of V is the $\mathbb{C}$-vector space $\mathsf{V}_{\mathbb{C}}$ defined by $\mathsf{V}_{\mathbb{C}} = \mathsf{V} \times \mathsf{V}$ with vector addition and scalar multiplication defined by

$$(u_1, u_2) + (v_1, v_2) = (u_1 + v_1, u_2 + v_2), \quad (a + \mathrm{i}b)(u, v) = (au - bv, av + bu).$$

We now indicate how to associate to every $\mathbb{R}$-linear map from U to V a $\mathbb{C}$-linear map from $\mathsf{U}_{\mathbb{C}}$ to $\mathsf{V}_{\mathbb{C}}$.

5.4.62 Definition (Complexification of a linear map) Let U and V be $\mathbb{R}$-vector spaces with $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$ their complexifications. If $\mathsf{L} \in \operatorname{Hom}_{\mathbb{R}}(\mathsf{U}; \mathsf{V})$ then the ***complexification*** of L is the element $\mathsf{L}_{\mathbb{C}} \in \operatorname{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$ defined by

$$\mathsf{L}_{\mathbb{C}}(u, v) = (\mathsf{L}(u), \mathsf{L}(v)). \qquad \bullet$$

Of course, one must verify that $\mathsf{L}_{\mathbb{C}}$ is actually $\mathbb{C}$-linear. To give a useful characterisation of $\mathsf{L}_{\mathbb{C}}$, let us recall from Definition 4.5.62 the canonical representation of $(u, v) \in \mathsf{V}_{\mathbb{C}}$ as $u + \mathrm{i}v$. In this case we simply have

$$\mathsf{L}_{\mathbb{C}}(u + \mathrm{i}v) = \mathsf{L}(u) + \mathrm{i}\mathsf{L}(v);$$

thus the definition of $\mathsf{L}_{\mathbb{C}}$ is that it acts as does L on both the real and imaginary parts of $\mathsf{V}_{\mathbb{C}}$. This representation of $\mathsf{L}_{\mathbb{C}}$ makes it straightforward to verify that $\mathsf{L}_{\mathbb{C}}$ is actually $\mathbb{C}$-linear:

$$\begin{aligned}\mathsf{L}_{\mathbb{C}}((a + \mathrm{i}b)(u + \mathrm{i}v)) &= \mathsf{L}_{\mathbb{C}}((au - bv) + \mathrm{i}(bu + av)) = \mathsf{L}(au - bv) + \mathrm{i}\mathsf{L}(bu + av)\\ &= (a + \mathrm{i}b)(\mathsf{L}(u) + \mathrm{i}\mathsf{L}(v)) = (a + \mathrm{i}b)\mathsf{L}_{\mathbb{C}}(u + \mathrm{i}v).\end{aligned}$$

The set $\mathsf{V}_{\mathbb{C}}$ has the structure of both a $\mathbb{C}$-vector space and a $\mathbb{R}$-vector space; see Proposition 4.5.61. Thus, given $\mathbb{R}$-vector spaces U and V, we may consider the sets $\operatorname{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$ and $\operatorname{Hom}_{\mathbb{R}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$ of $\mathbb{C}$-linear and $\mathbb{R}$-linear maps, respectively. We also have the subset of $\operatorname{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$ consisting of the complexification of elements of $\operatorname{Hom}_{\mathbb{R}}(\mathsf{U}; \mathsf{V})$. Let us denote this subset by

$$\operatorname{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})_{\mathbb{R}} = \{\mathsf{L}_{\mathbb{C}} \mid \mathsf{L} \in \operatorname{Hom}_{\mathbb{R}}(\mathsf{U}; \mathsf{V})\}.$$

To summarise, we have three sets of linear maps between $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$:

1. $\mathrm{Hom}_{\mathbb{R}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ is the set of $\mathbb{R}$-linear maps;
2. $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ is the set of $\mathbb{C}$-linear maps;
3. $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}}$ is the set of complexifications of $\mathbb{R}$-linear maps from U to V.

We wish to explore the relationships between these three sets of linear maps. To do so, we recall from Definition 4.5.62 the complex conjugation map $\sigma_{\mathsf{V}}\colon \mathsf{V}_{\mathbb{C}} \to \mathsf{V}_{\mathbb{C}}$. We also introduce the map $\mathrm{i}_{\mathsf{V}}\colon \mathsf{V}_{\mathbb{C}} \to \mathsf{V}_{\mathbb{C}}$ given by

$$\mathrm{i}_{\mathsf{V}}(u + \mathrm{i}v) = \mathrm{i}(u + \mathrm{i}v) = -v + \mathrm{i}u,$$

where we use the canonical representation of elements of $\mathsf{V}_{\mathbb{C}}$. One can verify that this map is both $\mathbb{R}$-linear and $\mathbb{C}$-linear (see Exercise 5.4.23).

We may complex conjugation to define the complex conjugate of a $\mathbb{C}$-linear map.

5.4.63 Definition (Complex conjugate of a $\mathbb{C}$-endomorphism) Let U and V be a $\mathbb{R}$-vector spaces with $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$ the complexifications. The ***complex conjugate*** of $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ is the element $\bar{\mathsf{L}} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ defined by $\bar{\mathsf{L}} = \sigma_{\mathsf{V}} \circ \mathsf{L} \circ \sigma_{\mathsf{U}}$. •

While σ_{U} and σ_{V} are *not* $\mathbb{C}$-linear, it is nonetheless true that $\bar{\mathsf{L}}$ is $\mathbb{C}$-linear; this is Exercise 5.4.25. Moreover, one may verify that the complex conjugate has the following properties.

5.4.64 Proposition (Properties of the complex conjugate) *If U, V, and W are $\mathbb{R}$-vector spaces with $\mathsf{U}_{\mathbb{C}}$, $\mathsf{V}_{\mathbb{C}}$, and $\mathsf{W}_{\mathbb{C}}$ their complexifications, then the following statements hold:*

(i) $\overline{\mathsf{L} + \mathsf{K}} = \bar{\mathsf{L}} + \bar{\mathsf{K}}$ *for all* $\mathsf{L}, \mathsf{K} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$*;*

(ii) $\overline{a\mathsf{L}} = \bar{a}\bar{\mathsf{L}}$ *for all* $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ *and* $a \in \mathbb{C}$*;*

(iii) $\overline{\mathsf{L} \circ \mathsf{K}} = \bar{\mathsf{L}} \circ \bar{\mathsf{K}}$ *for all* $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{V}_{\mathbb{C}};\mathsf{W}_{\mathbb{C}})$ *and* $\mathsf{K} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$*.*

Proof This is Exercise 5.4.26. ■

We may now clearly state the relationships between the various classes of linear maps between $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$.

5.4.65 Proposition (Characterisation of linear maps between $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$) *Let U and V be $\mathbb{R}$-vector spaces with $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$ the complexifications. Then the following statements hold:*

(i) the map $\mathsf{L}_{\mathbb{C}} \mapsto \mathsf{L} + \mathrm{i}\mathsf{L}$ *is a $\mathbb{R}$-monomorphism from the $\mathbb{R}$-subspace* $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}}$ *of* $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ *to the complexification of the $\mathbb{R}$-vector space* $\mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V})$*;*

(ii) $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ *is exactly the subset of linear maps* $\mathsf{L} \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ *for which the following diagram commutes:*

$$\begin{tikzcd}
\mathsf{U}_{\mathbb{C}} \arrow[r, "\mathsf{L}"] \arrow[d, "\mathrm{i}_{\mathsf{U}}"'] & \mathsf{V}_{\mathbb{C}} \arrow[d, "\mathrm{i}_{\mathsf{V}}"] \\
\mathsf{U}_{\mathbb{C}} \arrow[r, "\mathsf{L}"'] & \mathsf{V}_{\mathbb{C}}
\end{tikzcd}$$

(iii) $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}}$ *is exactly the subset of linear maps* $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ *for which the following diagram commutes:*

$$\begin{tikzcd}
\mathsf{U}_{\mathbb{C}} \arrow[r, "\mathsf{L}"] \arrow[d, "\sigma_{\mathsf{U}}"'] & \mathsf{V}_{\mathbb{C}} \arrow[d, "\sigma_{\mathsf{V}}"] \\
\mathsf{U}_{\mathbb{C}} \arrow[r, "\mathsf{L}"'] & \mathsf{V}_{\mathbb{C}}
\end{tikzcd}$$

Proof (i) The complexification of $\mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V})$, denoted by $\mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V})_{\mathbb{C}}$, is by definition the $\mathbb{C}$-vector space $\mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V}) \times \mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V})$ with vector addition and scalar multiplication defined by

$$(\mathsf{L}_1,\mathsf{L}_2) + (\mathsf{K}_1,\mathsf{K}_2) = (\mathsf{L}_1 + \mathsf{K}_1, \mathsf{L}_2 + \mathsf{K}_2), \quad (a + ib)(\mathsf{L},\mathsf{K}) = (a\mathsf{L} - b\mathsf{K}, b\mathsf{L} + a\mathsf{K}).$$

Using the canonical representation, we write elements in this complexification as $\mathsf{L}+\mathrm{i}\mathsf{K}$. It is then clear that the map $\mathsf{L}_{\mathbb{C}} \mapsto \mathsf{L} + \mathrm{i}\mathsf{L}$ is injective. To verify that it is $\mathbb{C}$-linear, we should first verify that $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}}$ is indeed a $\mathbb{C}$-subspace of $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$. To check this we note that if $\mathsf{L}_{\mathbb{C}}, \mathsf{K}_{\mathbb{C}} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}}$ and if $a \in \mathbb{R}$ then

$$(\mathsf{L}_{\mathbb{C}} + \mathsf{K}_{\mathbb{C}})(u + \mathrm{i}v) = \mathsf{L}_{\mathbb{C}}(u + \mathrm{i}v) + \mathsf{K}_{\mathbb{C}}(u + \mathrm{i}v) = (\mathsf{L} + \mathsf{K})(u) + \mathrm{i}(\mathsf{L} + \mathsf{K})(v),$$

so that $\mathsf{L}_{\mathbb{C}} + \mathsf{K}_{\mathbb{C}} = (\mathsf{L} + \mathsf{K})_{\mathbb{C}}$, and

$$(a\mathsf{L}_{\mathbb{C}})(u + \mathrm{i}v) = a(\mathsf{L}_{\mathbb{C}}(u + \mathrm{i}v)) = a(\mathsf{L}(u) + \mathrm{i}\mathsf{L}(v)) = (a\mathsf{L})(u) + \mathrm{i}(a\mathsf{L})(v),$$

so that $a\mathsf{L}_{\mathbb{C}} = (a\mathsf{L})_{\mathbb{C}}$. Thus $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}}$ is indeed a $\mathbb{R}$-subspace of $\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$. In the process of verifying this, we have also shown that the map $\mathsf{L}_{\mathbb{C}} \mapsto \mathsf{L} + \mathrm{i}\mathsf{L}$ is $\mathbb{R}$-linear, since from our above computations we have

$$(\mathsf{L}_{\mathbb{C}} + \mathsf{K}_{\mathbb{C}}) \mapsto (\mathsf{L} + \mathsf{K}) + \mathrm{i}(\mathsf{L} + \mathsf{K}) = (\mathsf{L} + \mathrm{i}\mathsf{L}) + (\mathsf{K} + \mathrm{i}\mathsf{K})$$

and

$$a\mathsf{L}_{\mathbb{C}} \mapsto a\mathsf{L} + \mathrm{i}a\mathsf{L} = a(\mathsf{L} + \mathrm{i}\mathsf{L}).$$

(ii) We write a general element $\mathsf{L} \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ as

$$\mathsf{L}(u + \mathrm{i}v) = (\mathsf{L}_{11}(u) + \mathsf{L}_{12}(v)) + \mathrm{i}(\mathsf{L}_{21}(u) + \mathsf{L}_{22}(v))$$

for $\mathsf{L}_{11}, \mathsf{L}_{12}, \mathsf{L}_{21}, \mathsf{L}_{22} \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V})$. The condition that the diagram in the statement of the proposition commute amounts to the conditions

$$\mathsf{L}_{11} = \mathsf{L}_{22}, \quad \mathsf{L}_{12} = -\mathsf{L}_{21}, \tag{5.25}$$

as may be verified by a direct computation. That these conditions are necessary for L to be $\mathbb{C}$-linear follows since the commuting of the diagram exactly means that $\mathsf{L}(\mathrm{i}(u+\mathrm{i}v)) = \mathrm{i}\mathsf{L}(u+\mathrm{i}v)$. Conversely, if the relations (5.25) hold, it can be directly verified by a computation that L is $\mathbb{C}$-linear.

(iii) By the preceding part of the result we can write an element $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ as

$$\mathsf{L}(u + \mathrm{i}v) = (\mathsf{L}_1(u) + \mathsf{L}_2(v)) + \mathrm{i}(-\mathsf{L}_2(u) + \mathsf{L}_1(v))$$

for $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$. The commuting of the given diagram then amounts to the condition that $\mathsf{L}_2 = -\mathsf{L}_2$, i.e., that $\mathsf{L}_2 = 0_{\mathsf{V}}$. The result follows directly. ■

5.4.66 Remarks (Characterisation of linear maps between $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$)

1. The following set inclusions hold:

$$\mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})_{\mathbb{R}} \subset \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}}) \subset \mathrm{Hom}_{\mathbb{R}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}}).$$

2. Part (ii) of the preceding result says that a $\mathbb{R}$-linear map L from $\mathsf{U}_{\mathbb{C}}$ to $\mathsf{V}_{\mathbb{C}}$ is $\mathbb{C}$-linear if and only if $\mathsf{L}(\mathrm{i}(u+\mathrm{i}v)) = \mathrm{i}\mathsf{L}(u+\mathrm{i}v)$. That is to say, to check $\mathbb{C}$-linearity one need only check $\mathbb{R}$-linearity and linearity with respect to multiplication by i.
3. By using the fact that $\sigma_{\mathsf{V}}^{-1} = \sigma_{\mathsf{V}}$, part (iii) of the preceding result can be seen to show that a $\mathbb{C}$-linear map L from $\mathsf{U}_{\mathbb{C}}$ to $\mathsf{V}_{\mathbb{C}}$ is the complexification of a $\mathbb{R}$-linear from U to V if and only if $\mathsf{L} = \bar{\mathsf{L}}$. Equivalently, $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}};\mathsf{V}_{\mathbb{C}})$ is the complexification of a $\mathbb{R}$ linear map from U to V if and only if L maps the subset $\{(u, 0_{\mathsf{U}}) \mid u \in \mathsf{U}\}$ of real vectors in U to the subset $\{(v, 0_{\mathsf{V}}) \mid v \in \mathsf{V}\}$ of real vectors in V; this is Exercise 5.4.27. •

The primary reason for complexifying a $\mathbb{R}$-vector space and then a $\mathbb{R}$-linear map is for the purpose of studying eigenvalues and spectral values of $\mathbb{R}$-endomorphisms. Thus we let V be a $\mathbb{R}$-vector space, let $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$, with $\mathsf{V}_{\mathbb{C}}$ and $\mathsf{L}_{\mathbb{C}}$ the associated complexifications. We are interested in studying how the eigenvalues and spectral values of L and $\mathsf{L}_{\mathbb{C}}$ are related. The following result gives the relationships we seek.

5.4.67 Theorem (Eigenvalues, spectral values, and eigenspaces of an endomorphism and its complexification) *If V is a $\mathbb{R}$-vector space and if $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$ with $\mathsf{L}_{\mathbb{C}} \in \mathrm{End}_{\mathbb{C}}(\mathsf{V}_{\mathbb{C}})$ its complexification, then the following statements hold:*

(i) $\lambda \in \mathbb{R}$ is an eigenvalue (resp. spectral value) for L if and only if λ is an eigenvalue (resp. spectral value) for $\mathsf{L}_{\mathbb{C}}$;

(ii) if $\lambda \in \mathbb{C}$ is an eigenvalue (resp. spectral value) for $\mathsf{L}_{\mathbb{C}}$, then $\bar{\lambda}$ is an eigenvalue (resp. spectral value) for $\mathsf{L}_{\mathbb{C}}$;

(iii) if $\lambda \in \mathbb{R}$ is an eigenvalue for L then

$$\mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}}) = \{(\mathrm{u}, \mathrm{v}) \mid \mathrm{u}, \mathrm{v} \in \mathsf{W}(\lambda, \mathsf{L})\};$$

(iv) if $\lambda \in \mathbb{R}$ is an eigenvalue for L then

$$\overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}}) = \{(\mathrm{u}, \mathrm{v}) \mid \mathrm{u}, \mathrm{v} \in \overline{\mathsf{W}}(\lambda, \mathsf{L})\};$$

(v) if $\lambda \in \mathbb{C}$ is an eigenvalue for $\mathsf{L}_{\mathbb{C}}$ then

$$\mathsf{W}(\bar{\lambda}, \mathsf{L}_{\mathbb{C}}) = \{(\mathrm{u}, \mathrm{v}) \in \mathsf{V}_{\mathbb{C}} \mid (\mathrm{u}, -\mathrm{v}) \in \mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}})\};$$

(vi) if $\lambda \in \mathbb{C}$ is an eigenvalue for $\mathsf{L}_{\mathbb{C}}$ then

$$\overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_{\mathbb{C}}) = \{(\mathrm{u}, \mathrm{v}) \in \mathsf{V}_{\mathbb{C}} \mid (\mathrm{u}, -\mathrm{v}) \in \overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}})\}.$$

Proof (i) First suppose that λ is an eigenvalue for L. Then L_λ is not injective, and we denote by $\mathsf{W}(\lambda, \mathsf{L}) \neq \{0_{\mathsf{V}}\}$ the eigenspace. We claim that

$$\ker(\mathsf{L}_{\mathbb{C},\lambda}) = \{(u, v) \in \mathsf{V}_{\mathbb{C}} \mid \ u, v \in \mathsf{W}(\lambda, \mathsf{L})\}.$$

Indeed, by definition of the complexification of a linear map, $\mathsf{L}_{\mathbb{C},\lambda}(u, v) = (0_{\mathsf{V}}, 0_{\mathsf{V}})$ if and only if $\mathsf{L}(u) = \lambda u$ and $\mathsf{L}(v) = \lambda v$. This shows that λ is an eigenvalue of $\mathsf{L}_{\mathbb{C}}$ and that the eigenspace is $\{(u, v) \in \mathsf{V}_{\mathbb{C}} \mid \ u, v \in \mathsf{W}(\lambda, \mathsf{L})\}$.

Now suppose that $\lambda \in \mathbb{R}$ is an eigenvalue for $\mathsf{L}_{\mathbb{C}}$ and let $\mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}})$ be the eigenspace. Thus, by definition of the complexification of a linear map we have

$$\ker(\mathsf{L}_{\mathbb{C},\lambda}) = \{(u, v) \in \mathsf{V}_{\mathbb{C}} \mid \ u, v \in \ker(\mathsf{L}_\lambda)\},$$

so giving λ as an eigenvalue for L and also giving

$$\mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}}) = \{(u, v) \in \mathsf{V}_{\mathbb{C}} \mid \ u, v \in \mathsf{W}(\lambda, \mathsf{L})\}.$$

That $\lambda \in \mathbb{R}$ is a spectral value for L if and only if it is a spectral value for $\mathsf{L}_{\mathbb{C}}$ follows from Exercise 5.4.22 and the definition of spectral value.

(ii) We may as well suppose that λ is not real. Thus we write $\lambda = a + ib$ for $a, b \in \mathbb{R}$ and with $b \neq 0$. We first claim that $\mathsf{L}_{\mathbb{C},\bar{\lambda}} = \bar{\mathsf{L}}_{\mathbb{C},\lambda}$. Indeed, using Proposition 5.4.64 and Proposition 5.4.65,

$$\bar{\mathsf{L}}_{\mathbb{C},\lambda} = \overline{\mathsf{L}_{\mathbb{C}} - \lambda\,\mathrm{id}_{\mathsf{V}}} = \bar{\mathsf{L}}_{\mathbb{C}} - \bar{\lambda}\,\bar{\mathrm{id}}_{\mathsf{V}} = \mathsf{L} - \bar{\lambda}\,\mathrm{id}_{\mathsf{V}} = \mathsf{L}_{\mathbb{C},\bar{\lambda}}.$$

The following lemma gives us a useful characterisation of the kernel and image of the complex conjugate of an linear map, and this characterisation will be used several times in the remainder of the proof.

1 **Lemma** *If* U *and* V *are* $\mathbb{R}$*-vector spaces and if* $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$*, then*

(i) $\ker(\bar{\mathsf{L}}) = \{(\mathrm{u}, \mathrm{v}) \in \mathsf{U}_{\mathbb{C}} \mid \ (\mathrm{u}, -\mathrm{v}) \in \ker(\mathsf{L})\}$ *and*

(ii) $\mathrm{image}(\bar{\mathsf{L}}) = \{(\mathrm{u}, \mathrm{v}) \in \mathsf{U}_{\mathbb{C}} \mid \ (\mathrm{u}, -\mathrm{v}) \in \mathrm{image}(\mathsf{L})\}$.

Proof As in the proof of part (ii) of Proposition 5.4.65, we may write

$$\mathsf{L}(u, v) = (\mathsf{L}_1(u) + \mathsf{L}_2(v), -\mathsf{L}_2(u) + \mathsf{L}_1(v))$$

for $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{U}; \mathsf{V})$. We then compute

$$\begin{aligned}
\ker(\bar{\mathsf{L}}) &= \ker(\sigma_{\mathsf{V}} \circ \mathsf{L} \circ \sigma_{\mathsf{U}}) \\
&= \{(u, v) \in \mathsf{U}_{\mathbb{C}} \mid \ \sigma_{\mathsf{V}} \circ \mathsf{L} \circ \sigma_{\mathsf{U}}(u, v) = (0_{\mathsf{V}}, 0_{\mathsf{V}})\} \\
&= \{(u, v) \in \mathsf{U}_{\mathbb{C}} \mid \ \sigma_{\mathsf{V}} \circ \mathsf{L}(u, -v) = (0_{\mathsf{V}}, 0_{\mathsf{V}})\} \\
&= \{(u, v) \in \mathsf{U}_{\mathbb{C}} \mid \ \sigma_{\mathsf{V}}(\mathsf{L}_1(u) - \mathsf{L}_2(v), -\mathsf{L}_2(u) - \mathsf{L}_1(v)) = (0_{\mathsf{V}}, 0_{\mathsf{V}})\} \\
&= \{(u, v) \in \mathsf{U}_{\mathbb{C}} \mid \ (\mathsf{L}_1(u) - \mathsf{L}_2(v), \mathsf{L}_2(u) + \mathsf{L}_1(v)) = (0_{\mathsf{V}}, 0_{\mathsf{V}})\} \\
&= \{(u, -v) \in \mathsf{U}_{\mathbb{C}} \mid \ (\mathsf{L}_1(u) + \mathsf{L}_2(v), \mathsf{L}_2(u) - \mathsf{L}_1(v)) = (0_{\mathsf{V}}, 0_{\mathsf{V}})\} \\
&= \{(u, -v) \in \mathsf{U}_{\mathbb{C}} \mid \ (\mathsf{L}_1(u) + \mathsf{L}_2(v), -\mathsf{L}_2(u) + \mathsf{L}_1(v)) = (0_{\mathsf{V}}, 0_{\mathsf{V}})\} \\
&= \{(u, v) \in \mathsf{U}_{\mathbb{C}} \mid \ (u, -v) \in \ker(\mathsf{L})\},
\end{aligned}$$

giving the first part of the lemma.

For the second part we write

$$\begin{aligned}\operatorname{image}(\bar{\mathsf{L}}) = &\ \operatorname{image}(\sigma_{\mathsf{V}} \circ \mathsf{L} \circ \sigma_{\mathsf{U}})\\ &= \{\sigma_{\mathsf{V}} \circ \mathsf{L} \circ \sigma_{\mathsf{U}}(u,v) \mid (u,v) \in \mathsf{U}_{\mathbb{C}}\}\\ &= \{\sigma_{\mathsf{V}} \circ \mathsf{L}(u,-v) \mid (u,v) \in \mathsf{U}_{\mathbb{C}}\}\\ &= \{\sigma_{\mathsf{V}}(\mathsf{L}_1(u) - \mathsf{L}_2(v), -\mathsf{L}_2(u) - \mathsf{L}_1(v)) \mid (u,v) \in \mathsf{U}_{\mathbb{C}}\}\\ &= \{(\mathsf{L}_1(u) - \mathsf{L}_2(v), \mathsf{L}_2(u) + \mathsf{L}_1(v)) \mid (u,v) \in \mathsf{U}_{\mathbb{C}}\}\\ &= \{(\mathsf{L}_1(u) + \mathsf{L}_2(v), \mathsf{L}_2(u) - \mathsf{L}_1(v)) \mid (u,-v) \in \mathsf{U}_{\mathbb{C}}\}\\ &= \{(\mathsf{L}_1(u) + \mathsf{L}_2(v), \mathsf{L}_2(u) - \mathsf{L}_1(v)) \mid (u,v) \in \mathsf{U}_{\mathbb{C}}\}\\ &= \{(u',v') \mid (u',-v') \in \operatorname{image}(\mathsf{L})\},\end{aligned}$$

so giving the second part of the lemma. ▼

Now we proceed with the proof. Let us first consider the case when λ is an eigenvalue for $\mathsf{L}_{\mathbb{C}}$. By the lemma we have

$$\ker(\mathsf{L}_{\mathbb{C},\bar{\lambda}}) = \{(u,-v) \mid (u,v) \in \ker(\mathsf{L}_{\mathbb{C},\lambda})\}.$$

Thus $\bar{\lambda}$ is an eigenvalue for $\mathsf{L}_{\mathbb{C}}$ and

$$\mathsf{W}(\bar{\lambda}, \mathsf{V}_{\mathbb{C}}) = \{(u,v) \in \mathsf{V}_{\mathbb{C}} \mid (u,-v) \in \mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}})\}.$$

Now suppose that λ is a spectral value for $\mathsf{L}_{\mathbb{C}}$. We may as well suppose that λ is not an eigenvalue, and so this means that $\mathsf{L}_{\mathbb{C},\lambda}$ is not surjective. By the lemma above, this also means that $\mathsf{L}_{\mathbb{C},\bar{\lambda}}$ is not surjective, and so $\bar{\lambda}$ is also a spectral value for L.

(iii) This was proved during the course of proving (i).

(iv) We have $\mathsf{L}^j_{\mathbb{C},\lambda}(u,v) = (\mathsf{L}^j_\lambda(u), \mathsf{L}^j_\lambda(v))$ for each $j \in \mathbb{Z}_{>0}$ and $(u,v) \in \mathsf{V}_{\mathbb{C}}$. Therefore,

$$\ker(\mathsf{L}^j_{\mathbb{C},\lambda}) = \{(u,v) \in \mathsf{V}_{\mathbb{C}} \mid u,v \in \ker(\mathsf{L}^j_\lambda)\}.$$

From this we infer that

$$\cup_{j\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^j_{\mathbb{C},\lambda}) = \left\{(u,v) \in \mathsf{V}_{\mathbb{C}} \;\middle|\; u,v \in \cup_{j\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^j_\lambda)\right\},$$

which is the desired result.

(v) This was proved during the course of the proof of part (ii).

(vi) Since $\mathsf{L}_{\mathbb{C},\bar{\lambda}} = \bar{\mathsf{L}}_{\mathbb{C},\lambda}$, it follows from Proposition 5.4.64 that $\mathsf{L}^j_{\mathbb{C},\bar{\lambda}} = \bar{\mathsf{L}}^j_{\mathbb{C},\lambda}$ for each $j \in \mathbb{Z}_{>0}$. From the lemma above we then conclude that, for each $j \in \mathbb{Z}_{>0}$,

$$\ker(\mathsf{L}^j_{\mathbb{C},\bar{\lambda}}) = \left\{(u,v) \in \mathsf{V}_{\mathbb{C}} \;\middle|\; (u,-v) \in \ker(\mathsf{L}^j_{\mathbb{C},\lambda})\right\}.$$

It follows that

$$\cup_{j\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^j_{\mathbb{C},\bar{\lambda}}) = \left\{(u,v) \in \mathsf{V}_{\mathbb{C}} \;\middle|\; (u,-v) \in \cup_{j\in\mathbb{Z}_{>0}} \ker(\mathsf{L}^j_{\mathbb{C},\lambda})\right\},$$

which is exactly the claim. ■

The theorem tells us that every eigenvalue or spectral value of L is also an eigenvalue or spectral value of $\mathsf{L}_{\mathbb{C}}$. Of course, it is not generally the case that eigenvalues or spectral values of $\mathsf{L}_{\mathbb{C}}$ are also eigenvalues or spectral values of L, since the former are allowed to be complex, whereas the latter are always real. Nonetheless, one can wonder what implications the existence of non-real eigenvalues for $\mathsf{L}_{\mathbb{C}}$ has on the structure of L. The following result addresses precisely this point. The essential idea is that eigenspaces for $\mathsf{L}_{\mathbb{C}}$ give rise to invariant subspaces for L of twice the dimension.

5.4.68 Theorem (Real invariant subspaces for complex eigenvalues) *Let V be a $\mathbb{R}$-vector space, let $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$, and let $\mathsf{V}_{\mathbb{C}}$ and $\mathsf{L}_{\mathbb{C}}$ be the corresponding complexifications. Suppose that $\lambda = \mathrm{a}+\mathrm{ib}$, $\mathrm{a}, \mathrm{b} \in \mathbb{R}$, $\mathrm{b} \neq 0$, is a complex eigenvalue for $\mathsf{L}_{\mathbb{C}}$ and let $\mathscr{B}_\lambda$ and $\overline{\mathscr{B}}_\lambda$ be bases for the eigenspace $\mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}})$ and the generalised eigenspace $\overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}})$, respectively. Then the following statements hold:*

(i) the sets

$$\begin{aligned} \mathscr{B}'_\lambda &= \{\mathrm{u} \in \mathsf{V} \mid (\mathrm{u},\mathrm{v}) \in \mathscr{B}_\lambda\} \cup \{\mathrm{v} \in \mathsf{V} \mid (\mathrm{u},\mathrm{v}) \in \mathscr{B}_\lambda\}, \\ \overline{\mathscr{B}}'_\lambda &= \{\mathrm{u} \in \mathsf{V} \mid (\mathrm{u},\mathrm{v}) \in \overline{\mathscr{B}}_\lambda\} \cup \{\mathrm{v} \in \mathsf{V} \mid (\mathrm{u},\mathrm{v}) \in \overline{\mathscr{B}}_\lambda\} \end{aligned} \tag{5.26}$$

are linearly independent;

(ii) if $(\mathrm{u},\mathrm{v}) \in \mathscr{B}_\lambda$ then

$$\mathsf{L}(\mathrm{u}) = \mathrm{au} - \mathrm{bv}, \quad \mathsf{L}(\mathrm{v}) = \mathrm{bu} + \mathrm{av},$$

and so, in particular, the two-dimensional subspace $\mathrm{span}_{\mathbb{R}}(\mathrm{u},\mathrm{v})$ is L-invariant;

(iii) the subspaces $\mathrm{span}_{\mathbb{R}}(\mathscr{B}'_\lambda)$ and $\mathrm{span}_{\mathbb{R}}(\overline{\mathscr{B}}'_\lambda)$ are L-invariant;

(iv) relative to the partition given in (5.26) for the basis $\mathscr{B}'_\lambda$, the restriction of L to $\mathrm{span}_{\mathbb{R}}(\mathscr{B}'_\lambda)$ has the matrix representative

$$\begin{bmatrix} \mathrm{a}\mathbf{I}_{\mathrm{I}} & \mathrm{b}\mathbf{I}_{\mathrm{I}} \\ -\mathrm{b}\mathbf{I}_{\mathrm{I}} & \mathrm{a}\mathbf{I}_{\mathrm{I}} \end{bmatrix},$$

where I is a set for which there exists a bijection $\phi\colon \mathrm{I} \to \mathscr{B}_\lambda$.

Proof (i) We shall prove the result for $\overline{\mathscr{B}}'_\lambda$, the proof for $\mathscr{B}'_\lambda$ being entirely similar. Let us define

$$\overline{\mathscr{B}}_{\bar{\lambda}} = \{(u, -v) \in \mathsf{V}_{\mathbb{C}} \mid (u, v) \in \overline{\mathscr{B}}_\lambda\},$$

noting by Proposition 5.4.60 that $\mathrm{span}_{\mathbb{C}}(\overline{\mathscr{B}}_\lambda) \cap \mathrm{span}_{\mathbb{C}}(\overline{\mathscr{B}}_{\bar{\lambda}}) = \{(0_{\mathsf{V}}, 0_{\mathsf{V}})\}$. Moreover, by Theorem 5.4.67(vi) we also know that $\overline{\mathscr{B}}_{\bar{\lambda}}$ is a basis for $\overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_{\mathbb{C}})$. These facts together ensure that $\overline{\mathscr{B}}_\lambda \cup \overline{\mathscr{B}}_{\bar{\lambda}}$ is a basis for $\overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}}) \oplus \overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_{\mathbb{C}})$. Now let I be an index set for which there exists a bijection $\phi\colon I \to \overline{\mathscr{B}}_\lambda$, define $J = I \mathbin{\mathring{\cup}} I$, and define a column finite matrix $\boldsymbol{P} \in \mathrm{Mat}_{(J \times J)}(\mathsf{F})$ in terms of the natural partition of $I \mathbin{\mathring{\cup}} I$ by

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{I}_I & \boldsymbol{I}_I \\ \boldsymbol{I}_I & -\boldsymbol{I}_I \end{bmatrix}.$$

This matrix is invertible as one can see by checking that it has an inverse given by

$$\boldsymbol{P}^{-1} = \frac{1}{2}\begin{bmatrix} \boldsymbol{I}_I & \boldsymbol{I}_I \\ \boldsymbol{I}_I & -\boldsymbol{I}_I \end{bmatrix}.$$

Thus $\boldsymbol{P}$ is a change of basis matrix from the basis $\overline{\mathscr{B}}_\lambda \cup \overline{\mathscr{B}}_{\bar{\lambda}}$ for $\overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}}) \oplus \overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_{\mathbb{C}})$ to another basis for $\overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}}) \oplus \overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_{\mathbb{C}})$. Using the definition of the change of basis matrix one can further check that this new basis is exactly

$$\{(u, 0_{\mathsf{V}}) \mid (u, v) \in \overline{\mathscr{B}}_\lambda\} \cup \{(0_{\mathsf{V}}, v) \mid (u, v) \in \overline{\mathscr{B}}_\lambda\}. \tag{5.27}$$

Using the fact that this is a basis, and so linearly independent, we now prove that $\overline{\mathscr{B}}'_\lambda$ is linearly independent. Let

$$u_1, \dots, u_k \in \{u \in \mathsf{V} \mid (u, v) \in \overline{\mathscr{B}}_\lambda\}, \quad v_1, \dots, v_l \in \{u \in \mathsf{V} \mid (u, v) \in \overline{\mathscr{B}}_\lambda\}$$

and suppose that

$$a_1 u_1 + \dots + a_k u_k + b_1 v_1 + \dots + b_l v_l = 0_{\mathsf{V}}$$

for $a_1, \dots, a_k, b_1, \dots, b_l \in \mathsf{F}$. Using the definition of scalar multiplication in $\mathsf{V}_{\mathbb{C}}$ this implies that

$$(a_1 + \mathrm{i}0)(u_1, 0_{\mathsf{V}}) + \dots + (a_k + \mathrm{i}0)(u_k, 0_{\mathsf{V}}) \\ + (b_1 + \mathrm{i}0)(0_{\mathsf{V}}, v_1) + \dots + (b_l + \mathrm{i}0)(0_{\mathsf{V}}, v_l) = (0_{\mathsf{V}}, 0_{\mathsf{V}}).$$

Since the set if (5.27) is linearly independent we must have $a_j + \mathrm{i}0 = 0 + \mathrm{i}0$ for $j \in \{1, \dots, k\}$ and $b_j + \mathrm{i}0 = 0 + \mathrm{i}0$ for $j \in \{1, \dots, l\}$. This gives linear independence of $\overline{\mathscr{B}}'_\lambda$.

(ii) If $(u, v) \in \mathscr{B}_\lambda$ then we have

$$(\mathsf{L}(u), \mathsf{L}(v)) = \mathsf{L}_{\mathbb{C}}(u, v) = (a + ib)(u, v) = (au - bv, bu + av),$$

as claimed. Since $\mathsf{L}(u), \mathsf{L}(v) \in \operatorname{span}_{\mathbb{R}}(u, v)$, it follows that $\operatorname{span}_{\mathbb{R}}(u, v)$ is L-invariant.

(iii) To prove this part of the result it is useful to employ a lemma that captures the essence of what is going on.

1 **Lemma** *Let V be a $\mathbb{R}$-vector space with complexification $\mathsf{V}_{\mathbb{C}}$ and let $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$ have complexification $\mathsf{L}_{\mathbb{C}}$. If U is a subspace of $\mathsf{V}_{\mathbb{C}}$ which is invariant under $\mathsf{L}_{\mathbb{C}}$ then*

(i) the subspace

$$\overline{\mathsf{U}} = \{(\mathrm{u}, -\mathrm{v}) \mid (\mathrm{u}, \mathrm{v}) \in \mathsf{U}\}$$

is $\mathsf{L}_{\mathbb{C}}$-invariant and

(ii) the subspaces

$$\{\mathrm{u} \in \mathsf{V} \mid (\mathrm{u}, \mathrm{v}) \in \mathsf{U} + \overline{\mathsf{U}}\}, \quad \{\mathrm{v} \in \mathsf{V} \mid (\mathrm{u}, \mathrm{v}) \in \mathsf{U} + \overline{\mathsf{U}}\}$$

of V are invariant under L.

Proof (i) Let $(u, -v) \in \overline{\mathsf{U}}$ for $(u, v) \in \mathsf{U}$. Then $(\mathsf{L}(u), \mathsf{L}(v)) \in \mathsf{U}$ since U is $\mathsf{L}_{\mathbb{C}}$-invariant. Therefore

$$\mathsf{L}_{\mathbb{C}}(u, -v) = (\mathsf{L}(u), -\mathsf{L}(v)) \in \overline{\mathsf{U}},$$

giving invariance of $\overline{\mathsf{U}}$ under $\mathsf{L}_{\mathbb{C}}$ as desired.

(ii) Let $u \in \{u' \in \mathsf{V} \mid (u', v') \in \mathsf{U} + \overline{\mathsf{U}}\}$ and let $v \in \mathsf{V}$ have the property that $(u, v) \in \mathsf{U} + \overline{\mathsf{U}}$. Then $(u, -v) \in \mathsf{U} + \overline{\mathsf{U}}$. Since $\mathsf{U} + \overline{\mathsf{U}}$ is $\mathsf{L}_{\mathbb{C}}$-invariant,

$$(\mathsf{L}(u), \mathsf{L}(v)), (\mathsf{L}(u), -\mathsf{L}(v)) \in \mathsf{U} + \overline{\mathsf{U}}.$$

Therefore $(2\mathsf{L}(u), 0_{\mathsf{V}}) \in \mathsf{U} + \overline{\mathsf{U}}$ and so $\mathsf{L}(u) \in \{u' \in \mathsf{V} \mid (u', v') \in \mathsf{U} + \overline{\mathsf{U}}\}$, giving invariance of $\{u' \in \mathsf{V} \mid (u', v') \in \mathsf{U} + \overline{\mathsf{U}}\}$ under L. A similar computation gives invariance of $\{v' \in \mathsf{V} \mid (u', v') \in \mathsf{U} + \overline{\mathsf{U}}\}$ under L. ▼

By applying the lemma with $\mathsf{U} = \mathsf{W}(\lambda, \mathsf{L}_{\mathbb{C}})$ and then with $\mathsf{U} = \overline{\mathsf{W}}(\lambda, \mathsf{L}_{\mathbb{C}})$, this part of the theorem follows.

(iv) Let us represent elements of $J = I \overset{\circ}{\cup} I$ by $(\{a\}, i)$ where $a \in \{1, 2\}$ and where $i \in I$. Thus $\{a\}$ indicates whether i is to be thought of as in the first or second copy of I. Let us then write

$$\mathscr{B}_\lambda = \{(u_i, v_i) \mid i \in I\},$$

so indexing the basis elements for $\mathscr{B}_\lambda$. The basis $\mathscr{B}'_\lambda$ can then be written as

$$\mathscr{B}'_\lambda = \{u_i \mid i \in I\}, \quad \{v_i \mid i \in I\}.$$

Let $\psi \colon J \to \mathscr{B}'_\lambda$ be defined by

$$\psi(\{a\}, i) = \begin{cases} u_i, & a = 1, \\ v_i, & a = 2, \end{cases}$$

this giving the bijection of J with $\mathscr{B}'_\lambda$ as per the partition. We then have

$$\mathsf{L}(\psi(\{1\}, i) = a\psi(\{1\}, i) - b\psi(\{2\}, i), \quad \mathsf{L}(\psi(\{2\}, i) = b\psi(\{1\}, i) + a\psi(\{2\}, i).$$

Using the definition of matrix representative, this then gives the matrix representative of $\mathsf{L}|\operatorname{span}_{\mathbb{R}}(\mathscr{B}'_\lambda)$ to be

$$\begin{bmatrix} a\boldsymbol{I}_I & b\boldsymbol{I}_I \\ -b\boldsymbol{I}_I & a\boldsymbol{I}_I \end{bmatrix},$$

as desired. ■

The idea is that, for every one-dimensional (over $\mathbb{C}$) subspace of $\mathsf{V}_{\mathbb{C}}$ that is invariant under $\mathsf{L}_{\mathbb{C}}$ there corresponds a two-dimensional (over $\mathbb{R}$) subspace of V that is invariant under L. Moreover, if as a basis for this two-dimensional real subspace one chooses the real and imaginary parts of the basis for the one-dimensional complex subspace, then the representation of L in this two-dimensional real subspace is related to the complex eigenvalue in a simple way (i.e., as in part (ii)). Some geometric intuition concerning this will form part of our understanding of linear ordinary differential equations in Section V-5.2.2.

5.4.11 Linear maps on vector spaces extended by scalars

In the preceding section we indicated how linear maps between $\mathbb{R}$-vector spaces could be extended to linear maps between the complexifications. In Section 4.5.8 we saw how the notion of complexification of a vector space could be generalised to arbitrary extensions of a field. In this section we perform the analogue of the preceding section for general field extensions. Thus we let U and V be F-vector spaces with $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$, we let K be an extension of F, and we indicate how to construct, in a natural way, a linear map $\mathsf{L}_{\mathsf{K}} \in \mathrm{Hom}_{\mathsf{K}}(\mathsf{U}_{\mathsf{K}};\mathsf{V}_{\mathsf{K}})$. This construction depends on understanding the tensor product definition of the extended vector spaces U_{K} and V_{K}. This construction was made in Section 4.5.8, and relies on the tensor product that we will not get to until Section 5.6.3.

Let us first indicate that a natural construction at least fits part of the bill.

5.4.69 Proposition (A linear map between extended vector spaces) *Let F be a field, let U and V be F-vector spaces, and let K be a field extension of F with $\mathsf{U}_{\mathsf{K}} = \mathsf{K} \otimes \mathsf{U}$ and $\mathsf{V}_{\mathsf{K}} = \mathsf{K} \otimes \mathsf{V}$ the corresponding K-vector spaces. If $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ then there exists a unique linear map $\mathsf{L}_{\mathsf{K}} \in \mathrm{Hom}_{\mathsf{K}}(\mathsf{U}_{\mathsf{K}};\mathsf{V}_{\mathsf{K}})$ satisfying $\mathsf{L}_{\mathsf{K}}(\mathrm{a} \otimes \mathrm{u}) = \mathrm{a} \otimes \mathsf{L}(\mathrm{u})$ for $\mathrm{a} \in \mathsf{K}$ and $\mathrm{u} \in \mathsf{U}$.*

Proof Define $\phi_{\mathsf{L}} : \mathsf{K} \times \mathsf{U} \to \mathsf{V}_{\mathsf{K}}$ by $\phi_{\mathsf{L}}(a, u) = a \otimes \mathsf{L}(u)$. Since this map is easily shown to be bilinear, there exists a unique homomorphism $\mathsf{L}_{\mathsf{K}}(\mathsf{U}_{\mathsf{K}};\mathsf{V}_{\mathsf{K}})$ satisfying $\mathsf{L}_{\mathsf{K}}(a \otimes u) = a \otimes \mathsf{L}(u)$. It remains to show that L_{K} is not just F-linear, but K-linear. To see this we compute

$$\mathsf{L}_{\mathsf{K}}(b(a \otimes u)) = \mathsf{L}_{\mathsf{K}}((ba) \otimes u) = (ba) \otimes \mathsf{L}(u) = b(a \otimes \mathsf{L}(u)) = b\mathsf{L}_{\mathsf{K}}(a \otimes u),$$

so L_{K} commutes with scalar multiplication in K. It is trivial that L_{K} commutes with vector addition. ■

Let us make a definition based on this result.

5.4.70 Definition (Extension of scalars for a linear map) Let F be a field, let U and V be F-vector spaces, and let K be a field extension of F with $\mathsf{U}_{\mathsf{K}} = \mathsf{K} \otimes \mathsf{U}$ and $\mathsf{V}_{\mathsf{K}} = \mathsf{K} \otimes \mathsf{V}$ the corresponding K-vector spaces. The linear map L_{K} of Proposition 5.4.69 is the ***extension*** of L by K. •

Let us illustrate that this idea does indeed generalise complexification of linear maps.

5.4.71 Example ($\mathsf{L}_{\mathbb{C}} = \mathsf{L}_{\mathbb{C}}$) Let U and V be $\mathbb{R}$-vector spaces and let $\mathsf{L} \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{U};\mathsf{V})$. We now have two possibly competing notions for the notation $\mathsf{L}_{\mathbb{C}}$: one from Section 5.4.10 and one from Definition 5.4.70. Let us show that these are the same, given the isomorphism $\iota_{\mathbb{C}}$ of Example 4.5.66. For the moment, in order to distinguish the two linear maps, let us denote by $\tilde{\mathsf{L}}_{\mathbb{C}}$ the complexification of L as in Section 5.4.10. Let us also denote by $\mathsf{U}_{\mathbb{C}}$ and $\mathsf{V}_{\mathbb{C}}$ the complexifications as in Section 4.5.7 and by $\mathbb{C} \otimes \mathsf{U}$ and $\mathbb{C} \otimes \mathsf{V}$ the extensions as in Definition 4.5.64. We claim that the following diagram commutes:

$$\begin{tikzcd}
\mathsf{U}_{\mathbb{C}} \arrow[r, "\tilde{\mathsf{L}}_{\mathbb{C}}"] \arrow[d, "\iota_{\mathbb{C}}"'] & \mathsf{V}_{\mathbb{C}} \arrow[d, "\iota_{\mathbb{C}}"] \\
\mathbb{C} \otimes \mathsf{U} \arrow[r, "\mathsf{L}_{\mathbb{C}}"'] & \mathbb{C} \otimes \mathsf{V}
\end{tikzcd}$$

Let us show this explicitly:

$$\iota_{\mathbb{C}} \circ \tilde{\mathsf{L}}_{\mathbb{C}}(u_1, u_2) = \iota_{\mathbb{C}}(\mathsf{L}(u_1), \mathsf{L}(u_2)) = 1 \otimes \mathsf{L}(u_1) + \mathrm{i} \otimes \mathsf{L}(u_2)$$

and

$$\mathsf{L}_{\mathbb{C}} \circ \iota_{\mathbb{C}}(u_1, u_2) = \mathsf{L}_{\mathbb{C}}(1 \otimes u_1 + \mathrm{i} \otimes u_2) = 1 \otimes \mathsf{L}(u_1) + \mathrm{i} \otimes \mathsf{L}(u_2),$$

as desired.

The reader may now feel free to think of $\mathsf{U}_{\mathbb{C}}$, $\mathsf{V}_{\mathbb{C}}$, and $\mathsf{L}_{\mathbb{C}}$ in whatever of the two ways they choose. •

Exercises

5.4.1 Let F be a field, let V be an F-vector space, let U be a subspace of V, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Show that, if W_1 and W_2 are two L-invariant subspaces containing U, then $\mathsf{W}_1 \cap \mathsf{W}_2$ is itself an L-invariant subspace containing U.

5.4.2 Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Show that if $\mathsf{U} \subseteq \mathsf{V}$ is an L-invariant subspace then U is an L^j-invariant subspace for $j \in \mathbb{Z}_{\geq 0}$.

5.4.3 Prove Proposition 5.4.9.

5.4.4 Prove Proposition 5.4.10.

5.4.5 Let F be a field and let $\mathsf{U}_1, \dots, \mathsf{U}_r, \mathsf{V}_1, \dots, \mathsf{V}_s$ be F-vector spaces. Show that if

$$\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}_1 \oplus \cdots \oplus \mathsf{U}_r; \mathsf{V}_1 \oplus \cdots \oplus \mathsf{V}_s),$$

then there exists unique linear maps $\mathsf{L}_{jk} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}_j; \mathsf{V}_k)$, $j \in \{1, \dots, r\}$, $k \in \{1, \dots, s\}$, such that, for each $u_j \in \mathsf{U}_j$, $j \in \{1, \dots, r\}$,

$$\mathsf{L}(u_1 + \cdots + u_r) = (\mathsf{L}_{11}(u_1) + \cdots + \mathsf{L}_{1r}(u_r)) + \cdots + (\mathsf{L}_{s1}(u_1) + \cdots + \mathsf{L}_{sr}(u_r)).$$

5.4.6 Let F be a field and let V be an F-vector space. A linear mapping $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V}; \mathsf{V})$ is ***idempotent*** if $\mathsf{L} \circ \mathsf{L} = \mathsf{L}$. Show that, if L is idempotent, then $\mathsf{V} = \mathrm{image}(\mathsf{L}) \oplus \ker(\mathsf{L})$.

5.4.7 Prove Proposition 5.4.16.

5.4.8 Let F be a field and let U and V be finite-dimensional F-vector spaces. Show that

$$\dim_{\mathsf{F}}(\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})) = \dim_{\mathsf{F}}(\mathsf{U}) \cdot \dim_{\mathsf{F}}(\mathsf{U}).$$

5.4.9 Find a field F, an F-vector space V, and endomorphisms $\mathsf{L}, \mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ which do not commute.

5.4.10 Let F be a field, let V be an F-vector space, and let $\mathsf{L}, \mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. For $m_{\mathsf{L}}, m_{\mathsf{K}} \in \mathbb{Z}_{>0}$ let $r_1, \dots, r_k, p_1, \dots, p_k \in \{0, 1, \dots, m_{\mathsf{L}}\}$ and $s_1, \dots, s_l, q_1, \dots, q_l \in \{0, 1, \dots, m_{\mathsf{K}}\}$ have the property that

$$\sum_{j=1}^{k} r_j = \sum_{j=1}^{l} p_j = m_{\mathsf{L}}, \quad \sum_{j=1}^{k} s_j = \sum_{j=1}^{l} q_j = m_{\mathsf{K}}.$$

Show that, if L and K commute, then

$$\mathsf{L}^{r_1} \circ \mathsf{K}^{s_1} \circ \dots \circ \mathsf{L}^{r_1} \circ \mathsf{K}^{s_k} = \mathsf{L}^{p_1} \circ \mathsf{K}^{q_1} \circ \dots \circ \mathsf{L}^{p_1} \circ \mathsf{K}^{q_k},$$

i.e., show that arbitrary powers of L and K commute.

5.4.11 Let F be a field, let V be an F-vector space, let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$, and let $\mathsf{U} \subseteq \mathsf{V}$ be an L-invariant subspace. Show that if $\mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ commutes with L, then U is K-invariant.

In the next exercise you will need the following definition.

Definition (Homothety) Let F be a field, let $\lambda \in \mathsf{F}$, and let V be an F-vector space. The ***homothety*** of ratio λ is the endomorphism of V given by $v \mapsto \lambda v$. •

5.4.12 Let F be a field and let V be a finite-dimensional F-vector space. Show that $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ commutes with *every* linear map $\mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ if and only if L is a homothety of ratio λ for some $\lambda \in \mathsf{F}$.

5.4.13 Let F be a field, let V be an F-vector space, and let $\mathsf{L}, \mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Show that, if L and K commute, then $\ker(\mathsf{L})$ is K-invariant and that $\ker(\mathsf{K})$ is L-invariant.

5.4.14 Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Show that there exists a basis $\mathscr{B}$ such that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is upper triangular if and only if there exists a basis $\mathscr{B}'$ such that $[\mathsf{L}]_{\mathscr{B}'}^{\mathscr{B}'}$ is lower triangular.

For the following two exercises the reader will wish to recall Definitions 5.1.61 and the definition preceding Exercise 5.3.4.

5.4.15 Let F be a field and let V be an F-vector space. On $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ define the product

$$[\mathsf{L}, \mathsf{K}] = \mathsf{L} \circ \mathsf{K} - \mathsf{K} \circ \mathsf{L}.$$

Answer the following questions.

(a) Show that $(\mathrm{End}_{\mathsf{F}}(\mathsf{V}), [\cdot, \cdot])$ is an F-Lie algebra.

(b) Show that if F does not have characteristic 2 then $[\mathsf{L}, \mathsf{K}] = -[\mathsf{K}, \mathsf{L}]$ for every $\mathsf{L}, \mathsf{K} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$.

5.4.16 Let F be a field and let V be an F-vector space.

(a) Show that the set of invertible endomorphisms of V is a group with group operation given by composition.

This group is called the ***general linear group*** of V and is denoted by $\mathsf{GL}(\mathsf{V})$.

(b) Is $\mathsf{GL}(\mathsf{V})$ a subalgebra of $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$?

5.4.17 Let F be a field and let V be an n-dimensional F-vector space. We consider $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ as a ring by Corollary 5.4.18.

(a) For what values of n is it true that $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is a commutative ring?

(b) For what values of n is it true that $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is an integral domain?

5.4.18 Let F be a field, let V be a finite-dimensional F-vector space, and consider the F-lie algebra $(\mathrm{End}_{\mathsf{F}}(\mathsf{V}), [\cdot,\cdot])$ from Exercise 5.4.15.

(a) Show that the set of endomorphisms of trace 0_{F} is a Lie subalgebra of $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$.

(b) Suppose that the map $m_n\colon v \mapsto nv$ (by nv we mean the n-fold sum of v with itself) is an isomorphism of V. Define $\mathrm{Tr}_0\colon \mathrm{End}_{\mathsf{F}}(\mathsf{V}) \to \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by

$$\mathrm{Tr}_0(\mathsf{L}) = \mathsf{L} - \mathrm{tr}(\mathsf{L}) m_n^{-1} \circ \mathrm{id}_{\mathsf{V}}\,.$$

Show that $\mathrm{tr}(\mathrm{Tr}_0(\mathsf{L})) = 0_{\mathsf{F}}$ for all $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$.

5.4.19 Let F be a field, let V be a finite-dimensional F-vector space, and recall from Exercise 5.4.16 that $\mathsf{GL}(\mathsf{V})$ denotes the group of invertible endomorphisms of V.

(a) Show that the subset of endomorphisms with determinant 1_{F} is a subgroup of $\mathsf{GL}(\mathsf{V})$.

This subgroup of invertible endomorphisms with determinant 1_{F} is denoted by $\mathsf{SL}(\mathsf{V})$ and is called the ***special linear group*** of V.

(b) Is $\mathsf{SL}(\mathsf{V})$ a subalgebra of $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$?

5.4.20 Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Show that, if there exists a basis $\mathscr{B}$ for V such that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is upper triangular, then the eigenvalues of L are the diagonal entries of the matrix $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$.

5.4.21 Let F be a field, let V be an F-vector space, let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ and let $k \in \mathbb{Z}_{\geq 0}$. Show that if $\lambda \in \mathsf{F}$ is an eigenvalue for L with eigenvector v, then λ^k is an eigenvalue of L^k with eigenvector v.

5.4.22 Let V be a $\mathbb{R}$-vector space and let $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$. Show that L is invertible if and only if $\mathsf{L}_{\mathbb{C}}$ is invertible, and that in case L is invertible, $\mathsf{L}_{\mathbb{C}}^{-1}(u, v) = (\mathsf{L}^{-1}(u), \mathsf{L}^{-1}(v))$.

5.4.23 Let V be a $\mathbb{R}$-vector space with $\mathsf{V}_{\mathbb{C}}$ its complexification. Show that the map i_{V} is both a $\mathbb{R}$-linear map and a $\mathbb{C}$-linear map of $\mathsf{V}_{\mathbb{C}}$.

5.4.24 Let V be a $\mathbb{R}$-vector space and let $\mathscr{B}$ be a basis for V.

(a) Show that

$$\mathscr{B}' = \{(u, 0_{\mathsf{V}}) \mid u \in \mathscr{B}\} \cup \{(0_{\mathsf{V}}, u) \mid u \in \mathscr{B}\}$$

is a basis for the $\mathbb{R}$-vector space $\mathsf{V}_{\mathbb{C}}$.

(b) Show that, with respect to the partition of the basis $\mathscr{B}'$ for the $\mathbb{R}$-vector space $\mathsf{V}_{\mathbb{C}}$, the matrix representative of σ_{V} is

$$[\sigma_{\mathsf{V}}]^{\mathscr{B}'}_{\mathscr{B}'} = \begin{bmatrix} \boldsymbol{I}_I & \boldsymbol{0}_{I\times I} \\ \boldsymbol{0}_{I\times I} & -\boldsymbol{I}_I \end{bmatrix},$$

where I is a set for which there exists a bijection $\phi\colon I \to \mathscr{B}$.

(c) Show that, with respect to the partition of the basis $\mathscr{B}'$ for the $\mathbb{R}$-vector space $\mathsf{V}_{\mathbb{C}}$, the matrix representative of σ_{V} is

$$[\sigma_{\mathsf{V}}]^{\mathscr{B}'}_{\mathscr{B}'} = \begin{bmatrix} \boldsymbol{0}_{I\times I} & -\boldsymbol{I}_I \\ \boldsymbol{I}_I & \boldsymbol{0}_{I\times I} \end{bmatrix},$$

where I is a set for which there exists a bijection $\phi\colon I \to \mathscr{B}$.

5.4.25 Let V be a $\mathbb{R}$-vector space with $\mathsf{V}_{\mathbb{C}}$ the complexification. If $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$, show that $\bar{\mathsf{L}} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$.

5.4.26 Prove Proposition 5.4.64.

5.4.27 Let U and V be $\mathbb{R}$-vector spaces. Show that $\mathsf{L} \in \mathrm{Hom}_{\mathbb{C}}(\mathsf{U}_{\mathbb{C}}; \mathsf{V}_{\mathbb{C}})$ is the complexification of a $\mathbb{R}$ linear map from U to V if and only if L maps the subset $\{(u, 0_{\mathsf{U}}) \mid u \in \mathsf{U}\}$ of real vectors in U to the subset $\{(v, 0_{\mathsf{V}}) \mid v \in \mathsf{V}\}$ of real vectors in V.

Section 5.5

Linear algebra over rings

In this section we will study linear algebra over rings, i.e., homomorphisms of modules. A general discussion here would take us very far afield, and indeed to places where many parts of the general theory are topics of current research. The theory of modules is significantly more complicated than the theory of vector spaces, and this is reflected when one studies homomorphisms in each case. Thus we will only adapt the results of Section 5.4 in cases where it is profitable or insightful for us to do so. We shall be more hasty in this section than we were in Section 5.4, and we assume that the reader already has reaped what benefit there is from reading the presentation of linear algebra over fields.

Do I need to read this section? This section can be omitted until the results are subsequently needed. •

5.5.1 Submodules and modules associated to a homomorphism

In this section we let R be a ring and let M and N be (left or right) R-modules. We will denote a typical element of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ by L.

As with linear maps between vector spaces, a homomorphism of modules has various submodules and modules associated to it. What is not true for homomorphisms of modules is that the dimensions of these submodules and modules can be defined. Thus we cannot generally define the notions of rank, nullity, or defect for homomorphisms.

5.5.1 Definition (Image, kernel, cokernel, coimage) Let R be a ring, let M and N be (left or right) R-modules, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$.

(i) The ***image*** of L is the submodule of N given by $\mathrm{image}(\mathsf{L}) = \{\mathsf{L}(x) \mid x \in \mathsf{M}\}$.

(ii) The ***kernel*** of L is the submodule of M given by $\ker(\mathsf{L}) = \{x \in \mathsf{M} \mid \mathsf{L}(x) = 0_{\mathsf{N}}\}$.

(iii) The ***cokernel*** of L is the R-module $\mathrm{coker}(\mathsf{L}) = \mathsf{N}/\,\mathrm{image}(\mathsf{L})$.

(iv) The ***coimage*** of L is the R-vector space $\mathrm{coimage}(\mathsf{L}) = \mathsf{M}/\ker(\mathsf{L})$. •

As with vector spaces, there are natural injective, surjective, and bijective homomorphisms associated with the various submodules associated with a homomorphism. The proof of the following theorem is just like its counterpart, Theorem 5.4.2, for vector spaces.

5.5.2 Theorem (Homomorphisms derived from a homomorphism) *Let R be a ring, let M and N be (left or right) R-modules, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$. Then the following statements hold:*

(i) *there exists a unique homomorphism* $\mathsf{L}_{\mathrm{inj}} \in \mathrm{Hom}_{\mathsf{R}}(\mathrm{coimage}(\mathsf{L}); \mathsf{N})$ *such that the diagram*

$$\begin{tikzcd}
\mathsf{M} \arrow[rr, "\mathsf{L}"] \arrow[dr, "\pi_{\mathrm{coimage}(\mathsf{L})}"'] & & \mathsf{N} \\
& \mathrm{coimage}(\mathsf{L}) \arrow[ur, "\mathsf{L}_{\mathrm{inj}}"'] &
\end{tikzcd}$$

commutes, where $\pi_{\mathrm{coimage}(\mathsf{L})} \colon \mathsf{M} \to \mathrm{coimage}(\mathsf{L})$ *is the canonical projection;*

(ii) *there exists a unique homomorphism* $\mathsf{L}_{\mathrm{srj}} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathrm{image}(\mathsf{L}))$ *such that the diagram*

$$\begin{tikzcd}
\mathsf{M} \arrow[rr, "\mathsf{L}"] \arrow[dr, "\mathsf{L}_{\mathrm{srj}}"'] & & \mathsf{N} \\
& \mathrm{image}(\mathsf{L}) \arrow[ur, "\mathrm{i}_{\mathrm{image}(\mathsf{L})}"'] &
\end{tikzcd}$$

commutes;

(iii) *there exists a unique homomorphism* $\mathsf{L}_{\mathrm{bij}} \in \mathrm{Hom}_{\mathsf{R}}(\mathrm{coimage}(\mathsf{L}); \mathrm{image}(\mathsf{L}))$ *such that the diagram*

$$\begin{tikzcd}
\mathsf{M} \arrow[r, "\mathsf{L}"] \arrow[d, "\pi_{\mathrm{coimage}(\mathsf{L})}"'] & \mathsf{U} \\
\mathrm{coimage}(\mathsf{L}) \arrow[r, "\mathsf{L}_{\mathrm{bij}}"'] & \mathrm{image}(\mathsf{L}) \arrow[u, "\mathrm{i}_{\mathrm{image}(\mathsf{L})}"']
\end{tikzcd}$$

commutes.

Moreover, $\mathsf{L}_{\mathrm{inj}}$ *is injective,* $\mathsf{L}_{\mathrm{srj}}$ *is surjective, and* $\mathsf{L}_{\mathrm{bij}}$ *is bijective.*

Analogously to Corollary 5.4.3 for linear maps between vector spaces, a homomorphism L is injective if and only if $\mathrm{coimage}(\mathsf{L}) = \{0\}$ and is surjective if and only if $\mathrm{coker}(\mathsf{L}) = \{0\}$.

5.5.2 The algebra of homomorphisms

In this section we investigate the algebraic structure of sets of homomorphisms and endomorphisms. The picture here is much like that for linear maps over fields given in Section 5.4.3.

5.5.3 Definition (Sum and scalar multiplication for homomorphisms) Let R be a ring, let M and N be left (resp. right) R-modules, and let $\mathsf{A}, \mathsf{B} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{N})$ and $r \in \mathsf{R}$.

(i) The ***sum*** of A and B is the element $\mathsf{A} + \mathsf{B}$ of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{N})$ defined by

$$(\mathsf{A} + \mathsf{B})(x) = \mathsf{A}(x) + \mathsf{B}(x).$$

(ii) ***Multiplication*** of A with r is the element $r\mathsf{A}$ (resp. $\mathsf{A}r$) of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{N})$ defined by $(r\mathsf{A})(x) = r(\mathsf{A}(x))$ (resp. $(\mathsf{A}r)(x) = (\mathsf{A}(x))r$). •

Of course, just like linear maps between vector spaces, homomorphisms of modules have a natural product given by composition. Let us indicate the relationships between the operations of addition, scalar multiplication, and product, using the notation $0_{\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})}, -\mathsf{A} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ defined by

$$0_{\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})}(x) = 0_{\mathsf{N}}, \quad -\mathsf{A}(x) = -(\mathsf{A}(x)), \qquad x \in \mathsf{M}.$$

We then have the following.

5.5.4 Proposition (Properties of sum and composition of homomorphisms) *Let* R *be a ring, let* $\mathsf{M}, \mathsf{N}, \mathsf{P}$*, and* Q *be left (resp. right)* R*-modules, and let* $\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$*,* $\mathsf{B}_1, \mathsf{B}_2 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N};\mathsf{P})$*,* $\mathsf{C}_1 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{P};\mathsf{Q})$*, and* $r_1, r_2 \in \mathsf{R}$*. Then the following equalities hold:*

(i) $\mathsf{A}_1 + \mathsf{A}_2 = \mathsf{A}_2 + \mathsf{A}_1$;

(ii) $(\mathsf{A}_1 + \mathsf{A}_2) + \mathsf{A}_3 = \mathsf{A}_1 + (\mathsf{A}_2 + \mathsf{A}_3)$;

(iii) $\mathsf{A}_1 + 0_{\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})} = \mathsf{A}_1$;

(iv) $\mathsf{A}_1 + (-\mathsf{A}_1) = 0_{\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})}$;

(v) $\mathsf{B}_1 \circ (\mathsf{A}_1 + \mathsf{A}_2) = \mathsf{B}_1 \circ \mathsf{A}_1 + \mathsf{B}_1 \circ \mathsf{A}_2$;

(vi) $(\mathsf{B}_1 + \mathsf{B}_2) \circ \mathsf{A}_1 = \mathsf{B}_1 \circ \mathsf{A}_1 + \mathsf{B}_2 \circ \mathsf{A}_1$;

(vii) $(\mathsf{C}_1 \circ \mathsf{A}_1) \circ \mathsf{B}_1 = \mathsf{C}_1 \circ (\mathsf{A}_1 \circ \mathsf{B}_1)$;

(viii) $r_1(r_2\mathsf{A}_1) = (r_1r_2)\mathsf{A}_1$ *(resp.* $(\mathsf{A}_1r_1)r_2 = \mathsf{A}_1(r_1r_2)$*)*;

(ix) $(r_1 + r_2)\mathsf{A}_1 = r_1\mathsf{A}_1 + r_2\mathsf{A}_1$ *(resp.* $\mathsf{A}_1(r_1 + r_2) = \mathsf{A}_1r_1 + \mathsf{A}_1r_2$*)*;

(x) $r_1(\mathsf{A}_1 + \mathsf{A}_2) = r_1\mathsf{A}_1 + r_1\mathsf{A}_2$ *(resp.* $(\mathsf{A}_1 + \mathsf{A}_2)r_1 = \mathsf{A}_1r_1 + \mathsf{A}_2r_1$*)*.

Proof This is Exercise 5.5.1. ■

This then gives the following consequences, analogous to Corollaries 5.4.17 and 5.4.18.

5.5.5 Corollary (Homomorphisms as elements of a module) *If* R *is a ring and if* M *and* N *are left (resp. right)* R*-modules, then* $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ *is a left (resp. right)* R*-module with addition given by the sum of homomorphisms and with multiplication being given by multiplication of a matrix by a scalar.*

5.5.6 Corollary (Homomorphisms as elements of an algebra) *If* R *is a ring and if* M *is a left (resp. right)* R*-module, then* $\mathrm{End}_{\mathsf{R}}(\mathsf{M})$ *is a left (resp. right)* R*-algebra with the module structure of Corollary 5.5.5 and with the product given by the composition of linear maps.*

5.5.3 Homomorphisms and matrices

Thus far in this section we have not seen much difference between linear algebra over rings as opposed to that over fields. However, given the caveats expressed in Sections 4.8 and 5.2, it should come as no surprise that there are substantial differences between linear algebra over fields and that over rings. In this section we experience one of these significant differences. For investigating linear maps between vector spaces, there is a perfect correspondence between such maps and matrices, as discussed in Section 5.4.4. For modules one cannot expect that to be true since modules are not generally free. In this section we say what *can* be said about the correspondence between homomorphisms and matrices.

First let us indicate the basic association we can make. We let R be a ring and let M and N be free left (resp. right) R-modules. We let $\mathscr{B}_{\mathsf{M}}$ and $\mathscr{B}_{\mathsf{N}}$ be bases for M and N, respectively. We do not worry about the fact that M and/or N may possess bases with different cardinalities; we just fix some basis and go with it. We let I and J be index sets for which there exists bijections $\phi_{\mathsf{M}}\colon I \to \mathscr{B}_{\mathsf{M}}$ and $\phi_{\mathsf{N}}\colon J \to \mathscr{B}_{\mathsf{N}}$. Let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$. For $x \in \mathscr{B}_{\mathsf{M}}$ denote $j = \phi_{\mathsf{M}}^{-1}(x) \in J$. Now, just as for our definition of the matrix for a linear map, we can write

$$\mathsf{L}(x) = \sum_{i\in I} r_{ij}\phi_{\mathsf{N}}(i) \quad \left(\text{resp. } \mathsf{L}(x) = \sum_{i\in I} \phi_{\mathsf{N}}(i)r_{ij}\right),$$

for suitable uniquely defined constants $r_{ij} \in \mathsf{R}$, $i \in I$, only finitely many of which are nonzero. This then gives a matrix $[L]^{:}_{\mathscr{B}_{\mathsf{M}}}\mathscr{B}_{\mathsf{N}} \to I \times J\mathsf{R}$ defined by $[\mathsf{L}]^{\mathscr{B}_{\mathsf{N}}}_{\mathscr{B}_{\mathsf{M}}}(i,j) = r_{ij}$. The following definition formally records this construction.

5.5.7 Definition (Matrix representative of homomorphism relative to bases) Let R be a ring, let M and N be (left or right) R-modules, let $\mathscr{B}_{\mathsf{M}}$ and $\mathscr{B}_{\mathsf{N}}$ be bases for M and N, respectively, and let I and J be sets for which there exist bijections $\phi_{\mathsf{M}}\colon J \to \mathscr{B}_{\mathsf{M}}$ and $\phi_{\mathsf{N}}\colon I \to \mathscr{B}_{\mathsf{N}}$. If $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ then the ***matrix representative*** of L with respect to the bases $\mathscr{B}_{\mathsf{M}}$ and $\mathscr{B}_{\mathsf{N}}$ and to the bijections ϕ_{M} and ϕ_{N} is the map $[\mathsf{L}]^{\mathscr{B}_{\mathsf{N}}}_{\mathscr{B}_{\mathsf{M}}}\colon I \times J \to \mathsf{R}$ as defined above. •

Let us now give an interpretation of the matrix representative. We suppose now that R is a ring with unit and we let M be a free left (resp. right) R-module. We let $\mathscr{B}$ be a basis for M and let I be an index set such that there exists a bijection $\phi_{\mathscr{B}}\colon I \to \mathscr{B}$. We define an isomorphism $\iota_{\mathscr{B}}$ of M with the left (resp. right) R-module R^I_0 as follows. For $x \in \mathsf{M}$ we write

$$x = c_1\phi_{\mathscr{B}}(i_1) + \cdots + c_k\phi_{\mathscr{B}}(i_k) \quad (\mathit{resp.}\ x = \phi_{\mathscr{B}}(i_1)c_1 + \cdots + \phi_{\mathscr{B}}(i_k)c_k),$$

for unique nonzero $c_1,\dots,c_k \in \mathsf{R}$ and $i_1,\dots,i_k \in I$. We then take

$$\iota_{\mathscr{B}}(x)(i) = \begin{cases} \phi_{\mathscr{B}}(i), & i \in \{1,\dots,i_k\}, \\ 0_{\mathsf{R}}, & i \notin \{1,\dots,i_k\}. \end{cases}$$

We can then describe the matrix representative as follows.

5.5.8 Theorem (Interpretation of matrix representative) *Let* R *be a unit ring, let* M *and* N *be free left (resp. right)* R*-modules, let* $\mathscr{B}_{\mathsf{M}}$ *and* $\mathscr{B}_{\mathsf{N}}$ *be bases for* M *and* N*, respectively, and let* I *and* J *be sets for which there exist bijections* $\phi_{\mathsf{M}}\colon J \to \mathscr{B}_{\mathsf{M}}$ *and* $\phi_{\mathsf{N}}\colon I \to \mathscr{B}_{\mathsf{N}}$*. If* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ *then, with the isomorphisms* $\iota_{\mathscr{B}_{\mathsf{M}}}\colon \mathsf{M} \to \mathsf{R}_0^J$ *and* $\iota_{\mathscr{B}_{\mathsf{N}}}\colon \mathsf{N} \to \mathsf{R}_0^I$ *as defined above, the following diagram commutes:*

$$\begin{tikzcd}
\mathsf{M} \arrow[r, "\mathsf{L}"] \arrow[d, "\iota_{\mathscr{B}_{\mathsf{M}}}"'] & \mathsf{N} \arrow[d, "\iota_{\mathscr{B}_{\mathsf{N}}}"] \\
\mathsf{R}_0^J \arrow[r, "{[\mathsf{L}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}}"'] & \mathsf{R}_0^I
\end{tikzcd}$$

In particular, the map $\mathsf{L} \mapsto \iota_{\mathscr{B}_{\mathsf{N}}} \circ \mathsf{L} \circ \iota_{\mathscr{B}_{\mathsf{M}}}^{-1}$ *is an isomorphism of the left (resp. right)* R*-modules* $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M},\mathsf{N})$ *and* $\mathrm{Mat}_{I\times J}(\mathsf{R})$*.*

Proof Let us give the proof for right modules, the proof for left modules being entirely similar.

Let $x \in \mathsf{M}$ and write

$$x = \phi_{\mathsf{M}}(j_1)c_1 + \cdots + \phi_{\mathsf{M}}(j_k)c_k$$

for unique nonzero $c_1,\dots,c_k \in \mathsf{R}$ and $j_1,\dots,j_k \in J$. Then, using the definition of $\iota_{\mathscr{B}_{\mathsf{M}}}$, we compute, for $i \in I$,

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}} \circ \iota_{\mathscr{B}_{\mathsf{M}}}(x)(i) = \sum_{l=1}^{k} [\mathsf{L}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}(i, j_l)c_l. \tag{5.28}$$

On the other hand, using the definition of the matrix representative,

$$\mathsf{L}(u) = \sum_{l=1}^{k} \mathsf{L}(\phi_{\mathsf{M}}(j_l))c_l = \sum_{i\in I}\sum_{l=1}^{k} \phi_{\mathsf{N}}(i)[\mathsf{L}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}(i, j_l)c_l. \tag{5.29}$$

The result now follows by comparing (5.28) and (5.29) and from the definition of $\iota_{\mathscr{B}_{\mathsf{N}}}$.

The final assertion of the theorem follows by noting that the given map is a bijection, and by directly checking that it is a homomorphism. ■

The next result tells us that the matrix representative of the composition of linear maps is the product of the matrix representatives.

5.5.9 Theorem (Matrix representative of a composition) *Let* R *be a unit ring, let* M*,* N*, and* P *be free (left or right)* R*-modules, let* $\mathscr{B}_{\mathsf{M}}$*,* $\mathscr{B}_{\mathsf{N}}$*, and* $\mathscr{B}_{\mathsf{P}}$ *be bases for* M*,* N*, and* P*, respectively, and let* I*,* J*, and* K *be index sets for which there exists bijections* $\phi_{\mathsf{M}}\colon I \to \mathscr{B}_{\mathsf{M}}$*,* $\phi_{\mathsf{N}}\colon J \to \mathscr{B}_{\mathsf{N}}$*, and* $\phi_{\mathsf{P}}\colon K \to \mathscr{B}_{\mathsf{P}}$*. If* $\mathsf{A} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ *and* $\mathsf{B} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N};\mathsf{P})$*, then*

$$[\mathsf{B}\circ\mathsf{A}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{P}}} = [\mathsf{B}]_{\mathscr{B}_{\mathsf{N}}}^{\mathscr{B}_{\mathsf{P}}}[\mathsf{A}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}.$$

Proof We shall give the proof for right modules; the proof for left modules follows in the same vein.

For $i \in I$ compute, using the definition of matrix representative and using linearity,

$$\begin{aligned}(\mathsf{B} \circ \mathsf{A})(\phi_{\mathsf{M}}(i)) &= \mathsf{B}\left(\sum_{j \in J} \phi_{\mathsf{N}}(j)[\mathsf{A}]^{\mathscr{B}_{\mathsf{N}}}_{\mathscr{B}_{\mathsf{M}}}(j,i)\right)\\ &= \sum_{j \in J} \mathsf{B}(\phi_{\mathsf{N}}(j))[\mathsf{A}]^{\mathscr{B}_{\mathsf{N}}}_{\mathscr{B}_{\mathsf{M}}}(j,i)\\ &= \sum_{j \in J}\sum_{k \in K} \phi_{\mathsf{P}}(k)[\mathsf{B}]^{\mathscr{B}_{\mathsf{P}}}_{\mathscr{B}_{\mathsf{N}}}(k,j)[\mathsf{A}]^{\mathscr{B}_{\mathsf{N}}}_{\mathscr{B}_{\mathsf{M}}}(j,i)\\ &= \sum_{k \in K} \phi_{\mathsf{P}}(k)[\mathsf{B} \circ \mathsf{A}]^{\mathscr{B}_{\mathsf{P}}}_{\mathscr{B}_{\mathsf{M}}}(k,i),\end{aligned}$$

which gives the result. ■

For endomorphisms of a module, the preceding result, combined with Theorem 5.5.8 has the following corollary.

5.5.10 Corollary (Matrix representatives of endomorphisms) *Let* R *be a unit ring, let* M *be a free (resp. right)* R*-module, let* $\mathscr{B}$ *be a basis for* M*, let* I *be an index set for which there exists a bijection* $\phi\colon \mathrm{I} \to \mathscr{B}$*, and let* $\iota_{\mathscr{B}}\colon \mathsf{M} \to \mathsf{R}^{\mathrm{I}}_0$ *be the isomorphism defined preceding the statement of Theorem 5.5.8. Then the map* $\mathsf{L} \mapsto \iota_{\mathscr{B}} \circ \mathsf{L} \circ \iota_{\mathscr{B}}^{-1}$ *is an isomorphism of the left (resp. right)* R *algebras* $\mathrm{End}_{\mathsf{R}}(\mathsf{M})$ *and* $\mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{R})$*.*

As with linear maps between vector spaces, the matrix representative *is* the matrix if the homomorphism is defined using a matrix (see Theorem 5.2.11).

5.5.11 Proposition (Matrix representative associated to standard bases) *Let* R *be a unit ring, let* I *and* J *be index sets, let* $\mathsf{M} = \mathsf{F}^{\mathrm{J}}_0$ *and* $\mathsf{N} = \mathsf{F}^{\mathrm{I}}_0$*, and let* $\mathscr{B}_{\mathsf{M}} = \{\mathbf{e}_{\mathrm{j}}\}_{\mathrm{j}\in\mathrm{J}}$ *and* $\mathscr{B}_{\mathsf{N}} = \{\mathbf{f}_{\mathrm{i}}\}_{\mathrm{i}\in\mathrm{I}}$ *be the standard bases for* $\mathsf{R}^{\mathrm{J}}_0$ *and* $\mathsf{R}^{\mathrm{I}}_0$*, respectively. Let* $\phi_{\mathsf{M}}\colon \mathrm{J} \to \mathscr{B}_{\mathsf{M}}$ *and* $\phi_{\mathsf{N}}\colon \mathrm{I} \to \mathscr{B}_{\mathsf{N}}$ *be defined by*

$$\phi_{\mathsf{M}}(\mathrm{j}) = \mathbf{e}_{\mathrm{j}},\ \mathrm{j} \in \mathrm{J}, \quad \phi_{\mathsf{N}}(\mathrm{i}) = \mathbf{f}_{\mathrm{i}},\ \mathrm{i} \in \mathrm{I}.$$

If $\mathbf{A} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{J}}(\mathsf{R})$ *is column finite and so defines an element of* $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ *by Theorem 5.2.11, then* $[\mathbf{A}]^{\mathscr{B}_{\mathsf{N}}}_{\mathscr{B}_{\mathsf{M}}} = \mathbf{A}$*.*

Proof This follows directly from the computations in the proof of Theorem 5.2.11 and the constructions preceding Definition 5.4.20. ■

5.5.4 Changing of bases

The change of basis rules for homomorphisms of free modules goes like that for linear maps of vector spaces. The biggest difference is that one needs to carefully account for the fact that the same module may possess bases of different cardinalities.

5.5.12 Proposition (Existence of change of basis matrix) *Let* R *be a unit ring, let* M *be a free left (resp. right)* R*-module, let* $\mathscr{B}$ *and* $\mathscr{B}'$ *be bases for* M *having the same cardinality, and let* I *be an index set for which there exist bijections* $\phi\colon \mathrm{I} \to \mathscr{B}$ *and* $\phi'\colon \mathrm{I} \to \mathscr{B}'$*. Then there exists a unique invertible column finite matrix* $\mathbf{P} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{R})$ *such that*

$$\phi(\mathrm{i}_0) = \sum_{\mathrm{i}\in\mathrm{I}} \mathbf{P}(\mathrm{i},\mathrm{i}_0)\phi'(\mathrm{i}) \quad \left(\textit{resp. } \phi(\mathrm{i}_0) = \sum_{\mathrm{i}\in\mathrm{I}} \phi'(\mathrm{i})\mathbf{P}(\mathrm{i},\mathrm{i}_0)\right)$$

for each $\mathrm{i}_0 \in \mathrm{I}$.

Proof We give the proof for right modules since the proof for left modules follows along similar lines.

Let $i_0 \in I$. Since $\mathscr{B}'$ is a basis there exists unique $i_1, \dots, i_k \in I$ and nonzero $c_1, \dots, c_k \in \mathsf{R}$ such that

$$\phi(i_0) = \phi'(i_1)c_1 + \cdots + \phi'(i_k)c_k.$$

One then defines $\boldsymbol{P}$ by asking that $\boldsymbol{P}(i, i_0) = c_j$ if $i = i_j$ for some $j \in \{1, \dots, k\}$, and that $\boldsymbol{P}(i, i_0) = 0_{\mathsf{R}}$ for $i \in I \setminus \{i_1, \dots, i_k\}$. Note that $\boldsymbol{P}$ thus defined is column finite. We next show that it is invertible. To do this, we construct its inverse. By swapping the rôles of $\mathscr{B}$ and $\mathscr{B}'$, our above argument gives the existence a column finite matrix $\boldsymbol{Q} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ such that

$$\phi'(i_0) = \sum_{i\in I} \phi(i)\boldsymbol{Q}(i, i_0)$$

for each $i_0 \in I$. We claim that $\boldsymbol{QP} = \boldsymbol{I}_I$. Let $i_0 \in I$ and note that

$$\phi(i_0) = \sum_{i\in I} \phi'(i)\boldsymbol{P}(i, i_0) = \sum_{i\in I}\sum_{i'\in I} \phi(i')\boldsymbol{Q}(i', i)\boldsymbol{P}(i, i_0) = \sum_{i'\in I} \phi(i')(\boldsymbol{QP})(i', i_0).$$

Since $\{\phi(i)\}_{i\in I}$ is a basis, and in particular is linearly independent, this implies that

$$(\boldsymbol{QP})(i', i_0) = \begin{cases} 1_{\mathsf{R}}, & i' = i_0, \\ 0_{\mathsf{R}}, & i' \neq i_0. \end{cases}$$

In other words, $\boldsymbol{QP} = \boldsymbol{I}_I$. In like manner we compute

$$\phi'(i_0) = \sum_{i\in I} \phi(i)\boldsymbol{Q}(i, i_0) = \sum_{i\in I}\sum_{i'\in I} \phi'(i')\boldsymbol{P}(i', i)\boldsymbol{Q}(i, i_0) = \sum_{i'\in I} \phi'(i')(\boldsymbol{PQ})(i', i_0),$$

which leads to the conclusion, using the fact that $\{\phi'(i)\}_{i\in I}$ is linearly independent, that $\boldsymbol{PQ} = \boldsymbol{I}_I$. Thus $\boldsymbol{Q}$ is the inverse of $\boldsymbol{P}$. ■

We introduce some notation and terminology associated with the matrix $\boldsymbol{P}$ of the preceding result.

5.5.13 Definition (Change of basis matrix) Let R be a unit ring, let M be a free (left or right) R-module, let $\mathscr{B}$ and $\mathscr{B}'$ be bases for M having the same cardinality, and let I be an index set for which there exist bijections $\phi\colon I \to \mathscr{B}$ and $\phi'\colon I \to \mathscr{B}'$. The matrix $\boldsymbol{P}$ of Proposition 5.5.12 is called the ***change of basis matrix*** from the basis $\mathscr{B}$ to the basis $\mathscr{B}'$, and is denoted by $\boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}$. •

As a corollary to Proposition 5.5.12 we have the following result.

5.5.14 Corollary (Inverse of change of basis matrix) *Let* R *be a unit ring, let* M *be a free (left or right)* R*-module, let* $\mathscr{B}$ *and* $\mathscr{B}'$ *be bases for* M *having the same cardinality, and let* I *be an index set for which there exist bijections* $\phi\colon \mathrm{I} \to \mathscr{B}$ *and* $\phi'\colon \mathrm{I} \to \mathscr{B}'$*. Then* $\mathbf{P}^{\mathscr{B}}_{\mathscr{B}'} = (\mathbf{P}^{\mathscr{B}'}_{\mathscr{B}})^{-1}$.

Proof This was shown during the course of the proof of Proposition 5.4.26. ■

If one has more than two bases, the change of basis matrices can be related in a simple way.

5.5.15 Proposition (Product of change of basis matrices is a change of basis matrix) *Let* R *be a unit ring, let* M *be a free (left or right)* R*-module, let* $\mathscr{B}$*,* $\mathscr{B}'$*, and* $\mathscr{B}''$ *be bases for* M *having the same cardinality, and let* I *be an index set for which there exist bijections* $\phi\colon \mathrm{I} \to \mathscr{B}$*,* $\phi'\colon \mathrm{I} \to \mathscr{B}'$*, and* $\phi''\colon \mathrm{I} \to \mathscr{B}''$*. Then*

$$\mathbf{P}^{\mathscr{B}''}_{\mathscr{B}} = \mathbf{P}^{\mathscr{B}''}_{\mathscr{B}'}\mathbf{P}^{\mathscr{B}'}_{\mathscr{B}}.$$

Proof We give the proof in the case of right modules.

Let $i_0 \in I$ and compute

$$\begin{aligned}\phi(i_0) &= \sum_{i\in I} \phi'(i)\boldsymbol{P}^{\mathscr{B}'}_{\mathscr{B}}(i, i_0)\\ &= \sum_{i\in I}\sum_{i'\in I} \phi''(i')\boldsymbol{P}^{\mathscr{B}''}_{\mathscr{B}'}(i', i)\boldsymbol{P}^{\mathscr{B}'}_{\mathscr{B}}(i, i_0)\\ &= \sum_{i'\in I} \phi''(i')(\boldsymbol{P}^{\mathscr{B}''}_{\mathscr{B}'}\boldsymbol{P}^{\mathscr{B}'}_{\mathscr{B}})(i', i_0),\end{aligned}$$

giving the result by definition of $\boldsymbol{P}^{\mathscr{B}''}_{\mathscr{B}}$. ■

Finally we show that every invertible matrix gives rise to a change of basis.

5.5.16 Proposition (Invertible matrices give rise to changes of basis) *Let* R *be a unit ring, let* V *be a free (left or right)* R*-module, let* $\mathscr{B}$ *be a basis for* M*, and let* I *be an index set such that there exists a bijection* $\phi\colon \mathrm{I} \to \mathscr{B}$*. Then, given an invertible column matrix* $\mathbf{P} \in \mathrm{Mat}_{\mathrm{I}\times\mathrm{I}}(\mathsf{R})$*, there exists a basis* $\mathscr{B}'$ *for* M *such that* $\mathbf{P} = \mathbf{P}^{\mathscr{B}'}_{\mathscr{B}}$.

Proof We give the proof for right modules.

We define $\mathscr{B}'$ by defining an injective map $\phi'\colon I \to \mathsf{M}$ and taking $\mathscr{B}' = \mathrm{image}(\phi)$. We define ϕ' by

$$\phi'(i_0) = \sum_{i\in I} \phi(i)\boldsymbol{P}^{-1}(i, i_0).$$

Let us show that $\{\phi'(i)\}_{i\in I}$ is a basis for M. First we prove linear independence. Suppose that $c_1, \dots, c_k \in \mathsf{R}$ and $i_1, \dots, i_k \in I$ satisfy

$$\phi'(i_1)c_1 + \dots + \phi'(i_k)c_k = 0_{\mathsf{M}}.$$

Then

$$\sum_{j=1}^{k}\sum_{i\in I} \phi(i)c_j\boldsymbol{P}^{-1}(i, i_j) = 0_{\mathsf{M}}.$$

Since $\{\phi(i)\}_{i\in I}$ is a basis we have

$$\sum_{j=1}^{k} c_j \boldsymbol{P}^{-1}(i,i_j) = 0_{\mathsf{R}}, \qquad i \in I. \tag{5.30}$$

Now, if we define $\boldsymbol{c} \in \mathsf{R}_0^I$ by

$$\boldsymbol{c}(i) = \begin{cases} c_j, & i = i_j,\ j \in \{1,\dots,k\}, \\ 0_{\mathsf{R}}, & i \notin \{i_1,\dots,i_k\}, \end{cases}$$

then (5.30) is simply $\boldsymbol{c}\boldsymbol{P} = \boldsymbol{0}_{\mathsf{R}_0^I}$, and so $\boldsymbol{c} = \boldsymbol{0}_{\mathsf{R}_0^I}$ by Exercise 4.8.3. Thus $c_1 = \cdots = c_k = 0_{\mathsf{R}}$, giving linear independence. Now we show that $\{\phi'(i)\}_{i\in I}$ generates V. Note that

$$\sum_{i'\in I} \phi(i')\boldsymbol{P}(i',i_0) = \sum_{i\in I}\sum_{i'\in I} \phi(i)\boldsymbol{P}^{-1}(i,i')\boldsymbol{P}(i',i_0) = \phi(i_0). \tag{5.31}$$

Therefore, every element in the basis $\{\phi(i)\}_{i\in I}$ is a finite linear combination of elements from $\{\phi'(i)\}_{i\in I}$. Since every element in M is a finite linear combination of elements from $\{\phi(i)\}_{i\in I}$, it then immediately follows that every element in M is a finite linear combination of vectors from $\{\phi'(i)\}_{i\in I}$. Thus $\mathscr{B}' = \{\phi'(i)\}_{i\in I}$ is a basis.

It follows (5.31) that $\boldsymbol{P} = \boldsymbol{P}_{\mathscr{B}}^{\mathscr{B}'}$. ■

We may now address the matter of how matrix representations change when one changes bases.

5.5.17 Theorem (Change of basis formula) *Let* R *be a unit ring, let* M *and* N *be free (left or right)* R*-modules, let* $\mathscr{B}_{\mathsf{M}}$ *and* $\mathscr{B}'_{\mathsf{M}}$ *be bases for* M *having the same cardinality, let* $\mathscr{B}_{\mathsf{N}}$ *and* $\mathscr{B}'_{\mathsf{N}}$ *be bases for* N *having the same cardinality, and let* I *and* J *be sets for which there exist bijections* $\phi_{\mathsf{M}}\colon \mathrm{J} \to \mathscr{B}_{\mathsf{M}}$, $\phi'_{\mathsf{M}}\colon \mathrm{J} \to \mathscr{B}'_{\mathsf{M}}$, $\phi_{\mathsf{N}}\colon \mathrm{I} \to \mathscr{B}_{\mathsf{N}}$, *and* $\phi'_{\mathsf{N}}\colon \mathrm{I} \to \mathscr{B}'_{\mathsf{N}}$. *If* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ *then*

$$[\mathsf{L}]_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}'_{\mathsf{N}}} = \mathbf{P}_{\mathscr{B}_{\mathsf{N}}}^{\mathscr{B}'_{\mathsf{N}}}[\mathsf{L}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}(\mathbf{P}_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}'_{\mathsf{M}}})^{-1}.$$

This relation is called the ***change of basis formula****.*

Proof We give the proof for right modules.

For $j_0 \in J$ we compute

$$\begin{aligned}
\mathsf{L}(\phi'_{\mathsf{M}}(j_0)) &= \mathsf{L}\left(\sum_{j\in J} \phi_{\mathsf{M}}(j)\boldsymbol{P}_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{M}}}(j,j_0)\right) \\
&= \sum_{j\in J} \mathsf{L}(\phi_{\mathsf{M}}(j))\boldsymbol{P}_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{M}}}(j,j_0) \\
&= \sum_{j\in J}\sum_{i\in I} \phi_{\mathsf{N}}(i)[L]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}(i,j)\boldsymbol{P}_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{M}}}(j,j_0) \\
&= \sum_{j\in J}\sum_{i\in I}\sum_{i'\in I} \phi'_{\mathsf{N}}(i')\boldsymbol{P}_{\mathscr{B}_{\mathsf{N}}}^{\mathscr{B}'_{\mathsf{N}}}(i',i)[L]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}(i,j)\boldsymbol{P}_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{M}}}(j,j_0) \\
&= \sum_{i'\in I} (\boldsymbol{P}_{\mathscr{B}_{\mathsf{N}}}^{\mathscr{B}'_{\mathsf{N}}}[L]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}}\boldsymbol{P}_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{M}}})(i',j_0)\phi'_{\mathsf{N}}(i').
\end{aligned}$$

The result now follows from the definition of the matrix representative and from Corollary 5.5.14. ■

5.5.5 Determinant and trace of an endomorphism

The entirety of Section 5.4.6 can be transplanted here, with "a field F" replaced by "a commutative unit ring R." We do not do this, but leave it as a mental exercise for the reader.

5.5.6 Equivalence of homomorphisms

In this section we say a few words about equivalence of homomorphisms of modules. This problem is too difficult to manage in any sort of generality, so we restrict ourselves to giving a basis-free version of Theorem 5.2.43 for finite matrices over principal ideal domains. The main theorem we state, Theorem 5.5.20, follows from Theorem 5.2.43, but we give a more elegant, but unconstructive proof.

We begin by giving the general definition of equivalence of homomorphisms. If we wish to have a useful form of generality—for example, for homomorphisms of modules that may not be free—we cannot use anything analogous to Definition 5.4.38.

5.5.18 Definition (Equivalence of homomorphisms) Let R be a ring and let M and N be free (left or right) modules. Maps $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{N})$ are ***equivalent*** if there exists isomorphisms $\mathsf{P} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{N}; \mathsf{N})$ and $\mathsf{Q} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{M})$ such that $\mathsf{L}_2 = \mathsf{P} \circ \mathsf{L}_2 \circ \mathsf{Q}$. •

It is not too difficult to show that equivalence of homomorphisms as defined above agrees with equivalence as in Definition 5.4.38 in the case of vector spaces over fields. The reader can prove this as Exercise 5.5.4.

According to Corollary 5.4.42, along with Corollary 4.5.48, two linear maps $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ between vector spaces U and V are equivalent if and only if the following pairs of vector spaces are isomorphic:

1. $\mathrm{image}(\mathsf{L}_1)$ and $\mathrm{image}(\mathsf{L}_2)$;
2. $\ker(\mathsf{L}_1)$ and $\ker(\mathsf{L}_2)$;
3. $\mathrm{coker}(\mathsf{L}_1)$ and $\mathrm{coker}(\mathsf{L}_2)$.

Moreover, if U and V are finite-dimensional then it is enough that $\mathrm{rank}(\mathsf{L}_1) = \mathrm{rank}(\mathsf{L}_2)$. This is not true for rings, even for principal ideal domains as the following simple example shows.

5.5.19 Example (Nonequivalent homomorphisms) We let $\mathsf{R} = \mathbb{Z}$ and consider homomorphisms of $\mathbb{Z}^2$ with itself defined by the following 2×2 matrices:

$$\boldsymbol{A}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}.$$

Note that image(A_1) and image(A_2) are isomorphic via the isomorphism $(j, 2k) \mapsto (j, 3k)$. Also, ker(A_1) and ker(A_2) are both equal to $\{(0, 0)\}$. However, A_1 and A_2 are not equivalent. •

The equivalence problem for homomorphisms of free, finitely generated modules over principal ideal domains is characterised as follows.

5.5.20 Theorem (Equivalence of homomorphisms for principal ideal domains) *Let* R *be a principal ideal domain, let* M *and* N *be free, finitely generated* R*-modules, and let* $\mathsf{L} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}; \mathsf{N})$*. Then there exists bases* $\mathscr{B}_{\mathsf{M}}$ *and* $\mathscr{B}_{\mathsf{N}}$ *for* M *and* N*, respectively, such that*

$$[\mathsf{L}]_{\mathscr{B}_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{N}}} = \left[\begin{array}{c|c} \mathbf{D}_{\mathrm{r}} & \mathbf{0}_{\mathrm{r}\times(\mathrm{n}-\mathrm{r})} \\ \hline \mathbf{0}_{(\mathrm{m}-\mathrm{r})\times\mathrm{r}} & \mathbf{0}_{(\mathrm{m}-\mathrm{r})\times(\mathrm{n}-\mathrm{r})} \end{array}\right]$$

with $\mathbf{D}_{\mathrm{r}}$ *being the diagonal matrix*

$$\mathbf{D}_{\mathrm{r}} = \begin{bmatrix} \mathrm{d}_1 & 0_{\mathsf{R}} & \cdots & 0_{\mathsf{R}} \\ 0_{\mathsf{R}} & \mathrm{d}_2 & \cdots & 0_{\mathsf{R}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{R}} & 0_{\mathsf{R}} & \cdots & \mathrm{d}_{\mathrm{r}} \end{bmatrix},$$

and where $\mathrm{d}_1 | \cdots | \mathrm{d}_{\mathrm{r}}$*. Moreover, the ideals* $(\mathrm{d}_1), \dots, (\mathrm{d}_{\mathrm{r}})$ *are uniquely determined by* L*.*

Proof Suppose that rank(M) $= n$ and rank(N) $= n$. Let $\mathscr{B}'_{\mathsf{M}}$ and $\mathscr{B}'_{\mathsf{N}}$ be bases for M and N, respectively. By Theorem 5.2.43 there exists invertible matrices $\boldsymbol{P} \in \mathrm{Mat}_{m\times m}(\mathsf{R})$ and $\boldsymbol{Q} \in \mathrm{Mat}_{n\times n}(\mathsf{R})$ such that

$$\boldsymbol{P}[\mathsf{L}]_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}'_{\mathsf{N}}}\boldsymbol{Q} = \left[\begin{array}{c|c} \boldsymbol{D}_r & \mathbf{0}_{r\times(n-r)} \\ \hline \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{array}\right].$$

The result now follows by taking $\mathscr{B}_{\mathsf{M}}$ to be the basis for which $\boldsymbol{P}_{\mathscr{B}'_{\mathsf{M}}}^{\mathscr{B}_{\mathsf{M}}} = \boldsymbol{Q}^{-1}$ and by taking $\mathscr{B}_{\mathsf{N}}$ to be the basis for which $\boldsymbol{P}_{\mathscr{B}'_{\mathsf{N}}}^{\mathscr{B}_{\mathsf{N}}} = \boldsymbol{P}$. ■

5.5.7 Notes

A fairly general discussion of linear algebra over commutative rings may be found in the book of McDonald [1984]. In particular, much attention is paid in this volume to homomorphisms of so-called projective modules.

Exercises

5.5.1 Prove Proposition 5.5.4.

5.5.2 Let R be a commutative unit ring and let M be a finitely-generated, free R-module.

(a) Show that the set of invertible endomorphisms of M is a group with group operation given by composition.

This group is called the ***general linear group*** of M and is denoted by GL(M).

(b) Is $\mathsf{GL}(\mathsf{M})$ a subalgebra of $\mathrm{End}_{\mathsf{R}}(\mathsf{M})$?

5.5.3 Let R be a commutative unit, let M be a finitely-generated, free R-module, and recall from Exercise 5.5.2 that $\mathsf{GL}(\mathsf{M})$ denotes the group of invertible endomorphisms of M.

(a) Show that the subset of endomorphisms with determinant 1_{R} is a subgroup of $\mathsf{GL}(\mathsf{M})$.

This subgroup of invertible endomorphisms with determinant 1_{R} is denoted by $\mathsf{SL}(\mathsf{M})$ and is called the ***special linear group*** of M.

(b) Is $\mathsf{SL}(\mathsf{M})$ a subalgebra of $\mathrm{End}_{\mathsf{R}}(\mathsf{M})$?

5.5.4 Prove the following result.

Proposition *Let F be a field and let U and V be F-vector spaces. Then $\mathsf{L}_1, \mathsf{L}_2 \in$ $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ are equivalent in the sense of Definition 5.4.38 if and only if they are equivalent in the sense of Definition 5.5.18.*

Section 5.6

Multilinear algebra

Multilinear algebra, not surprisingly, generalises linear algebra to maps which have multiple vectors as their arguments. This is a subject of significant importance in algebra, geometry, and in some parts of mathematical physics. However, our uses of multilinear algebra will generally be a little mundane. Nonetheless, for purposes of presentation, it is helpful to organise all of our facts about multilinear algebra in one place, and to do so in a way that indicates that our mundane usage is part of a bigger story.

Do I need to read this section? This section should be read when it is either needed, or the reader is ready to appreciate it. •

5.6.1 Multilinear maps

In the following definition we use a basic fact about the symmetric group; see Example 4.1.5–11 for the definition of the symmetric group and Section 4.1.6 for more details. Namely we use the fact that every element of the symmetric group is a product of a finite number of transpositions (i.e., swappings of elements), and that, while the number of transpositions needed is not uniquely defined, the evenness or oddness of the number of transpositions *is* well-defined. Thus, given $\sigma \in \mathfrak{S}_k$, we define $\operatorname{sign}(\sigma)$ to be $+1$ if σ is a composition of an even number of transpositions (these are ***even*** permutations) and -1 if σ is a composition of an odd number of transpositions (these are ***odd*** permutations).

5.6.1 Definition (Multilinear map) Let R be a commutative unit ring and let $\mathsf{M}_1, \dots, \mathsf{M}_k, \mathsf{N}$ be R-modules. A map

$$\phi \colon \mathsf{M}_1 \times \cdots \times \mathsf{M}_k \to \mathsf{N}$$

is **R-*multilinear*** if for each $j_0 \in \{1, \dots, k\}$ and for each $x_j \in \mathsf{M}_j$, $j \in \{1, \dots, k\} \setminus \{j_0\}$, the map

$$x_{j_0} \mapsto \phi(x_1, \dots, x_{j_0}, \dots, x_n)$$

is an element of $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_{j_0}; \mathsf{N})$. The set of R-multilinear maps from $\mathsf{M}_1 \times \cdots \times \mathsf{M}_k$ to N is denoted by $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})$. If $\mathsf{M}_1 = \cdots = \mathsf{M}_k = \mathsf{M}$ for some R-modules M then we denote $\mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M}; \mathsf{N}) = \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}, \dots, \mathsf{M}; \mathsf{N})$. •

It is easy to see that $\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})$ is an R-module.

5.6.2 Proposition ($\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})$ is an R-module) *Let* R *be a commutative unit ring and let* $\mathsf{M}_1, \dots, \mathsf{M}_k, \mathsf{N}$ *be* R*-modules. If we define addition and scalar multiplication* *by*

$$(\phi_1 + \phi_2)(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k}) = \phi_1(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k}) + \phi_2(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k}),$$
$$(\mathrm{r}\phi)(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k}) = \mathrm{r}(\phi(\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{k})),$$

for $\phi, \phi_1, \phi_2 \in \mathrm{Hom}_\mathsf{R}(\mathsf{M}_1, \dots, \mathsf{M}_\mathrm{k}; \mathsf{N})$, $\mathrm{r} \in \mathsf{R}$, *and* $\mathrm{x}_\mathrm{j} \in \mathsf{M}_\mathrm{j}$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$, *then* $\mathrm{Hom}_\mathsf{R}(\mathsf{M}_1, \dots, \mathsf{M}_\mathrm{k}; \mathsf{N})$ *is an* R*-module.*

Proof This is left as Exercise 5.6.1. ■

Let us now consider some important classes of multilinear maps.

5.6.3 Definition (Symmetric, skew-symmetric, and alternating multilinear maps) Let R be a commutative unit ring, let M and N be R-modules, let $k \in \mathbb{Z}_{>0}$, and let $\phi \in \mathrm{Hom}^k_\mathsf{R}(\mathsf{M}; \mathsf{N})$. Then

(i) ϕ is ***symmetric*** if

$$\phi(x_{\sigma(1)}, \dots, x_{\sigma(k)}) = \phi(x_1, \dots, x_k)$$

for every $x_j \in \mathsf{M}_j$, $j \in \{1, \dots, k\}$, and for every $\sigma \in \mathfrak{S}_k$,

(ii) ϕ is ***skew-symmetric*** if

$$\phi(x_{\sigma(1)}, \dots, x_{\sigma(k)}) = \mathrm{sign}(\sigma)\phi(x_1, \dots, x_k)$$

for every $x_j \in \mathsf{M}_j$, $j \in \{1, \dots, k\}$, and for every $\sigma \in \mathfrak{S}_k$, and

(iii) ϕ is ***alternating*** if $\phi(x_1, \dots, x_k) = 0_\mathsf{N}$ whenever $x_i = x_j$ for some $i, j \in \{1, \dots, k\}$.

We denote by $S^k(\mathsf{M}; \mathsf{N})$ (resp. $\bigwedge^k(\mathsf{M}; \mathsf{N})$) the set of symmetric (resp. skew-symmetric) multilinear maps from M^k to N. •

Let us make sure to note that $S^k(\mathsf{M}; \mathsf{N})$ and $\bigwedge^k(\mathsf{M}; \mathsf{N})$ are R-modules.

5.6.4 Proposition (Sets of symmetric and skew-symmetric maps are modules) *If* R *is a commutative unit ring, if* M *and* N *are* R*-modules, and if* $\mathrm{k} \in \mathbb{Z}_{>0}$, *then* $\mathrm{S}^\mathrm{k}(\mathsf{M}; \mathsf{N})$ *and* $\bigwedge^\mathrm{k}(\mathsf{M}; \mathsf{N})$ *are submodules of* $\mathrm{Hom}^\mathrm{k}_\mathsf{R}(\mathsf{M}; \mathsf{N})$.

Proof This is left as Exercise 5.6.2. ■

It is easy to show that alternating multilinear maps are skew-symmetric.

5.6.5 Proposition (Alternating multilinear maps are skew-symmetric) *Let* R *be a commutative unit ring and let* M *and* N *be* R*-modules. Then a multilinear map* $\phi\colon \mathsf{M}^\mathrm{k} \to \mathsf{N}$ *is skew-symmetric if it is alternating.*

Proof Since $\mathrm{sign}(\sigma_1 \circ \sigma_2) = \mathrm{sign}(\sigma_1)\,\mathrm{sign}(\sigma_2)$ it suffices to show that

$$\phi(x_{\sigma(1)}, \dots, x_{\sigma(k)}) = \mathrm{sign}(\sigma)\phi(x_1, \dots, x_k)$$

when σ is a transposition. Thus suppose that σ swaps the ith and jth entries with $i < j$. We then compute,

$$\begin{aligned} 0_{\mathsf{N}} &= \phi(x_1, \dots, \underbrace{x_i + x_j}_{i\text{th spot}}, \dots, \underbrace{x_i + x_j}_{j\text{th spot}}, \dots, x_k) \\ &= \phi(x_1, \dots, x_i, \dots, x_i, \dots, x_k) + \phi(x_1, \dots, x_i, \dots, x_j, \dots, x_k) \\ &\quad + \phi(x_1, \dots, x_j, \dots, x_i, \dots, x_k) + \phi(x_1, \dots, x_j, \dots, x_j, \dots, x_k) \\ &= \phi(x_1, \dots, x_i, \dots, x_j, \dots, x_k) + \phi(x_1, \dots, x_j, \dots, x_i, \dots, x_k), \end{aligned}$$

giving

$$\phi(x_1, \dots, x_i, \dots, x_j, \dots, x_k) = -\phi(x_1, \dots, x_j, \dots, x_i, \dots, x_k),$$

as desired. ■

It is not true that skew-symmetric multilinear maps are alternating, however, as the following example shows.

5.6.6 Example (Skew-symmetric but non-alternating multilinear map) Let us consider the ring $\mathbb{Z}_2 = \mathbb{Z}/2\mathbb{Z}$ and the multilinear map $\phi\colon \mathbb{Z}_2^2 \to \mathbb{Z}_2$ given by

$$\begin{aligned} \phi((j_1 + 2\mathbb{Z}, k_1 + 2\mathbb{Z}), (j_2 + 2\mathbb{Z}, k_2 + 2\mathbb{Z})) = (j_1 + 2\mathbb{Z})(j_2 + 2\mathbb{Z}) + (j_1 + 2\mathbb{Z})(k_2 + 2\mathbb{Z}) \\ + (k_1 + 2\mathbb{Z})(j_2 + 2\mathbb{Z}) + (k_1 + 2\mathbb{Z})(k_2 + 2\mathbb{Z}). \end{aligned}$$

One can directly check that ϕ is skew-symmetric since $1+2\mathbb{Z} = -(1+2\mathbb{Z})$. However,

$$\phi((1 + 2\mathbb{Z}, 0 + 2\mathbb{Z}), (1 + 2\mathbb{Z}, 0 + 2\mathbb{Z})) = 1 + 2\mathbb{Z},$$

and so ϕ is not alternating. •

Let us now explore some connections between linear maps and multilinear maps. The following result will be important in our discussion of derivatives in Section II-1.4.

5.6.7 Proposition (Linear functions of multilinear maps) *Let* R *be a commutative unit ring and let* $\mathsf{M}_0, \mathsf{M}_1, \dots, \mathsf{M}_k$ *and* N *be* R*-modules. The map*

$$\Psi\colon \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0; \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})) \to \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0, \mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})$$

given by

$$\Psi(\phi)(\mathsf{x}_0, \mathsf{x}_1, \dots, \mathsf{x}_k) = \phi(\mathsf{x}_0) \cdot (\mathsf{x}_1, \dots, \mathsf{x}_k)$$

is an isomorphism of R*-modules. (We use the "·" to separate the arguments in order to eliminate a proliferation of parentheses.)*

Proof If $\phi_1, \phi_2 \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0; \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N}))$ then

$$\begin{aligned} \Psi(\phi_1 + \phi_2)(x_0, x_1, \dots, x_k) &= (\phi_1 + \phi_2)(x_0) \cdot (x_1, \dots, x_k) \\ &= (\phi_1(x_0) + \phi_2(x_0)) \cdot (x_1, \dots, x_k) \\ &= \phi_1(x_0) \cdot (x_1, \dots, x_k) + \phi_2(x_0) \cdot (x_1, \dots, x_k) \\ &= \Psi(\phi_1)(x_0) \cdot (x_1, \dots, x_k) + \Psi(\phi_2)(x_0) \cdot (x_1, \dots, x_k). \end{aligned}$$

If $r \in \mathsf{R}$ and $\phi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0; \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N}))$ then

$$\begin{aligned}\Psi(r\phi)(x_0, x_1, \dots, x_k) &= (r\phi)(x_0) \cdot (x_1, \dots, x_k)\\ &= (r(\phi(x_0))) \cdot (x_1, \dots, x_k)\\ &= r(\phi(x_0) \cdot (x_1, \dots, x_k))\\ &= r(\Psi(\phi)(x_0, x_1, \dots, x_k)).\end{aligned}$$

This shows that Ψ is linear.

Suppose that $\Psi(\phi) = 0_{\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0, \mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})}$. Then

$$\Psi(\phi)(x_0, x_1, \dots, x_k) = 0_{\mathsf{N}}$$

for all $x_j \in \mathsf{M}_j$, $j \in \{0, 1, \dots, k\}$. Thus

$$\phi(x_0) \cdot (x_1, \dots, x_k) = 0_{\mathsf{N}}$$

for all $x_0 \in \mathsf{M}_0$ and for all $x_j \in \mathsf{M}_j$, $j \in \{1, \dots, k\}$. Thus $\phi(x_0) = 0_{\mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k)}$ for all $x_0 \in \mathsf{M}_0$. Thus $\phi = 0$ and so Ψ is injective by Exercise 4.5.23.

Now let $\psi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0, \mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N})$ and define

$$\phi \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_0; \mathrm{Hom}_{\mathsf{R}}(\mathsf{M}_1, \dots, \mathsf{M}_k; \mathsf{N}))$$

by

$$\phi(x_0) \cdot (x_1, \dots, x_k) = \psi(x_0, x_1, \dots, x_k).$$

We should show that ϕ is linear. For $y_0, x_0 \in \mathsf{M}_0$ we compute

$$\begin{aligned}\phi(y_0 + x_0) \cdot (x_1, \dots, x_k) &= \psi(y_0 + x_0, x_1, \dots, x_k)\\ &= \psi(y_0, x_1, \dots, x_k) + \psi(x_0, x_1, \dots, x_k)\\ &= \phi(y_0) \cdot (x_1, \dots, x_k) + \phi(x_0) \cdot (x_1, \dots, x_k)\\ &= (\phi(y_0) + \phi(x_0)) \cdot (x_1, \dots, x_k).\end{aligned}$$

For $r \in \mathsf{R}$ and $x_0 \in \mathsf{M}_0$ we similarly have

$$\begin{aligned}\phi(rx_0) \cdot (x_1, \dots, x_k) &= \psi(rx_0, x_1, \dots, x_k)\\ &= r\psi(x_0, x_1, \dots, x_k)\\ &= r(\phi(x_0) \cdot (x_1, \dots, x_k))\\ &= (r\phi(x_0)) \cdot (x_1, \dots, x_k).\end{aligned}$$

This shows that ϕ is indeed linear, and so also shows that Ψ is surjective since we clearly have $\Psi(\phi) = \psi$. ■

5.6.2 Bases for sets of multilinear maps

It will be convenient to represent multilinear maps in bases, and in this section we indicate how this is done. For general modules the matter of representing a multilinear map in a convenient way is a rather complicated issue. However, we shall be interested only in the simplest of cases where the modules are free (note that Theorem 4.8.25 implies that the rank of such modules is well-defined).

We begin by defining a collection of multilinear maps associated to bases for the modules under consideration. We let R be a commutative unit ring and let M and N be unity R-modules. We let $(e_i)_{i\in I}$ and $(f_j)_{j\in J}$ be basis for M and N, respectively. We shall define a collection of elements of $\mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ that we will show form a basis for this set of multilinear maps. For $x_1,\dots,x_k\in\mathsf{M}$ and for each $l\in\{1,\dots,k\}$, write $x_l=\sum_{i\in I}a^i_l e_i$ for uniquely defined coefficients $a^i_l\in\mathsf{R}$, $i\in I$, only finitely many of which are nonzero. For $i_1,\dots,i_k\in I$ and for $j\in J$ define $E^{i_1\cdots i_k}_j\in\mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ by

$$(E^{i_1\cdots i_k}_j(x_1,\dots,x_k))(j')=\begin{cases}a^{i_1}_1\cdots a^{i_k}_k, & j=j',\\ 0_{\mathsf{N}}, & j\neq j'.\end{cases}$$

Thus

$$E^{i_1\cdots i_k}_j(x_1,\dots,x_k)=a^{i_1}_1\cdots a^{i_k}_k f_j.$$

Let us define a related notion.

5.6.8 Definition (Components of a multilinear map) Let R be a commutative unit ring and let M and N be unity R-modules with bases $(e_i)_{i\in I}$ and $(f_j)_{j\in J}$, respectively. For $\phi\in\mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ the ***components*** of ϕ are defined by

$$\phi^j_{i_1\cdots i_k}=\phi(e_{i_1},\dots,e_{i_k})(j). \qquad \bullet$$

The following result puts the preceding notation together.

5.6.9 Proposition (Bases for sets of multilinear maps) *Let* R *be a commutative unit ring and let* M *and* N *be unity* R*-modules with bases* $(\mathrm{e_i})_{\mathrm{i\in I}}$ *and* $(\mathrm{f_j})_{\mathrm{j\in J}}$*, respectively. Then*

$$\left\{\mathrm{E}^{\mathrm{i_1\cdots i_k}}_{\mathrm{j}}\;\middle|\;\mathrm{i_1,\dots,i_k\in I,\ j\in J}\right\}$$

is a basis for $\mathrm{Hom}^{\mathrm{k}}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$*. Moreover, if* $\phi\in\mathrm{Hom}^{\mathrm{k}}_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ *then*

$$\phi=\sum_{\mathrm{i_1,\dots,i_k\in I}}\sum_{\mathrm{j\in J}}\phi^{\mathrm{j}}_{\mathrm{i_1\cdots i_k}}\mathrm{E}^{\mathrm{i_1\cdots i_k}}_{\mathrm{j}}.$$

Proof Suppose that there exist coefficients $c^j_{i_1\cdots i_k}\in\mathsf{R}$, $i_1,\dots,i_k\in I$, $j\in J$, only finitely many of which are nonzero, such that

$$\sum_{i_1,\dots,i_k\in I}\sum_{j\in J}c^j_{i_1\cdots i_k}E^{i_1\cdots i_k}_j=0_{\mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})}.$$

Then, for every $i'_1,\dots,i'_k\in I$ we have

$$0_{\mathsf{N}}=\sum_{i_1,\dots,i_k\in I}\sum_{j\in J}c^j_{i_1\cdots i_k}E^{i_1\cdots i_k}_j(e_{i'_1},\dots,e_{i'_k})=\sum_{j\in J}c^j_{i'_1\cdots i'_k}f_j,$$

using the definition of $E_j^{i_1\cdots i_k}$. Since $(f_j)_{j\in J}$ is linearly independent this means that $c^j_{i'_1\cdots i'_k} = 0_{\mathsf{R}}$ for every $i'_1,\dots,i'_k \in I$ and $j \in J$. This gives linear independence of

$$\left\{E_j^{i_1\cdots i_k} \;\middle|\; i_1,\dots,i_k \in I,\ j \in J\right\}.$$

Let $\phi \in \mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ and for $x_1,\dots,x_k \in \mathsf{M}$ write $x_l = \sum_{i_l\in I} a_l^{i_l} e_{i_l}$. Then we compute

$$\begin{aligned}\phi(x_1,\dots,x_k) &= \sum_{i_1,\dots,i_k\in I} \phi(a_1^{i_1}e_{i_1},\dots,a_k^{i_k}e_{i_k})\\ &= \sum_{i_1,\dots,i_k\in I}\sum_{j\in J} a_1^{i_1}\cdots a_k^{i_k}(\phi(e_{i_1},\dots,e_{i_k})(j))f_j\\ &= \sum_{i_1,\dots,i_k\in I}\sum_{j\in J} a_1^{i_1}\cdots a_k^{i_k}\phi^j_{i_1\cdots i_k}f_j\\ &= \sum_{i_1,\dots,i_k\in I}\sum_{j\in J} \phi^j_{i_1\cdots i_k}E_j^{i_1\cdots i_k}(x_1,\dots,x_k),\end{aligned}$$

giving the final assertion of the result, and in particular the fact that the set

$$\left\{E_j^{i_1\cdots i_k} \;\middle|\; i_1,\dots,i_k \in I,\ j \in J\right\}$$

generates $\mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$. ■

5.6.10 Remark (Warning about bilinear maps) Let R be a commutative unit ring and let M be a unity R-module with basis $(e_i)_{i\in I}$. Let us make some observations about the set $\mathrm{Hom}^2_{\mathsf{R}}(\mathsf{M};\mathsf{R})$ of scalar-valued bilinear maps. Since R has the obvious basis $\{1_{\mathsf{R}}\}$, from Proposition 5.6.9 the components of $\mathsf{B} \in \mathrm{Hom}_{\mathsf{R}}(\mathsf{M};\mathsf{R})$ are naturally thought of as defining an matrix $\boldsymbol{B} \in \mathrm{Mat}_{I\times I}(\mathsf{R})$ by $\boldsymbol{B}(i_1,i_2) = \mathsf{B}(e_{i_1},e_{i_2})$. Because of this, there can be a tendency to confound the notions of linear maps and bilinear maps since both are represented in bases with matrices. However, it is an error to do this; bilinear maps and linear maps are very obviously different things. However, when R is a field and M is thus a vector space, as we shall see in Section 5.7.6, there *is* a natural way of thinking of B as a linear map, albeit one whose image is the algebraic dual M'. •

5.6.3 Tensor products of vector spaces

In this section we give a brief overview of tensor products. Tensor products are notoriously difficult to get one's hands on, and since we shall make only limited use of them, this section can certainly be skipped until the reader feels compelled to understand tensor products by the need to understand vector spaces over field extensions (see Section 4.6) and the related concepts that will arise in our discussion of linear algebra on finite-dimensional vector spaces as described in Section 5.8.9.

There is no easy definition of the tensor product. The definition we give is the one that is most useful in practice, since it provides that property of the tensor product that captures their essence. For simplicity we restrict ourselves to vector spaces rather than general modules. We also simply define the tensor product of two vector spaces. This construction can be used to inductively define tensor products of any finite collection of vector spaces, and it is also possible to directly define—making fairly obvious modifications to our constructions—tensor products of finitely many vector spaces. However, since we shall not have occasion to use these constructions, we do not present them here.

We let F be a field and let U, V, and W be F-vector spaces. Recall from the discussion above that $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U},\mathsf{V};\mathsf{W})$ denotes the set of multilinear maps from $\mathsf{U}\times\mathsf{V}$ to W. Such multilinear maps we call ***bilinear*** since there are only two components to the domain.

Now we can give the definition of a tensor product.

5.6.11 Definition (Tensor product) Let F be a field and let U and V be F-vector spaces. A ***tensor product*** of U and V is an F-vector space $\mathsf{U}\otimes\mathsf{V}$ and a bilinear map $\iota\colon \mathsf{U}\times\mathsf{V}\to\mathsf{U}\otimes\mathsf{V}$ such that, for any F-vector space W and any bilinear map $\phi\in\mathrm{Hom}_{\mathsf{F}}(\mathsf{U},\mathsf{V};\mathsf{W})$, there exists a unique linear map $\mathsf{L}_\phi\in\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}\otimes\mathsf{V};\mathsf{W})$ such that the diagram

$$\begin{tikzcd}
\mathsf{U}\times\mathsf{V} \arrow[r, "\phi"] \arrow[d, "\iota"'] & \mathsf{W} \\
\mathsf{U}\otimes\mathsf{V} \arrow[ur, "\mathsf{L}_\phi"'] &
\end{tikzcd}$$

commutes. •

The idea is that associated with a *bilinear* map from two vector spaces is a *linear* map from their tensor product. What is true is that the notion of a tensor product will take some time to absorb, so the reader should be prepared to accept a few moments of discomfort before acquiring a facility in using tensor products. We shall get at this by exploring some of the properties of tensor products.

We begin by showing that tensor products exist and are essentially unique. The proof of the theorem is constructive, in principle, although perhaps the most concrete realisation of tensor products will come when we give bases for them in terms of bases for the vector spaces involved.

5.6.12 Theorem (Existence and uniqueness of tensor products) *If F is a field and if U and V are two F-vector spaces, then there exists a tensor product of U and V. Moreover, two tensor products of U and V are unique.*

Proof The explicit construction of the tensor product is quite abstract. Let $\mathsf{F}_0^{\mathsf{U}\times\mathsf{V}}$ be the vector space generated by the index set $\mathsf{U}\times\mathsf{V}$. Thus

$$\mathsf{F}_0^{\mathsf{U}\times\mathsf{V}}=\bigoplus_{(u,v)\in\mathsf{U}\times\mathsf{V}}\mathsf{F}.$$

We represent an element of $F_0^{U\times V}$ as a map $\phi\colon U\times V\to F$ which is zero except for a finite number of elements of its argument. Consider the subspace X of $F_0^{U\times V}$ generated by elements of the form

$$(u_1+u_2,v)-(u_1,v)-(u_2,v),\quad (u,v_1+v_2)-(u,v_1)-(u,v_2),$$
$$(au,v)-a(u,v),\quad (u,av)-a(u,v)$$

for $u,u_1,u_2\in U$, $v,v_1,v_2\in V$, and $a\in F$. Let us denote $U\otimes V=F_0^{U\times V}/X$.

We claim that $U\otimes V$ is a tensor product if we define $\iota\colon U\times V\to U\otimes V$ by $\iota(u,v)=(u,v)+X$. Let us abbreviate $\iota(u,v)=u\otimes v$. It is evident that the set $\{u\otimes v \mid u\in U,\ v\in V\}$ generates $U\otimes V$ since $U\times V$ generates $F_0^{U\times V}$. Thus, given $\phi\in\mathrm{Hom}_F(U,V;W)$, to define $L_\phi\in\mathrm{Hom}_F(U\otimes V;W)$ it suffices to define $L_\phi(u\otimes v)$ for $(u,v)\in U\times V$. We define $L_\phi(u\otimes v)=\phi(u,v)$.

For this to define a homomorphism, we should show that it is well-defined. Thus suppose that

$$\tilde{u}\otimes\tilde{v}=u\otimes v,$$

i.e., suppose that

$$(\tilde{u},\tilde{v})-(u,v)\in X.$$

Then, by definition of X and since ϕ is bilinear, it follows that

$$\phi((\tilde{u},\tilde{v})-(u,v))=0,$$

so that

$$L_\phi(\tilde{u}\otimes\tilde{v})=L_\phi(u\otimes v).$$

Thus, our definition of L_ϕ on elements of the form $u\otimes v$ makes sense. It is clear by mere definition of L_ϕ that the diagram of Definition 5.6.11 commutes.

To show uniqueness of the tensor product, let T and T' be two tensor products with associated bilinear maps $\iota\colon U\times V\to T$ and $\iota'\colon U\times V\to T'$. By definition of the tensor product there exists a unique $L_\iota\in\mathrm{Hom}_F(T';T)$ such that $L_\iota\circ\iota'=\iota$ and a unique $L_{\iota'}\in\mathrm{Hom}_F(T;T')$ such that $L_{\iota'}\circ\iota=\iota'$. Thus

$$L_\iota\circ L_{\iota'}\circ\iota=L_\iota\circ\iota'=\iota.$$

Thus the diagram

$$\begin{tikzcd}
U\times V \arrow[r, "\iota"] \arrow[d, "\iota"'] & T \\
T \arrow[ur, "L_\iota\circ L_{\iota'}"'] &
\end{tikzcd}$$

commutes and so uniquely defines $L_\iota\circ L_{\iota'}=\mathrm{id}_T$. Similarly we get $L_{\iota'}\circ L_\iota=\mathrm{id}_T$, and so L_ι is an isomorphism with inverse $L_{\iota'}$. ■

5.6.13 Notation ($u\otimes v$) In the proof we denoted by $u\otimes v\in U\otimes V$ the image of $(u,v)\in U\times V$ under the bilinear map $\iota\colon U\times V\to U\otimes V$. We also noted in the proof that the set $\{u\otimes v \mid u\in U,\ v\in V\}$ generates $U\otimes V$. We shall use this notation and this fact often. •

The following result gives perhaps the best toehold into understanding the tensor product.

5.6.14 Theorem (Bases for tensor products) *Let* F *be a field, let* U *and* V *be* F*-vector spaces, and let* $\{e_i\}_{i\in I}$ *and* $\{f_j\}_{j\in J}$ *be bases for* U *and* V*, respectively. Then*

$$\{e_i \otimes f_j \mid i \in I,\ j \in J\}$$

is a basis for $\mathsf{U}\otimes\mathsf{V}$*.*

Proof Let $w \in \mathsf{U}\otimes\mathsf{V}$. Then

$$w = u_1 \otimes v_1 + \cdots + u_k \otimes v_k.$$

We also write $u_1,\dots,u_k$ and $v_1,\dots,v_k$ as finite linear combinations from $\{e_i\}_{i\in I}$ and $\{f_j\}_{j\in J}$, respectively. This gives w as a finite linear combination of elements of the form $e_i \otimes f_j$, $i \in I$, $j \in J$. Now suppose that

$$\sum_{i\in I}\sum_{j\in J} c_{ij} e_i \otimes f_j = 0_{\mathsf{U}\otimes\mathsf{V}}$$

is a linear combination summing to zero with only finitely many of the c_{ij}'s being nonzero. Define a bilinear map $\phi\colon \mathsf{U}\times\mathsf{V}\to\mathsf{F}$ on basis elements by

$$\phi(e_i, f_j) = c_{ij}, \qquad i \in I,\ j \in J,$$

and then extending to $\mathsf{U}\times\mathsf{V}$ by bilinearity. Thus if

$$u = \sum_{i\in I} u_i e_i \in \mathsf{U}, \quad v = \sum_{j\in J} v_j f_j \in \mathsf{V},$$

then

$$\phi(u, v) = \sum_{i\in I}\sum_{j\in J} c_{ij} u_i v_j.$$

By Proposition 5.6.9 this means that

$$\phi = \sum_{i\in I}\sum_{j\in J} c_{ij} E_{ij}.$$

Associated to this bilinear map is the linear map $\mathsf{L}_\phi \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}\otimes\mathsf{V};\mathsf{F})$ which satisfies

$$\begin{aligned}
\mathsf{L}_\phi(u\otimes v) = \phi(u,v) &= \sum_{i\in I}\sum_{j\in J} c_{ij} u_i v_j \\
&= \sum_{i\in I}\sum_{j\in J}\sum_{k\in I}\sum_{l\in J} c_{ij} E_{kl}(e_i, f_j) u_k v_l \\
&= \sum_{i\in I}\sum_{j\in J}\sum_{k\in I}\sum_{l\in J} c_{ij} \mathsf{L}_{E_{kl}}(e_i \otimes f_j) u_k v_l \\
&= \sum_{k\in I}\sum_{l\in J} \mathsf{L}_{E_{kl}}\left(\sum_{i\in I}\sum_{j\in J} c_{ij} e_i \otimes f_j\right) u_k v_l = 0_{\mathsf{F}}.
\end{aligned}$$

Thus $\phi = 0_{\mathrm{Hom}_{\mathsf{F}}(\mathsf{U},\mathsf{V};\mathsf{F})}$ and so $c_{ij} = 0_{\mathsf{F}}$, $i \in I$, $j \in J$, by Proposition 5.6.9. ■

When both U and V are finite-dimensional, there is a more concrete representation of the tensor product as follows.

5.6.15 Proposition (Sometimes $\mathsf{U} \otimes \mathsf{V} = \mathrm{Hom}_{\mathsf{F}}(\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}, \mathsf{V}; \mathsf{F}); \mathsf{F})$) *Let F be a field and let U and V be finite-dimensional F-vector spaces. Then there exists an isomorphism $\Phi_{\mathsf{U},\mathsf{V}}$ of the F-vector spaces $\mathsf{U} \otimes \mathsf{V}$ and $\mathrm{Hom}_{\mathsf{F}}(\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}, \mathsf{V}; \mathsf{F}); \mathsf{F})$ that satisfies*

$$\Phi_{\mathsf{U},\mathsf{V}}(\mathrm{u} \otimes \mathrm{v})(\phi) = \phi(\mathrm{u}, \mathrm{v}), \qquad \mathrm{u} \in \mathsf{U},\ \mathrm{v} \in \mathsf{V}.$$

Proof Let $\{e_1, \dots, e_n\}$ be a basis for U and let $\{f_1, \dots, f_m\}$ be a basis for V. Let $\{E^{ij} \mid i \in \{1, \dots, n\},\ j \in \{1, \dots, m\}\}$ be the basis for $\mathrm{Hom}_{\mathsf{F}}(\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}, \mathsf{V}; \mathsf{F}); \mathsf{F})$ dual to the basis $\{E_{ij} \mid i \in \{1, \dots, n\},\ j \in \{1, \dots, m\}\}$ for $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}, \mathsf{V}; \mathsf{F})$. We define $\Phi_{\mathsf{U},\mathsf{V}}$ by asking that $\Phi_{\mathsf{U},\mathsf{V}}(e_i \otimes e_j) = E^{ij}$, and then extend to $\mathsf{U} \otimes \mathsf{V}$ by linearity since $\{e_i \otimes f_j \mid i \in \{1, \dots, n\},\ j \in \{1, \dots, m\}\}$ is a basis. It is evident that $\Phi_{\mathsf{U},\mathsf{V}}$ is an isomorphism since it maps a basis for $\mathsf{U} \otimes \mathsf{V}$ to a basis for $\mathrm{Hom}_{\mathsf{F}}(\mathrm{Hom}_{\mathsf{F}}(\mathsf{U}, \mathsf{V}; \mathsf{F}); \mathsf{F})$. Moreover,

$$\Phi_{\mathsf{U},\mathsf{V}}\left(\left(\sum_{i=1}^{n} u_i e_i\right) \otimes \left(\sum_{j=1}^{m} v_j f_j\right)\right)(E_{kl}) = \sum_{i=1}^{n}\sum_{j=1}^{m} u_i v_j E^{kl}(E_{ij}) = u_k v_l = E_{kl}\left(\sum_{i=1}^{n} u_i e_i, \sum_{j=1}^{m} v_j f_j\right),$$

giving $\Phi_{\mathsf{U},\mathsf{V}}(u, v)(E_{kl}) = E_{kl}(u, v)$. Therefore, by linearity, the relation in the statement of the proposition holds. ■

In the next section we shall see an application of the tensor product that we shall encounter at various points in the text. However, to try to get some intuition into the tensor product, let us look at an example.

5.6.16 Example ($\mathsf{F}[\xi_1, \xi_2] = \mathsf{F}[\xi_1] \otimes \mathsf{F}[\xi_2]$) Let us consider two copies of the ring of polynomials over a field, using two different symbols for the indeterminate: $\mathsf{F}[\xi_1]$ and $\mathsf{F}[\xi_2]$. These are vector spaces over F with bases $\{\xi_1^j\}_{j \in \mathbb{Z}_{\geq 0}}$ and $\{\xi_2^j\}_{j \in \mathbb{Z}_{\geq 0}}$: see Example 4.5.2–6 and Example 4.5.28–5. Thus

$$\{\xi_1^{j_1} \otimes \xi_2^{j_2} \mid j_1, j_2 \in \mathbb{Z}_{\geq 0}\}$$

is a basis for $\mathsf{F}[\xi_1] \otimes \mathsf{F}[\xi_2]$. By mapping $\xi_1^{j_1} \otimes \xi_2^{j_2}$ to the monomial $\xi_1^{j_1} \xi_2^{j_2}$ we establish a natural isomorphism from $\mathsf{F}[\xi_1] \otimes \mathsf{F}[\xi_2]$ to $\mathsf{F}[\xi_1, \xi_2]$. Thus we see that this is a natural instance where the notion of the tensor product corresponds to taking products of elements of the two vector spaces. There are other interpretations of the tensor product. It is a surprisingly useful device. •

If one has linear maps on the components of a tensor product, then there is a naturally induced linear map on the tensor product, as the following result records.

5.6.17 Proposition (Linear maps on tensor products) *Let F be a field and let U, V, W, and X be F-vector spaces. If $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{W})$ and $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}; \mathsf{X})$, then there exists a unique $\mathsf{L} \otimes \mathsf{M} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U} \otimes \mathsf{V}; \mathsf{W} \otimes \mathsf{X})$ satisfying*

$$\mathsf{L} \otimes \mathsf{M}(\mathrm{u} \otimes \mathrm{v}) = \mathsf{L}(\mathrm{u}) \otimes \mathsf{M}(\mathrm{v}), \qquad \mathrm{u} \in \mathsf{U},\ \mathrm{v} \in \mathsf{V}.$$

Proof Note that the map $(u, v) \mapsto \mathsf{L}(u) \otimes \mathsf{M}(v)$ is bilinear. The result now follows from the definition of tensor product. ■

5.6.4 Classification of $\mathbb{R}$-valued bilinear maps

There is a particular sort of multilinear map that is quite important in applications, and we investigate these here. The multilinear maps we are concerned with are those for vector spaces defined over $\mathbb{R}$. In this case the natural total order on $\mathbb{R}$ allows for some interesting additional structure.

5.6.18 Definition (Definiteness of symmetric bilinear maps) Let V be a $\mathbb{R}$-vector space, let $\mathsf{B} \in S^2(\mathsf{V};\mathbb{R})$, and define $Q_{\mathsf{B}}\colon V \to \mathbb{R}$ by $Q_{\mathsf{B}}(v) = B(v,v)$. We say that B is

(i) ***positive-semidefinite*** if $\operatorname{image}(Q_{\mathsf{B}}) \subseteq \mathbb{R}_{\geq 0}$, is

(ii) ***positive-definite*** if it is positive-semidefinite and $Q_{\mathsf{B}}^{-1}(0) = \{0_{\mathsf{V}}\}$, is

(iii) ***negative-semidefinite*** if $-\mathsf{B}$ is positive-semidefinite, is

(iv) ***negative-definite*** if $-\mathsf{B}$ is positive-definite, is

(v) ***semidefinite*** if it is either positive- or negative-semidefinite, is

(vi) ***definite*** if it is either positive- or negative-definite, and is

(vii) ***indefinite*** if it is neither positive- nor negative-semidefinite. •

It is important to be able to check when a symmetric bilinear map is positive-definite, and the following result offers a simple means of doing this when V is finite-dimensional. We recall the definition of a minor of a matrix from Definition 5.3.15.

5.6.19 Theorem (Test for positive-definiteness) *Let $n \in \mathbb{Z}_{>0}$ and let V be an n-dimensional $\mathbb{R}$-vector space with basis $\mathscr{B} = \{e_1,\dots,e_n\}$. For $\mathsf{B} \in S^2(\mathsf{V};\mathbb{R})$ let $[\mathsf{B}]_{\mathscr{B}} \in \operatorname{Mat}_{n\times n}(\mathbb{R})$ be defined by*

$$[\mathsf{B}]_{\mathscr{B}}(i,j) = \mathsf{B}(e_i,e_j), \qquad i,j \in \{1,\dots,n\}.$$

Then the following statements are equivalent:

(i) B *is positive-definite;*

(ii) for each $k \in \{1,\dots,n\}$ the (K,K)th minor of $[\mathsf{B}]_{\mathscr{B}}$ is positive, where $K = \{1,\dots,k\}$.

Proof The following lemma will be useful.

1 Lemma *Let $\mathbf{A} \in \operatorname{Mat}_{n\times N}(\mathbb{R})$ be a matrix satisfying $\mathbf{A}(i,j) = \mathbf{A}(j,i)$ for each $i,j \in \{1,\dots,n\}$ and write*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}' & \mathbf{a} \\ \mathbf{a}^T & a' \end{bmatrix}$$

for $\mathbf{A}' \in \operatorname{Mat}_{(n-1)\times(n-1)}(\mathbb{R})$, $\mathbf{a} \in \operatorname{Mat}_{(n-1)\times 1}(\mathbb{R})$, and $a' \in \mathbb{R}$. If $\mathbf{A}'$ is invertible then there exists an invertible matrix $\mathbf{P} \in \operatorname{Mat}_{n\times n}(\mathbb{R})$ such that

$$\mathbf{P}\mathbf{A}\mathbf{P}^T = \begin{bmatrix} \mathbf{A}' & \mathbf{0}_{(n-1)\times 1} \\ \mathbf{0}_{1\times(n-1)} & a'' \end{bmatrix}.$$

Moreover, if $\det \mathbf{A}' > 0$ and $\det \mathbf{A} > 0$ then $a'' > 0$.

Proof Since $\boldsymbol{A}'$ is invertible it follows that the columns $\boldsymbol{c}(\boldsymbol{A}',1),\dots,\boldsymbol{c}(\boldsymbol{A},n-1)$ of $\boldsymbol{A}'$ are a basis for $\mathbb{R}^{n-1}$. Thus, thinking of $\boldsymbol{a}$ as an element of $\mathbb{R}^n$ let us write

$$\boldsymbol{a} = \alpha_1\boldsymbol{c}(\boldsymbol{A}',1) + \dots + \alpha_{n-1}\boldsymbol{c}(\boldsymbol{A}',n-1)$$

for $\alpha_1,\dots,\alpha_{n-1} \in \mathbb{R}$. If we define

$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -\alpha_1 \\ 0 & 1 & \cdots & 0 & -\alpha_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -\alpha_{n-1} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

then $\boldsymbol{P}$ is invertible by Exercise 5.3.7. Then a direct computation shows that

$$\boldsymbol{PAP}^T = \begin{bmatrix} \boldsymbol{A}' & \boldsymbol{0}_{(n-1)\times 1} \\ \boldsymbol{0}_{1\times(n-1)} & a'' \end{bmatrix}$$

for some $a'' \in \mathbb{R}$. By Proposition 5.3.3 and Exercise 5.3.8 we have $(\det \boldsymbol{P})^2 \det \boldsymbol{A} = a'' \det \boldsymbol{A}''$ which gives $a'' > 0$ if $\det \boldsymbol{A} > 0$ and $\det \boldsymbol{A}' > 0$. ▼

Now suppose that B is positive-definite and let us abbreviate $\boldsymbol{B} = [\mathsf{B}]_{\mathscr{B}}$. We have $\boldsymbol{B}(i,j) = \boldsymbol{B}(j,i)$ for $i,j \in \{1,\dots,n\}$. By Exercise 5.6.3 we have

$$\mathsf{B}(v,v) = \sum_{i,j=1}^{n} \boldsymbol{B}(i,j)\boldsymbol{v}(i)\boldsymbol{v}(j)$$

if $\boldsymbol{v}(1),\dots,\boldsymbol{v}(n)$ are the components of $v \in \mathsf{V}$. Thus we have

$$\sum_{i,j=1}^{n} \boldsymbol{B}(i,j)\boldsymbol{v}(i)\boldsymbol{v}(j) > 0 \tag{5.32}$$

for every $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. To show that all (K,K)th minors of $\boldsymbol{B}$ are positive for $K = \{1,\dots,k\}$ and for each $k \in \{1,\dots,n\}$ we proceed by induction on n. The assertion is obviously true for $n = 1$. Suppose that the assertion is true for $n = m-1$ and let $\boldsymbol{B} \in \mathrm{Mat}_{m\times m}(\mathbb{R})$ satisfy $\boldsymbol{B}(i,j) = \boldsymbol{B}(j,i)$ for $i,j \in \{1,\dots,m\}$ and (5.32). We write

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{B}' & \boldsymbol{b} \\ \boldsymbol{b}^T & b' \end{bmatrix} \tag{5.33}$$

as in the lemma above. By the lemma above let $\boldsymbol{P}$ be such that

$$\boldsymbol{PBP}^T = \begin{bmatrix} \boldsymbol{B}' & \boldsymbol{0}_{(m-1)\times 1} \\ \boldsymbol{0}_{1\times(m-1)} & b'' \end{bmatrix}. \tag{5.34}$$

Let $\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_m\}$ be the standard basis for $\mathbb{R}^m$. We have, using (5.34),

$$\sum_{i,j=1}^{m} \boldsymbol{B}(i,j)\boldsymbol{P}(k,i)\boldsymbol{e}_m(k)\boldsymbol{P}(l,j)\boldsymbol{e}_m(l) = b'' > 0$$

Also using (5.34) we have

$$\sum_{i,j=1}^{m} \boldsymbol{B}(i,j)\boldsymbol{P}(k,i)\boldsymbol{v}(k)\boldsymbol{P}(l,j)\boldsymbol{v}(l) = \sum_{i,j=1}^{m-1} \boldsymbol{B}'(i,j)\boldsymbol{v}(i)\boldsymbol{v}(j)$$

whenever $\boldsymbol{v} \in \operatorname{span}_{\mathbb{R}}(\boldsymbol{e}_1,\dots,\boldsymbol{e}_{m-1})$. Thus, by the induction hypothesis, all (K,K)th minors of $\boldsymbol{B}'$ are positive for $K = \{1,\dots,k\}$ and $k \in \{1,\dots,m-1\}$. Moreover,

$$\det(\boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^T) = (\det \boldsymbol{P})^2 \det \boldsymbol{B} = b'' \det \boldsymbol{B}' > 0.$$

Thus $\det \boldsymbol{B} > 0$. Since the (K,K)th minors of $\boldsymbol{B}$ are the same as those of $\boldsymbol{B}'$ for $K = \{1,\dots,k\}$ whenever $k \in \{1,\dots,m-1\}$, it follows that all (K,K)th minors of $\boldsymbol{B}$ are positive for $K = \{1,\dots,k\}$ and $k \in \{1,\dots,m\}$.

Now suppose that (ii) holds. Then we prove (i) by induction on n. For $n = 1$ this is obvious. Suppose that (i) holds for $n = m-1$ and let $\boldsymbol{B} \in \mathrm{Mat}_{m\times m}(\mathbb{R})$ satisfy $\boldsymbol{B}(i,j) = \boldsymbol{B}(j,i)$ for $i,j = \{1,\dots,m\}$ and satisfy (ii). Let us write $\boldsymbol{B}$ as (5.33) and by the lemma above let $\boldsymbol{P}$ be such that (5.34) holds. Then, for any $\boldsymbol{v} \in \mathbb{R}^n$ we have

$$\sum_{i,j,k,l=1}^{m} \boldsymbol{B}(i,j)\boldsymbol{P}(k,i)\boldsymbol{v}(k)\boldsymbol{P}(l,j)\boldsymbol{v}(l) = \sum_{i,j=1}^{m-1} \boldsymbol{B}'(i,j)\boldsymbol{v}(i)\boldsymbol{v}(j) + b''\boldsymbol{v}(m)^2.$$

By the induction hypothesis the first term on the right is positive. By assumption and using the lemma we also have $b'' > 0$ and so

$$\sum_{i,j,k,l=1}^{m} \boldsymbol{B}(i,j)\boldsymbol{P}(k,i)\boldsymbol{v}(k)\boldsymbol{P}(l,j)\boldsymbol{v}(l) > 0$$

for every $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$. However, since $\boldsymbol{P}$ is invertible that this is equivalent to

$$\sum_{i,j=1}^{m} \boldsymbol{B}(i,j)\boldsymbol{v}(i)\boldsymbol{v}(j) > 0,$$

giving the result. ■

Exercises

5.6.1 Prove Proposition 5.6.2.

5.6.2 Prove Proposition 5.6.4.

5.6.3 Let R be a commutative unit ring and let M and N be unity R-modules with bases $(e_i)_{i\in I}$ and $(f_j)_{j\in J}$, respectively. Let $\phi \in \mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ and let $x_1,\dots,x_k \in \mathsf{M}$. Let $\phi^j_{i_1\cdots i_k}$, $j \in J$, and $i_1,\dots,i_k \in I$, be the components of ϕ and let x^i_l, $i \in I$, be the components of x_l for $l \in \{1,\dots,k\}$. Show that

$$\phi(x_1,\dots,x_k) = \sum_{j\in J} \sum_{i_1,\dots,i_k\in I} \phi^j_{i_1\cdots i_k} x_1^{i_1} \cdots x_k^{i_k} f_j.$$

5.6.4 Let R be a commutative unit ring and let M and N be unity R-modules with bases $(e_i)_{i\in I}$ and $(f_j)_{j\in J}$, respectively. Show that $\phi \in \mathrm{Hom}^k_{\mathsf{R}}(\mathsf{M};\mathsf{N})$ is symmetric if and only if its components, $\phi^j_{i_1\cdots i_k}$, $j \in J, i_1,\dots,i_k \in I$, satisfy

$$\phi^j_{i_{\sigma(1)}\cdots i_{\sigma(k)}} = \phi^j_{i_1\cdots i_k}$$

for every $\sigma \in \mathfrak{S}_k$.

5.6.5 Let F be a field and let V be an F-vector space.

(a) Show that $\bigwedge^k(\mathsf{F};\mathsf{V})$ is the trivial vector space for $k \in \mathbb{Z}_{>0}$.

(b) Show that $\mathrm{Hom}^k_{\mathsf{F}}(\mathsf{F};\mathsf{V}) = S^k(\mathsf{F};\mathsf{V})$ for every $k \in \mathbb{Z}_{>0}$.

(c) Show that $\mathrm{Hom}^k_{\mathsf{F}}(\mathsf{F};\mathsf{V})$ is naturally isomorphic to V.

5.6.6 If V is a two-dimensional $\mathbb{R}$-vector space with basis $\mathscr{B} = \{e_1, e_2\}$ and if $\mathsf{B} \in S^2(\mathsf{V};\mathbb{R})$, show that B is positive-definite if and only if $\mathrm{tr}([\mathsf{B}]_{\mathscr{B}}) > 0$ and $\det([\mathsf{B}]_{\mathscr{B}}) > 0$.

Section 5.7

The algebraic dual

The notion of duality is one that is of great importance in algebra and analysis. In this section we focus on the algebraic structure of duality. The notion of duality can be confusing because it is so simple in finite dimensions. There is an inclination to disregard the importance of the dual space in finite dimensions since they are finite dimensional vector spaces with the same dimension, and so are isomorphic. However, they are not *naturally* isomorphic. And the fact is that there are times when the mathematics demands that an object be an element in the dual space. So one should be sure to understand the dual space on its own terms.

Coming with the notion of the algebraic dual is the notion of the dual of a linear map. It is here where the connection with the transpose of a matrix come up. The reader might have noticed that the transpose is completely absent from Section 5.4. This is because the transpose in linear algebra is connected with the dual of a linear map.

The reader should be warned that there is another notion of duality, namely topological duality which is discussed in Sections III-3.9, III-4.2.1, and III-6.4.1. The two notions are decidedly different (although related in the obvious way that the topological dual is a subspace of the algebraic dual) and should not be confused.

Do I need to read this section? The reader should certainly know what a dual space is, just as a matter of course. Similarly, the notion of the dual of a linear map should be understood. Both of these things are quite elementary. Much of the technical material in this section is related to infinite-dimensional complications. Much of this can be sidestepped at a first reading. •

5.7.1 Definitions and first examples

Note that while we only consider vector spaces in this section, many of the constructions apply to modules over rings, although not all results extend to this case. Since we are only interested in duals of vector spaces, we do not deal with the more general situation of modules.

Let us get started with the definition.

5.7.1 Definition (Algebraic dual) Let F be a field and let V be an F-vector space. The ***algebraic dual*** of V is $\mathsf{V}' = \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{F})$. •

5.7.2 Notation (Pairing of element of V′ with element of V) Since an element $\alpha \in \mathsf{V}'$ is a map from V to F, the most natural way to write the image of $v \in \mathsf{V}$ under this map is as $\alpha(v)$. However, sometimes other notation is used, and we shall use this other notation as well. We may, for example, use any of the symbols

$$\alpha(v), \quad \alpha \cdot v, \quad \langle \alpha; v \rangle$$

to represent the same thing. Note that $\langle \cdot ; \cdot \rangle$ is not an inner product since it takes elements from different vector spaces as its argument. It is sometimes called the ***natural pairing*** of V with V'. Similarly, "$\cdot$" is not the "dot product." This will not be confusing since we will never ever under any circumstances use "$\cdot$" as the dot product. •

From Corollary 5.4.17 it follows that V' is an F-vector space with addition defined by

$$(a\alpha)(v) = a(\alpha(v)), \quad (\alpha_1 + \alpha_2)(v) = \alpha_1(v) + \alpha_2(v),$$

for $a \in \mathsf{F}$, $v \in \mathsf{V}$, and $\alpha, \alpha_1, \alpha_2 \in \mathsf{V}'$. We shall use various forms of notation for how elements of V' act on elements of V. Thus we shall use

$$\alpha(v),\ \langle \alpha; v \rangle,\ \alpha \cdot v, \qquad \alpha \in \mathsf{V}',\ v \in \mathsf{V}$$

to represent the same thing. Do not confuse $\alpha \cdot v$ for dot product. We will never use this symbol for the dot product, so there can be no confusion over this.

Since V' is an F-vector space it has a dual. We denote the algebraic dual of V' is denoted by V''. We shall have more to say about the dual of the dual as we go along.

Let us look at some simple examples of elements of the algebraic dual.

5.7.3 Examples (Algebraic dual)

1. Let V be an F-vector space with basis $\{e_i\}_{i \in I}$. Thus every element $v \in \mathsf{V}$ is written as

$$v = \sum_{i \in I} v_i e_i$$

where all but finitely many of the v_i, $i \in I$, are zero. For $i_0 \in I$ we then define $\alpha_{i_0} \in \mathsf{V}'$ by

$$\alpha_{i_0}\left(\sum_{i \in I} v_i e_i\right) = v_{i_0}.$$

Thus α_{i_0} returns the i_0th component of a vector. It is easy to see that this map is indeed linear, and so is in V'.

2. Let $I \subseteq \mathbb{R}$ be an interval and let $\mathsf{C}^0(I; \mathbb{R})$ be the $\mathbb{R}$-vector space of continuous functions on I; see Example 4.5.2–9. For $t_0 \in I$ define $\delta_{t_0} \in \mathsf{C}^0(I; \mathbb{R})'$ by

$$\delta_{t_0}(f) = f(t_0).$$

It is evident that δ_{t_0} is linear and so is an element of the algebraic dual. The map δ_{t_0} is the "Dirac delta-function" about which we shall have a great deal to say in Chapter IV-3.

3. We let $\mathsf{C}^0([a,b];\mathbb{R})$ be the $\mathbb{R}$-vector space of continuous functions on the compact interval $[a,b]$. We define $\alpha \in \mathsf{C}^0([a,b];\mathbb{R})'$ by

$$\alpha(f) = \int_a^b f(x)\,\mathrm{d}x.$$

Since all continuous functions on $[a,b]$ are Riemann integrable by Corollary 3.4.12, this definition makes sense. Moreover, since the integral is linear by Proposition 3.4.22, α is indeed an element of the algebraic dual. •

We close this section with a simple version of an idea that will be of great importance when we discuss normed and topological vector spaces in Chapters III-3 and III-6. The result has to do with extending linear maps if one knows their value on a single vector, and such results fall into the category of "Hahn–Banach Theorems."

5.7.4 Proposition (Algebraic Hahn[1]–Banach Theorem) *Let F be a field and let V be an F-vector space. Then the following statements hold:*

(i) if $\mathsf{U} \subseteq \mathsf{V}$ is a subspace and if $\alpha \in \mathsf{U}'$ then there exists $\beta \in \mathsf{V}'$ such that $\beta(\mathrm{u}) = \alpha(\mathrm{u})$ for each $\mathrm{u} \in \mathsf{U}$;

(ii) if $\mathsf{U} \subseteq \mathsf{V}$ is a subspace, if $\mathrm{v}_0 \in \mathsf{V} \setminus \mathsf{U}$, and if $\mathrm{a} \in \mathsf{F}$, then there exists $\alpha \in \mathsf{V}'$ such that $\alpha(\mathrm{v}_0) = \mathrm{a}$ and $\alpha(\mathrm{u}) = 0_{\mathsf{F}}$ for each $\mathrm{u} \in \mathsf{U}$;

(iii) if $\mathrm{a} \in \mathsf{F}$ and if $\mathrm{v}_0 \in \mathsf{V} \setminus \{0_{\mathsf{V}}\}$ then there exists $\alpha \in \mathsf{V}'$ such that $\alpha(\mathrm{v}_0) = \mathrm{a}$.

Proof (i) Let $\mathscr{B}_{\mathsf{U}}$ be a basis for U and, by Theorem 4.5.26, let $\mathscr{B}_{\mathsf{V}}$ be a basis for V containing $\mathscr{B}_{\mathsf{U}}$. Define $\phi\colon \mathscr{B}_{\mathsf{V}} \to \mathsf{F}$ by

$$\phi(v) = \begin{cases} \alpha(v), & v \in \mathsf{U}, \\ 0_{\mathsf{F}}, & \text{othwerwise.} \end{cases}$$

Then let $\alpha_\phi \in \mathsf{V}'$ be the unique linear map guaranteed by Theorem 4.5.24. Taking $\beta = \alpha_\phi$ gives this part of the result.

(ii) Let $\mathscr{B}_{\mathsf{U}}$ be a basis for U and note that $\mathscr{B}_{\mathsf{U}} \cup \{v_0\}$ is linearly independent. By Theorem 4.5.26 let $\mathscr{B}_{\mathsf{V}}$ be a basis for V containing $\mathscr{B}_{\mathsf{U}} \cup \{v_0\}$. Then define $\phi\colon \mathscr{B}_{\mathsf{V}} \to \mathsf{F}$ by

$$\phi(v) = \begin{cases} a, & v = v_0, \\ 0_{\mathsf{F}}, & \text{otherwise.} \end{cases}$$

The unique element $\alpha_\phi \in \mathsf{V}'$ defined by Theorem 4.5.24 then has the desired properties.

(iii) This follows from part (ii) by taking $\mathsf{U} = \{0_{\mathsf{V}}\}$. ■

Of course, there is no uniqueness in the previous statement since there are generally many ways to construct an element in the dual that extends the map $v \mapsto a$.

[1] Hans Hahn (1879–1934) was an Austrian mathematician who made fundamental contributions to set theory and the then burgeoning field of functional analysis.

5.7.2 Bases and dimension for the algebraic dual

Let us first give a representation for the algebraic dual of the "canonical" F-vector space, F^I_0. In the statement of the result we recall that elements of F^I (and so elements of F^I_0 since $\mathsf{F}^I_0 \subseteq \mathsf{F}^I$) are maps from I to F.

5.7.5 Proposition (The algebraic dual of $\mathsf{F}^{\mathrm{I}}_0$ is F^{I}) *Let F be a field and let I is an index set, and define $\iota\colon \mathsf{F}^{\mathrm{I}} \to (\mathsf{F}^{\mathrm{I}}_0)'$ by*

$$\iota(\mathrm{g})(\mathrm{f}) = \sum_{\mathrm{i}\in\mathrm{I}} \mathrm{g}(\mathrm{i})\mathrm{f}(\mathrm{i})$$

(this sum making sense since it is finite). Then ι is an isomorphism of F-vector spaces.

Proof It is easy to see that ι is linear, so we do not explicitly show this. To see that ι is injective, suppose that $\iota(g) = 0_{(\mathsf{F}^I_0)'}$. Then, for every $f \in \mathsf{F}^I_0$,

$$\iota(g)(f) = \sum_{i\in I} g(i)f(i) = 0_{\mathsf{F}},$$

and this particular holds when f is the ith standard basis vector for F^I_0. But this then gives $g(i) = 0_{\mathsf{F}}$ for every $i \in I$, whence ι is injective by Exercise 4.5.23. To show that ι is surjective, let $\alpha \in (\mathsf{F}^I_0)'$ and define $g_\alpha \in \mathsf{F}^I$ by $g_\alpha(i) = \alpha(\boldsymbol{e}_i)$, where $\{\boldsymbol{e}_i\}_{i\in I}$ denotes the standard basis. Then

$$\iota(g_\alpha)(f) = \sum_{i\in I} g_\alpha(i)f(i) = \sum_{i\in I} \alpha(\boldsymbol{e}_i)f(i) = \alpha(f),$$

giving $\iota(g_\alpha) = \alpha$, as desired. ■

The previous result gives a very handy and concrete way of viewing the algebraic dual. We shall often simply identify F^I with the algebraic dual of F^I_0 since this is so convenient.

As we saw in Example 5.7.31, if iV is an F-vector space with basis $\{e_i\}_{i\in I}$ then there exists a subset $\{\alpha_i\}_{i\in I}$ of V' where $\alpha_i \in \mathsf{V}'$ is defined by $\alpha_i(e_{i'}) = 1_{\mathsf{V}}$ when $i = i'$ and $\alpha_i(e_{i'}) = 0_{\mathsf{F}}$ when $i \neq i'$. The following result gives some properties of this subset.

5.7.6 Theorem (Dual bases for algebraic duals) *Let F be a field and let V be an F-vector space with basis $\{\mathrm{e}_\mathrm{i}\}_{\mathrm{i}\in\mathrm{I}}$. Let $\{\alpha_\mathrm{i}\}_{\mathrm{i}\in\mathrm{I}} \subseteq \mathsf{V}'$ be as defined in Example 5.7.31. Then the following statements hold:*

(i) the family $(\alpha_\mathrm{i})_{\mathrm{i}\in\mathrm{I}}$ is linearly independent;

(ii) the set $\{\alpha_\mathrm{i}\}_{\mathrm{i}\in\mathrm{I}}$ is a basis for V' if and only if V is finite-dimensional.

Proof To see that $(\alpha_i)_{i\in I}$ is linearly independent, suppose that $i_1, \dots, i_k \in I$ and that

$$c_1\alpha_{i_1} + \cdots + c_k\alpha_{i_k} = 0_{\mathsf{V}'}.$$

Then, for any $i \in I$,

$$c_1\alpha_{i_1}(e_i) + \cdots + c_k\alpha_{i_k}(e_i) = 0_{\mathsf{V}'}.$$

Choosing in particular $i \in \{i_1, \dots, i_k\}$ gives $c_j = 0_{\mathsf{F}}$ for $j \in \{1, \dots, k\}$. This gives linear independence of $(\alpha_i)_{i\in I}$.

Now suppose that V is finite-dimensional and let $\beta \in \mathsf{V}'$. Then take $\beta_j = \beta(e_j)$ for $j \in \{1, \dots, n\}$. Then

$$\begin{aligned}\beta(v_1e_1 + \dots + v_ne_n) &= v_1\beta_1 + \dots + v_n\beta_n \\ &= \beta_1\alpha_1(v_1e_1 + \dots + v_ne_n) + \dots + \beta_n\alpha_n(v_1e_1 + \dots + v_ne_n),\end{aligned}$$

giving

$$\beta = \beta_1\alpha_1 + \dots + \beta_n\alpha_n,$$

and so $\{\alpha_1, \dots, \alpha_n\}$ spans V', and so is a basis.

Finally, suppose that V is infinite-dimensional and define $\beta \in \mathsf{V}'$ by $\beta(e_i) = 1_{\mathsf{F}}$ for $i \in I$. This uniquely defines β by Theorem 4.5.24. We claim that β is not in $\mathrm{span}_{\mathsf{F}}(\alpha_i \mid\ i \in I)$. Indeed, if $\beta \in \mathrm{span}_{\mathsf{F}}(\alpha_i \mid\ i \in I)$ then we have

$$\beta = \beta_1\alpha_{i_1} + \dots + \beta_k\alpha_{i_k}$$

for some $k \in \mathbb{Z}_{>0}$, $i_1, \dots, i_k \in I$, and $\beta_1, \dots, \beta_k \in \mathsf{F}$. However, for $i \notin \{i_1, \dots, i_k\}$ we would then have $\beta(e_i) = 0_{\mathsf{F}}$, giving a contradiction. ■

In the case when V is finite-dimensional the proposition tells us that there is a natural basis for V' associated with every basis for V. This basis has a name.

5.7.7 Definition (Dual basis) If F is a field and if V is a finite-dimensional F-vector space with basis $\{e_1, \dots, e_n\}$, the set $\{\alpha_1, \dots, \alpha_n\}$ as in Example 5.7.31 is the ***dual basis*** for V. •

Let us work this out in the simplest example.

5.7.8 Example (The dual basis for F^n) The standard basis for F^n is denoted by $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$. If we denote the dual basis by $\{\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n\}$ then this basis is defined by

$$\boldsymbol{\alpha}_j(v_1, \dots, v_j, \dots, v_n) = v_j, \qquad j \in \{1, \dots, n\}.$$

It is common to represent elements of F^n by "column vectors" and then elements of $(\mathsf{F}^n)'$ are represented as $1 \times n$-matrices, i.e., as "row vectors." (cf. (5.3)). One must resist at all costs the temptation to think of "column vectors" and "row vectors" as being the same thing, but "transposed." This is gauche at best and wrong at worst. •

By Theorem 5.7.6 we know that, corresponding to a basis $\{e_i\}_{i\in I}$ for V, there is a linearly independent family $(\alpha_i)_{i\in I}$ in V', and that this is a basis when V is finite-dimensional. Thus we can immediately conclude that $\dim_{\mathsf{F}}(\mathsf{V}) \le \dim_{\mathsf{F}}(\mathsf{V}')$ with equality when V is finite-dimensional. The following result says that this is the only occasion when we have equality.

5.7.9 Theorem ($\dim_{\mathsf{F}}(\mathsf{V}) < \dim_{\mathsf{F}}(\mathsf{V}')$ if V is infinite-dimensional) *If* F *is a field and if* V *is an* F*-vector space then*

(i) $\dim_{\mathsf{F}}(\mathsf{V}') = \dim_{\mathsf{F}}(\mathsf{V})$ *if* V *is finite-dimensional and*

(ii) $\dim_{\mathsf{F}}(\mathsf{V}') = \mathrm{card}(\mathsf{F})^{\dim_{\mathsf{F}}(\mathsf{V})}$ *if* V *is infinite-dimensional.*

Proof The first assertion follows from Theorem 5.7.6, so we prove the second assertion. We let I be a set such that $\mathrm{card}(I) = \dim_{\mathsf{F}}(\mathsf{V})$. Thus V is isomorphic to F_0^I by Theorem 4.5.45. By Proposition 5.7.5 it suffices to show that $\dim_{\mathsf{F}}(\mathsf{F}^I) = \mathrm{card}(\mathsf{F})^{\mathrm{card}(I)}$. Let I' be a set such that $\mathrm{card}(I') = \dim_{\mathsf{F}}(\mathsf{V}')$ and let $\{\alpha_a\}_{a\in I'}$ be a basis for F^I. An element of F^I is thus of the form

$$c_1\alpha_{a_1} + \cdots + c_k\alpha_{a_k} \tag{5.35}$$

for some unique $k \in \mathbb{Z}_{\geq 0}, a_1, \ldots, a_k \in I'$, and $c_1, \ldots, c_k \in \mathsf{F}^*$. Thus $\mathrm{card}(\mathsf{F}^I)$ is the number of such linear combinations. To count the number of such linear combinations, consider the linear combinations of the form (5.35) with $k = 1$. Such a linear combination is of the form $c\alpha_a$ for $c \in \mathsf{F}^*$ and $a \in I'$. Thus there are

$$\mathrm{card}(I') \cdot \mathrm{card}(\mathsf{F}^*) = \mathrm{card}(I') \cdot (\mathrm{card}(\mathsf{F}) - 1)$$

elements of the form (5.35) with $k = 1$. In like manner there are $(\mathrm{card}(I') \cdot (\mathrm{card}(\mathsf{F}) - 1))^2$ elements of the form (5.35) with $k = 2$. Including the zero vector and using Theorem 1.7.17(ii) this shows that

$$\mathrm{card}(\mathsf{F}^I) = 1 + \sum_{k=1}^{\infty}(\mathrm{card}(I') \cdot (\mathrm{card}(\mathsf{F}) - 1))^k = \sum_{k=1}^{\infty}(\mathrm{card}(I') \cdot (\mathrm{card}(\mathsf{F}) - 1))^k.$$

By Theorem 1.7.17(iii) we have $\mathrm{card}(I')^k = \mathrm{card}(I')$ for $k \in \mathbb{Z}_{>0}$. Thus

$$\mathrm{card}(\mathsf{F}^I) = \mathrm{card}(I')\sum_{k=1}^{\infty}(\mathrm{card}(\mathsf{F}) - 1)^k.$$

If $\mathrm{card}(\mathsf{F})$ is finite then

$$\sum_{k=1}^{\infty}(\mathrm{card}(\mathsf{F}) - 1)^k = \mathrm{card}(\mathbb{Z}_{>0})$$

(using the rules of cardinal algebra) since the countable union of finite sets is countable by Proposition 1.7.16. Then we have

$$\mathrm{card}(I')\sum_{k=1}^{\infty}(\mathrm{card}(\mathsf{F}) - 1)^k = \mathrm{card}(I')\,\mathrm{card}(\mathbb{Z}_{>0}) = \mathrm{card}(I') = \mathrm{card}(I')\,\mathrm{card}(\mathsf{F})$$

by Corollary 1.7.18. If $\mathrm{card}(\mathsf{F})$ is infinite then

$$\sum_{k=1}^{\infty}(\mathrm{card}(\mathsf{F}) - 1)^k = \sum_{k=1}^{\infty}\mathrm{card}(\mathsf{F}) = \mathrm{card}(\mathsf{F})$$

by Theorem 1.7.17. In this case we again have

$$\operatorname{card}(I')\sum_{k=1}^{\infty}(\operatorname{card}(\mathsf{F})-1)^k = \operatorname{card}(I')\operatorname{card}(\mathsf{F}).$$

This shows that $\operatorname{card}(\mathsf{F}^I) = \operatorname{card}(I')\operatorname{card}(\mathsf{F})$. But by the definition of cardinal arithmetic we have $\operatorname{card}(F^I) = \operatorname{card}(\mathsf{F})^{\operatorname{card}(I)}$. Thus we have

$$\operatorname{card}(\mathsf{F})^{\operatorname{card}(I)} = \operatorname{card}(I')\operatorname{card}(\mathsf{F}).$$

We now prove a lemma.

1 **Lemma** card(F) ≤ card(I′).

Proof It obviously suffices to prove this for the smallest infinite-dimensional F-vector space. That is, it suffices to prove the lemma when $I = \mathbb{Z}_{>0}$. Thus we still denote by $\{\alpha_a\}_{a\in I'}$ a basis for $\mathsf{F}^{\mathbb{Z}_{>0}}$, and now note that these basis elements are sequences in F. Let us suppose that $\operatorname{card}(I') < \operatorname{card}(\mathsf{F})$. We write $\alpha_a = (\alpha_{aj})_{j\in\mathbb{Z}_{>0}}$. Let $S_\alpha = \{\alpha_{aj} \mid a \in I',\ j \in \mathbb{Z}_{>0}\}$. Let K_0 be the smallest subfield of F containing S_α and note from Exercise 4.6.1 that $\mathsf{K}_0 = \mathsf{F}_0(S_\alpha)$ where F_0 is the prime field of F. By Theorem 4.6.6 elements of K_0 are rational functions with indeterminate taken from S_α and with coefficients taken from F_0. Since the prime field F_0 is countable, it follows that $\operatorname{card}(\mathsf{K}_0) \le \operatorname{card}(\mathbb{Z}_{>0})\operatorname{card}(I') = \operatorname{card}(I')$, using Corollary 1.7.18. As we are assuming that $\operatorname{card}(I') < \operatorname{card}(\mathsf{F})$ there exists $a_1 \in \mathsf{F} \setminus \mathsf{K}_0$. Let $\mathsf{K}_1 = \mathsf{K}_0(a_1)$ and note from Theorem 4.6.6 that elements of K_1 are rational functions in indeterminate a_1 with coefficients from K_0. Since each such rational function involves finitely many coefficients from K_0 we have $\operatorname{card}(\mathsf{K}_1) = \operatorname{card}(\mathbb{Z}_{>0})\operatorname{card}(\mathsf{K}_0)$ and so $\operatorname{card}(\mathsf{K}_1) \le \operatorname{card}(I')$. Continuing in this way we arrive at a sequence

$$\mathsf{K}_0 \subset \mathsf{K}_1 \subset \cdots \subset \mathsf{K}_j \subset \cdots$$

of subfields of F with $\mathsf{K}_j = \mathsf{K}_{j-1}(a_j)$ and with $\operatorname{card}(\mathsf{K}_j) \le \operatorname{card}(I')$ for $j \in \mathbb{Z}_{>0}$.

Let $\boldsymbol{a} \in \mathsf{F}^I$ be defined by $\boldsymbol{a}(j) = a_j$ and write

$$\boldsymbol{a} = c_1\alpha_{a_1} + \cdots + c_k\alpha_{a_k} \tag{5.36}$$

for $k \in \mathbb{Z}_{>0}$, $a_1,\dots,a_k \in I'$, and $c_1,\dots,c_k \in \mathsf{F}^*$. Consider the linear map from F^k to F^I defined by

$$(c_1,\dots,c_k) \mapsto c_1\alpha_{a_1} + \cdots + c_k\alpha_{a_k} \tag{5.37}$$

Since $\{\alpha_a\}_{a\in I'}$ is a basis, this map is injective. The matrix representative for this map relative to the standard basis for F^k and the basis $\{\alpha_a\}_{a\in I'}$ for F^I is the matrix $\boldsymbol{A} \in \operatorname{Mat}_{I'\times\{1,\dots,k\}}(\mathsf{K}_0)$ defined by

$$\boldsymbol{A}(a,j) = \begin{cases} \alpha_{a_l j}, & a = a_l \text{ for } l \in \{1,\dots,k\}, \\ 0_{\mathsf{F}}, & a \notin \{a_1,\dots,a_k\} \end{cases}$$

(Note that this matrix *is* over K_0 as asserted.) Thus injectivity of the map (5.37) implies that there exists $j_1,\dots,j_k \in \mathbb{Z}_{>0}$ such that the matrix $\boldsymbol{A}' \in \operatorname{Mat}_{k\times k}(\mathsf{K}_0)$ defined by $\boldsymbol{A}'(r,s) = \boldsymbol{A}(a_r, j_s)$ is invertible. The equation (5.36) then gives

$$a_{j_s} = \boldsymbol{a}(j_s) = \sum_{r=1}^{k} \boldsymbol{A}'(r,s)c_r.$$

Since the entries in $\boldsymbol{A}'$ are in $\boldsymbol{K}_0$ and since $a_{j_s} \in \mathsf{K}_N$ where $N = \max\{j_1, \dots, j_k\}$, it follows by solving the preceding equation for $c_1, \dots, c_k$ that $c_1, \dots, c_k \in \mathsf{K}_N$. But then (5.36) implies that

$$a_j = \boldsymbol{a}(j) = c_1\alpha_{a_1 j} + \cdots + c_k\alpha_{a_k j},$$

giving $a_j \in \mathsf{K}_N$ for all $j \in \mathbb{Z}_{>0}$. This is a contradiction, and so our assumption that $\operatorname{card}(I') < \operatorname{card}(\mathsf{F})$ must be invalid. ▼

Now we use the lemma to complete the proof of the theorem. Prior to the lemma we showed that $\operatorname{card}(\mathsf{F})^{\operatorname{card}(I)} = \operatorname{card}(I')\operatorname{card}(\mathsf{F})$. If $\operatorname{card}(\mathsf{F}) \leq \operatorname{card}(I')$ as asserted by the lemma then we have

$$\operatorname{card}(I')\operatorname{card}(\mathsf{F}) \leq \operatorname{card}(I')^2 = \operatorname{card}(I')$$

by Theorem 1.7.17(iii). Thus $\operatorname{card}(I')\operatorname{card}(\mathsf{F}) = \operatorname{card}(I')$ and so $\operatorname{card}(I') = \operatorname{card}(\mathsf{F})^{\operatorname{card}(I)}$ as desired. ■

For the cases of most interest to us, this gives the following result.

5.7.10 Corollary (Dimension of algebraic duals over $\mathbb{R}$ and $\mathbb{C}$) *Let $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and let V be an infinite-dimensional $\mathbb{F}$-vector space. Then $\dim_{\mathsf{F}}(\mathsf{V}') = 2^{\dim_{\mathsf{F}}(\mathsf{V})}$.*

Proof By Exercise 1.7.5 we have $\operatorname{card}(\mathbb{R}) = 2^{\operatorname{card}(\mathbb{Z}_{>0})}$ and, since $\mathbb{C}$ is a two-dimensional $\mathbb{R}$-vector space, by Corollary 4.5.47 we have $\operatorname{card}(\mathbb{C}) = \operatorname{card}(\mathbb{R}) = 2^{\operatorname{card}(\mathbb{Z}_{>0})}$. Then

$$\dim_{\mathsf{F}}(\mathsf{V}') = (2^{\operatorname{card}(\mathbb{Z})})^{\dim_{\mathsf{F}}(\mathsf{V})} = 2^{\operatorname{card}(\mathbb{Z})\dim_{\mathsf{F}}(\mathsf{V})} = 2^{\dim_{\mathsf{F}}(\mathsf{V})}$$

by Corollary 1.7.18. ■

This shows that, for $\mathbb{R}$- and $\mathbb{C}$-vector spaces, the dimension of the algebraic dual is significantly larger than that of the vector space.

5.7.3 Algebraic reflexivity

There is a natural map from V to V'' which we denote by ι_{V} and define by

$$\iota_{\mathsf{V}}(v) \cdot \alpha = \alpha(v).$$

This map is injective.

5.7.11 Proposition (V is a subspace of V'') *Let F and let V be an F-vector space. The map ι_{V} is injective so that $\operatorname{image}(\iota_{\mathsf{V}})$ is a subspace of V'' isomorphic to V. Moreover, ι_{V} is surjective if and only if V is finite-dimensional.*

Proof If $\iota_{\mathsf{V}}(v) = 0_{\mathsf{V}''}$ then $\alpha(v) = 0_{\mathsf{F}}$ for every $\alpha \in \mathsf{V}'$. We claim that this implies that $v = 0_{\mathsf{V}}$. Suppose that $v \neq 0_{\mathsf{V}}$. Then, by Proposition 5.7.4, there exists $\alpha \in \mathsf{V}'$ such that $\alpha(v) = 1_{\mathsf{F}}$. This then precludes having $\alpha(v) = 0_{\mathsf{F}}$ for every $\alpha \in \mathsf{V}'$. This then gives injectivity of ι_{V}.

Now suppose that V is finite-dimensional. Then $\dim_{\mathsf{F}}(\mathsf{V}') = \dim_{\mathsf{F}}(\mathsf{V})$ by Theorem 5.7.6. This also gives $\dim_{\mathsf{F}}(\mathsf{V}'') = \dim_{\mathsf{F}}(\mathsf{V})$ and so ι_{V} is an isomorphism by Corollary 5.4.44.

Finally, suppose that V is infinite-dimensional and let $\{e_i\}_{i\in I}$ be a basis for V. Let $\{\alpha_i\}_{i\in I}$ be the subset of V' as in Example 5.7.3–1. Now let $\beta \in \mathsf{V}'$ be defined (using Theorem 4.5.24) by $\beta(e_i) = 1_{\mathsf{F}}$ for $i \in I$. Note that $\beta \notin \operatorname{span}_{\mathsf{F}}(\alpha_i|\ i \in I)$ as we saw in the proof of Theorem 5.7.6. Thus the set $\{\alpha_i\}_{i\in I} \cup \{\beta\}$ is linearly independent. Let $\{\beta_j\}_{j\in J}$ be a basis for V' that contains $\{\alpha_i\}_{i\in I} \cup \{\beta\}$. Next let $\{\omega_j\}_{j\in J}$ be the subset of V'' defined by $\omega_j(\beta_{j'}) = 1_{\mathsf{F}}$ if $j = j'$ and $\omega_j(\beta_{j'}) = 0_{\mathsf{F}}$ otherwise. Thus $\{\omega_j\}_{j\in J}$ corresponds to the basis $\{\beta_j\}_{j\in J}$ as in Example 5.7.3–1, and so is linearly independent by Theorem 5.7.6. We claim that $\iota_{\mathsf{V}}(\{e_i\}_{i\in I}) \subseteq \{\omega_j\}_{j\in J}$. Since $\{\alpha_i\}_{i\in I} \subseteq \{\beta_j\}_{j\in J}$, for each $i \in I$ there exists $j_i \in J$ such that $\beta_{j_i} = \alpha_i$. Then, for any $j \in J$ we directly have $\omega_{j_i}(\beta_j) = \beta_j(e_{j_i})$, so giving $\iota_{\mathsf{V}}(e_i) = \omega_{j_i}$. Now let $j_0 \in J$ be such that $\omega_{j_0}(\beta) = 1_{\mathsf{F}}$ and $\omega_{j_0}(\beta_j) = 0_{\mathsf{F}}$ for all j such that $\beta_j \neq 0_{\mathsf{F}}$. Then

$$\omega_{j_0} \notin \operatorname{span}_{\mathsf{F}}(\omega_{j_i}|\ i \in I) = \operatorname{image}(\iota_{\mathsf{V}}),$$

so showing that ι_{V} is not surjective. ■

Let us give some terminology associated with the map ι_{V}.

5.7.12 Definition (Algebraic reflexivity) Let F be a field and let V be an F-vector space. If $\iota_{\mathsf{V}} \in \operatorname{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V}'')$ is an isomorphism then V is ***algebraically reflexive***. •

The following result is then simple a rephrasing of Proposition 5.7.11.

5.7.13 Corollary (Only finite-dimensional vector spaces are algebraically reflexive) *If* F *is a field, an* F*-vector space is algebraically reflexive if and only if* V *is finite-dimensional.*

In the algebraic setup, algebraic reflexivity is simply not very interesting. However, we present the idea in order to contrast it with the more interesting notion of topological reflexivity that we will consider in Sections III-3.9.1 and III-6.4.2.

5.7.4 Annihilators and coannihilators

It is often interesting to examine elements of the algebraic dual that evaluate to zero on subsets of the vector space. The terminology and notation for organising this is as follows.

5.7.14 Definition (Annihilator, coannihilator) Let F be a field and let V be an F-vector space with $S \subseteq \mathsf{V}$ and $\Lambda \subseteq \mathsf{V}'$.

(i) The ***annihilator*** of S is the subset

$$\operatorname{ann}(S) = \{\alpha \in \mathsf{V}' \mid \alpha(v) = 0 \text{ for all } v \in S\}$$

of V'.

(ii) The ***coannihilator*** of Λ is the subset

$$\operatorname{coann}(\Lambda) = \{v \in \mathsf{V} \mid \alpha(v) = 0 \text{ for all } \alpha \in \Lambda\}$$

of V. •

Let us record some of the properties of annihilators and coannihilators.

5.7.15 Proposition (Properties of annihilators and coannihilators) *Let* F *be a field and let* V *be an* F*-vector space with* $\mathsf{S}, \mathsf{T} \subseteq \mathsf{V}$ *and* $\Lambda, \Theta \subseteq \mathsf{V}'$*. Then the following statements hold:*

(i) $\mathrm{ann}(\mathsf{S})$ *is a subspace of* V'*;*

(ii) $\mathrm{coann}(\Lambda)$ *is a subspace of* V*;*

(iii) if $\mathsf{S} \subseteq \mathsf{T}$ *then* $\mathrm{ann}(\mathsf{T}) \subseteq \mathrm{ann}(\mathsf{S})$*;*

(iv) if $\Lambda \subseteq \Theta$ *then* $\mathrm{coann}(\Theta) \subseteq \mathrm{coann}(\Lambda)$*;*

(v) $\mathsf{S} \subseteq \mathrm{coann}(\mathrm{ann}(\mathsf{S}))$*;*

(vi) $\Lambda \subseteq \mathrm{ann}(\mathrm{coann}(\Lambda))$*.*

Proof (i) and (ii) These are Exercise 5.7.1.

(iii) If $\alpha \in \mathrm{ann}(T)$ then $\alpha(v) = 0_{\mathsf{F}}$ for every $v \in T$. In particular, $\alpha(v) = 0_{\mathsf{F}}$ for every $v \in S$ and so $\alpha \in \mathrm{ann}(S)$.

(iv) If $v \in \mathrm{coann}(\Theta)$ then $\alpha(v) = 0_{\mathsf{F}}$ for every $\alpha \in \Theta$. In particular, $\alpha(v) = 0_{\mathsf{F}}$ for every $\alpha \in \Lambda$ and so $v \in \mathrm{coann}(\Lambda)$.

(v) Let $v \in S$ and $\alpha \in \mathrm{ann}(S)$. Then $\alpha(v) = 0_{\mathsf{F}}$. As this holds for every $\alpha \in \mathrm{ann}(S)$ it follows that $v \in \mathrm{coann}(\mathrm{ann}(S))$.

(vi) Let $\alpha \in \Lambda$ and let $v \in \mathrm{coann}(\Lambda)$. Then $\alpha(v) = 0_{\mathsf{F}}$. As this holds for every $v \in \mathrm{coann}(\Lambda)$ it follows that $\alpha \in \mathrm{ann}(\mathrm{coann}(\Lambda))$. ■

Clearly there is no hope that $S = \mathrm{coann}(\mathrm{ann}(S))$ (resp. $\Lambda = \mathrm{ann}(\mathrm{coann}(\Lambda))$) unless S (resp. Λ) is a subspace. One can then ask whether this holds when S (resp. Λ) *is* a subspace. Here is some terminology associated with this question.

5.7.16 Definition (ann-closed, coann-closed) Let F be a field and let V be an F-vector space.

(i) A subspace $\mathsf{U} \subseteq \mathsf{V}$ is ***ann-closed*** if $\mathsf{U} = \mathrm{coann}(\mathrm{ann}(\mathsf{U}))$.

(ii) A subspace $\Lambda \subseteq \mathsf{V}'$ is ***coann-closed*** if $\Lambda = \mathrm{ann}(\mathrm{coann}(\Lambda))$. •

As the following result indicates, the notion of ann-closed is simple, but the notion of coann-closed can be complicated in infinite-dimensions.

5.7.17 Theorem (Characterisations of ann- and coann-closed subspaces) *Let* F *be a field and let* V *be an* F*-vector space. Then the following statements hold:*

(i) every subspace of V *is* ann*-closed;*

(ii) every subspace of V' *is* coann*-closed if and only if* V *is finite-dimensional;*

(iii) if Λ *is a finite-dimensional subspace of* V' *then* Λ *is* coann*-closed.*

Proof (i) By Proposition 5.7.15 we have $\mathsf{U} \subseteq \mathrm{coann}(\mathrm{ann}(\mathsf{U}))$. Now let $v \in \mathsf{V} \setminus \mathsf{U}$. By Proposition 5.7.4 let $\alpha \in \mathsf{V}'$ have the property that $\alpha(u) = 0_{\mathsf{F}}$ for every $u \in \mathsf{U}$ and that $\alpha(v) = 1_{\mathsf{F}}$. Thus $\alpha \in \mathrm{ann}(\mathsf{U})$ but $\alpha(v) \neq 0_{\mathsf{F}}$. In other words, $v \notin \mathrm{coann}(\mathrm{ann}(\mathsf{U}))$. Thus $\mathsf{V} \setminus \mathsf{U} \subseteq \mathsf{V} \setminus \mathrm{coann}(\mathrm{ann}(\mathsf{U}))$, i.e., $\mathrm{coann}(\mathrm{ann}(\mathsf{U})) \subseteq \mathsf{U}$.

(ii) First suppose that V is finite-dimensional so that V is finite-dimensional with the same dimension as V by Theorem 5.7.6. Let $\{\alpha_1, \dots, \alpha_n\}$ be a basis for V' such that $\{\alpha_1, \dots, \alpha_k\}$ is a basis for Λ. Denote the dual basis for V'' by $\{\omega_1, \dots, \omega_n\}$ and let

$e_j = \iota_{\mathsf{V}}^{-1}(\omega_j)$, $j \in \{1,\dots,n\}$, so that $\{e_1,\dots,e_n\}$ is a basis for V. We claim that $\{\omega_{k+1},\dots,\omega_n\}$ is a basis for $\operatorname{ann}(\Lambda)$. It is clear that $\omega_j(\alpha_l) = 0_{\mathsf{F}}$ for $j \in \{k+1,\dots,n\}$ and $l \in \{1,\dots,k\}$. Thus

$$\operatorname{span}_{\mathsf{F}}(\omega_{k+1},\dots,\omega_n) \subseteq \operatorname{ann}(\Lambda).$$

Now suppose that

$$\theta = c_1\omega_1 + \dots + c_n\omega_n \in \operatorname{ann}(\Lambda).$$

Then $\theta(\alpha_j) = c_j$, $j \in \{1,\dots,n\}$, and so θ being in $\operatorname{ann}(\Lambda)$ implies that $c_1 = \dots = c_k = 0_{\mathsf{V}}$. Thus θ is a linear combination of $\omega_{k+1},\dots,\omega_n$, giving $\{\omega_{k+1},\dots,\omega_n\}$ as a basis for $\operatorname{ann}(\Lambda)$, as desired. We now claim that ι_{V}^{-1} maps $\operatorname{ann}(\Lambda)$ isomorphically to $\operatorname{coann}(\Lambda)$. First of all, for $j \in \{k+1,\dots,n\}$ and $l \in \{1,\dots,k\}$ we have

$$\alpha_l(\iota_{\mathsf{V}}^{-1}(\omega_j)) = \omega_j(\alpha_l) = 0_{\mathsf{F}},$$

giving $\iota_{\mathsf{V}}^{-1}(\omega_j) \in \operatorname{coann}(\Lambda)$ for $j \in \{k+1,\dots,n\}$. Therefore, $\iota_{\mathsf{V}}^{-1}(\operatorname{ann}(\Lambda)) \subseteq \operatorname{coann}(\Lambda)$. Now suppose that $u \in \operatorname{coann}(\Lambda)$ and write

$$u = c_1e_1 + \dots + c_ne_n.$$

Then, for $j \in \{1,\dots,k\}$, $\alpha_j(u) = c_j = 0_{\mathsf{F}}$ since $u \in \operatorname{coann}(\Lambda)$. Thus u is a linear combination of $e_{k+1},\dots,e_n$. Thus $u \in \operatorname{image}(\iota_{\mathsf{V}}^{-1})$, giving $\operatorname{coann}(\Lambda) = \iota_{\mathsf{V}}^{-1}(\operatorname{ann}(\Lambda))$, as desired. Finally we show that $\operatorname{ann}(\operatorname{coann}(\Lambda)) = \Lambda$ by showing that $\{\alpha_1,\dots,\alpha_k\}$ is a basis for $\operatorname{ann}(\operatorname{coann}(\Lambda))$. First of all, $\alpha_j(e_l) = 0_{\mathsf{F}}$ for $j \in \{1,\dots,k\}$ and $l \in \{k+1,\dots,n\}$. Thus $\alpha_1,\dots,\alpha_k \in \operatorname{ann}(\operatorname{coann}(\Lambda))$. If

$$\beta = c_1\alpha_1 + \dots + c_n\alpha_n \in \operatorname{ann}(\operatorname{coann}(\Lambda))$$

then we have $\beta(e_j) = 0_{\mathsf{F}}$ for $j \in \{k+1,\dots,n\}$, showing that β is a linear combination of $\alpha_1,\dots,\alpha_k$. Thus $\{\alpha_1,\dots,\alpha_k\}$ is a basis for $\operatorname{ann}(\operatorname{coann}(\Lambda))$ as desired.

Now suppose that V is infinite-dimensional. By choosing a basis we suppose without loss of generality that $\mathsf{V} = \mathsf{F}_0^I$ and that $\mathsf{V}' = \mathsf{F}^I$ for some set I (see Proposition 5.7.5). In this case $\mathsf{F}_0^I \subseteq \mathsf{F}^I$. We claim that $\operatorname{coann}(\mathsf{F}_0^I) = \{0_{\mathsf{F}_0^I}\}$. Indeed, let $f\colon I \to \mathsf{F}$ be an element of $\operatorname{coann}(\mathsf{F}_0^I)$. Think of the standard basis vector $\boldsymbol{e}_i$ for F_0^I as being an element of F^I. Since $f \in \operatorname{coann}(\mathsf{F}_0^I)$ we have $\boldsymbol{e}_i(f) = f(i) = 0_{\mathsf{F}}$. Since this is true for every $i \in I$ we have $f = 0_{\mathsf{F}_0^I}$, as claimed. Now it immediately follows that $\operatorname{ann}(\operatorname{coann}(\mathsf{F}_0^I)) = \mathsf{F}^I$ and so $\mathsf{F}_0^I \subset \operatorname{ann}(\operatorname{coann}(\mathsf{F}_0^I))$, giving this part of the theorem.

(iii) We let Λ be a finite-dimensional subspace of V' with basis $\{\alpha_1,\dots,\alpha_k\}$.

We claim that there are vectors $e_1,\dots,e_k \in \mathsf{V}$ such that $\alpha_j(e_j) = 1_{\mathsf{F}}$, $j \in \{1,\dots,k\}$, and such that $\alpha_j(e_l) = 0_{\mathsf{F}}$ wherever $j,l \in \{1,\dots,k\}$ satisfy $j \neq l$. We prove this by induction on k. Since α_1 is nonzero there exists $e_1 \in \mathsf{V}$ such that $\alpha_1(e_1) = 1_{\mathsf{F}}$. Now suppose that there exists $e_1,\dots,e_{k-1} \in \mathsf{V}$ having the claimed properties. Then, for $v \in \mathsf{V}$, write

$$v = \langle\alpha_1; v\rangle e_1 + \dots + \langle\alpha_{k-1}; v\rangle e_{k-1} + v' \tag{5.38}$$

where

$$v' = v - (\langle\alpha_1; v\rangle e_1 + \dots + \langle\alpha_{k-1}; v\rangle e_{k-1}).$$

Then

$$\alpha_j(v') = \alpha_j(v) - \langle\alpha_j; v\rangle = 0_{\mathsf{F}}, \qquad j \in \{1,\dots,k-1\}.$$

We now claim that there exists some $v \in \mathsf{V}$ such that $\alpha_k(v') = 1_{\mathsf{F}}$, where v' is as defined above. Suppose that $\alpha_k(v') = 0_{\mathsf{F}}$ for all $v \in \mathsf{V}$. Then this implies that

$$\left\langle \alpha_k - \sum_{j=1}^{k-1} \langle \alpha_k; e_j \rangle \alpha_j; v \right\rangle = 0_{\mathsf{V}}$$

for all $v \in \mathsf{V}$. But this contradicts the linear independence of $\{\alpha_1, \dots, \alpha_k\}$. Thus let $e_k \in \mathsf{V}$ satisfy $\alpha_k(e_k) = 1_{\mathsf{F}}$ and also define

$$\tilde{e}_j = e_j - \langle \alpha_k; e_j \rangle e_k, \qquad j \in \{1, \dots, k-1\}.$$

Direct computation then gives

$$\begin{aligned} \alpha_j(\tilde{e}_j) &= 1_{\mathsf{F}}, && j \in \{1, \dots, k-1\}, \\ \alpha_j(\tilde{e}_l) &= 0_{\mathsf{F}}, && j, l \in \{1, \dots, k-1\},\ j \neq l, \\ \alpha_k(\tilde{e}_j) &= 0_{\mathsf{F}}, && j \in \{1, \dots, k-1\}. \end{aligned}$$

The set $\{\tilde{e}_1, \dots, \tilde{e}_{k-1}, e_k\}$ now has the asserted properties, and our claim follows.

The vectors $\{e_1, \dots, e_k\}$ are easily shown to be linearly independent. We claim that $\mathsf{V} = \operatorname{coann}(\Lambda) \oplus \operatorname{span}_{\mathsf{F}}(e_1, \dots, e_k)$. Indeed, if $v \in \mathsf{V}$ we can write, as in (5.38),

$$v = \langle \alpha_1; v \rangle e_1 + \dots + \langle \alpha_k; v \rangle e_k + v'$$

where

$$v' = v - (\langle \alpha_1; v \rangle e_1 + \dots + \langle \alpha_k; v \rangle e_k).$$

Just as above we have $\alpha_j(v') = 0_{\mathsf{F}}$, $j \in \{1, \dots, k\}$, so that $v' \in \operatorname{coann}(\Lambda)$. Thus $\mathsf{V} = \operatorname{coann}(\Lambda) + \operatorname{span}_{\mathsf{F}}(e_1, \dots, e_k)$. Moreover, if $v \in \operatorname{coann}(\Lambda) \cap \operatorname{span}_{\mathsf{F}}(e_1, \dots, e_n)$ we have

$$v = c_1 e_1 + \dots + c_k e_k$$

giving $\alpha_j(v) = v_j = 0_{\mathsf{F}}$, $j \in \{1, \dots, k\}$. Thus $v = 0_{\mathsf{F}}$ which gives our claim.

Now suppose that $\beta \in \operatorname{ann}(\operatorname{coann}(\Lambda))$. For $v \in \mathsf{V}$ write $v = v' + v''$ with $v' \in \operatorname{coann}(\Lambda)$ and $v'' \in \operatorname{span}_{\mathsf{F}}(e_1, \dots, e_k)$. Let us additionally write

$$v'' = c_1 e_1 + \dots + c_k e_k.$$

We then have, using the fact that $\Lambda \subseteq \operatorname{ann}(\operatorname{coann}(\Lambda))$,

$$\begin{aligned} \left\langle \beta - \sum_{j=1}^{k} \langle \beta; e_j \rangle \alpha_k; v \right\rangle &= \left\langle \beta - \sum_{j=1}^{k} \langle \beta; e_j \rangle \alpha_k; v' \right\rangle + \left\langle \beta - \sum_{j=1}^{k} \langle \beta; e_j \rangle \alpha_k; v'' \right\rangle \\ &= \left\langle \beta; \sum_{l=1}^{k} c_l e_l \right\rangle - \left\langle \sum_{j=1}^{k} \langle \beta; e_j \rangle \alpha_k; \sum_{l=1}^{k} c_l e_l \right\rangle \\ &= \sum_{l=1}^{k} c_l \langle \beta; e_l \rangle - \sum_{l=1}^{k} c_l \langle \beta; e_l \rangle = 0_{\mathsf{F}}. \end{aligned}$$

Therefore,

$$\beta = \sum_{j=1}^{k} \langle \beta; e_j \rangle \alpha_k$$

and so $\mathrm{ann}(\mathrm{coann}(\Lambda)) \subseteq \Lambda$. ■

The theorem illustrates nicely why the algebraic dual in infinite-dimensions is so much more difficult to deal with than in finite-dimensions. The following result gives an additional illustration of this.

5.7.18 Proposition ((Co)annihilators of (co)annihilators) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{U} \subseteq \mathsf{V}$ *be a subspaces. Then the following statements hold:*

(i) $\mathrm{ann}(\mathrm{ann}(\mathsf{U})) \cap \iota_{\mathsf{V}}(\mathsf{V}) = \iota_{\mathsf{V}}(\mathsf{U})$*;*

(ii) $\iota_{\mathsf{V}}(\mathsf{U}) = \mathrm{ann}(\mathrm{ann}(\mathsf{U}))$ *if and only if* V *is finite-dimensional.*

Proof (i) Let $\alpha \in \mathrm{ann}(\mathsf{U})$ and let $u \in \mathsf{U}$. Then $\langle \iota_{\mathsf{V}}(u); \alpha \rangle = \langle \alpha; u \rangle = 0_{\mathsf{F}}$ giving $\iota_{\mathsf{V}}(u) \in \mathrm{ann}(\mathrm{ann}(\mathsf{U}))$. Thus $\iota_{\mathsf{V}}(\mathsf{U}) \subseteq \mathrm{ann}(\mathrm{ann}(\mathsf{U})) \cap \iota_{\mathsf{V}}(\mathsf{V})$. Now let $\omega \in \mathrm{ann}(\mathrm{ann}(\mathsf{U})) \cap \iota_{\mathsf{V}}(\mathsf{V})$. Then $\omega = \iota_{\mathsf{V}}(v)$ for some $v \in \mathsf{V}$. For $\alpha \in \mathrm{ann}\,\mathsf{U}$ we then have $\langle \omega; \alpha \rangle = \langle \alpha; v \rangle = 0_{\mathsf{F}}$. This shows that $v \in \mathrm{coann}(\mathrm{ann}(\mathsf{U}))$ and so $v \in \mathsf{U}$ by Theorem 5.7.17(i).

(ii) Suppose that V is finite-dimensional. During the course of the proof of part (ii) of Theorem 5.7.17 we showed that if V is finite-dimensional and if Λ is a subspace of V', then ι_{V} maps $\mathrm{coann}(\Lambda)$ isomorphically onto $\mathrm{ann}(\Lambda)$. Taking $\Lambda = \mathrm{ann}(\mathsf{U})$ shows that ι_{V} maps $\mathrm{coann}(\mathrm{ann}(\mathsf{U})) = \mathsf{U}$ isomorphically onto $\mathrm{ann}(\mathrm{ann}(\mathsf{U}))$ which is one part of this part of the result.

Now suppose that V is infinite-dimensional. By choosing a basis we can assume that $\mathsf{V} = \mathsf{F}_0^I$ for some appropriate index set I. Then $\mathsf{V}' = \mathsf{F}^I$ by Proposition 5.7.5. Taking $\mathsf{U} = \mathsf{F}_0^I$ we claim that $\mathrm{ann}(\mathsf{U}) = \{0_{\mathsf{F}^I}\}$. Indeed, if $\alpha \in \mathrm{ann}(\mathsf{U})$ then $\alpha(\boldsymbol{e}_i) = 0_{\mathsf{F}}$ for every $i \in I$, here $\{\boldsymbol{e}_i\}_{i \in I}$ is the standard basis. But this implies that $\alpha(i) = 0_{\mathsf{F}}$ for each $i \in I$, as desired. But now $\mathrm{ann}(\mathsf{U}) = (\mathsf{F}^I)'$. However, $\iota_{\mathsf{F}_0^I}(\mathsf{U}) = \mathrm{image}(\iota_{\mathsf{F}_0^I})$ is a strict subspace of $(\mathsf{F}^I)'$ since F_0^I is not algebraically reflexive. ■

5.7.5 Duals of linear maps

We close this section by showing that the algebraic dual provides the proper setup for understand the transpose of a matrix. The key idea is the following.

5.7.19 Definition (Dual of a linear map) Let F be a field and let U and V be F-vector spaces. The ***dual*** of $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ is the linear map $\mathsf{L}' \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}'; \mathsf{U}')$ defined by

$$\langle \mathsf{L}'(\alpha); u \rangle = \langle \alpha; \mathsf{L}(u) \rangle, \qquad \alpha \in \mathsf{V}',\ u \in \mathsf{U}.$$ •

One should first understand that the definition does actually define a map from V' to U'. That is to say, the definition gives $\mathsf{L}'(\alpha)$ as an element of U'. This is straightforward, but requires a moment's thought. It is also easy to see that L' is linear. Indeed, for $\alpha_1, \alpha_2 \in \mathsf{V}'$ and for $u \in \mathsf{U}$,

$$\begin{aligned} \langle \mathsf{L}'(\alpha_1 + \alpha_2); u \rangle &= \langle \alpha_1 + \alpha_2; \mathsf{L}(u) \rangle = \langle \alpha_1; \mathsf{L}(u) \rangle + \langle \alpha_2; \mathsf{L}(u) \rangle \\ &= \langle \mathsf{L}'(\alpha_1); u \rangle + \langle \mathsf{L}'(\alpha_2); u \rangle = \langle \mathsf{L}'(\alpha_1) + \mathsf{L}'(\alpha_2); u \rangle, \end{aligned}$$

giving $\mathsf{L}'(\alpha_1+\alpha_2) = \mathsf{L}'(\alpha_1)+\mathsf{L}'(\alpha_2)$. An altogether similar computation gives $\mathsf{L}'(a\alpha) = a\mathsf{L}'(\alpha)$ for $a \in \mathsf{F}$ and $\alpha \in \mathsf{V}'$.

Let us consider the basic properties of the dual of a linear map.

5.7.20 Proposition (Algebraic properties of the algebraic dual) *Let F be a field and let U, V, and W be F-vector spaces. The following statements hold:*

(i) the map $\mathsf{L} \mapsto \mathsf{L}'$ is an injective linear map from $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ to $\mathrm{Hom}_{\mathsf{F}}(\mathsf{V}';\mathsf{U}')$, and is an isomorphism if and only if V is finite-dimensional;

(ii) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ and $\mathsf{K} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{W})$ then $(\mathsf{K}\circ\mathsf{L})' = \mathsf{L}'\circ\mathsf{K}'$;

(iii) if $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ is an isomorphism then $(\mathsf{L}')^{-1} = (\mathsf{L}^{-1})'$.

Proof (i) It is easy to show that

$$(\mathsf{L}_1+\mathsf{L}_2)' = \mathsf{L}_1' + \mathsf{L}_2', \quad (a\mathsf{L})' = a\mathsf{L}'$$

for $a \in \mathsf{F}$ and $\mathsf{L}, \mathsf{L}_1, \mathsf{L}_2 \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$.

We now show that the map $\mathsf{L} \mapsto \mathsf{L}'$ is injective. Suppose that $\mathsf{L}' = 0_{\mathrm{Hom}_{\mathsf{F}}(\mathsf{V}';\mathsf{U}')}$ so $\mathsf{L}'(\alpha) = 0_{\mathsf{U}'}$ for every $\alpha \in \mathsf{V}'$. Thus

$$\langle \mathsf{L}'(\alpha); u\rangle = \langle \alpha; \mathsf{L}(u)\rangle = 0_{\mathsf{F}}, \qquad \alpha \in \mathsf{V}',\ u \in \mathsf{U}. \tag{5.39}$$

We claim that this implies that $\mathsf{L}(u) = 0_{\mathsf{V}}$ for every $u \in \mathsf{U}$. If not, let $\mathsf{L}(u_0) \neq 0_{\mathsf{V}}$. By Proposition 5.7.4 it follows that there exists $\alpha_0 \in \mathsf{V}'$ such that $\langle \alpha_0; \mathsf{L}(u_0)\rangle = 1_{\mathsf{F}}$, in contradiction with (5.39). Thus $\mathsf{L}(u) = 0_{\mathsf{V}}$ for every $u \in \mathsf{U}$ so $\mathsf{L} = 0_{\mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})}$. Therefore, the map $\mathsf{L} \mapsto \mathsf{L}'$ is injective by Exercise 4.5.23.

Now suppose that V is finite-dimensional. Let us choose bases $\mathscr{B}_{\mathsf{U}} = \{e_j\}_{j\in J}$ and $\mathscr{B}_{\mathsf{V}} = \{f_i\}_{i\in I}$ for U and V, respectively. Since I is finite let us take $I = \{1,\dots,m\}$. We first claim that the map $\boldsymbol{A} \mapsto \boldsymbol{A}^T$ from the set of column finite matrices in $\mathrm{Mat}_{I\times J}(\mathsf{F})$ to $\mathrm{Mat}_{J\times I}(\mathsf{F})$ is surjective. Indeed, let $\boldsymbol{B} \in \mathrm{Mat}_{J\times I}(\mathsf{F})$ so that $\boldsymbol{B}^T \in \mathrm{Mat}_{I\times J}(\mathsf{F})$. Since I is finite, $\boldsymbol{B}^T$ is column finite. Taking $\boldsymbol{A} = \boldsymbol{B}^T$ gives $\boldsymbol{A} \mapsto \boldsymbol{B}$, giving the desired surjectivity. Now let $\mathsf{K} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{V}';\mathsf{U}')$. Since V' and U' are isomorphic to F^I and F^J, respectively, via the isomorphisms from V to F_0^I and from V to F_0^J, respectively (cf. part (iii)), it follows that K gives rise to a homomorphism from F^I to F^J. By Theorem 5.1.13(iv) it follows that there exists a row finite matrix $\boldsymbol{B} \in \mathrm{Mat}_{J\times I}(\mathsf{F})$ for which

$$\boldsymbol{x} \ni \mathsf{F}^I \mapsto \boldsymbol{B}\boldsymbol{x} \in \mathsf{F}^J$$

is the homomorphism associated to K. Now, by our previous claim, there exists a column finite matrix $\boldsymbol{A} \in \mathrm{Mat}_{I\times J}(\mathsf{F})$ such that $\boldsymbol{A}^T = \boldsymbol{B}$. Let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U};\mathsf{V})$ be associated with $\boldsymbol{A}$ as in Theorem 5.4.21. By Theorem 5.7.22(i) it follows that $\mathsf{L}' = \mathsf{K}$. This shows that the map $\mathsf{L} \mapsto \mathsf{L}'$ is surjective, as desired.

Finally suppose that V is infinite-dimensional. Similarly with our argument in the preceding paragraph, it suffices to prove that there is a homomorphism of F^I to F^J that does not correspond to a row finite matrix. However, this is exactly what we showed in Theorem 5.1.13(v).

(ii) For $\beta \in \mathsf{W}'$ and $u \in \mathsf{U}$ we have

$$\langle (\mathsf{K}\circ\mathsf{L})'(\beta); u\rangle = \langle \beta; \mathsf{K}\circ\mathsf{L}(u)\rangle = \langle \mathsf{K}'(\beta); \mathsf{L}(u)\rangle = \langle \mathsf{L}'\circ\mathsf{K}'(\beta); u\rangle,$$

which gives $(\mathsf{K} \circ \mathsf{L})' = \mathsf{L}' \circ \mathsf{K}'$ as desired.

(iii) For $\alpha \in \mathsf{U}'$ and $u \in \mathsf{U}$ we have

$$\langle \alpha; u \rangle = \langle \alpha; \mathsf{L}^{-1} \circ \mathsf{L}(u) \rangle = \langle (\mathsf{L}^{-1})'(\alpha); \mathsf{L}(u) \rangle = \langle \mathsf{L}' \circ (\mathsf{L}^{-1})'(\alpha); u \rangle,$$

giving $\mathsf{L}' \circ (\mathsf{L}^{-1})' = \mathrm{id}_{\mathsf{U}'}$. In a similar manner we get

$$\langle \beta; v \rangle = \langle (\mathsf{L}^{-1})' \circ \mathsf{L}'(\beta); v \rangle, \qquad \beta \in \mathsf{V}',\ v \in \mathsf{V}.$$

Thus $(\mathsf{L}^{-1})' \circ \mathsf{L}' = \mathrm{id}_{\mathsf{V}'}$, and so L' is invertible with inverse $(\mathsf{L}^{-1})'$. ■

Part (iii) has the following corollary.

5.7.21 Corollary (The algebraic dual is an isomorphism invariant) *If F is a field and if U and V are isomorphic F-vector spaces, then U' and V' are isomorphic.*

Now we consider matrices associated to duals of linear maps. In finite-dimensions this is straightforward, but in infinite-dimensions the situation is not complicated, but does require some interpretation. This is because, given a basis for a vector space there is only a natural choice of basis for the algebraic dual when the vector space is finite-dimensional; see Theorem 5.7.6. The result one can prove is the following, whose statement relies on Proposition 5.7.5 and Proposition 5.7.20(iii) for its statement.

5.7.22 Theorem (Matrices and duals) *Let F be a field, let U and V be F-vector spaces, and let $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$. Let $\mathscr{B}_{\mathsf{U}} = \{\mathrm{e}_j\}_{j \in J}$ and $\mathscr{B}_{\mathsf{V}} = \{\mathrm{f}_i\}_{i \in I}$ be bases for U and V, respectively, with $\{\alpha_j\}_{j \in J}$ and $\{\beta_i\}_{i \in I}$ the linearly independent subsets of U' and V' as in Example 5.7.3–1. Let $\phi_J \colon \mathsf{U} \to \mathsf{F}_0^J$ and $\phi_I \colon \mathsf{V} \to \mathsf{F}_0^I$ be the isomorphisms associated with the bases as in Theorem 4.5.45. Then the following statements hold:*

(i) the following diagram commutes:

$$\begin{tikzcd}
\mathsf{V}' \arrow[r, "\mathsf{L}'"] \arrow[d, "(\phi_I^{-1})'"'] & \mathsf{U}' \arrow[d, "(\phi_J^{-1})'"] \\
\mathsf{F}^I \arrow[r, "([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T"'] & \mathsf{F}^J
\end{tikzcd}$$

(ii) $([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T(j, i) = \langle \mathsf{L}'(\beta_i); \mathrm{e}_j \rangle$;

(iii) if U and V are finite-dimensional then the matrix representative of L' with respect to the dual bases associated to $\mathscr{B}_\mathsf{U}$ and $\mathscr{B}_\mathsf{V}$ is $([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T$.

Proof (i) By Theorem 5.4.21 the diagram

$$\begin{tikzcd}
\mathsf{U} \arrow[r, "\mathsf{L}"] \arrow[d, "\phi_J"'] & \mathsf{V} \arrow[d, "\phi_I"] \\
\mathsf{F}_0^J \arrow[r, "{[\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}}}"'] & \mathsf{F}_0^I
\end{tikzcd}$$

commutes. Taking the dual of the diagram and using Proposition 5.7.5 and parts (ii) and (iii) of Proposition 5.7.20 gives the desired commutative diagram.

(ii) Let $\{\boldsymbol{e}_j\}_{j\in J}$ and $\{\boldsymbol{f}_i\}_{i\in I}$ be the standard bases for F^J_0 and F^I_0, respectively. Note that for $j \in J$ and $i \in I$ the definition of matrix-vector multiplication gives

$$\langle([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T(\boldsymbol{f}_i);\boldsymbol{e}_j\rangle = (([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T\boldsymbol{f}_i)(j) = \sum_{i'\in I}([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T(j,i')\boldsymbol{f}_i(i') = ([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T(j,i).$$

Now we note that $\boldsymbol{f}_i = (\phi_I^{-1})'(\beta_i)$, $i \in I$, and $\boldsymbol{e}_j = \phi_J(e_j)$, $j \in J$, where, in the first instance, we think of $\boldsymbol{f}_i$ as an element of F^I since $\mathsf{F}^I_0 \subseteq \mathsf{F}^I$. Thus, for $i \in I$ and $j \in J$, we compute

$$\begin{aligned}\langle\mathsf{L}'(\beta_i);e_j\rangle &= \langle\mathsf{L}'\circ\phi'_I(\boldsymbol{f}_i);\phi_J^{-1}(\boldsymbol{e}_j)\rangle = \langle(\phi_J^{-1})'\circ\mathsf{L}'\circ\phi'_I(\boldsymbol{f}_i);\boldsymbol{e}_j\rangle\\ &= \langle([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T(\boldsymbol{f}_i);\boldsymbol{e}_j\rangle = ([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^T(j,i),\end{aligned}$$

as desired, where we have used part (i).

(iii) Let us write the bases for U and V as $\{e_1,\dots,e_n\}$ and $\{f_1,\dots,f_m\}$, respectively, with the dual bases denoted by $\{\alpha_1,\dots,\alpha_n\}$ and $\{\beta_1,\dots,\beta_m\}$, respectively. Then write

$$\mathsf{L}'(\beta_i) = c_{1i}\alpha_1 + \cdots + c_{ni}\alpha_n.$$

We then have

$$\langle\mathsf{L}'(\beta_i);e_j\rangle = c_{ji}, \qquad j \in \{1,\dots,m\}.$$

Therefore,

$$\mathsf{L}'(\beta_i) = \sum_{j=1}^m \langle\mathsf{L}'(\beta_i);e_j\rangle\alpha_j,$$

and the result follows from part (ii) and the definition of the matrix representative. ■

Note that the result cannot be said to say that, "The matrix representative of the dual is the transpose of the matrix representative," except in finite-dimensions. Nonetheless, the result does establish a concrete interpretation of the matrix representative of the dual, and moreover makes clear that the matrix transpose should be thought of as taking the dual of the linear map associated with a matrix.

Let us reformulate part (iii) of the preceding result in order to highlight its simple character.

5.7.23 Corollary (The matrix representative of the dual in finite-dimensions) *Let* F *be a field, let* U *and* V *be* F*-vector spaces with bases* $\mathscr{B}_\mathsf{U} = \{\mathrm{e}_1,\dots,\mathrm{e}_\mathrm{n}\}$ *and* $\mathscr{B}_\mathsf{V} = \{\mathrm{f}_1,\dots,\mathrm{f}_\mathrm{m}\}$*. Let* $\mathscr{B}'_\mathsf{U} = \{\alpha_1,\dots,\alpha_\mathrm{n}\}$ *and* $\mathscr{B}'_\mathsf{V} = \{\lambda_1,\dots,\lambda_\mathrm{m}\}$ *be the dual bases for* U' *and* V'*, respectively. If* $\mathsf{L} \in \mathrm{Hom}_\mathsf{F}(\mathsf{U};\mathsf{V})$ *then*

$$[\mathsf{L}']^{\mathscr{B}'_\mathsf{U}}_{\mathscr{B}'_\mathsf{V}} = ([\mathsf{L}]^{\mathscr{B}_\mathsf{V}}_{\mathscr{B}_\mathsf{U}})^\mathrm{T}.$$

5.7.6 Linear maps from a vector space to its algebraic dual

In this section we address the matter of the algebraic dual of a finite-dimensional vector space being isomorphic, though not in a natural way, with the vector space. As we commented in Example 5.7.8, it can be a temptation to believe that a vector space and its dual are the "same thing." They are not, although they are isomorphic. What we do in this section is examine linear maps, in particular isomorphisms, from a vector space to its dual. As we shall see, every such linear map corresponds to a particular bit of structure, namely a bilinear map on the vector space. By showing that such bilinear maps arise from isomorphisms of a vector space and its dual, we hope to convince the reader that there is no *natural* isomorphism of a vector space and its dual. The reader will benefit from a reading of Section 5.6.3.

We recall that $\mathrm{Hom}_{\mathsf{F}}(\mathsf{U},\mathsf{V};\mathsf{W})$ denotes the set of bilinear maps from $\mathsf{U}\times\mathsf{V}$ to W. Let us define a map $\Psi_{\mathsf{V}}\colon \mathrm{Hom}_{\mathsf{F}}(\mathsf{V},\mathsf{V};\mathsf{F})\to\mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V}')$ by

$$\langle\Psi_{\mathsf{V}}(\mathsf{B})(v);u\rangle=\mathsf{B}(u,v),\qquad u,v\in\mathsf{V}.$$

With this notation we have the following result.

5.7.24 Theorem (Bilinear maps and linear maps between V and V′) *If* F*is a field and* V *is an* F*-vector space, then the map* Ψ_{V} *is an isomorphism of the vector spaces* $\mathrm{Hom}_{\mathsf{F}}(\mathsf{V},\mathsf{V};\mathsf{F})$ *and* $\mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V}')$.

Proof That the map $\mathsf{B}\mapsto\mathsf{L}_{\mathsf{B}}$ is linear is shown as follows:

$$\begin{aligned}\langle\Psi_{\mathsf{V}}(\mathsf{B}_1+\mathsf{B}_2)(v);u\rangle&=(\mathsf{B}_1+\mathsf{B}_2)(u,v)=\mathsf{B}_1(u,v)+\mathsf{B}_2(u,v)\\&=\langle\Psi_{\mathsf{V}}(\mathsf{B}_1)(v);u\rangle+\langle\Psi_{\mathsf{V}}(\mathsf{B}_2)(v);u\rangle\\&=\langle(\Psi_{\mathsf{V}}(\mathsf{B}_1)+\Psi_{\mathsf{V}}(\mathsf{B}_2))(v);u\rangle\end{aligned}$$

and

$$\langle\Psi_{\mathsf{V}}(a\mathsf{B})(v);u\rangle=(a\mathsf{B})(u,v)=a(\mathsf{B}(u,v)=\langle a\Psi_{\mathsf{V}}(\mathsf{B})(v);u\rangle$$

for $u,v\in\mathsf{V}$, $a\in\mathsf{F}$, $\mathsf{B},\mathsf{B}_1,\mathsf{B}_2\in\mathrm{Hom}_{\mathsf{F}}(\mathsf{V},\mathsf{V};\mathsf{F})$. This gives

$$\Psi_{\mathsf{V}}(\mathsf{B}_1+\mathsf{B}_2)=\Psi_{\mathsf{V}}(\mathsf{B}_1)+\Psi_{\mathsf{V}}(\mathsf{B}_2),\quad\Psi_{\mathsf{V}}(a\mathsf{B})=a\Psi_{\mathsf{V}}(\mathsf{B}).$$

To show that the Ψ_{V} is injective, suppose that $\Psi_{\mathsf{V}}(\mathsf{B})=0_{\mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V}')}$. Then, for any $u,v\in\mathsf{V}$, we have

$$\langle\Psi_{\mathsf{V}}(\mathsf{B})(v);u\rangle=\mathsf{B}(u,v)=0_{\mathsf{F}}.$$

Thus B is zero, giving injectivity of Ψ_{V} by Exercise 4.5.23.

To show that Ψ_{V} is surjective let $\mathsf{L}\in\mathrm{Hom}_{\mathsf{F}}(\mathsf{V};\mathsf{V}')$ and define $\mathsf{B}_{\mathsf{L}}\in\mathrm{Hom}_{\mathsf{F}}(\mathsf{V},\mathsf{V};\mathsf{F})$ by

$$\mathsf{B}_{\mathsf{L}}(u,v)=\langle\mathsf{L}(v);u\rangle,\qquad u,v\in\mathsf{V}.$$

Then we have

$$\langle\Psi_{\mathsf{V}}(\mathsf{B}_{\mathsf{L}})(v);u\rangle=\mathsf{B}_{\mathsf{L}}(u,v)=\langle\mathsf{L}(v);u\rangle,\qquad u,v\in\mathsf{V}.$$

Thus $\Psi_{\mathsf{V}}(\mathsf{B}_{\mathsf{L}})=\mathsf{L}$, as desired. ■

The essential point is this: Any linear map between V and V' is associated uniquely to a bilinear F-valued map on V. In particular, if V is finite-dimensional then any isomorphism of V with V' defines and is defined by a bilinear F-valued map on V. Let us examine this in an example.

5.7.25 **Example (The canonical identification of F^n with $(\mathsf{F}^n)'$)** As we saw in Example 5.7.8, if one naturally thinks of elements of F^n as "column vectors," then elements of $(\mathsf{F}^n)'$ can be naturally thought of as "row vectors." This then gives an obvious linear map from F^n to $(\mathsf{F}^n)'$ given as

$$\begin{bmatrix} v_1 \\ \cdots \\ v_n \end{bmatrix} \mapsto \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}.$$

By Theorem 5.7.24, corresponding to this isomorphism is a unique bilinear F-valued map on V. It is a straightforward exercise to show that the bilinear map is

$$\mathsf{B}((u_1, \dots, u_n), (v_1, \dots, v_n)) = u_1 v_1 + \cdots + u_n v_n.$$

Indeed, to verify that this is indeed the correct bilinear map one need only check the obvious identity

$$\begin{bmatrix} u_1 & \cdots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \mathsf{B}((u_1, \dots, u_n), (v_1, \dots, v_n)).$$

The bilinear map B is a rather simple one, reflecting the fact that the isomorphism is correspondingly simple. •

Exercises

5.7.1 Let F be a field, let V be an F-vector space, and let $S \subseteq \mathsf{V}$ and $\Lambda \subseteq \mathsf{V}'$. Show that $\operatorname{ann}(S)$ is a subspace of V' and that $\operatorname{coann}(\Lambda)$ is a subspace of V.

5.7.2 Let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$. Show that $\mathsf{U} \subseteq \mathsf{V}$ is L-invariant if and only if $\operatorname{ann}(\mathsf{U}) \subseteq \mathsf{V}'$ is L'-invariant.

Section 5.8

The structure of linear maps on finite-dimensional vector spaces

In this section we focus on something rather specific: the structure of an endomorphism of a finite-dimensional vector space. By being more specific, we ought to be able to say more, and indeed we shall see that it is possible to give a quite descriptive characterisation of an endomorphism in the guise of the Jordan canonical form. This will tell us that, at least if the field is algebraically closed, if you look at things in the right way (more precisely, choose an appropriate basis) then an endomorphism behaves in a rather simple way. For fields that are not algebraically closed, the picture is more complicated. However, for $\mathbb{R}$, the relationship with the algebraic closure $\mathbb{C}$ is sufficiently understandable that one can give an analogue of the Jordan canonical form in this case.

The crucial element in understanding the structure of an endomorphism involves polynomials in a way which seems initially surprising. To be more specific, given an endomorphism of a finite-dimensional vector space, one can place a natural finitely generated module structure on the vector space over the ring of polynomials. This places one in a position to use the results of Section 4.9 concerning the structure of finitely generated modules over principal ideal domains. Indeed, doing so makes it rather easy to get at the most important results concerning the classification of endomorphisms. Moreover, this approach is intrinsically satisfying since one feels that the essence of the problem is being understood by making it a specific instance of something more general. The approach also gives one an opportunity to use ones "geometric intuition" by relating the structure to that of finitely generated Abelian groups, i.e., finitely generated $\mathbb{Z}$-modules. The difficulty, of course, is that one must stray fairly far afield from basic linear algebra. For this reason we essentially offer two approaches to understanding the structure of an endomorphism, one relying on the module structure over the polynomial ring, and the other independent (in some way) of this. We shall try to indicate as we go along what parts of the text are associated with which approach.

Do I need to read this section? Certainly the results in this section are of great importance, and so should be understood. In particular the structure of endomorphisms on finite-dimensional $\mathbb{R}$-vector spaces given in Theorem 5.8.74 will be essential in understanding ordinary differential equations in Section V-5.2.2. It is certainly possible that many of the details of the treatment here could be omitted until the reader feels as if they are suffering by not understanding them. However, even in these cases, many of the constructions are so revealing that we recommend that the entirety of this section be at least skimmed. •

5.8.1 Similarity

In this section we introduce the problem that we solve in this section. The problem is quite analogous to that for equivalence of linear maps between vector spaces as studied in Section 5.4.7. There the problem is to find bases for vector spaces U and V for which the matrix representative of a linear map $\mathsf{L} \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{U}; \mathsf{V})$ has the "simplest possible form." This simplest form is given in Theorem 5.4.41. For *endomorphisms*, of course, one could do the same trick. That is to say, one could find bases $\mathscr{B}_1$ and $\mathscr{B}_2$ for V such that the matrix representative of an endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ relative to these bases has the form given in Theorem 5.4.41. However, this is pretty clearly not the problem one wants to solve. The reasonable thing is that one should try to find a *single* basis for V such that the matrix representative of L in this basis has the simplest possible form.

Given the preceding discussion, the following adaptations of Definitions 5.1.38 and 5.4.38 now hopefully seem natural.

5.8.1 Definition (Similar matrices) Let F be a field and let I be an index set. Column finite matrices $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ are ***equivalent*** if there exists an invertible column finite matrix $\boldsymbol{P} \in \mathrm{Mat}_{I\times I}(\mathsf{F})$ such that $\boldsymbol{A}_2 = \boldsymbol{P}\boldsymbol{A}_1\boldsymbol{P}^{-1}$. •

5.8.2 Definition (Similar endomorphisms) Let F be a field and let V be an F-vector space. Endomorphisms $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ are ***equivalent*** if there exists bases $\mathscr{B}_1$ and $\mathscr{B}_2$ for V such that $[\mathsf{L}_1]_{\mathscr{B}_1}^{\mathscr{B}_1} = [\mathsf{L}_2]_{\mathscr{B}_2}^{\mathscr{B}_2}$. •

Of course, similarity is an equivalence relation, as the reader may show in Exercise 5.8.1. The problem in this section is to arrive at a useful characterisation of the equivalence classes under this equivalence relation in the case when V is finite-dimensional. It is perhaps not obvious, although hopefully it is also not surprising, that classification of endomorphisms by similarity is fundamentally different than classification by equivalence.

Let us first relate the notions of similarity for matrices and for endomorphisms. The following result mirrors Proposition 5.4.39 for equivalence, and may be proved in the same way.

5.8.3 Proposition (Similar endomorphisms and similar matrices) *Let F be a field, let V be an F-vector space, and let $\mathscr{B}$ be a basis for V. Then the following statements are equivalent:*

(i) the endomorphisms $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ are similar;

(ii) $[\mathsf{L}_1]_{\mathscr{B}}^{\mathscr{B}}$ and $[\mathsf{L}_2]_{\mathscr{B}}^{\mathscr{B}}$ are similar;

(iii) there exists an invertible endomorphism $\mathsf{P} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ such that $\mathsf{L}_2 = \mathsf{P} \circ \mathsf{L}_1 \circ \mathsf{P}^{-1}$.

Proof This is Exercise 5.8.2. ■

5.8.2 The $\mathsf{F}[\xi]$-module structure on a vector space induced by an endomorphism

As mentioned in the preamble to this section, one of the more useful realisations in studying an endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ of a finite-dimensional vector space is that there is a natural $\mathsf{F}[\xi]$-module structure induced on V. Since $\mathsf{F}[\xi]$ is a principal ideal domain (indeed, it is a Euclidean domain; Corollary 4.4.14), we can then wonder whether the results of Section 4.9 are useful; indeed they are.

We begin by defining the module structure. Let us first set the table; we let F be a field, let V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Let $P \in \mathsf{F}[\xi]$ and write

$$P = \sum_{j=0}^{k} a_j \xi^j.$$

Let us define $\mathrm{Ev}_{\mathsf{F}}(P)\colon \mathrm{End}_{\mathsf{F}}(\mathsf{V}) \to \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by

$$\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = \sum_{j=0}^{k} a_j \mathsf{L}^j. \tag{5.40}$$

Thus $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L})$ is the polynomial P with L in place of the indeterminate. This notation is actually a special case of the notation introduced in Proposition 4.4.9, as we now explain. First note that $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is a ring (Corollary 5.4.18) and so let us abbreviate this as R to make the notation simpler. The ring R gives rise to the polynomial ring $\mathsf{R}[\xi]$; these are polynomials whose coefficients are endomorphisms. The map $a \mapsto a\,\mathrm{id}_{\mathsf{V}}$ is then a ring monomorphism of F into R. This then gives rise to a monomorphism from the ring $\mathsf{F}[\xi]$ into the ring $\mathsf{R}[\xi]$ given by

$$\sum_{j=0}^{k} a_j \xi^j \mapsto \sum_{j=0}^{k} a_j\,\mathrm{id}_{\mathsf{V}}\,\xi^j.$$

The polynomial ring $\mathsf{R}[\xi]$ possesses its evaluation homomorphism $\mathrm{Ev}_{\mathsf{R}}\colon \mathsf{R}[\xi] \to \mathsf{R}^{\mathsf{R}}$, just as in Proposition 4.4.9. The map $\mathrm{Ev}_{\mathsf{F}}\colon \mathsf{F}[\xi] \to \mathsf{R}^{\mathsf{R}}$ defined in (5.40) is simply the restriction of Ev_{R} to the image of $\mathsf{F}[\xi]$ in $\mathsf{R}[\xi]$.

We can now define the product of $P \in \mathsf{F}[\xi]$ and $v \in \mathsf{V}$ by

$$P \cdot v = (\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}))(v). \tag{5.41}$$

Somewhat more explicitly,

$$\left(\sum_{j=0}^{k} a_j \xi^j\right) \cdot v = \sum_{j=0}^{k} a_j \mathsf{L}^j(v).$$

Note that this product depends on a choice of L. The following result records the properties of this product in the case when V is finite-dimensional.

5.8.4 Proposition (The $\mathsf{F}[\xi]$-module structure) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Then the set V with addition as usual and with scalar product given by (5.41) is a finitely generated torsion module over $\mathsf{F}[\xi]$.*

Proof We leave it as an easy exercise for the reader to verify that V is an $\mathsf{F}[\xi]$-module, and just show that it is a finitely generated torsion module.

If $\{e_1, \dots, e_n\}$ is a basis for V as an F-vector space, we claim that this set also generated V as an $\mathsf{F}[\xi]$-module. Indeed, if $v \in \mathsf{V}$ then we can write

$$v = c_1 e_1 + \cdots + c_n e_n$$

for $c_1, \dots, c_n \in \mathsf{F}$. However, this means that

$$v = c_1 \mathsf{L}^0(e_1) + \cdots + c_n \mathsf{L}^0(e_n)$$

since $\mathsf{L}^0 = \mathrm{id}_{\mathsf{V}}$. Thus

$$v = P_1 \cdot e_1 + \cdots + P_n \cdot e_n$$

where $P_j = c_j \xi^0$, $j \in \{1, \dots, n\}$. This shows that the set $\{e_1, \dots, e_n\}$ generated V as an $\mathsf{F}[\xi]$-module.

To show that V is a torsion module we claim that there exists a nonzero polynomial $P \in \mathsf{F}[\xi]$ such that $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Let $n = \dim_{\mathsf{F}}(\mathsf{V})$. Since $\dim_{\mathsf{F}}(\mathrm{End}_{\mathsf{F}}(\mathsf{V})) = n^2$ by Exercise 5.1.3, it follows that the family of endomorphisms $(\mathrm{id}_{\mathsf{V}}, \mathsf{L}, \dots, \mathsf{L}^{n^2})$ is linearly dependent by Lemma 1 in the proof of Theorem 4.5.25. Therefore there exists $a_0, a_1, \dots, a_{n^2} \in \mathsf{F}^*$ such that

$$a_0 \,\mathrm{id}_{\mathsf{V}} + a_1 \mathsf{L} + \cdots + a_{n^2} \mathsf{L}^{n^2} = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}.$$

This means that $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Therefore, $P \cdot v = 0_{\mathsf{V}}$ for every $v \in \mathsf{V}$, and so $\mathrm{ann}(v) \neq \{0_{\mathsf{F}[\xi]}\}$. Since $\mathsf{F}[\xi]$ is an integral domain, this implies that all elements of V are torsion elements. ■

Now the vector space V has two algebraic structures, its natural vector space structure and the $\mathsf{F}[\xi]$-module structure induced by L. It is sometimes convenient to notationally distinguish between these two structures.

5.8.5 Definition (The $\mathsf{F}[\xi]$-module) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The $\mathsf{F}[\xi]$-module V will be denoted by V_{L}. •

This is fine. But in order to make something of there, there should be some relationship between the $\mathsf{F}[\xi]$-module V_{L} and the structure of the endomorphism L. There is, in fact, a strong connection between these things, and the following result begins our understanding of this.

5.8.6 Theorem (Isomorphisms of V as an $\mathsf{F}[\xi]$-module) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Then the $\mathsf{F}[\xi]$-modules $\mathsf{V}_{\mathsf{L}_1}$ and $\mathsf{V}_{\mathsf{L}_2}$ are isomorphic if and only if L_1 and L_2 are similar.*

Proof Suppose that there exists an isomorphism $\mathsf{P} \in \mathrm{Hom}_{\mathsf{F}[\xi]}(\mathsf{V}_{\mathsf{L}_1}; \mathsf{V}_{\mathsf{L}_2})$ of $\mathsf{F}[\xi]$-modules. Firstly, P is obviously a bijection of the set V. Secondly,

$$\begin{aligned} \mathsf{P}(v_1 + v_2) &= \mathsf{P}(v_1) + \mathsf{P}(v_1), \qquad v_1, v_2 \in \mathsf{V}, \\ \mathsf{P}(Q \cdot v) &= Q \cdot \mathsf{P}(v), \qquad v \in \mathsf{V},\ Q \in \mathsf{F}[\xi]. \end{aligned}$$

Applying the second of these relations for constant polynomials, $Q = a\xi^0$ for $a \in \mathsf{F}$, shows that $\mathsf{P} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V}; \mathsf{V})$. Thus P is an isomorphism of F-vector spaces. Now taking $Q = \xi$ in the relation above gives

$$\mathsf{P}(\mathsf{L}_1(v)) = \mathsf{L}_2(\mathsf{P}(v)), \qquad v \in \mathsf{V}.$$

Thus $\mathsf{L}_2 = \mathsf{P} \circ \mathsf{L}_1 \circ \mathsf{P}^{-1}$, and so L_1 and L_2 are similar by Proposition 5.8.3.

Before we proceed with the next part of the proof we give a simple lemma.

1 Lemma *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then, for* $\mathrm{Q} \in \mathsf{F}[\xi]$ *and for* $\mathsf{P} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *an isomorphism,*

$$\mathrm{Ev}_{\mathsf{F}}(\mathrm{Q})(\mathsf{P} \circ \mathsf{L} \circ \mathsf{P}^{-1}) = \mathsf{P} \circ (\mathrm{Ev}_{\mathsf{F}}(\mathrm{Q})(\mathsf{L})) \circ \mathsf{P}^{-1}.$$

Proof This follows since, for every $j \in \mathbb{Z}_{\geq 0}$,

$$(\mathsf{P} \circ \mathsf{L} \circ \mathsf{P}^{-1})^j = \underbrace{\mathsf{P} \circ \mathsf{L} \circ \mathsf{P}^{-1} \circ \cdots \mathsf{P} \circ \mathsf{L} \circ \mathsf{P}^{-1}}_{j \text{ times}} = \mathsf{P} \circ \mathsf{L}^j \circ \mathsf{P}^{-1}. \qquad \blacktriangledown$$

Now suppose that L_1 and L_2 are similar, with $\mathsf{L}_2 = \mathsf{P} \circ \mathsf{L}_1 \circ \mathsf{P}^{-1}$ for an invertible endomorphism $\mathsf{P} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. We claim that P is also an isomorphism of $\mathsf{F}[\xi]$-modules. Since P is a bijection and since $\mathsf{P}(v_1 + v_2) = \mathsf{P}(v_1) + \mathsf{P}(v_2)$ for every $v_1, v_2 \in \mathsf{V}$, it only remains to show that $\mathsf{P}(Q \cdot v) = Q \cdot \mathsf{P}(v)$ for every $Q \in \mathsf{F}[\xi]$. Suppose that

$$Q = \sum_{j=0}^{k} a_j \xi^j.$$

Then

$$\begin{aligned} \mathsf{P}(Q \cdot v) = \mathsf{P}\left(\left(\sum_{j=0}^{k} a_j \xi^j\right) \cdot v\right) &= \mathsf{P}\left(\sum_{j=0}^{k} a_j \mathsf{L}_1^j(v)\right) \\ &= \sum_{j=0}^{k} a_j \mathsf{L}_2^j(\mathsf{P}(v)) = Q \cdot \mathsf{P}(v), \end{aligned}$$

using the lemma above. Thus P is also an isomorphism of $\mathsf{F}[\xi]$-modules. ■

The preceding result is obviously an important one, since it tells us that classifying endomorphisms under similarity is the same as classifying the $\mathsf{F}[\xi]$-module structures on V under isomorphism.

Let us explore the relationship between the module and vector space structures of V further. The following result characterises the L-invariant subspaces of V. It will become increasingly apparent as things move along that the invariant subspaces of L play a crucial rôle in understanding its structure.

5.8.7 Proposition (Submodules of V_L) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_\mathsf{F}(\mathsf{V})$*. Then a subset* $\mathsf{U} \subseteq \mathsf{V}$ *is a submodule of the* $\mathsf{F}[\xi]$*-module* V_L *if and only if it is an* L*-invariant subspace.*

Proof First suppose that U is a submodule of V_L. Then

$$\begin{aligned} u_1 + u_2 &\in \mathsf{U}, &\quad& u_1, u_2 \in \mathsf{U}, \\ Q \cdot u &\in \mathsf{U}, && v \in \mathsf{V},\ Q \in \mathsf{F}[\xi]. \end{aligned}$$

In particular, letting $Q = a\xi^0$ for $a \in \mathsf{F}$ shows that U is a subspace. Now letting $Q = \xi$ shows that $\mathsf{L}(u) \in \mathsf{U}$ for all $u \in \mathsf{U}$, and so U is L-invariant.

Now suppose that U is an L-invariant subspace. Since $u_1 + u_2 \in \mathsf{U}$ for all $u_1, u_2 \in \mathsf{U}$ we only need to show that $Q \cdot v \in \mathsf{U}$ for each $Q \in \mathsf{F}[\xi]$ and $u \in \mathsf{U}$. Let us denote

$$Q = \sum_{j=0}^{k} a_j \xi^j$$

and compute

$$Q \cdot u = \sum_{j=0}^{k} a_j \mathsf{L}^j(u),$$

from which the result follows since $\mathsf{L}^j(u) \in \mathsf{U}$ by Exercise 5.4.2. ■

5.8.3 Another $\mathsf{F}[\xi]$-module

In the preceding section we introduced the structure of an $\mathsf{F}[\xi]$-module on V induced by an endomorphism L. When V is finite-dimensional this module is a finitely generated torsion module. Now we investigate *another* $\mathsf{F}[\xi]$-module structure whose construction is based on V. The module we construct in this way will allow us to give a more elegant characterisation of the characteristic polynomial than we were able to give in Definition 5.4.55. Moreover, the construction does not depend on an endomorphism, and so is intrinsic to the vector space. Our construction is related to the notion of the tensor product between vector spaces as considered in Section 5.6.3. We ask the reader to provide this connection in Exercise 5.8.3.

Let us first try to motivate our more general construction by looking at a specific case where things are easier to understand. We begin with the n-dimensional F-vector space F^n. In this vector space, vectors are multiplied by scalars from the field F. Suppose that we wish to allow vectors to by multiplied, not by elements of F, but by polynomials from the ring $\mathsf{F}[\xi]$. That is to say, one wishes to extend the F-vector space structure to an $\mathsf{F}[\xi]$-module structure. In this case there is an obvious way to do this: merely consider the module $\mathsf{F}[\xi]^n$. That is to say, just "replace" the entries from F in F^n with entries from $\mathsf{F}[\xi]$. If all one were interested in were F^n, then this would be fine. If one were interested in a general finite-dimensional

F-vector space, one might choose a basis to reduce to the case of F^n. However, a more enlightened view is preferable, and it is this that we now describe. The reader might be well-advised to at this time revisit the construction of the polynomial ring $\mathsf{F}[\xi]$, as the constructions we make have a great deal in common with that.

We let F be a field and let V be an F-vector space. We denote by $\mathsf{V}[\xi]$ the set of maps $\Phi\colon \mathbb{Z}_{\geq 0} \to \mathsf{V}$ such that $\Phi(j) = 0$ except for a finite number of $j \in \mathbb{Z}_{\geq 0}$, i.e., as a set, $\mathsf{V}[\xi]$ is the direct sum $\oplus_{j\in\mathbb{Z}_{\geq 0}}\mathsf{V}$. We wish to endow $\mathsf{V}[\xi]$ with the structure of a module over $\mathsf{F}[\xi]$. Let us for the moment think of $\mathsf{F}[\xi]$ as the direct sum $\oplus_{j\in\mathbb{Z}_{\geq 0}}\mathsf{F}$, this being the definition given in Definition 4.4.1. Now let and $\phi \in \mathsf{F}[\xi]$ and $\Phi, \Phi_1, \Phi_2 \in \mathsf{V}[\xi]$. Define $\Phi_1 + \Phi_2, \phi\Phi \in \mathsf{V}[\xi]$ by

$$(\Phi_1 + \Phi_2)(k) = \Phi_1(k) + \Phi_2(k),$$
$$(\phi\Phi)(k) = \sum_{j=0}^{k} \phi(j)\Phi(k-j).$$

The reader should compare these definitions to the definitions for sum and product of polynomials. These definitions of sum and product make $\mathsf{V}[\xi]$ into an $\mathsf{F}[\xi]$-module.

5.8.8 Theorem ($\mathsf{V}[\xi]$ is an $\mathsf{F}[\xi]$-module) *If F is a field and if V is an F-vector space, then $\mathsf{V}[\xi]$ is an $\mathsf{F}[\xi]$-module with the above definitions of sum and scalar multiplication.*

Proof The proof follows rather closely the proof of Theorem 4.4.2, so let us just give a sample proof of one of the module axioms. We let $\phi, \psi \in \mathsf{F}[\xi]$ and $\Phi \in \mathsf{V}[\xi]$, and compute

$$\begin{aligned}
\psi(\phi\Phi)(k) &= \sum_{j=0}^{k} \psi(j)\left(\sum_{l=0}^{k-j} \phi(l)\Phi(k-j-l)\right) \\
&= \sum_{\substack{j,l,m \\ j,l,m\geq 0, j+l+m-k}} \psi(j)(\phi(l)\Phi(m)) \\
&= \sum_{\substack{j,l,m \\ j,l,m\geq 0, j+l+m-k}} (\psi(j)\phi(l))\Phi(m) \\
&= \sum_{j=0}^{k}\left(\sum_{l=0}^{j} \psi(j)\phi(k-j)\right)\Phi(k-j).
\end{aligned}$$

We leave the remainder of the verifications to the reader. ■

Now let us adapt the indeterminate notation for polynomials to the module $\mathsf{V}[\xi]$. For $v \in \mathsf{V}$ and $k \in \mathbb{Z}_{\geq 0}$ let $\xi^k \cdot v \in \mathsf{V}[\xi]$ be given by

$$(\xi^k \cdot v)(j) = \begin{cases} v, & j = k, \\ 0_{\mathsf{V}}, & \text{otherwise.} \end{cases}$$

This is analogous to the definition of the indeterminate ξ in $\mathsf{F}[\xi]$, cf. Definition 4.4.4. With this notation, every element $\Phi \in \mathsf{V}[\xi]$ can be uniquely expressed as

$$\Phi = \sum_{j=0}^{\infty} \xi^j \cdot v_j$$

for vectors $v_j \in \mathsf{V}$, $j \in \mathbb{Z}_{\geq 0}$, only finitely many of which are nonzero. With this expression, the module operations on $\mathsf{V}[\xi]$ can then be easily figured out by using the usual rules of associativity and distributivity for polynomials.

Let us see how this construction works out in a particular case, namely the case we started our discussion with, and the case that represents what is of most interest to us.

5.8.9 Proposition ($\mathsf{F}^{\mathrm{m}}[\xi] = \mathsf{F}[\xi]^{\mathrm{m}}$) *For a field* F*, the map*

$$\xi^{\mathrm{k}}\mathbf{v}_{\mathrm{k}} + \dots + \xi \cdot \mathbf{v}_1 + \mathbf{v}_0 \mapsto \begin{bmatrix} \mathbf{v}_{\mathrm{k}}(1)\xi^{\mathrm{k}} + \dots + \mathbf{v}_1(1)\xi + \mathbf{v}_0(1) \\ \dots \\ \mathbf{v}_{\mathrm{k}}(\mathrm{n})\xi^{\mathrm{k}} + \dots + \mathbf{v}_1(\mathrm{n})\xi + \mathbf{v}_0(\mathrm{n}) \end{bmatrix}$$

is an isomorphism of the $\mathsf{F}[\xi]$*-modules* $\mathsf{F}^{\mathrm{n}}[\xi]$ *and* $\mathsf{F}[\xi]^{\mathrm{n}}$*.*

Proof This is simply a matter of checking that the given map is a homomorphism and is invertible. Both of these things are easy to check directly, and we leave it to the doubt-filled reader to do this. ■

Let us close this section by relating the two $\mathsf{F}[\xi]$-module structures we have at hand. We let F be a field, V be an F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. We then define $\mathsf{L}[\xi] \in \mathrm{End}_{\mathsf{F}[\xi]}(\mathsf{V}[\xi])$ as follows. Let $V \in \mathsf{V}[\xi]$ and write

$$V = \xi^k \cdot v_k + \dots + \xi \cdot v_1 + v_0,$$

for $v_0, v_1, \dots, v_k \in \mathsf{V}$. We then define

$$\mathsf{L}[\xi](V) = \xi^k \cdot \mathsf{L}(v_k) + \dots + \xi \cdot \mathsf{L}(v_1) + \mathsf{L}(v_0).$$

For $V \in \mathsf{V}[\xi]$ expressed as above let us define $\phi_{\mathsf{L}} \colon \mathsf{V}[\xi] \to \mathsf{V}_{\mathsf{L}}$ by

$$\phi_{\mathsf{L}}(V) = \mathsf{L}^k(v_k) + \dots + \mathsf{L}(v_1) + v_0.$$

One can verify directly that ϕ_{L} is a homomorphism of $\mathsf{F}[\xi]$-modules. Let us also define $\psi_{\mathsf{L}} \colon \mathsf{V}[\xi] \to \mathsf{V}[\xi]$ by

$$\psi_{\mathsf{L}}(V) = (\xi\, \mathrm{id}_{\mathsf{V}[\xi]} - \mathsf{L}[\xi])(V).$$

With these maps, we have the following result.

5.8.10 Proposition (Relating the two $\mathsf{F}[\xi]$-modules) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then the following*

$$\begin{tikzcd} \{0\} \arrow[r] & \mathsf{V}[\xi] \arrow[r, "\psi_{\mathsf{L}}"] & \mathsf{V}[\xi] \arrow[r, "\phi_{\mathsf{L}}"] & \mathsf{V}_{\mathsf{L}} \arrow[r] & \{0\} \end{tikzcd}$$

is an exact sequence of $\mathsf{F}[\xi]$*-modules.*

Proof It is clear that ϕ_{L} is surjective; let us show that ψ_{L} is injective. We write

$$V = \sum_{j=0}^{\infty} \xi^j v_j$$

where only finitely many of the v_j's are nonzero. Then

$$\psi_{\mathsf{L}}(V) = \sum_{j=0}^{\infty} \xi^j (v_{j-1} - \mathsf{L}(v_j)),$$

with the convention that $v_{-1} = 0_{\mathsf{V}}$. If $\psi_{\mathsf{L}}(V) = 0_{\mathsf{V}[\xi]}$, then $v_{j-1} = \mathsf{L}(v_j)$ for all $j \in \mathbb{Z}_{\geq 0}$. We claim that this implies that $v_j = 0_{\mathsf{V}}$ for all $j \in \mathbb{Z}_{\geq 0}$. If not, let k be the largest integer such that $v_k \neq 0_{\mathsf{V}}$. We then have $v_k = \mathsf{L}(v_{k+1}) = 0_{\mathsf{V}}$, which is a contradiction. Thus $V = 0_{\mathsf{V}[\xi]}$, and so ψ_{L} is injective by Exercise 4.8.3.

We now show that $\mathrm{image}(\psi_{\mathsf{L}}) = \ker(\phi_{\mathsf{L}})$. To see this, we first claim that the following diagram commutes:

$$\begin{tikzcd} \mathsf{V}[\xi] \arrow[r, "{\mathsf{L}[\xi]}"] \arrow[d, "\phi_{\mathsf{L}}"'] & \mathsf{V}[\xi] \arrow[d, "\phi_{\mathsf{L}}"] \\ \mathsf{V}_{\mathsf{L}} \arrow[r, "\mathsf{L}"'] & \mathsf{V}_{\mathsf{L}} \end{tikzcd}$$

Indeed, we have

$$\phi_{\mathsf{L}} \circ \mathsf{L}[\xi]\left(\sum_{j\in\mathbb{Z}_{\geq 0}} \xi^j v_j\right) = \phi_{\mathsf{L}}\left(\sum_{j\in\mathbb{Z}_{\geq 0}} \xi^j \mathsf{L}(v_j)\right) = \sum_{j\in\mathbb{Z}_{\geq 0}} \mathsf{L}^{j+1}(v_j)$$

and

$$\mathsf{L} \circ \phi_{\mathsf{L}}\left(\sum_{j\in\mathbb{Z}_{\geq 0}} \xi^j v_j\right) = \mathsf{L}\left(\sum_{j\in\mathbb{Z}_{\geq 0}} \mathsf{L}^j(v_j)\right) = \sum_{j\in\mathbb{Z}_{\geq 0}} \mathsf{L}^{j+1}(v_j),$$

as desired. For $V \in \mathsf{V}[\xi]$ we have

$$\phi_{\mathsf{L}}(\xi \cdot V) = \xi \cdot \phi_{\mathsf{L}}(V) = \mathsf{L} \circ \phi_{\mathsf{L}}(V),$$

using first the fact that ϕ_{L} is a homomorphism of $\mathsf{F}[\xi]$-modules, and using second the definition of the F-module structure on V_{L}. Therefore, by commutativity of the above diagram,

$$\phi_{\mathsf{L}} \circ \xi = \mathsf{L} \circ \phi_{\mathsf{L}} = \phi_{\mathsf{L}} \circ \mathsf{L}[\xi].$$

Thus $\phi_{\mathsf{L}} \circ (\xi - \mathsf{L}[\xi])$ is zero, and so $\operatorname{image}(\psi_{\mathsf{L}}) \subseteq \ker(\phi_{\mathsf{L}})$. Now let

$$\sum_{j \in \mathbb{Z}_{\geq 0}} \xi^j v_j \in \ker(\phi_{\mathsf{L}}) \quad \Longrightarrow \quad \sum_{j \in \mathbb{Z}_{\geq 0}} \mathsf{L}^j(v_j) = 0_{\mathsf{V}}.$$

Therefore

$$\sum_{j \in \mathbb{Z}_{\geq 0}} \xi^j v_j = \sum_{j \in \mathbb{Z}_{\geq 0}} (\xi^j v_j - \mathsf{L}^j(v_j)).$$

Now we have

$$\begin{aligned}\xi^j \operatorname{id}_{\mathsf{V}[\xi]} - \mathsf{L}[\xi]^j &= (\xi \operatorname{id}_{\mathsf{V}[\xi]} - \mathsf{L}[\xi])(\xi^{j-1} \operatorname{id}_{\mathsf{V}} + \mathsf{L}^{j-1}) \\ &= (\xi \operatorname{id}_{\mathsf{V}[\xi]} - \mathsf{L}[\xi]) \circ \sum_{l=0}^{j-1} \xi^l \operatorname{id}_{\mathsf{V}[\xi]} \circ \mathsf{L}[\xi]^{j-l-1},\end{aligned}$$

using the fact that $\operatorname{id}_{\mathsf{V}[\xi]}$ and $\mathsf{L}[\xi]$ commute and using the Binomial Formula. Thus we have

$$\sum_{j \in \mathbb{Z}_{\geq 0}} \xi^j v_j = (\xi \operatorname{id}_{\mathsf{V}[\xi]} - \mathsf{L}[\xi]) \circ \left(\sum_{j \in \mathbb{Z}_{\geq 0}} \sum_{l=0}^{j-1} \xi^l \operatorname{id}_{\mathsf{V}[\xi]} \circ \mathsf{L}[\xi]^{j-l-1}(v_j) \right),$$

giving $\ker(\phi_{\mathsf{L}}) \subseteq \operatorname{image}(\psi_{\mathsf{L}})$. ■

5.8.4 The minimal and characteristic polynomials

We continue with a field F, a finite-dimensional F-vector space V, and $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$. Since V_{L} is a *finitely generated* torsion $\mathsf{F}[\xi]$-module by Proposition 5.8.4, it follows that $\operatorname{ann}(\mathsf{V})$ is a nonzero ideal of $\mathsf{F}[\xi]$ (that V is finite-dimensional is essential here; see Exercise 5.8.9 and cf. Exercise 4.9.1). Since $\mathsf{F}[\xi]$ is a principal ideal domain, it follows that $\operatorname{ann}(\mathsf{V}) = (P)$ for some polynomial $P \in \mathsf{F}[\xi]$. Moreover, since polynomials generated the same ideal differ only by multiplication by a unit in $\mathsf{F}[\xi]$ (i.e., by a nonzero constant polynomial by Exercise 4.4.3), it follows that P is uniquely specified once one asks that it be monic. This leads to the following definition.

5.8.11 Definition (Minimal polynomial) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$. The ***minimal polynomial*** of L is the unique monic polynomial $M_{\mathsf{L}} \in \mathsf{F}[\xi]$ such that $\operatorname{ann}(\mathsf{V}) = (M_{\mathsf{L}})$. •

Let us give an characterisation of the minimal polynomial that is equivalent to the definition, but stated in perhaps less intimidating language.

5.8.12 Proposition (Characterisation of minimal polynomial) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$. Then the minimal polynomial M_{L} is the unique monic polynomial with the property that, if $\mathrm{P} \in \mathsf{F}[\xi]$ satisfies $\mathrm{Ev}_{\mathsf{F}}(\mathrm{P})(\mathsf{L}) = 0_{\operatorname{End}_{\mathsf{F}}(\mathsf{V})}$, then $\mathrm{M}_{\mathsf{L}}|\mathrm{P}$.*

Proof Let us denote by $\tilde{M}_{\mathsf{L}}$ the unique monic polynomial such that, if $P \in \mathsf{F}[\xi]$ satisfies $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$, then $\tilde{M}_{\mathsf{L}}|P$.

First let us show that this definition of $\tilde{M}_{\mathsf{L}}$ makes sense. From the proof of Proposition 5.8.4 the set

$$\{P \in \mathsf{F}[\xi] \mid \mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}\}$$

is a nonempty, nonzero ideal; it is the ideal $\mathrm{ann}(\mathsf{V})$. Let

$$k = \min\{\deg(P) \mid P \in \mathrm{ann}(\mathsf{V}) \setminus \{0_{\mathsf{F}[\xi]}\}$$

and suppose that $M \in \mathrm{ann}(\mathsf{V})$ has degree k. Then, by the Euclidean Algorithm, if $P \in \mathrm{ann}(\mathsf{V})$ we can write $P = Q \cdot M + R$ where $\deg(R) < \deg(M)$. Since $R = P - Q \cdot M$, since $M, P \in \mathrm{ann}(\mathsf{V})$, and since $\mathrm{ann}(\mathsf{V})$ is an ideal, it follows that $R \in \mathrm{ann}(\mathsf{V})$. But M has smallest positive degree of the polynomials in $\mathrm{ann}(\mathsf{V})$, and so $R = 0_{\mathsf{F}[\xi]}$. Thus $M|P$, and so there exists a polynomial M such that $M|P$ for every $P \in \mathrm{ann}(\mathsf{V})$. There then exists a monic polynomial $\tilde{M}_{\mathsf{L}}$ with this property.

Now we claim that $\tilde{M}_{\mathsf{L}}$ is uniquely defined by its being monic and by its dividing every polynomial in $\mathrm{ann}(\mathsf{V})$. Let $\tilde{M}'_{\mathsf{L}}$ be another such polynomial. Then write $\tilde{M}'_{\mathsf{L}} = Q \cdot \tilde{M}_{\mathsf{L}} + R$ for $\deg(R) < \deg(\tilde{M}_{\mathsf{L}})$. Just as we argued above, $R = 0_{\mathsf{F}[\xi]}$. Since $\deg(\tilde{M}_{\mathsf{L}}) = \deg(\tilde{M}'_{\mathsf{L}})$ the polynomial Q must be a unit. Since both $\tilde{M}_{\mathsf{L}}$ and $\tilde{M}'_{\mathsf{L}}$ are both monic, $Q = 1_{\mathsf{F}}$. Thus the definition of $\tilde{M}_{\mathsf{L}}$ makes sense.

To prove the result, it remains to show that M_{L} is a monic polynomial of least degree in $\mathrm{ann}(\mathsf{V})$. This, however, follows since M_{L} generates $\mathrm{ann}(\mathsf{V})$, cf. Proposition 4.2.61. ■

In other words, the minimal polynomial is the lowest degree monic polynomial which, when L is substituted for the indeterminate, the result is zero.

As we shall see, the endomorphism L is uniquely characterised up to similarity by its minimal polynomial. But there is some work to be done before we can prove this. A key idea in this development is the following result.

5.8.13 Proposition (Eigenvalues are roots of the minimal polynomial) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Then $\lambda \in \mathsf{F}$ is an eigenvalue for L if and only if it is a root of M_{L}.*

Proof Let $\lambda \in \mathsf{F}$ be a root of M_{L} so that $M_{\mathsf{L}} = (\xi - \lambda) \cdot P$ for some polynomial P by Proposition 4.4.25. Since $\deg(P) < \deg(M_{\mathsf{L}})$ and by Proposition 5.8.12, $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) \neq 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Therefore, there exists $v \in \mathsf{V}$ such that $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) \cdot v \neq 0_{\mathsf{V}}$. Then

$$0_{\mathsf{V}} = \mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}})(L)\dot{v} = (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}}) \circ \mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) \cdot v.$$

Therefore, letting $v' = \mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) \cdot v$, $\mathsf{L}(v') = \lambda v'$, and so λ is an eigenvalue.

Conversely, suppose that λ is an eigenvalue of L and let $v \in \mathsf{V} \setminus \{0_{\mathsf{V}}\}$ be such that $\mathsf{L}(v) = \lambda v$. Let us write

$$M_{\mathsf{L}} = \xi^k + a_{k-1}\xi^{k-1} + \cdots + a_1\xi + a_0.$$

We then have

$$\begin{aligned}0_{\mathsf{V}} &= \mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}})(\mathsf{L}) \cdot v \\ &= (\mathsf{L}^k + a_{k-1}\mathsf{L}^{k-1} + \cdots + a_1\mathsf{L} + a_0\,\mathrm{id}_{\mathsf{V}})(v) \\ &= (\lambda^k + a_{k-1}\lambda^{k-1} + \cdots + a_1\lambda + a_0)v \\ &= \mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}})(\lambda),\end{aligned}$$

so showing that λ is a root of M_{L}. ■

Despite the preceding result, the eigenvalues alone are not enough to characterise the minimal polynomial, as the following example illustrates.

5.8.14 Examples (Eigenvalues and the minimal polynomial)

1. We take $\mathsf{V} = \mathsf{F}^2$ and consider the two endomorphisms, in this case these are simply matrices, given by

$$\mathsf{L} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}, \quad \mathsf{L}' = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

Since these matrices are upper triangular, their eigenvalues are simply the diagonal entries; see Exercise 5.4.20. Thus both endomorphisms have 1_{F} as their only eigenvalue. Thus the only root of the minimal polynomial is 1_{F}. Thus the minimal polynomial must be of the form $(\xi - 1_{\mathsf{F}})^k \cdot P$ for some $k \in \mathbb{Z}_{>0}$ and for some polynomial P. Let us define $M = \xi - 1_{\mathsf{F}}$ and $M' = (\xi - 1_{\mathsf{F}})^2$. We then have

$$\mathrm{Ev}_{\mathsf{F}}(M)(\mathsf{L}) = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{bmatrix}, \quad \mathrm{Ev}_{\mathsf{F}}(M')(\mathsf{L}') = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{bmatrix}, \quad \mathrm{Ev}_{\mathsf{F}}(M')(\mathsf{L}') = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{bmatrix}.$$

From this we conclude that $M_{\mathsf{L}} = \xi - 1_{\mathsf{F}}$ and $M_{\mathsf{L}'} = (\xi - 1_{\mathsf{F}})^2$. Thus, although L and L' have the same eigenvalues, they do not have the same minimal polynomials.

2. Let $\mathsf{F} = \mathbb{R}$, let $\mathsf{V} = \mathbb{R}^2$, and consider the endomorphism given as the 2×2-matrix

$$\mathsf{L} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

We note that if $P = \xi^2 - 2\xi + 2$ then $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Therefore, the minimal polynomial divides P. However, as an element of $\mathbb{R}[\xi]$, P is irreducible. Thus we must have $M_{\mathsf{L}} = \xi^2 - 2\xi + 2$. This polynomial has no roots, reflecting the fact that L has no eigenvalues. •

The following result shows that the minimal polynomial is a function of the equivalence classes under similarity.

5.8.15 Proposition (Minimal polynomial is invariant under similarity) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then* $\mathrm{M}_{\mathsf{L}_1} = \mathrm{M}_{\mathsf{L}_2}$ *if* L_1 *and* L_2 *are similar.*

Proof Let $\mathsf{P} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ be such that $\mathsf{L}_2 = \mathsf{P} \circ \mathsf{L}_1 \circ \mathsf{P}^{-1}$. By Lemma 1 of Theorem 5.8.6 it holds that

$$\mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}_2})(\mathsf{L}_2) = \mathsf{P} \circ (\mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}_2})(\mathsf{L}_1)) \circ \mathsf{P}^{-1}.$$

Therefore, $\mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}_2})(\mathsf{L}_1) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$, and reversing the argument gives $\mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}_1})(\mathsf{L}_2) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. By Proposition 5.8.12, $M_{\mathsf{L}_1} | M_{\mathsf{L}_2}$ and $M_{\mathsf{L}_2} | M_{\mathsf{L}_1}$, and so $M_{\mathsf{L}_2} = aM_{\mathsf{L}_1}$ for $a \in \mathsf{F}^*$. Since both minimal polynomials are monic, the result follows. ■

The converse of the result is not true. That is to say, the minimal polynomial is not enough to distinguish equivalence classes under similarity.

5.8.16 Example (Nonsimilar matrices with the same minimal polynomial) We consider a field F and take $\mathsf{V} = \mathsf{F}^4$. Define two endomorphisms by the matrices

$$\mathsf{L} = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}, \qquad \mathsf{L}' = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

One can check that $M_{\mathsf{L}} = M_{\mathsf{L}'} = (\xi - 1_{\mathsf{F}})^2$ (we will see as we go along how one can *prove* that this is the minimal polynomial). However, L and L' are not similar, as will be shown in Example 5.8.37 below. •

Now let us turn our attention to the characteristic polynomial. In Definition 5.4.55 we gave a definition of the characteristic polynomial based on a construction that involved first choosing a basis $\mathscr{B}$ for V, then constructing the matrix $\xi \boldsymbol{I}_n - [\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$, the taking the determinant of this matrix, to give a polynomial. This construction does not depend on basis since, if $\mathscr{B}'$ is another basis, we have

$$\begin{aligned} \det(\xi \boldsymbol{I}_n - [\mathsf{L}]_{\mathscr{B}'}^{\mathscr{B}'}) &= \det(\xi \boldsymbol{P}\boldsymbol{P}^{-1} - \boldsymbol{P}[\mathsf{L}]_{\mathscr{B}'}^{\mathscr{B}'}\boldsymbol{P}) \\ &= \det \boldsymbol{P} \det(\xi \boldsymbol{I}_n - [\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}) \det \boldsymbol{P}^{-1} \\ &= \det(\xi \boldsymbol{I}_n - [\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}), \end{aligned}$$

using the properties of determinant, and if $\boldsymbol{P}$ is the change of basis matrix. Thus this construction is completely unambiguous and well-defined. However, there is something ungratifying about the fact that one must use a basis in the definition, even though in practice one would always choose a basis to do the computation. To get a more gratifying description of the characteristic polynomial, we use the $\mathsf{F}[\xi]$-module $\mathsf{V}[\xi]$ defined in Section 5.8.3.

We may now give a basis free description of the characteristic polynomial C_{L}, recalling from Section 5.8.3 the notation $\mathsf{L}[\xi]$. This will simply be the characterisation many readers with some elementary linear algebra under their belt already know, but presented in fancier language.

5.8.17 Proposition (Intrinsic definition of characteristic polynomial) *If* F *is a field,* V *is a finite-dimensional* F*-vector space, and* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*, then* $\mathsf{C}_{\mathsf{L}} = \det(\xi\,\mathrm{id}_{\mathsf{V}}[\xi] - \mathsf{L}[\xi])$.

Proof Since we may determine C_{L} by choosing a basis as in Definition 5.4.55, we may as well suppose that $\mathsf{V} = \mathsf{F}^n$ and that L is naturally regarded as an $n \times n$ matrix. Thus we shall employ matrix notation and replace L with $\boldsymbol{A}$. As in Proposition 5.8.9 we have a natural isomorphism of $\mathsf{F}^n[\xi]$ with $\mathsf{F}[\xi]^n$; let us denote this isomorphism by ϕ. Let $\boldsymbol{B} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ be a general matrix. Since $\mathsf{F} \subseteq \mathsf{F}[\xi]$, the matrix $\boldsymbol{B}$ has associated with it in the usual manner an endomorphism of $\mathsf{F}[\xi]^n$. Let us denote this endomorphism by $\boldsymbol{B}'$. One can then directly check that the diagram

$$\begin{tikzcd}
\mathsf{F}^n[\xi] \arrow[r, "\boldsymbol{B}[\xi]"] \arrow[d, "\phi"'] & \mathsf{F}^n[\xi] \arrow[d, "\phi"] \\
\mathsf{F}[\xi]^n \arrow[r, "\boldsymbol{B}'"'] & \mathsf{F}[\xi]^n
\end{tikzcd}$$

commutes. That is to say, $\boldsymbol{B}[\xi]$ and $\boldsymbol{B}'$ are "the same" up to the isomorphism ϕ. Moreover, since ϕ maps the basis vector $\xi^0 \cdot \boldsymbol{e}_j$ for $\mathsf{F}^n[\xi]$ to the basis vector $\boldsymbol{e}_j$ for $\mathsf{F}[\xi]^n$, $j \in \{1, \dots, n\}$, it holds that

$$\det(\Phi) = \det(\phi^{-1} \circ \Phi \circ \phi)$$

for $\Phi \in \mathrm{End}_{\mathsf{F}[\xi]}(\mathsf{F}^n[\xi])$. Putting all of this together we obtain

$$\det(\xi\boldsymbol{I}_n[\xi] - \boldsymbol{A}[\xi]) = \det(\xi\boldsymbol{I}_n - \boldsymbol{A}),$$

as desired. ■

Perhaps the most difficult thing about the preceding proposition is believing that there is something to prove. Moreover, the result is of limited practical value, since, as mentioned previously, when one computes the characteristic polynomial in practice, one always chooses a basis and then computes $\det(\xi\boldsymbol{I}_n - \boldsymbol{A})$. However, the result does give a direct, basis free definition of the characteristic polynomial. Moreover, understanding the abstract construction in this specific case will be helpful in Section V-7.3 when we consider generalisations of the endomorphism $\xi\,\mathrm{id}_{\mathsf{V}}[\xi] - \mathsf{L}[\xi]$ to polynomial systems.

Having defined the characteristic polynomial, one would like to say something interesting about it. And, indeed, there are many interesting things to be said. However, it turns out that these interesting features of the characteristic polynomial are not so easy to get at, and so will only fall out after some more general development. For the moment, let us simply recall from Proposition 5.4.54 that the eigenvalues of L are exactly the roots of the characteristic polynomial. We also know from Proposition 5.8.13 that eigenvalues are roots of the minimal polynomial. The big advantage of the characteristic polynomial over the minimal polynomial in this respect is that it easy to compute.

An interesting feature of the characteristic polynomial is the following.

5.8.18 Proposition (Determinant, trace, and the characteristic polynomial) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. If*

$$\mathsf{C}_{\mathsf{L}} = \xi^{\mathsf{n}} + \mathsf{a}_{\mathsf{n}-1}\xi^{\mathsf{n}-1} + \cdots + \mathsf{a}_1\xi + \mathsf{a}_0,$$

then $\mathsf{a}_0 = (-1)^{\mathsf{n}} \det \mathsf{L}$ *and* $\mathsf{a}_{\mathsf{n}-1} = -\operatorname{tr} \mathsf{L}$*.*

Proof Without loss of generality we suppose that $\mathsf{V} = \mathsf{F}^n$ and that L is then an $n \times n$-matrix; let us denote $\mathsf{L} = \boldsymbol{A}$. We have

$$a_0 = \mathrm{Ev}_{\mathsf{F}}(C_{\boldsymbol{A}})(0_{\mathsf{F}}) = \det(0_{\mathsf{F}}\boldsymbol{I}_n - \boldsymbol{A}) = (-1)^n \det \boldsymbol{A}.$$

We prove that $a_{n-1} = -\operatorname{tr} \boldsymbol{A}$ by induction on n. It is trivial when $n = 1$, so suppose that the assertion is true for $n \in \{1, \dots, k-1\}$ and let $\boldsymbol{A} \in \mathrm{Mat}_{k \times n}(\mathsf{F})$. To compute the characteristic polynomial, let us expand $\det(\xi \boldsymbol{I}_k - \boldsymbol{A})$ along (say) the first column. Letting $\boldsymbol{A}' \in \mathrm{Mat}_{(k-1)\times(k-1)}(\mathsf{F})$ be the matrix obtained by deleting the first row and column from $\boldsymbol{A}$, we have

$$\det(\xi \boldsymbol{I}_k - \boldsymbol{A}) = (\xi - \boldsymbol{A}(1,1)) \det(\xi \boldsymbol{I}_{k-1} - \boldsymbol{A}') + P,$$

where P represents the terms coming from the expansion corresponding to the second through kth rows. We claim that $\deg(P) \leq k-2$. To see this note that P will be a linear combination of terms of the form $a \det \boldsymbol{B}$ where $a \in \mathsf{F}$ and where $\boldsymbol{B}$ is a $(k-1) \times (k-1)$ matrix with entries in $\mathsf{F}[\xi]$. Indeed, after a moments thought, one can see that $\boldsymbol{B}$ will contain $k-2$ entries that are of degree one, and all other entries will be degree zero. Thus $\deg(\det \boldsymbol{B}) \leq k-2$. Therefore, since we are interested in the coefficient of ξ^{k-1} in the characteristic polynomial, this will be the coefficient of ξ^{k-1} in the polynomial

$$(\xi - \boldsymbol{A}(1,1)) \det(\xi \boldsymbol{I}_{k-1} - \boldsymbol{A}').$$

Let us write

$$\det(\xi \boldsymbol{I}_{k-1} - \boldsymbol{A}') = \xi^{k-1} + a'_{k-2}\xi^{k-2} + \cdots + a'_1\xi + a'_0$$

so that

$$\begin{aligned}(\xi - \boldsymbol{A}(1,1)) &\det(\xi \boldsymbol{I}_{k-1} - \boldsymbol{A}') \\ &= \xi^k + (a'_{k-2} - \boldsymbol{A}(1,1))\xi^{k-1} + \cdots + (a'_0 - \boldsymbol{A}(1,1)a'_1)\xi + \boldsymbol{A}(1,1)a_0.\end{aligned}$$

By the induction hypothesis we have

$$a'_{k-2} = -\boldsymbol{A}(2,2) - \cdots - \boldsymbol{A}(k,k),$$

giving the result. ■

The reader is asked to explore this characterisation of the characteristic polynomial for 2 × 2-matrices in Exercise 5.8.8.

As a final bit of business concerning the characteristic polynomial, let us prove the Cayley–Hamilton[2] Theorem. This turns out to be a very important theorem, so we shall give alternative proofs as Corollaries 5.8.36 and 5.8.62 below.

[2] Arthur Cayley (1821–1895) was a British mathematician whose principal contributions were to geometry and linear algebra. He is regarded as the inventor of matrices. William Rowan Hamilton (1805–1865) was Irish, and his mathematical work was mainly in the area of algebra. He was knighted, becoming Sir William, in 1835.

5.8.19 Theorem (Cayley–Hamilton Theorem) *If* F *is a field, if* V *is a finite-dimensional* F*-vector space, and if* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*, then* $\mathrm{Ev}_{\mathsf{F}}(\mathsf{C}_{\mathsf{L}})(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$*, i.e.,* L *satisfies its own characteristic polynomial.*

Proof Let us denote

$$\mathsf{C}_{\mathsf{L}} = \xi^n + a_{n-1}\xi^{n-1} + \cdots + a_1\xi + a_0.$$

For $v \in \mathsf{V}$ define $\mathsf{A}_v \in \mathrm{Hom}_{\mathsf{F}}(\mathsf{F}^{n+1}; \mathsf{V})$ by

$$\mathsf{A}_v(c_0, c_1, \dots, c_n) = c_0 v + c_1\mathsf{L}(v) + \cdots + c_n\mathsf{L}^n(v).$$

We will show that

$$A_v(a_0, a_1, \dots, a_n) = 0_{\mathsf{V}}$$

for all $v \in \mathsf{V}$, noting that this proves the theorem.

Note that the family $(v, \mathsf{L}(v), \dots, \mathsf{L}^n(v))$ must be linearly independent since it consists of $n+1$ vectors in the n-dimensional vector space V. Let $k \in \mathbb{Z}_{\geq 0}$ be the least integer for which the set $\{v, \mathsf{L}(v), \dots, \mathsf{L}^{k-1}(v)\}$ is linearly independent. Note that this implies that

$$\mathsf{L}^j(v) \in \mathrm{span}_{\mathsf{F}}(v, \mathsf{L}(v), \dots, \mathsf{L}^{k-1}(v))$$

for every $j \in \{1, \dots, n\}$. (According to our terminology in Definition 5.8.21, the subspace spanned by $\{v, \mathsf{L}(v), \dots, \mathsf{L}^{k-1}(v)\}$ is a cyclic subspace.) Let $b_0, b_1, \dots, b_{k-1} \in \mathsf{F}$ be such that

$$\mathsf{L}^k(v) = b_0 v + b_1\mathsf{L}(v) + \cdots + \mathsf{L}^{k-1}(v).$$

Let us give a lemma which provides the form for $\ker(\mathsf{L}_v)$.

1 Lemma *Let* $\{\mathrm{e}_1, \dots, \mathrm{e}_{\mathrm{n}+1}\}$ *be the standard basis for* $\mathsf{F}^{\mathrm{n}+1}$ *and define*

$$\mathrm{u}_{\mathrm{j}} = -\mathrm{b}_0\mathrm{e}_{\mathrm{j}+1} - \mathrm{b}_1\mathrm{e}_{\mathrm{j}+2} - \cdots - \mathrm{b}_{\mathrm{k}-1}\mathrm{e}_{\mathrm{j}+\mathrm{k}} + \mathrm{e}_{\mathrm{j}+\mathrm{k}+1}, \qquad \mathrm{j} \in \{1, \dots, \mathrm{n}-\mathrm{k}+1\}.$$

Then $\{\mathrm{u}_1, \dots, \mathrm{u}_{\mathrm{n}-\mathrm{k}+1}\}$ *is a basis for* $\ker(\mathsf{L}(\mathrm{v}))$*.*

Proof First note that $\{u_1, \dots, u_{n-k+1}\}$ is linearly independent (why?). Next note that, byCorollary 5.4.4, $\dim(\ker(\mathsf{A}_v)) = n-k+1$ since $\mathrm{rank}(\mathsf{A}_v) = k$ ($\{v, \mathsf{L}(v), \dots, \mathsf{L}^{k-1}(v)\}$ is a basis for $\mathrm{image}(\mathsf{A}_v)$). Therefore, we need only show that $u_j \in \ker(\mathsf{A}_v)$ for $j \in \{1, \dots, n-k+1\}$. But we have

$$\begin{aligned}\mathsf{A}_v(u_j) = &-b_0\mathsf{L}^j(v) - b_1\mathsf{L}^{j+1}(v) - \cdots - b_{k-1}\mathsf{L}^{j+k-1}(v) + \mathsf{L}^{j+k}(v)\\ &= \mathsf{L}^j(-b_0 v - b_1\mathsf{L}(v) - \cdots - b_{k-1}\mathsf{L}^{k-1}(v) + \mathsf{L}^k(v)) = 0_{\mathsf{V}},\end{aligned}$$

giving the result. ▼

Now take a basis $\mathscr{B} = \{f_1, \dots, f_n\}$ such that $f_j = \mathsf{L}^{j-1}(v)$ for $j \in \{1, \dots, k\}$. Then, using the definition of the matrix representative, one readily sees that

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{0}_{(n-k)\times k} & \boldsymbol{A}_{22} \end{bmatrix},$$

where

$$\boldsymbol{A}_{11} = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -b_0 \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -b_1 \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -b_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -b_{k-2} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 1_{\mathsf{F}} & -b_{k-1} \end{bmatrix},$$

cf. the proof of Theorem 5.8.20. By Exercise 5.3.8 it follows that $C_{\mathsf{L}} = C_{\boldsymbol{A}_{11}} C_{\boldsymbol{A}_{22}}$. As we shall see below in Proposition 5.8.26,

$$C_{\boldsymbol{A}_{11}} = b_0 + b_1\xi + \cdots + b_{k-1}\xi^{k-1} + \xi^k.$$

If

$$C_{\boldsymbol{A}_{22}} = \alpha_0 + \alpha_1\xi + \cdots + \alpha_{n-k-1}\xi^{n-k-1} + \xi^{n-k}$$

that

$$C_{\mathsf{L}} = \alpha_0 C_{\boldsymbol{A}_{11}} + \alpha_1\xi C_{\boldsymbol{A}_{11}} + \cdots + \alpha_{n-k-1}\xi^{n-k-1} C_{\boldsymbol{A}_{11}} + \xi^{n-k} C_{\boldsymbol{A}_{11}}.$$

Matching the coefficients ξ^j, $j \in \{0, 1, \dots, k\}$, on each side of the equation gives

$$(a_0, a_1, \dots, a_n) = \alpha_0 u_1 + \cdots + \alpha_{n-k-1} u_{n-k} + u_{n-k+1},$$

where $\{u_1, \dots, u_{n-k+1}\} \subseteq \mathsf{F}^{n+1}$ is the basis for $\ker(\mathsf{A}_v)$ from the lemma above. Thus shows that $(a_0, a_1, \dots, a_n) \in \ker(\mathsf{A}_v)$, as desired. ■

5.8.5 Cyclic modules and cyclic vector spaces

In the structural theory of modules over principal ideal domains in Section 4.9—and V_{L} is just such an object—the building blocks are cyclic submodules. This, of course, motivates us to understand cyclic submodules of V_{L}. These, indeed, possess a very nice structure, as we shall see.

5.8.20 Theorem (When V_{L} is a cyclic module) *Let F be a field, let V be a finite-dimensional F-module, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Suppose that V_{L} is a cyclic $\mathsf{F}[\xi]$-module, supposing that*

$$\mathsf{V}_{\mathsf{L}} = \{\mathrm{P} \cdot \mathrm{v}_0 \mid \mathrm{P} \in \mathsf{F}[\xi]\}$$

for $\mathrm{v}_0 \in \mathsf{V}_{\mathsf{L}}$ and with $\mathrm{P}_0 \in \mathsf{F}[\xi]$ the unique monic polynomial for which V_{L} is isomorphic to $\mathsf{F}[\xi]/(\mathrm{P}_0)$ (cf. Proposition 4.9.7). Let $\mathrm{k} = \deg(\mathrm{P}_0)$. Then

$$\mathscr{B} = \{\mathrm{v}_0, \mathsf{L}(\mathrm{v}_0), \dots, \mathsf{L}^{\mathrm{k}-1}(\mathrm{v}_0)\}$$

is a basis (as an F-vector space) for V. Moreover, if

$$\mathrm{P}_0 = \xi^{\mathrm{k}} + \mathrm{a}_{\mathrm{k}-1}\xi^{\mathrm{k}-1} + \cdots + \mathrm{a}_1\xi + \mathrm{a}_0, \qquad \mathrm{a}_0, \mathrm{a}_1, \dots, \mathrm{a}_{\mathrm{k}-1} \in \mathsf{F},$$

then

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \begin{bmatrix} 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -\mathsf{a}_0 \\ 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -\mathsf{a}_1 \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -\mathsf{a}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & -\mathsf{a}_{\mathsf{k}-2} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 1_{\mathsf{F}} & -\mathsf{a}_{\mathsf{k}-1} \end{bmatrix}.$$

Proof Let $v \in \mathsf{V}$ and write $v = P \cdot v_0$ for $P \in \mathsf{F}[\xi]$, this being possible since v_0 generates V_{L} as an $\mathsf{F}[\xi]$-module. By the Euclidean Algorithm write $P = QP_0 + R$ where $\deg(R) < \deg(P_0)$. We then have $P \cdot v_0 = R \cdot v_0$ since $P_0 \cdot v_0 = 0_{\mathsf{V}}$. This shows that

$$v = P \cdot v_0 \in \mathrm{span}_{\mathsf{F}}(v_0, \xi \cdot v_0, \dots, \xi^{k-1} \cdot v_0).$$

In other words, the set $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{k-1}(v_0)\}$ generates V as an F-vector space.

Now suppose that

$$c_0 v_0 + c_1 \mathsf{L}(v_0) + \cdots + c_k \mathsf{L}^{k-1}(v_0) = 0_{\mathsf{V}}$$

for $c_0, c_1, \dots, c_{k-1} \in \mathsf{F}$. This means that $R \cdot v_0 = 0_{\mathsf{V}}$ where

$$R = c_{k-1}\xi^{k-1} + \cdots + c_1\xi + c_0.$$

Thus $R \in \mathrm{ann}(v_0)$. Since $\mathrm{ann}(v_0) = (P_0)$ and since $P_0 \nmid R$ it follows that $R = 0_{\mathsf{F}[\xi]}$, and so that $c_0 = c_1 = \cdots = c_{k-1} = 0_{\mathsf{F}}$. Thus $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{k-1}(v_0)\}$ is linearly independent, and so a basis.

For the final assertion of the theorem let us denote $e_j = \mathsf{L}^j(v_0)$, $j \in \{1, \dots, k\}$. Then

$$\begin{aligned} \mathsf{L}(e_0) &= \mathsf{L}(v_0) = 0_{\mathsf{F}} e_0 + 1_{\mathsf{F}} e_1 + 0_{\mathsf{F}} e_2 + \cdots + 0_{\mathsf{F}} e_{k-1}, \\ \mathsf{L}(e_1) &= \mathsf{L}^2(v_0) = 0_{\mathsf{F}} e_0 + 0_{\mathsf{F}} e_1 + 1_{\mathsf{F}} e_2 + \cdots + 0_{\mathsf{F}} e_{k-1}, \\ &\vdots \\ \mathsf{L}(e_{k-2}) &= \mathsf{L}^{k-1}(v_0) = 0_{\mathsf{F}} e_0 + 0_{\mathsf{F}} e_1 + 0_{\mathsf{F}} e_2 + \cdots + 1_{\mathsf{F}} e_{k-1}, \\ \mathsf{L}(e_{k-1}) &= \mathsf{L}^k(v_0) = -a_0 e_0 - a_1 e_1 - a_2 e_2 - \cdots - a_{k-1} e_{k-1}. \end{aligned}$$

The theorem now follows from the definition of the matrix representative. ■

Based on the theorem we make the following definition.

5.8.21 Definition (Cyclic vector space) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Then V is ***L-cyclic*** if there exists $v_0 \in \mathsf{V}$ and $k \in \mathbb{Z}_{>0}$ such that $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{k-1}(v_0)\}$ is a basis for V. The vector v_0 is a ***generator*** for the L-cyclic vector space V. •

It is useful to be able to characterise a vector $v_0 \in \mathsf{V}$ for which $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{k-1}(v_0)\}$ is a basis of a cyclic vector space. The following result tells us such a criterion in a special case to which the general case can be reduced, as we shall see.

5.8.22 Proposition (Generators for a cyclic vector space) *Let* F *be a field, let* V *be a* k*-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *be such that* $\mathsf{C}_{\mathsf{L}} = (\xi - \lambda)^{\mathrm{k}}$*. If* V *is* L*-cyclic, then the following statements are equivalent ror a vector* $\mathrm{v}_0 \in \mathsf{V}$*:*

(i) v_0 *is a generator for* V *as an* L*-cyclic vector space;*

(ii) $\mathsf{L}^{\mathrm{k}}(\mathrm{v}_0) = 0_{\mathsf{V}}$ *and* $\mathsf{L}^{\mathrm{k}-1}(\mathrm{v}_0) \neq 0_{\mathsf{V}}$.

Proof Let us assemble a few facts that hold under the hypotheses of the proposition. First of all, by Proposition 5.8.27 below, $M_{\mathsf{L}} = C_{\mathsf{L}}$, and so there exists $v_0 \in \mathsf{V}$ such that $(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{k-1}(v_0) \neq 0$. We also note that v_0 is a generator for the L-cyclic vector space V if and only if $\mathrm{ann}(v_0) = (C_{\mathsf{L}})$, thinking of v_0 as an element of the $\mathsf{F}[\xi]$-module V_{L}. This follows from Theorem 5.8.20 and from Proposition 5.8.26 below.

(i) $\Longrightarrow$ (ii) Suppose that v_0 is a generator for the L-cyclic vector space V. Then it holds that $\mathrm{ann}(v_0) = (C_{\mathsf{L}})$. Suppose that $(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{k-1}(v_0) = 0_{\mathsf{V}}$. Then we have $(\xi - \lambda)^{k-1} \in \mathrm{ann}(v_0)$, but $C_{\mathsf{L}} \nmid (\xi - \lambda)^{k-1}$ and so $(\xi - \lambda)^{k-1} \notin (C_{\mathsf{L}})$. This contradiction gives $(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{k-1}(v_0) \neq 0_{\mathsf{V}}$.

(ii) $\Longrightarrow$ (i) Suppose that $(\mathsf{L} - \mathrm{id}_{\mathsf{V}})^{k-1}(v_0) \neq 0_{\mathsf{V}}$. Suppose that $P \in \mathrm{ann}(v_0)$ and, by Corollary 4.4.16, write

$$P = A_0 + A_1(\xi - \lambda) + \cdots + A_m(\xi - \lambda)^m$$

where $m = \deg(P)$, and for $A_0, A_1, \dots, A_m \in \mathsf{F}[\xi]$ of degree 0, i.e., scalars. Then, since $(\mathsf{A} - \lambda\,\mathrm{id}_{\mathsf{V}})^j = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$ for $j \geq k$, we have

$$P \cdot v_0 = A_0 \cdot v_0 + A_1(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})(v_0) + \cdots + A_{k-1}(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{k-1}(v_0) = 0_{\mathsf{V}}.$$

Then

$$P(\xi - \lambda)^{k-1}(v_0) = A_0(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{k-1}(v_0) = 0_{\mathsf{V}}.$$

Thus, since A_0 is a scalar, we must have $A_0 = 0$. One continues in this way, multiplying by successively lower powers of $\xi - \lambda)$, that $A_0 = A_1 = \cdots = A_{k-1} = 0_{\mathsf{F}}$. But, in this case, $P \in (C_{\mathsf{L}})$. Thus $\mathrm{ann}(v_0) \subseteq (C_{\mathsf{L}})$. By the Cayley–Hamilton Theorem, $(C_L) \subseteq \mathrm{ann}(v_0)$, and so we conclude that v_0 is a generator for V as an L-cyclic vector space. ■

While it is evident from Theorem 5.8.20 that if V_{L} is cyclic then V is L-cyclic, the converse is also true.

5.8.23 Proposition (Equivalence of notions of cyclic) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then* V_{L} *is a cyclic* $\mathsf{F}[\xi]$*-module if and only if* V *is an* L*-cyclic vector space.*

Moreover, if V *is* L*-cyclic with basis* $\{\mathrm{v}_0, \mathsf{L}(\mathrm{v}_0), \dots, \mathsf{L}^{\mathrm{k}-1}(\mathrm{v}_0)\}$*, and if*

$$\mathsf{L}^{\mathrm{k}}(\mathrm{v}_0) = -\mathrm{a}_0 - \mathrm{a}_1\xi - \cdots - \mathrm{a}_{\mathrm{k}-1}\xi^{\mathrm{k}-1},$$

then V_{L} *is isomorphic to* $\mathsf{F}[\xi]/(\mathrm{P}_0)$ *where*

$$\mathrm{P}_0 = \xi^{\mathrm{k}} + \mathrm{a}_{\mathrm{k}-1}\xi^{\mathrm{k}-1} + \cdots + \mathrm{a}_1\xi + \mathrm{a}_0.$$

Proof From Theorem 5.8.20 it follows that if V_{L} is a cyclic $\mathsf{F}[\xi]$-module then it is an L-cyclic vector space. So suppose that $v_0 \in \mathsf{V}$ and $k \in \mathbb{Z}_{>0}$ are such that $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{k-1}(v_0)\}$ is a basis for V. Now let $v \in \mathsf{V}$ and write

$$v = c_0 v_0 + c_1 \mathsf{L}(v_0) + \dots + c_{k-1}\mathsf{L}^{k-1}(v_0).$$

Thus $v = R \cdot v_0$ where

$$R = c_{k-1}\xi^{k-1} + \dots + c_1\xi + c_0.$$

Thus v_0 generates the $\mathsf{F}[\xi]$-module V_{L}, meaning that this module is cyclic. This gives the first part of the result.

Now write

$$\mathsf{L}^k(v_0) = -a_{k-1}\mathsf{L}^{k-1}(v_0) - \dots - a_1\mathsf{L}(v_0) - a_0 v_0$$

for some $a_0, a_1, \dots, a_{k-1} \in \mathsf{F}$ and define

$$P_0 = \xi^k + a_{k-1}\xi^{k-1} + \dots + a_1\xi + a_0.$$

It follows that $P_0 \cdot v_0 = 0_{\mathsf{V}}$ and so $P_0 \in \operatorname{ann}(v_0)$. If $\operatorname{ann}(v_0) = (Q_0)$ for some polynomial Q_0 (as will be the case since $\mathsf{F}[\xi]$ is a principal ideal domain), then we must have $(P_0) \subseteq (Q_0)$ and so $Q_0|P_0$ by Proposition 4.2.61(i). In particular, this implies that $\deg(Q_0) \leq \deg(P_0)$. Thus we may write

$$Q_0 = b_k\xi^k + \dots + b_1\xi + b_0$$

for some $b_0, b_1, \dots, b_k \in \mathsf{F}$. Then, since $Q_0 \cdot v_0 = 0_{\mathsf{V}}$, we have

$$\begin{aligned}
& b_k\mathsf{L}^k(v_0) + b_{k-1}\mathsf{L}^{k-1}(v_0) + \dots + b_1\mathsf{L}(v_0) + b_0 v_0 = 0_{\mathsf{V}} \\
\implies \quad & b_k(-a_{k-1}\mathsf{L}^{k-1}(v_0) - \dots - a_1\mathsf{L}(v_0) - a_0 v_0) + b_{k-1}\mathsf{L}^{k-1}(v_0) + \\
& + \dots + b_1\mathsf{L}(v_0) + b_0 v_0 = 0_{\mathsf{V}} \\
\implies \quad & b_j = b_k a_j, \qquad j \in \{1, \dots, k-1\}.
\end{aligned}$$

Thus Q_0 is an associate of P_0, and so $(Q_0) = (P_0)$ and so $\mathsf{V} = \mathsf{F}[\xi]/(P_0)$ by Proposition 4.9.7. ■

While the form for the matrix representative of L given in Theorem 5.8.20 is the most natural from the point of view of the structure of V as a cyclic $\mathsf{F}[\xi]$-module, it will turn out that for us there is an alternative matrix representative that will be more revealing. While we shall not see the benefits of this here, let us give this alternative form here.

5.8.24 Proposition (Alternative matrix representative for L-cyclic vector spaces) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$. Suppose that V_{L} is a cyclic $\mathsf{F}[\xi]$-module, supposing that*

$$\mathsf{V}_{\mathsf{L}} = \{\mathrm{P} \cdot \mathrm{v}_0 \mid \mathrm{P} \in \mathsf{F}[\xi]\}$$

for $\mathrm{v}_0 \in \mathsf{V}_{\mathsf{L}}$ and with $\mathrm{P}_0 \in \mathsf{F}[\xi]$ the unique monic polynomial for which V_{L} is isomorphic to $\mathsf{F}[\xi]/(\mathrm{P}_0)$. Write

$$\mathrm{P}_0 = \xi^{\mathrm{k}} + \mathrm{a}_{\mathrm{k}-1}\xi^{\mathrm{k}-1} + \dots + \mathrm{a}_1\xi + \mathrm{a}_0.$$

Then there exists a basis $\mathscr{B}$ *for* V *such that*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 1_{\mathsf{F}} \\ -\mathrm{a}_0 & -\mathrm{a}_1 & -\mathrm{a}_2 & -\mathrm{a}_3 & \cdots & -\mathrm{a}_{\mathrm{k}-1} \end{bmatrix}.$$

Proof Let $e_j = \mathsf{L}^{k-j}(v_0)$, $j \in \{1,\dots,k\}$ so that $\mathscr{B}' = \{e_k,\dots,e_1\}$ is the basis giving the matrix representative of Theorem 5.8.20. In particular, $\{e_1,\dots,e_k\}$ is a basis. Define $\boldsymbol{T} \in \mathrm{Mat}_{k\times k}(\mathsf{F})$ by

$$\boldsymbol{T} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ a_{k-1} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ a_{k-2} & a_{k-1} & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_1 & a_2 & a_3 & \cdots & 1_{\mathsf{F}} \end{bmatrix}.$$

Since $\det \boldsymbol{T} = 1_{\mathsf{F}}$ by Proposition 5.3.3(iv), $\boldsymbol{T}$ is invertible by Theorem 5.3.10. Therefore, if we define

$$f_j = \sum_{l=1}^{k} \boldsymbol{T}(l,j)e_l, \qquad j \in \{1,\dots,k\},$$

then $\{f_1,\dots,f_k\}$ is a basis for V. Rather than explicitly computing the inverse of $\boldsymbol{T}$ to determine the matrix representative using the change of basis formula, let us take a more indirect route. Define polynomials $Q_0, Q_1, \dots, Q_k \in \mathsf{F}[\xi]$ by

$$Q_j = \sum_{l=0}^{k-j} a_{l+j}\xi^l,$$

where we let $a_k = 1_{\mathsf{F}}$. Thus

$$Q_0 = P_0,\ Q_1 = a_1 + a_2\xi + \cdots + a_k\xi^{k-1},\ \dots,\ Q_k = a_k.$$

Note that

$$\begin{aligned} \xi Q_j = \sum_{l=0}^{k-j} a_{l+j}\xi^{l+1} &= \sum_{l=0}^{k-j} a_{l+j}\xi^{l+1} + a_{j-1} - a_{j-1} = \sum_{l=-1}^{k-j} a_{l+j}\xi^{l+1} - a_{j-1}Q_k \\ &= \sum_{l'=0}^{k-(j-1)} a_{l'+(j-1)}\xi^{l'} - a_{j-1}Q_k = Q_{j-1} - a_{j-1}Q_k, \end{aligned}$$

for each $j \in \{0,1,\dots,k\}$. One directly sees that

$$f_j = Q_j \cdot v_0, \qquad j \in \{1,\dots,k\},$$

and so, by our preceding computation,

$$\mathsf{L}(f_j) = f_{j-1} - a_{j-1} f_k.$$

Therefore, taking $f_0 = Q_0 \cdot v_0 = 0_{\mathsf{V}}$,

$$\begin{aligned}
\mathsf{L}(f_1) &= 0_{\mathsf{F}} f_1 + 0_{\mathsf{F}} f_2 + \cdots + 0_{\mathsf{F}} f_{k-1} - a_0 f_k,\\
\mathsf{L}(f_2) &= 1_{\mathsf{F}} f_1 + 0_{\mathsf{F}} f_2 + \cdots + 0_{\mathsf{F}} f_{k-1} - a_1 f_k,\\
&\ \ \vdots\\
\mathsf{L}(f_{k-1}) &= 0_{\mathsf{F}} f_1 + 0_{\mathsf{F}} f_1 + \cdots + 0_{\mathsf{F}} f_{k-1} - a_{k-1} f_k,\\
\mathsf{L}(f_k) &= 0_{\mathsf{F}} f_1 + 0_{\mathsf{F}} f_1 + \cdots + 1_{\mathsf{F}} f_{k-1} - a_k f_k.
\end{aligned}$$

Taking $\mathscr{B} = \{f_1, \dots, f_k\}$, the result now follows by the definition of the matrix representative. ■

The matrix representative of the preceding result we shall give a name and some notation.

5.8.25 Definition (Companion matrix) Let F be a field, let $k \in \mathbb{Z}_{>0}$, let $P \in \mathsf{F}[\xi]$ be a monic polynomial of degree k, and write

$$P = \xi^k + a_{k-1}\xi^{k-1} + \cdots + a_1\xi + a_0.$$

The matrix

$$\boldsymbol{C}(P) \triangleq \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 1_{\mathsf{F}} \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{k-1} \end{bmatrix}$$

is the ***companion matrix*** associated to the polynomial P. •

Some authors call the matrix of Theorem 5.8.20 the companion matrix. There are also other possible forms, but they are all characterised by having the coefficients of the polynomial in the first or last row or column, and by having 1_{F} in the entries just above or just below the diagonal.

It will be useful to know the characteristic polynomial of L when V is L-cyclic.

5.8.26 Proposition (Characteristic polynomial corresponding to cyclic vector spaces) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Suppose that $\mathrm{P}_0 \in \mathsf{F}[\xi]$ is the unique monic polynomial for which V_{L} is isomorphic to $\mathsf{F}[\xi]/(\mathrm{P}_0)$. Then $\mathrm{C}_{\mathsf{L}} = \mathrm{P}_0$.*

Proof It suffices to prove that the characteristic polynomial for the companion matrix $\boldsymbol{C}(P_0)$ is P_0. We prove this by induction on $\dim_{\mathsf{F}}(\mathsf{V})$. For $k = 1$ we have $\boldsymbol{C}(P_0) = [-a_0]$ if $P_0 = \xi + a_0$. The characteristic polynomial of this 1×1 matrix is $\xi - a_0$, giving the result in this case. Now suppose that the result is true for $j \times j$ companion matrices for $j \in \{1, \dots, k-1\}$ and let $\boldsymbol{C}(P_0)$ be the companion matrix for the polynomial

$$P_0 = \xi^k + a_{k-1}\xi^{k-1} + \cdots + a_1\xi + a_0.$$

Let

$$P_0' = \xi^{k-1} + a_{k-1}\xi^{k-2} + \cdots + a_2\xi + a_1.$$

Expanding the determinant for $\xi\boldsymbol{I}_n - \boldsymbol{C}(P_0)$ about the first column yields

$$\det(\xi\boldsymbol{I}_n - \boldsymbol{C}(P_0)) = \xi \det(\xi\boldsymbol{I}_{n-1} - \boldsymbol{C}(P_0')) + (-1)^{k+1}a_0 \det \boldsymbol{A}',$$

where $\boldsymbol{A}'$ is the lower triangular $(k-1) \times (k-1)$ matrix with -1_{F}'s on the diagonal and ξ's below the diagonal. Thus $\det \boldsymbol{A}' = (-1_{\mathsf{F}})^{k-1}$ by Proposition 5.3.3(iv). By the induction hypothesis we then have

$$\det(\xi\boldsymbol{I}_n - \boldsymbol{C}(P_0)) = \xi P_0' + a_0 = P_0,$$

as desired. ■

As a mildly interesting corollary we note that to every polynomial there corresponds a matrix with that polynomial as its characteristic polynomial.

There is more one can say in terms of characterising the characteristic polynomial for cyclic modules. Indeed, in this special case, and only if this special case, the minimal polynomial and the characteristic polynomial agree.

5.8.27 Proposition (The minimal and characteristic polynomials agree for cyclic vector spaces) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Then the following two statements are equivalent:*

(i) V *is* L*-cyclic;*

(ii) $\mathrm{M}_{\mathsf{L}} = \mathrm{C}_{\mathsf{L}}$.

Proof Our proof relies on Theorem 5.8.33 and Proposition 5.8.35 below.

The following general lemma about modules over a principal ideal domain is helpful.

1 Lemma *Let* R *be a principal ideal domain and let* M *be a torsion* R*-module. Then* M *is cyclic if and only if the elementary divisors of* M *are of the form* $\mathrm{p}_1^{\mathrm{l}_1}, \dots, \mathrm{p}_\mathrm{k}^{\mathrm{l}_\mathrm{k}}$ *for nonassociate primes* $\mathrm{p}_1, \dots, \mathrm{p}_\mathrm{k}$ *and for* $\mathrm{l}_1, \dots, \mathrm{l}_\mathrm{k} \in \mathbb{Z}_{>0}$*.*

Proof If M is cyclic, then the form of the elementary divisors follows from Proposition 4.9.9(ii) and Theorem 4.9.18. So suppose that the elementary divisors are $p_1^{l_1}, \dots, p_k^{l_k}$ as stated. Then the primary-cyclic decomposition gives an isomorphism of M with

$$\mathsf{R}/(p_1^{l_1}) \oplus \cdots \oplus \mathsf{R}/(p_k^{l_1}).$$

However, by Proposition 4.9.9(ii) this means that M is isomorphic to $\mathsf{R}/(r)$ with $r = p_1^{l_1} \dots p_k^{l_k}$. Thus M is cyclic. ▼

The proposition now follows directly from frefprop:min/char-poly since $M_{\mathsf{L}} = C_{\mathsf{L}}$ if and only if the elementary divisors are of the form $P_1^{l_1}, \cdots, P_k^{l_k}$ for distinct, monic, irreducible polynomials $P_1, \dots, P_k$ and $l_1, \dots, l_k \in \mathbb{Z}_{>0}$. ■

5.8.6 The primary decomposition of V_L

We shall now give a canonical form for an endomorphism based on the decomposition theorems of Section 4.9 for modules over principal ideal domains. However, our presentation in this section will be made independently of the generalities from Section 4.9 in order that we have the presentation be as self-contained as possible.

First let us give some definitions that adapt those from the general setup.

5.8.28 Definition (Elementary divisor, invariant factor) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_\mathsf{F}(\mathsf{V})$. A monic polynomial $P \in \mathsf{F}[\xi]$ is an ***elementary divisor*** (resp. ***invariant factor***) for L if it is an elementary divisor (resp. invariant factor) for the $\mathsf{F}[\xi]$-module V_L. •

Note that if $P \in \mathsf{F}[\xi]$ is an elementary divisor or invariant factor for V_L, then there exists a unique monic polynomial P' such that $P' = aP$ for some $a \in \mathsf{F}^*$. Thus the elementary divisors or invariant factors for V_L are uniquely specified once one asks that they be monic polynomials. Moreover, once one also fixes this convention that the elementary divisors and invariant factors be monic, the invariant factors of L and V_L are precisely the same.

We begin our decomposition of V_L by describing the primary decomposition in terms of the minimal polynomial.

5.8.29 Theorem (The primary decomposition of V_L) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_\mathsf{F}(\mathsf{V})$. Write the minimal polynomial as*

$$\mathrm{M}_\mathsf{L} = \mathrm{P}_1^{\mathrm{l}_1} \cdots \mathrm{P}_\mathrm{k}^{\mathrm{l}_\mathrm{k}},$$

where the polynomials $\mathrm{P}_1, \dots, \mathrm{P}_\mathrm{k} \in \mathsf{F}[\xi]$ are distinct, monic, and irreducible, and where $\mathrm{l}_1, \dots, \mathrm{l}_\mathrm{k} \in \mathbb{Z}_{>0}$. Denote

$$\mathsf{P}_\mathrm{j}(\mathsf{L}) = \ker(\mathrm{Ev}_\mathsf{F}(\mathrm{P}_\mathrm{j}^{\mathrm{l}_\mathrm{j}})(\mathsf{L})), \qquad \mathrm{j} \in \{1, \dots, \mathrm{k}\}.$$

Then $\dim_\mathsf{F}(\mathsf{P}_\mathrm{j}(\mathsf{L})) > 0$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$, and the primary decomposition of V_L is

$$\mathsf{V}_\mathsf{L} = \mathsf{P}_1(\mathsf{L}) \oplus \cdots \oplus \mathsf{P}_\mathrm{k}(\mathsf{L}).$$

Proof This follows from the arguments of Theorem 4.9.14, however, we shall give an explicit and independent proof here for convenience.

Write $M_\mathsf{L} = P_1^{l_1} \cdots P_k^{l_k}$ as in the statement of the proposition. For $j \in \{1, \dots, k\}$ define

$$Q_j = P_1^{l_1} \cdots P_{j-1}^{l_{j-1}} P_{j+1}^{l_{j+1}} \cdots P_k^{l_k}.$$

Since the polynomials $Q_1, \dots, Q_k$ are coprime, by Corollary 4.4.36, there exist polynomials $R_1, \dots, R_k$ such that

$$R_1 Q_1 + \cdots + R_k Q_k = 1_\mathsf{F}.$$

Since

$$\mathrm{Ev}_{\mathsf{F}}(P_j^{l_j}R_jQ_j)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}, \qquad j \in \{1,\dots,k\},$$

$\mathrm{Ev}_{\mathsf{F}}(R_jQ_j)(\mathsf{L}) \in \mathsf{V}_{\mathsf{L}}(P_j)$ for each $j \in \{1,\dots,k\}$. Therefore, for each $v \in \mathsf{V}_{\mathsf{L}}$,

$$v = 1_{\mathsf{F}} \cdot v = (R_1Q_1) \cdot v + \dots + (R_kQ_k) \cdot v.$$

This gives $\mathsf{V}_{\mathsf{L}} = \sum_{j=1}^{k} \mathsf{V}_{\mathsf{L}}(P_j)$ as a sum of primary modules.

To show that the sum is direct, suppose that $v \in \mathsf{V}_{\mathsf{L}}(P_{j_1}) \cap \mathsf{V}_{\mathsf{L}}(P_{j_2})$ for $j_1 \neq j_2$. Thus $P_{j_1}^{k_1} \cdot v = P_{j_2}^{k_2} \cdot v = 0_{\mathsf{V}}$ for some $k_1, k_2 \in \mathbb{Z}_{>0}$. By Corollary 4.4.36 there exists polynomials R_1 and R_2 such that

$$R_1P_{j_1}^{k_1} + R_2P_{j_2}^{k_2} = 1_{\mathsf{F}}.$$

But then

$$v = 1_{\mathsf{F}} \cdot v = (R_1P_{j_1}^{k_1} + R_2P_{j_2}^{k_2}) \cdot v = (R_1P_{j_1}^{k_1}) \cdot v + (R_2P_{j_2}^{k_2}) \cdot v = 0_{\mathsf{V}}.$$

Thus $\mathsf{V}_{\mathsf{L}} = \oplus_{j=1}^{k} \mathsf{V}_{\mathsf{L}}(P_j)$.

Next we show that $\mathsf{V}_{\mathsf{L}}(P_j) = \ker(\mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L}))$ for $j \in \{1,\dots,k\}$. First we claim that

$$\ker(\mathrm{Ev}_{\mathsf{F}}(P_j^{r})(\mathsf{L}) = \mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L})$$

for every $r > l_j$. Indeed, suppose otherwise. Then there exists $v \in \ker(\mathrm{Ev}_{\mathsf{F}}(P_j^{r})(\mathsf{L}))$ such that $\mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L}) \cdot v \neq 0_{\mathsf{V}}$. Note that $v \in \mathsf{V}_{\mathsf{L}}(P_j)$ and so $P_j^{l_j} \cdot v \in \mathsf{V}_{\mathsf{L}}(P_j)$ since $\mathsf{V}_{\mathsf{L}}(P_j)$ is a submodule. Therefore,

$$M_{\mathsf{L}} \cdot v = (P_1^{l_1} \cdots P_{j-1}^{l_{j-1}}P_{j+1}^{l_{j+1}} \cdots P_k^{l_k}P_j^{l_j}) \cdot v \neq 0_{\mathsf{V}},$$

contradicting the definition of M_{L}. Now, by definition of $\mathsf{V}_{\mathsf{L}}(P_j)$, we have

$$\begin{aligned}\mathsf{V}_{\mathsf{L}}(P_j) =& \{v \in \mathsf{V}_{\mathsf{L}} \mid P_j^{r} \cdot v = 0_{\mathsf{V}} \text{ for some } r \in \mathbb{Z}_{>0}\}\\ =& \cup_{r\in\mathbb{Z}_{>0}} \ker(\mathrm{Ev}_{\mathsf{F}}(P_j^{r})(\mathsf{L}) = \ker(\mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L})),\end{aligned}$$

as desired.

Finally we show that $\dim_{\mathsf{F}}(\mathsf{P}_j(\mathsf{L})) > 0$. Suppose that $\dim_{\mathsf{F}}(\mathsf{P}_j(\mathsf{L})) = 0$. This means that $\mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L})$ is invertible. Define

$$P_j' = P_1^{l_1} \cdots P_{j-1}^{l_{j-1}}P_{j+1}^{l_{j+1}} \cdots P_k^{l_k}.$$

Then

$$0_{\mathrm{End}_{\mathsf{F}}(\mathsf{L})} = \mathrm{Ev}_{\mathsf{F}}(M_{\mathsf{L}})(\mathsf{L}) = \mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L}) \circ \mathrm{Ev}_{\mathsf{F}}(P_j')(\mathsf{L}),$$

which gives $\mathrm{Ev}_{\mathsf{F}}(P_j')(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$ since $\mathrm{Ev}_{\mathsf{F}}(P_j^{l_j})(\mathsf{L})$ is invertible. But this contradicts the fact that M_{L} is the least degree polynomial in $\mathrm{ann}(\mathsf{V})$. ■

One way of understanding the preceding theorem is that the minimal polynomial serves to determine the primary decomposition of V_L. However, the primary decomposition does not uniquely characterise V_L as a module over $\mathsf{F}[\xi]$, just as is the case of the general constructions of Section 4.9. This is reflected by the fact that the minimal polynomial does not uniquely determine an endomorphism up to similarity. Let us illustrate this with an example.

5.8.30 Example (Example 5.8.16 cont'd) We consider L and L' as given in Example 5.8.16:

$$\mathsf{L} = \begin{bmatrix} 1_\mathsf{F} & 1_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} \\ 0_\mathsf{F} & 1_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} \\ 0_\mathsf{F} & 0_\mathsf{F} & 1_\mathsf{F} & 1_\mathsf{F} \\ 0_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} & 1_\mathsf{F} \end{bmatrix}, \qquad \mathsf{L}' = \begin{bmatrix} 1_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} \\ 0_\mathsf{F} & 1_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} \\ 0_\mathsf{F} & 0_\mathsf{F} & 1_\mathsf{F} & 1_\mathsf{F} \\ 0_\mathsf{F} & 0_\mathsf{F} & 0_\mathsf{F} & 1_\mathsf{F} \end{bmatrix}.$$

We have already indicated that both endomorphisms have minimal polynomial $M_\mathsf{L} = M_{\mathsf{L}'} = (\xi - 1_\mathsf{F})^2$. However, the elementary divisors for L are $\{(\xi - 1_\mathsf{F})^2, (\xi - 1_\mathsf{F})^2\}$, whereas the elementary divisors for L' are $\{\xi - 1_\mathsf{F}, \xi - 1_\mathsf{F}, (\xi - 1_\mathsf{F})^2\}$. That these are indeed the elementary divisors may not be clear right now, but this will be evident shortly. •

The primary decomposition is also related to the notion of a generalised eigenspace as given in Definition 5.4.57.

5.8.31 Proposition (Minimal polynomial and generalised eigenspace) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_\mathsf{F}(\mathsf{V})$*. If* $\lambda \in \mathsf{F}$ *is an eigenvalue for* L *with algebraic multiplicity* k*, then*

$$\overline{\mathsf{W}}(\lambda, \mathsf{L}) = \ker(\mathrm{Ev}_\mathsf{F}((\xi - \lambda)^{\mathrm{k}})(\mathsf{L})).$$

Proof Let $\mathsf{L}_\lambda = \lambda\,\mathrm{id}_\mathsf{V} - \mathsf{L}$. Since

$$\overline{\mathsf{W}}(\lambda, \mathsf{L}) = \bigcup_{j \in \mathbb{Z}_{>0}} \ker(\mathsf{L}_\lambda^j),$$

we immediately have

$$\ker(\mathrm{Ev}_\mathsf{F}((\xi - \lambda)^k)(\mathsf{L})) \subseteq \overline{\mathsf{W}}(\lambda, \mathsf{L}).$$

Since λ is an eigenvalue, by Proposition 5.8.12 we have $(\xi - \lambda) | M_\mathsf{L}$. More specifically we can write $M_\mathsf{L} = (\xi - \lambda)^m P$ for some $m \in \mathbb{Z}_{>0}$ and for some $P \in \mathsf{F}[\xi]$ having the property that $(\xi - \lambda) \nmid P$. By Theorem 5.8.29 we have

$$\mathsf{V} = \ker(\mathrm{Ev}_\mathsf{F}((\xi - \lambda)^m)(\mathsf{L}) \oplus \ker(\mathrm{Ev}_\mathsf{F}(P)(\mathsf{L})),$$

and this decomposition is one of L-invariant subspaces. Moreover, since $(\xi - \lambda) \nmid P$, λ is not an eigenvalue of $\mathsf{L} | \ker(\mathrm{Ev}_\mathsf{F}(P)(\mathsf{L}))$. Thus the generalised eigenspace $\overline{\mathsf{W}}(\lambda, \mathsf{L})$ is a subspace of $\ker(\mathrm{Ev}_\mathsf{F}((\xi - \lambda)^m)(\mathsf{L})$. Moreover, $m \leq k$ by definition of the minimal polynomial. Thus

$$\overline{\mathsf{W}}(\lambda, \mathsf{L}) \subseteq \ker(\mathrm{Ev}_\mathsf{F}((\xi - \lambda)^k)(\mathsf{L}),$$

giving the result. ■

Th result has the following important corollary.

5.8.32 Corollary (Generalised eigenspace decomposition) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *be such that* M_{L} *splits in* F*. If the distinct eigenvalues of* L *are* $\lambda_1, \dots, \lambda_k \in \mathsf{F}$*, then*

$$\mathsf{V} = \overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \oplus \cdots \oplus \overline{\mathsf{W}}(\lambda_k, \mathsf{L}).$$

Proof Since M_{L} splits we must have

$$M_{\mathsf{L}} = (\xi - \lambda_1)^{l_1} \cdots (\xi - \lambda_k)^{l_k}$$

as the decomposition of M_{L} into products of powers of primes. The result now follows from Theorem 5.8.29 and Proposition 5.8.31. ■

5.8.7 The rational canonical form

We are now in a position to give a representation for an endomorphism that is fully adapted to the primary-cyclic decomposition of a module over a principal domain as given by Theorem 4.9.18. Indeed, we have the following theorem.

5.8.33 Theorem (Rational canonical form) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. Let* $\mathrm{P}_1, \dots, \mathrm{P}_k \in \mathsf{F}[\xi]$ *be distinct, monic, irreducible polynomials, let* $\mathrm{m}_1, \dots, \mathrm{m}_k \in \mathbb{Z}_{>0}$*, and let* $\mathrm{l}_{11}, \dots, \mathrm{l}_{1\mathrm{m}_1}, \dots, \mathrm{l}_{k1}, \dots, \mathrm{l}_{k\mathrm{m}_k} \in \mathbb{Z}_{>0}$ *be such that*

$$\mathrm{l}_{11} \leq \cdots \leq \mathrm{l}_{1\mathrm{m}_1}, \dots, \mathrm{l}_{k1} \leq \cdots \leq \mathrm{l}_{k\mathrm{m}_k}$$

and be such that the polynomials

$$\mathrm{P}_1^{\mathrm{l}_{11}}, \dots, \mathrm{P}_1^{\mathrm{l}_{1\mathrm{m}_1}}, \dots, \mathrm{P}_k^{\mathrm{l}_{k1}}, \dots, \mathrm{P}_k^{\mathrm{l}_{k\mathrm{m}_k}}$$

are the elementary divisors of V_{L}*, cf. Theorem 4.9.18. Then there exists* L*-invariant subspaces* $\mathsf{U}_{11}, \dots, \mathsf{U}_{1\mathrm{m}_1}, \dots, \mathsf{U}_{k1}, \dots, \mathsf{U}_{k\mathrm{m}_k}$ *of* V *such that*

(i) $\mathsf{V} = \mathsf{U}_{11} \oplus \cdots \oplus \mathsf{U}_{1\mathrm{m}_1} \oplus \cdots \oplus \mathsf{U}_{k1} \oplus \cdots \oplus \mathsf{U}_{k\mathrm{m}_k}$ *and such that*

(ii) $\mathsf{L}|\mathsf{U}_{jr}$ *is cyclic with characteristic polynomial* $\mathrm{P}_j^{\mathrm{l}_{jr}}$ *for each* $j \in \{1, \dots, k\}$ *and* $r \in \{1, \dots, \mathrm{m}_j\}$.

Moreover, there exists a basis $\mathscr{B}$ *for* V *such that*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \mathrm{diag}(\mathbf{C}(\mathrm{P}_1^{\mathrm{l}_{11}}), \dots, \mathbf{C}(\mathrm{P}_1^{\mathrm{l}_{1\mathrm{m}_1}}), \dots, \mathbf{C}(\mathrm{P}_k^{\mathrm{l}_{k1}}), \dots, \mathbf{C}(\mathrm{P}_k^{\mathrm{l}_{k\mathrm{m}_k}})).$$

Finally, if

$$[\mathsf{L}]_{\mathscr{B}'}^{\mathscr{B}'} = \mathrm{diag}(\mathbf{C}(\mathrm{Q}_1^{\mathrm{r}_{11}}), \dots, \mathbf{C}(\mathrm{Q}_1^{\mathrm{r}_{1\mathrm{n}_1}}), \dots, \mathbf{C}(\mathrm{Q}_s^{\mathrm{r}_{p1}}), \dots, \mathbf{C}(\mathrm{Q}_p^{\mathrm{r}_{p\mathrm{n}_p}}))$$

is a matrix representative of L *in another basis, then* $\mathrm{p} = \mathrm{k}$ *and there exists a permutation* $\sigma \in \mathfrak{S}_k$ *such that* $\mathrm{Q}_j = \mathrm{P}_\sigma(j)$*,* $\mathrm{n}_j = \mathrm{m}_{\sigma(j)}$*, and* $\mathrm{r}_{ja} = \mathrm{l}_{\sigma(j)a}$ *for* $j \in \{1, \dots, k\}$ *and* $a \in \{1, \dots, \mathrm{m}_j\}$.

Proof Let $\mathsf{V} = \mathsf{P}_1(\mathsf{L}) \oplus \cdots \oplus \mathsf{P}_k(\mathsf{L})$ be the primary decomposition as per Theorem 5.8.29, noting that $\mathsf{P}_j(\mathsf{L}) = \mathsf{V}_{\mathsf{L}}(P_j)$ for $j \in \{1, \dots, k\}$. According to Theorem 4.9.16, for each $j \in \{1, \dots, k\}$ we have a decomposition of $\mathsf{P}_j(\mathsf{L})$ into cyclic modules

$$\mathsf{P}_j(\mathsf{L}) = \mathsf{U}_{j1} \oplus \cdots \oplus \mathsf{U}_{jm_j}$$

where the submodule U_{jr} has order $P_j^{l_{jr}}$ for $r \in \{1, \dots, m_j\}$. According to Proposition 5.8.7, these submodules of V_{L} are L-invariant subspaces. Moreover, by Proposition 5.8.26, the characteristic polynomial of $\mathsf{L}|\mathsf{U}_{jr}$ is $P_j^{l_{jr}}$, just as asserted in the statement of the first part of the result.

Now, as in Proposition 5.8.24, for each of the subspaces U_{jr}, $j \in \{1, \dots, k\}$, $r \in \{1, \dots, m_j\}$, there is a basis $\mathscr{B}_{jr}$ such that $[\mathsf{L}|\mathsf{U}_{jr}]_{\mathscr{B}_{jr}}^{\mathscr{B}_{jr}} = \boldsymbol{C}(P_j^{l_{jr}})$. Taking

$$\mathscr{B} = \mathscr{B}_{11} \cup \cdots \cup \mathscr{B}_{1m_1} \cup \cdots \cup \mathscr{B}_{k1} \cup \cdots \cup \mathscr{B}_{km_k}$$

gives the second part of the result by Proposition 5.4.10.

The final assertion follows by the uniqueness of the primary-cyclic decomposition of V_{L}. ■

The matrix representative of the preceding theorem has a name.

5.8.34 Definition (Rational canonical form) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The matrix representative of Theorem 5.8.33 is the ***rational canonical form*** for L. •

One of the most useful features of the rational canonical form is that it allows us to easily characterise the minimal and characteristic polynomials.

5.8.35 Proposition (Elementary divisors and the minimal and characteristic polynomials) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Let $\mathrm{P}_1, \dots, \mathrm{P}_\mathrm{k} \in \mathsf{F}[\xi]$ be distinct, monic, irreducible polynomials, let $\mathrm{m}_1, \dots, \mathrm{m}_\mathrm{k} \in \mathbb{Z}_{>0}$, and let $\mathrm{l}_{11}, \dots, \mathrm{l}_{1\mathrm{m}_1}, \dots, \mathrm{l}_{\mathrm{k}1}, \dots, \mathrm{l}_{\mathrm{km}_\mathrm{k}} \in \mathbb{Z}_{>0}$ be such that*

$$\mathrm{l}_{11} \leq \cdots \leq \mathrm{l}_{1\mathrm{m}_1}, \dots, \mathrm{l}_{\mathrm{k}1} \leq \cdots \leq \mathrm{l}_{\mathrm{km}_\mathrm{k}}$$

and be such that the polynomials

$$\mathrm{P}_1^{\mathrm{l}_{11}}, \dots, \mathrm{P}_1^{\mathrm{l}_{1\mathrm{m}_1}}, \dots, \mathrm{P}_\mathrm{k}^{\mathrm{l}_{\mathrm{k}1}}, \dots, \mathrm{P}_\mathrm{k}^{\mathrm{l}_{\mathrm{km}_\mathrm{k}}}$$

are the elementary divisors of V_{L}. Then

$$\begin{aligned} \mathrm{M}_{\mathsf{L}} &= \mathrm{P}_1^{\mathrm{l}_{1\mathrm{m}_1}} \cdots \mathrm{P}_\mathrm{k}^{\mathrm{l}_{\mathrm{km}_\mathrm{k}}}, \\ \mathrm{C}_{\mathsf{L}} &= \mathrm{P}_1^{\mathrm{l}_{11}} \cdots \mathrm{P}_1^{\mathrm{l}_{1\mathrm{m}_1}} \cdots \mathrm{P}_\mathrm{k}^{\mathrm{l}_{\mathrm{k}1}} \cdots \mathrm{P}_\mathrm{k}^{\mathrm{l}_{\mathrm{km}_\mathrm{k}}}. \end{aligned}$$

Proof Let $\mathsf{V} = \mathsf{U}_{11} \oplus \cdots \oplus \mathsf{U}_{1m_1} \oplus \cdots \oplus \mathsf{U}_{k1} \oplus \cdots \oplus \mathsf{U}_{km_k}$ be the decomposition of Theorem 5.8.33. For $v \in \mathsf{V}$ let us write

$$v = v_{11} + \cdots + v_{1m_1} + \cdots + v_{k1} + \cdots + v_{km_k}$$

as the decomposition of v associated with the direct sum decomposition. For $j \in \{1, \dots, k\}$ and $r \in \{1, \dots, m_j\}$ define $\mathsf{L}_{jr} = \mathsf{L}|\mathsf{U}_{jr}$. Since

$$\mathsf{L} = \mathsf{L}_{11} \oplus \cdots \oplus \mathsf{L}_{1m_1} \oplus \cdots \oplus \mathsf{L}_{k1} \oplus \cdots \oplus \mathsf{L}_{km_k},$$

it follows that $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$ if and only if $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}_{jr}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{U}_{jr})}$ for each $j \in \{1, \dots, k\}$ and $r \in \{1, \dots, m_j\}$. Now let $\tilde{M}_{\mathsf{L}} = P_1^{l_{1m_1}} \cdots P_k^{l_{km_k}}$. Since $P_j^{l_{jr}}$ is an order for the cyclic module U_{jr} and since $l_{jm_j} \geq l_{jr}$ it follows that $\mathrm{Ev}_{\mathsf{F}}(P_j^{l_{jm_j}})(\mathsf{L}_{jr}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{U}_{jr})}$ for each $j \in \{1, \dots, k\}$ and $r \in \{1, \dots, m_j\}$. It, therefore, follows that $\mathrm{Ev}_{\mathsf{F}}(\tilde{M}_{\mathsf{L}})(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Thus $M_{\mathsf{L}}|\tilde{M}_{\mathsf{L}}$ by Proposition 5.8.12.

Now let P be any polynomial for which $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Then $\mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L}_{jr}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{U}_{jr})}$ for each $j \in \{1, \dots, k\}$ and $r \in \{1, \dots, m_j\}$. Therefore, since $P_j^{l_{jr}}$ is an order for U_{jr}, it follows that $P_j^{l_{jr}}|P$. Since this holds for every $j \in \{1, \dots, k\}$ and $r \in \{1, \dots, m_j\}$ it follows that $\tilde{M}_{\mathsf{L}}|P$. In particular, $\tilde{M}_{\mathsf{L}}|M_{\mathsf{L}}$ and so $M_{\mathsf{L}} = \tilde{M}_{\mathsf{L}}$ since both polynomials are monic.

Note that

$$\begin{aligned} C_{\mathsf{L}} = \det \operatorname{diag}(&\xi \boldsymbol{I}_{d_{11}} - \boldsymbol{C}(P_1^{l_{11}}), \dots, \xi \boldsymbol{I}_{d_{1m_1}} - \boldsymbol{C}(P_1^{l_{1m_1}}), \dots, \\ &\xi \boldsymbol{I}_{d_{k1}} - \boldsymbol{C}(P_k^{l_{k1}}), \dots, \xi \boldsymbol{I}_{d_{km_k}} - \boldsymbol{C}(P_1^{l_{km_k}})), \end{aligned}$$

where $d_{jr} = \deg(P_j^{l_{jr}})$, $j \in \{1, \dots, k\}$, $r \in \{1, \dots, m_j\}$. The determinant of a block diagonal matrix is the product of the determinants of the blocks by Exercise 5.3.8. That C_{L} is the product of the elementary divisors then follows from Proposition 5.8.26. ■

As a consequence of this characterisation we have the following rather important result.

5.8.36 Corollary (Cayley–Hamilton Theorem again) *If* F *is a field, if* V *is a finite-dimensional* F*-vector space, and if* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*, then* $\mathrm{Ev}_{\mathsf{F}}(\mathsf{C}_{\mathsf{L}})(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$*, i.e.,* L *satisfies its own characteristic polynomial.*

Let us illustrate the rational canonical form with an example.

5.8.37 Example (Rational canonical form) We consider L and L' as given in Example 5.8.16:

$$\mathsf{L} = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}, \qquad \mathsf{L}' = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

Let us first define subspaces

$$\begin{aligned}
\mathsf{U}_{11} &= \operatorname{span}_{\mathsf{F}}((1_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}}),(0_{\mathsf{F}},1_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}})),\\
\mathsf{U}_{12} &= \operatorname{span}_{\mathsf{F}}((0_{\mathsf{F}},0_{\mathsf{F}},1_{\mathsf{F}},0_{\mathsf{F}}),(0_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}},1_{\mathsf{F}})),\\
\mathsf{U}'_{11} &= \operatorname{span}_{\mathsf{F}}((1_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}})),\\
\mathsf{U}'_{12} &= \operatorname{span}_{\mathsf{F}}((0_{\mathsf{F}},1_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}})),\\
\mathsf{U}'_{13} &= \operatorname{span}_{\mathsf{F}}((0_{\mathsf{F}},0_{\mathsf{F}},1_{\mathsf{F}},0_{\mathsf{F}}),(0_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}},1_{\mathsf{F}})),
\end{aligned}$$

Noting that

$$\mathsf{V} = \mathsf{U}_{11} \oplus \mathsf{U}_{12} = \mathsf{U}_{11} \oplus \mathsf{U}_{12} \oplus \mathsf{U}_{13}.$$

One can now directly check that the endomorphisms

$$\mathsf{L}|\mathsf{U}_{11},\ \mathsf{L}|\mathsf{U}_{12},\ \mathsf{L}'|\mathsf{U}'_{11},\ \mathsf{L}'|\mathsf{U}'_{12},\ \mathsf{L}|\mathsf{U}'_{13}$$

are cyclic. Moreover, the characteristic polynomials of these endomorphisms are

$$(\xi-1_{\mathsf{F}})^2,\ (\xi-1_{\mathsf{F}})^2,\ \xi-1_{\mathsf{F}}, \xi-1_{\mathsf{F}},\ (\xi-1_{\mathsf{F}})^2,$$

respectively. Therefore, these are the elementary divisors for L and L' are the multisets $\{(\xi-1_{\mathsf{F}})^2,(\xi-1_{\mathsf{F}})^2\}$ and $\{\xi-1_{\mathsf{F}},\xi-1_{\mathsf{F}},(\xi-1_{\mathsf{F}})^2\}$, respectively. In particular, this gives the minimal and characteristic polynomials as

$$M_{\mathsf{L}_1} = M_{\mathsf{L}_2} = (\xi-1_{\mathsf{F}})^2, \qquad P_{\mathsf{L}_1} = P_{\mathsf{L}_2} = (\xi-1_{\mathsf{F}})^4,$$

just as we had previously written down.

The rational canonical forms for L and L' are

$$\left[\begin{array}{cc|cc}
0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}}\\
-1_{\mathsf{F}} & 2_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}}\\
\hline
0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}}\\
0_{\mathsf{F}} & 0_{\mathsf{F}} & -1_{\mathsf{F}} & 2_{\mathsf{F}}
\end{array}\right], \qquad
\left[\begin{array}{c|c|cc}
1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}}\\
\hline
0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}}\\
\hline
0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}}\\
0_{\mathsf{F}} & 0_{\mathsf{F}} & -1_{\mathsf{F}} & 2_{\mathsf{F}}
\end{array}\right]$$

respectively. •

5.8.8 The invariant factor canonical form

The rational canonical form of Theorem 5.8.33 corresponds to the primary-cyclic of the $\mathsf{F}[\xi]$-module V_{L}. One also has a canonical form associated with the invariant factor decomposition. This decomposition is not as often used as the rational canonical form, but let us give it nonetheless. It will be helpful in our characterisation of similar endomorphisms over the algebraic closure of a field.

We merely state and prove the theorem.

5.8.38 Theorem (Invariant factor canonical form) *Let* F *be a field, let* V *be an* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{L}}(\mathsf{V})$*. Then there exists monic polynomials* $\mathrm{Q}_1, \dots, \mathrm{Q}_m \in \mathsf{F}[\xi]$ *and* L*-invariant subspaces* $\mathsf{W}_1, \dots, \mathsf{W}_m$ *such that*

(i) $\mathrm{Q}_j | \mathrm{Q}_{j+1}$, $j \in \{1, \dots, m-1\}$,

(ii) $\mathsf{V} = \mathsf{W}_1 \oplus \cdots \oplus \mathsf{W}_k$, *and*

(iii) the minimal polynomial of $\mathsf{L}|\mathsf{W}_j$ *is* Q_j, $j \in \{1, \dots, k\}$.

Moreover, the polynomials $\mathrm{Q}_1, \dots, \mathrm{Q}_m$ *are uniquely determined by* L *and by the above conditions.*

Proof The existence of the polynomials and the L-invariant subspaces follows from Theorem 4.9.21 and Proposition 5.8.7. The final uniqueness assertion also follows from Theorem 4.9.21. It remains to show that the minimal polynomial of $\mathsf{L}|\mathsf{W}_j$ is Q_j. To see this, we recall from the proof of Theorem 4.9.21 the manner in which one constructs the polynomials $Q_1, \dots, Q_m$. We let $P_1, \dots, P_k$ be the distinct, monic, irreducible polynomials of Theorem 5.8.33 which give the elementary divisors as

$$P_1^{l_1}, \dots, P_1^{l_{1m_1}}, \dots, P_k^{l_{k1}}, \dots, P_k^{l_{km_k}}.$$

Then the polynomials $Q_1, \dots, Q_m$ are of the form

$$Q_j = P_1^{\alpha_1} \cdots P_k^{\alpha_k}, \qquad j \in \{1, \dots, m\},$$

where $\alpha_s \in \{l_{s1}, \dots, l_{sm_s}\}$ for $s \in \{1, \dots, k\}$. This gives

$$\mathsf{W}_j = \ker(\mathrm{Ev}_{\mathsf{F}}(Q_j)(\mathsf{L})) = \ker(\mathrm{Ev}_{\mathsf{F}}(P_1^{\alpha_1})(\mathsf{L})) \oplus \cdots \oplus \ker(\mathrm{Ev}_{\mathsf{F}}(P_k^{\alpha_k})(\mathsf{L})).$$

The characteristic polynomial of $\mathsf{L}|\mathsf{W}_j$ is thus Q_j. Moreover, since the polynomials $P_1^{\alpha_1}, \dots, P_k^{\alpha_k}$ are coprime, the subspace W_j is L-cyclic. Therefore, by Proposition 5.8.27 it follows that the minimal polynomial of $\mathsf{L}|\mathsf{W}_j$ is Q_j. ■

Note that the polynomials $Q_1, \dots, Q_m$ are not generally irreducible. It is for this reason that the invariant factor canonical form is not as useful as the rational canonical form. However, it is also this feature of the invariant canonical form that will allow us to prove Theorem 5.8.40 below.

Let us give an example of the invariant factor canonical form.

5.8.39 Example (Invariant factor canonical form) We consider a field F, take $\mathsf{V} = \mathsf{F}^4$, and define two endomorphisms by the matrices

$$\mathsf{L} = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}, \qquad \mathsf{L}' = \begin{bmatrix} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{bmatrix}.$$

In Example 5.8.37 we determined that the elementary divisors of L and L' are the multisets $\{(\xi - 1_{\mathsf{F}})^2, (\xi - 1_{\mathsf{F}})^2\}$ and $\{\xi - 1_{\mathsf{F}}, \xi - 1_{\mathsf{F}}, (\xi - 1_{\mathsf{F}})^2\}$, respectively. Therefore, the invariant factors for L are the multiset

$$\{(\xi - 1_{\mathsf{F}})^2, (\xi - 1_{\mathsf{F}})^2\}$$

and the invariant factors for L' are the multiset

$$\{\xi - 1_{\mathsf{F}}, (\xi - 1_{\mathsf{F}})(\xi - 1_{\mathsf{F}})^2\}.$$

(These are determined using the rules given in the proof of Theorem 4.9.21 and illustrated in Example 4.9.23.) Thus the invariant factor canonical form is formed by a block diagonal matrix with the blocks having characteristic polynomials given by the invariant factors. In this example, because it is somewhat degenerate having all elementary divisors being a power of the same irreducible polynomial, the invariant factor canonical form is the same as the rational canonical form. •

5.8.9 The rôle of field extensions

The development thus far in this section, culminating in the rational canonical form for an endomorphism, is valuable in that it is general: it is valid for any endomorphism of any finite-dimensional vector space, and for arbitrary fields. However, one might hope that if the field or if the endomorphism has particular properties, one may be able to say more about the structure of the equivalence classes under similarity. This is indeed the case, and the most commonly held structure is that the minimal polynomial (or, equivalently characteristic polynomial) of the endomorphism splits, i.e., it is a product of polynomials of degree 1. This will happen, for example, if the field is algebraically closed. Therefore, any additional information one obtains for the case when the minimal polynomial splits will hold, in particular, for endomorphisms of $\mathbb{C}$-vector spaces. However, for endomorphisms of, say, $\mathbb{R}$-vector spaces it might not happen that the minimal polynomial splits in $\mathbb{R}[\xi]$, even though it splits in $\mathbb{C}[\xi]$. Moreover, it is useful to not necessarily consider algebraically closed fields, but simply fields which contain the eigenvalues for the endomorphism. The point is this: It is useful to think of an endomorphism of an F-vector space as being an endomorphism of an K-vector space where K is a field extension of F (see Section 4.6 for details about field extensions).

To allow for endomorphisms to be defined over a field extension, we must use vector spaces over the extended field and endomorphisms of these extended vector spaces. This is discussed in Sections 4.5.7 and 5.4.10 for the extension from $\mathbb{R}$ to $\mathbb{C}$. For general extensions, using the notion of tensor product, the reader should refer to Sections 4.5.8 and 5.4.11. Since the extension from $\mathbb{R}$ to $\mathbb{C}$ will be the most useful extension for us, the reader may very well want to restrict in their mind to this case. This should be generally possible without too much difficulty.

Our main interest lies in the following theorem.

5.8.40 Theorem (Independence of similarity with respect to extension) *Let F be a field, let V be an F-vector space, and let $\mathsf{L}_1, \mathsf{L}_2 \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Also let K be an extension of F, and let V_{K}, and $\mathsf{L}_{\mathsf{K},1}, \mathsf{L}_{\mathsf{K},2} \in \mathrm{End}_{\mathsf{K}}(\mathsf{V}_{\mathsf{K}})$ be extensions to K. Then L_1 and L_2 are similar if and only if $\mathsf{L}_{\mathsf{F},1}$ and $\mathsf{L}_{\mathsf{F},2}$ are similar.*

Proof We let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ and we will show that the invariant factors of L uniquely determine the elementary divisors of L_{K}. This will give the result by Theorem 5.8.38.

Let us first consider a few F-modules.

1. We denote by $(\mathsf{V}_{\mathsf{K}})_{\mathsf{L}_{\mathsf{K}}}$ the $\mathsf{K}[\xi]$ module associated with the endomorphism $\mathsf{L}_{\mathsf{K}} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V}_{\mathsf{F}})$.
2. Note that V_{L} is an $\mathsf{F}[\xi]$-module with scalar multiplication given by

$$(P \cdot v)(j) = a_k \mathsf{L}^k(j,l)v(l) + \cdots + a_1\mathsf{L}(j,l)v(l) + a_0 v(l),$$

where

$$P = a_k\xi^k + \cdots + a_1\xi + a_0 \in \mathsf{F}[\xi]$$

(here we are using the fact that $\mathsf{V} = \mathsf{F}^n$). This $\mathsf{F}[\xi]$-module structure can be extended to a $\mathsf{K}[\xi]$-module structure by defining the product by

$$(\bar{P} \cdot v)(j) = \bar{a}_k \mathsf{L}^k(j,l)v(l) + \cdots + \bar{a}_1\mathsf{L}(j,l)v(l) + \bar{a}_0 v(l),$$

where

$$\bar{P} = \bar{a}_k\xi^k + \cdots + \bar{a}_1\xi + \bar{a}_0 \in \mathsf{K}[\xi].$$

Let us denote this $\mathsf{K}[\xi]$-module by $(\mathsf{V}_{\mathsf{L}})_{\mathsf{K}[\xi]}$.

It is clear that the $\mathsf{K}[\xi]$-modules $(\mathsf{V}_{\mathsf{K}})_{\mathsf{L}_{\mathsf{K}}}$ and $(\mathsf{V}_{\mathsf{L}})_{\mathsf{K}[\xi]}$ are isomorphic; they are, in fact, equal.

Let us denote by $\bar{Q}_1, \dots, \bar{Q}_m$ the inclusions of $Q_1, \dots, Q_m$ in $\mathsf{F}[\xi]$. Note that $\bar{Q}_j|\bar{Q}_{j+1}$ for $k \in \{1, \dots, m-1\}$. Let $\mathsf{W}_1, \dots, \mathsf{W}_m \subseteq \mathsf{V}$ be the L-invariant subspaces corresponding to $Q_1, \dots, Q_m$. Let $\bar{\mathsf{W}}_1, \dots, \bar{\mathsf{W}}_m \subseteq \mathsf{V}_{\mathsf{K}}$ be the L_{K}-invariant subspaces corresponding to $\bar{Q}_1, \dots, \bar{Q}_m$. Note that the subspaces $\bar{\mathsf{W}}_j$, $j \in \{1, \dots, m\}$, are L_{F}-cyclic for the same reason (see the proof of Theorem 5.8.38) that the subspaces W_j, $j \in \{1, \dots, m\}$, are L-cyclic. Moreover, from the definition of L_{F} it follows that the characteristic polynomial of $\mathsf{L}_{\mathsf{F}}|\bar{\mathsf{W}}_j$ is $\bar{Q}_j$. Thus $\bar{Q}_1, \dots, \bar{Q}_m$ are the invariant factors for $(\mathsf{V}_{\mathsf{F}})_{\mathsf{L}_{\mathsf{F}}}$. Thus the invariant factors for L_{F} are uniquely determined by those for L. ■

5.8.10 From the rational canonical form to the Jordan canonical form

With the preceding section as backdrop, in this section we consider the implications of assuming that the minimal (or equivalently, characteristic) polynomial splits, i.e., is a product of factors of degree one. We do this in this section by proceeding directly from the rational canonical form. In the next sections we shall explore this situation in a manner that does not rely explicitly on the rational canonical form.

Let us look at the easiest situation first, that where V is L-cyclic, and the minimal polynomial splits, having a single root of multiplicity $\dim_{\mathsf{F}}(\mathsf{V})$.

5.8.41 Proposition (Cyclic vector spaces and Jordan blocks) *Let F be a field and let V be a finite-dimensional F-vector space. Suppose that $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ has the following properties:*

(i) V is L-cyclic;

(ii) $\mathrm{M}_{\mathsf{L}} = (\xi - \lambda)^{\mathrm{k}}$ for $\lambda \in \mathsf{F}$ and where $\mathrm{k} = \dim_{\mathsf{F}}(\mathsf{V})$.

Let $\mathsf{L}_\lambda = \lambda\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}$. Then the following statements hold:

(iii) $\dim_{\mathsf{F}}(\ker(\mathsf{L}_\lambda)) = 1$;

(iv) *if* v_0 *generates the* $\mathsf{F}[\xi]$*-module* V_{L} *then*

$$\mathscr{B} = \{\mathrm{v}_0, (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})(\mathrm{v}_0), \dots, (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{\mathrm{k}-1}(\mathrm{v}_0)\}$$

is a basis for V;

(v) *we have*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \begin{bmatrix} \lambda & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 1_{\mathsf{F}} & \lambda & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & \lambda & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 1_{\mathsf{F}} & \lambda \end{bmatrix}.$$

Proof (iii) Note that λ is an eigenvalue for L by Proposition 5.8.13. Thus $\dim_{\mathsf{F}}(\ker(\mathsf{L}_\lambda)) \geq 1$. If $\dim_{\mathsf{F}}(\mathsf{V}) = 1$ then this part of the result follows, so we suppose that $\dim_{\mathsf{F}}(\mathsf{V}) \geq 2$. Let $\mathsf{W}(\lambda, \mathsf{L}) = \ker(\mathsf{L}_\lambda)$ and suppose that $\dim_{\mathsf{F}}(\mathsf{W}(\lambda, \mathsf{L})) \geq 2$. Then there are linearly independent vectors $\{v_1, v_2\}$ such that $\mathsf{U}_a \triangleq \mathrm{span}_{\mathsf{F}}(v_a)$, $a \in \{1, 2\}$, are L-invariant subspaces for which the characteristic polynomial of $\mathsf{L}|\mathsf{U}_a$ is $\xi - \lambda$. Let $\mathsf{U} = \mathsf{U}_1 \oplus \mathsf{U}_2$, noting that this subspace is L-invariant. Thus U is a submodule of V_{L} and so cyclic by Proposition 4.9.9(i). However, the elementary divisors of $\mathsf{L}|\mathsf{U}$ are the multiset $\{\xi - \lambda, \xi - \lambda\}$. This contradicts $\mathsf{L}|\mathsf{U}$ being cyclic by Proposition 5.8.35.

(iv) From Theorem 5.8.20 we know that

$$\mathscr{B}' = \{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{k-1}(v_0)\}$$

is a basis for V. Denote $\mathscr{B} = \{e_1, \dots, e_k\}$ and $\mathscr{B}' = \{e'_1, \dots, e'_k\}$. Using the Binomial Theorem in the form of Proposition 4.2.11, and using the fact that L and $\lambda\,\mathrm{id}_{\mathsf{V}}$ commute, we have

$$(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^m = \sum_{j=0}^{m} B_{m,j}(-1)^j \lambda^j \mathsf{L}^{m-j}.$$

A direct computation then shows that

$$e_j = \sum_{l=1}^{j} \boldsymbol{P}(l, j) e'_l$$

where $\boldsymbol{P}$ is lower triangular with 1_{F}'s on the diagonal. Thus this matrix has determinant 1_{F} by Proposition 5.3.3(iv), and so is invertible. Thus $\mathscr{B}$ is a basis by (iv).

(v) We have

$$\begin{aligned} \mathsf{L}(e_j) &= \mathsf{L} \circ (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{j-1}(v_0) \\ &= (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}} + \lambda\,\mathrm{id}_{\mathsf{V}}) \circ (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{j-1}(v_0) \\ &= (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{j}(v_0) + \lambda(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{j-1}(v_0) \\ &= e_{j+1} + \lambda e_j, \end{aligned}$$

for $j \in \{1, \dots, k-1\}$. Using a similar computation we have

$$\mathsf{L}(e_k) = (\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^k(v_0) + \lambda(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^{k-1}(v_0) = \lambda e_k$$

since $(\mathsf{L} - \lambda\,\mathrm{id}_{\mathsf{V}})^k = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. Putting this all together gives

$$\begin{aligned}
\mathsf{L}(e_1) &= \lambda e_1 + 1_{\mathsf{F}}e_2 + 0_{\mathsf{F}}e_3 + \dots + 0_{\mathsf{F}}e_{k-1} + 0_{\mathsf{F}}e_k,\\
\mathsf{L}(e_2) &= 0_{\mathsf{F}}e_1 + \lambda e_2 + 1_{\mathsf{F}}e_3 + \dots + 0_{\mathsf{F}}e_{k-1} + 0_{\mathsf{F}}e_k,\\
&\vdots\\
\mathsf{L}(e_{k-1}) &= 0_{\mathsf{F}}e_1 + 0_{\mathsf{F}}e_2 + 0_{\mathsf{F}}e_3 + \dots + \lambda e_{k-1} + 0_{\mathsf{F}}e_k,\\
\mathsf{L}(e_k) &= 0_{\mathsf{F}}e_1 + 0_{\mathsf{F}}e_2 + 0_{\mathsf{F}}e_3 + \dots + 1_{\mathsf{F}}e_{k-1} + \lambda e_k,
\end{aligned}$$

which gives the result by definition of the matrix representative. ■

As we did with the companion matrix associated with a L-cyclic vector space, we shall use an alternative matrix representation for the situation represented by the previous result. We describe this as follows.

5.8.42 Proposition (An alternative matrix representation) *Let* F *be a field and let* V *be a finite-dimensional* F*-vector space. Suppose that* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *has the following properties:*

(i) V *is* L*-cyclic;*

(ii) $M_{\mathsf{L}} = (\xi - \lambda)^{\mathrm{k}}$ *for* $\lambda \in \mathsf{F}$ *and where* $\mathrm{k} = \dim_{\mathsf{F}}(\mathsf{V})$*.*

Then there exists a basis $\mathscr{B}$ *for* V *such that*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \begin{bmatrix} \lambda & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & \lambda & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & \lambda & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & \lambda \end{bmatrix}.$$

Proof Let $\mathscr{B}' = \{e_1', \dots, e_k'\}$ be the basis from Proposition 5.8.41 and let $\mathscr{B} = \{e_1, \dots, e_k\}$ with $e_j = e_{k-j+1}'$, $j \in \{1, \dots, k\}$. Then

$$\mathsf{L}(e_1) = \mathsf{L}(e_k') = \lambda e_k' = \lambda e_1$$

and

$$\mathsf{L}(e_j) = \mathsf{L}(e_{k-j+1}') = e_{k-j+2}' + \lambda e_{k-j+1}' = e_{j-1} + \lambda e_j, \qquad j \in \{2, \dots, k\}.$$

It now follows from an application of the definition of the matrix representative that $\mathscr{B}$ gives the desired form for $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$. ■

The matrix representative in the preceding result is important enough to have a name.

5.8.43 Definition (Jordan block, Jordan arrangement) Let F be a field and let $\lambda \in \mathsf{F}$.

(i) For $k \in \mathbb{Z}_{>0}$, the matrix

$$\boldsymbol{J}(\lambda, k) \triangleq \begin{bmatrix} \lambda & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & \lambda & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & \lambda & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & \lambda \end{bmatrix}$$

is the ***Jordan block*** associated with k and λ.

(ii) For $r \in \mathbb{Z}_{>0}$ and $\boldsymbol{k} = (k_1, \dots, k_r) \in \mathbb{Z}^r_{>0}$, the matrix

$$\boldsymbol{J}(\ell, \boldsymbol{k}) = \begin{bmatrix} \boldsymbol{J}(\ell, k_1) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}(\ell, k_2) & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{J}(\ell, k_r) \end{bmatrix}$$

is the ***Jordan arrangement*** associated with $\boldsymbol{k}$ and λ. •

One can now adapt the preceding discussion to each of the invariant subspaces associated with the elementary divisors of L and arrive at the following theorem.

5.8.44 Theorem (Jordan canonical form) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. The following statements are equivalent:*

(i) M_{L} *(or, equivalently,* C_{L}*) splits in* F*;*

(ii) there exists

(a) $\mathrm{k} \in \mathbb{Z}_{>0}$,

(b) distinct $\lambda_{\mathrm{j}} \in \mathsf{F}$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$,

(c) $\mathrm{p_j} \in \mathbb{Z}_{>0}$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$, *and*

(d) $\mathbf{l}_{\mathrm{j}} \in \mathbb{Z}^{\mathrm{p_j}}_{>0}$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$,

such that the elementary divisors of L *are the multiset*

$$\{(\xi - \lambda_1)^{\mathrm{l}_{11}}, \dots, (\xi - \lambda_1)^{\mathrm{l}_{1\mathrm{m}_1}}, \dots, (\xi - \lambda_{\mathrm{k}})^{\mathrm{l}_{\mathrm{k}1}}, \dots, (\xi - \lambda_{\mathrm{k}})^{\mathrm{l}_{\mathrm{km_k}}}\}.$$

Moreover, if either statement holds then there exists a basis $\mathscr{B}$ for V *such that*

$$[\mathsf{L}]^{\mathscr{B}}_{\mathscr{B}} = \mathrm{diag}(\mathbf{J}(\lambda_1, \mathbf{l}_1), \dots, \mathbf{J}(\lambda_{\mathrm{k}}, \mathbf{l}_{\mathrm{k}}), \dots, \mathbf{J}(\lambda_{\mathrm{k}}, \mathrm{l}_{\mathrm{km_k}})).$$

Finally, if

$$[\mathsf{L}]^{\mathscr{B}'}_{\mathscr{B}'} = \mathrm{diag}(\mathbf{J}(\mu_1, \mathrm{r}_{11}), \dots, \mathbf{J}(\mu_1, \mathrm{r}_{1\mathrm{n}_1}), \dots, \mathbf{J}(\mu_{\mathrm{s}}, \mathrm{r}_{\mathrm{p}1}), \dots, \mathbf{J}(\mu_{\mathrm{p}}, \mathrm{r}_{\mathrm{pn_p}}))$$

is a matrix representative of L *in another basis, then* $\mathrm{p} = \mathrm{k}$ *and there exists a permutation* $\sigma \in \mathfrak{S}_{\mathrm{k}}$ *such that* $\mu_{\mathrm{j}} = \lambda_\sigma(\mathrm{j})$, $\mathrm{n_j} = \mathrm{m}_{\sigma(\mathrm{j})}$, *and* $\mathrm{r_{ja}} = \mathrm{l}_{\sigma(\mathrm{j})\mathrm{a}}$*for* $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$ *and* $\mathrm{a} \in \{1, \dots, \mathrm{m_j}\}$.

Proof Suppose that M_{L} splits. Let the elementary divisors for L be the multiset

$$\{P_1^{l_{11}}, \dots, P_1^{l_{1m_1}}, \dots, P_k^{l_{k1}}, \dots, P_k^{l_{km_k}}\}.$$

Since $M_{\mathsf{L}} = P_1^{l_{1m_1}} \cdots P_1^{l_{km_k}}$ by Proposition 5.8.35 and since M_{L} splits, it follows that each of the polynomials $P_1, \dots, P_k$ must split.

The second statement obviously implies the first since M_{L} is a product of powers of $(\xi - \lambda_1), \dots, (\xi - \lambda_k)$. This gives the fist part of the theorem.

Now let

$$\mathsf{V} = \mathsf{U}_{11} \oplus \cdots \oplus \mathsf{U}_{1m_1} \oplus \cdots \oplus \mathsf{U}_{k1} \oplus \cdots \oplus \mathsf{U}_{km_k}$$

be the decomposition of V into L-invariant subspaces as per Theorem 5.8.33. By Proposition 5.8.42 and since $\mathsf{L}|\mathsf{U}_{jr}$ is cyclic, there exists a basis $\mathscr{B}_{jr}$ for U_{jr} such that $[\mathsf{L}|\mathsf{U}_{jr}]_{\mathscr{B}_{jr}}^{\mathscr{B}_{jr}} = \boldsymbol{J}(\lambda_j, l_{jr})$ for each $j \in \{1, \dots, k\}$ and $r \in \{1, \dots, m_j\}$. Taking

$$\mathscr{B} = \mathscr{B}_{11} \cup \cdots \cup \mathscr{B}_{1m_1} \cup \cdots \cup \mathscr{B}_{k1} \cup \cdots \cup \mathscr{B}_{km_k}$$

gives the desired matrix representative.

The uniqueness assertion follows from the uniqueness part of Theorem 5.8.33. ■

The matrix representative of the theorem has a name.

5.8.45 Definition (Jordan canonical form) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ have a minimal polynomial that splits. The matrix representative of Theorem 5.8.44 is the ***Jordan canonical form*** for L. •

Let us consider our ongoing example in terms of the Jordan canonical form.

5.8.46 Example (Jordan canonical form) We consider L and L' as given in Example 5.8.16:

$$\mathsf{L} = \left[\begin{array}{cc|cc} 1_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \hline 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{array}\right], \qquad \mathsf{L}' = \left[\begin{array}{c|c|cc} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \hline 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \hline 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \end{array}\right].$$

These matrices are already in Jordan canonical form, and we have shown them divided up so as to illustrate the Jordan blocks in each case. •

5.8.11 Diagonalisation

The next few sections will be devoted to understanding the additional conclusions one can come to in the important special case when the minimal polynomial splits.

We begin with two entirely similar definitions.

5.8.47 Definition (Diagonalisable matrix) Let F be a field and let $n \in \mathbb{Z}_{>0}$. A matrix $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$ is ***diagonalisable*** if it is similar to a diagonal matrix. •

5.8.48 Definition (Diagonalisable endomorphism) Let F be a field and let V be a finite-dimensional F-vector space. An endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is ***diagonalisable*** if there exists a basis $\mathscr{B}$ for V such that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is diagonal. •

The next result follows immediately from the two definitions of diagonalisability and similarity.

5.8.49 Proposition (Diagonalisable matrices and endomorphisms) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathscr{B}$ be a basis for V. Then $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is diagonalisable if and only if $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is diagonalisable.*

From the preceding result it follows that we can simply talk about diagonalisable endomorphisms since diagonalisable matrices are a special case of this.

The following result gives a precise characterisation of diagonalisability.

5.8.50 Proposition (Characterisation of diagonalisability) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{L}}(\mathsf{V})$. Then the following statements are equivalent:*

(i) L is diagonalisable;

(ii) there exists a basis of V comprised of eigenvectors of L;

(iii) the minimal polynomial of L has the form

$$\mathrm{M}_{\mathsf{L}} = (\xi - \lambda_1)\cdots(\xi - \lambda_{\mathrm{k}})$$

for distinct $\lambda_1, \dots, \lambda_{\mathrm{k}} \in \mathsf{F}$.

Proof (i) $\Longrightarrow$ (ii) Let $\mathscr{B} = \{e_1, \dots, e_n\}$ be a basis such that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is diagonal. Then, by definition of the matrix representative, for each $j \in \{1, \dots, n\}$ we have $\mathsf{L}(e_j) = \lambda_j e_j$ for some $\lambda_j \in \mathsf{F}$. Thus e_j is an eigenvector with eigenvalue λ_j.

(ii) $\Longrightarrow$ (iii) Let $\mathscr{B} = \{e_1, \dots, e_n\}$ be a basis of eigenvectors and define $\lambda_1, \dots, \lambda_n \in \mathsf{F}$ (not necessarily distinct) by $\mathsf{L}(e_j) = \lambda_j e_j$, $j \in \{1, \dots, n\}$. Let us suppose, without loss of generality, let $\lambda_1, \dots, \lambda_k$ be the distinct elements of the multiset $\{\lambda_1, \dots, \lambda_n\}$. Define

$$\tilde{M}_{\mathsf{L}} = (\xi - \lambda_1)\cdots(\xi - \lambda_k).$$

Since $\lambda_1, \dots, \lambda_k$ are eigenvalues, by Proposition 5.8.13 we have $\tilde{M}_{\mathsf{L}} | M_{\mathsf{L}}$. For $v \in \mathsf{V}$ write

$$v = v_1 e_1 + \cdots + v_n e_n$$

be the representation of v in the basis $\mathscr{B}$. For $j \in \{1, \dots, n\}$ let $l(j) \in \{1, \dots, k\}$ be uniquely defined so that $\lambda_j = \lambda_{l(j)}$. Then

$$\begin{aligned}\mathrm{Ev}_{\mathsf{F}}(\tilde{M}_{\mathsf{L}})(\mathsf{L}) \cdot v &= (\lambda_1 \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) \circ \cdots \circ (\lambda_k \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})(v_1 e_1 + \cdots + v_n e_n) \\ &= \sum_{j=1}^{n} \left(\prod_{l \neq l(j)} (\lambda_l \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) \right) (\lambda_{l(j)} \,\mathrm{id}_{\mathsf{V}} - \mathsf{L})(v_j e_j) = 0_{\mathsf{V}},\end{aligned}$$

using the fact that the endomorphisms $\lambda_j \,\mathrm{id}_{\mathsf{V}} - \mathsf{L}$, $j \in \{1, \dots, k\}$, commute, and using the fact that e_j is an eigenvector for the eigenvalue $\lambda_{l(j)}$. This shows that $M_{\mathsf{L}} | \tilde{M}_{\mathsf{L}}$ by definition of the minimal polynomial. Therefore, $M_{\mathsf{L}} = \tilde{M}_{\mathsf{L}}$ since both polynomials are monic.

(iii) $\implies$ (i) Let $M_{\mathsf{L}} = (\xi - \lambda_1)\cdots(\xi-\lambda_k)$ for $\lambda_1,\dots,\lambda_k \in \mathsf{F}$ distinct. By Theorem 5.8.29 we have

$$\mathsf{V} = \ker(\lambda_1 \operatorname{id}_{\mathsf{V}} - \mathsf{L}) \oplus \cdots \oplus \ker(\lambda_k \operatorname{id}_{\mathsf{V}} - \mathsf{L}).$$

Since $\ker(\lambda_j - \mathsf{L})$ is the eigenspace for the eigenvalue λ_j, $j \in \{1,\dots,k\}$, it follows that if $\mathscr{B}_j = \{e_{j1},\dots,e_{jm_j}\}$ is a basis for $\ker(\lambda_j \operatorname{id}_{\mathsf{V}} - \mathsf{L})$, then $\mathsf{L}(e_{jl}) = \lambda_j e_{jl}$ for $j \in \{1,\dots,k\}$, $l \in \{1,\dots,m_j\}$, where $m_j = \dim_{\mathsf{F}}(\ker(\lambda_j \operatorname{id} \mathsf{V} - \mathsf{L}))$. Therefore, if $\mathscr{B} = \mathscr{B}_1 \cup \cdots \cup \mathscr{B}_k$, we have that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is diagonal. ■

An example illustrates the proposition.

5.8.51 Example (Diagonalisable endomorphisms) We let $\mathsf{V} = \mathsf{F}^3$ and consider endomorphisms

$$\mathsf{L} = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 2_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} \end{bmatrix}, \qquad \mathsf{L}' = \begin{bmatrix} 1_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 2_{\mathsf{F}} \\ 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}}. \end{bmatrix}$$

One can verify that L and L' both have as eigenvalues the multiset $\{1_{\mathsf{F}}, 1_{\mathsf{F}}, 2_{\mathsf{F}}\}$. We also have

$$\begin{aligned} \ker(1_{\mathsf{F}} \operatorname{id}_{\mathsf{V}} - \mathsf{L}) &= \operatorname{span}_{\mathsf{F}}((1_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}}),(0_{\mathsf{F}},0_{\mathsf{F}},1_{\mathsf{F}})), \\ \ker(2_{\mathsf{F}} \operatorname{id}_{\mathsf{V}} - \mathsf{L}) &= \operatorname{span}_{\mathsf{F}}((0_{\mathsf{F}},1_{\mathsf{F}},0_{\mathsf{F}})), \\ \ker(1_{\mathsf{F}} \operatorname{id}_{\mathsf{V}} - \mathsf{L}') &= \operatorname{span}_{\mathsf{F}}((1_{\mathsf{F}},0_{\mathsf{F}},0_{\mathsf{F}})), \\ \ker(2_{\mathsf{F}} \operatorname{id}_{\mathsf{V}} - \mathsf{L}') &= \operatorname{span}_{\mathsf{F}}((0_{\mathsf{F}},1_{\mathsf{F}},0_{\mathsf{F}})). \end{aligned}$$

In particular, L has a basis of eigenvectors whereas L' does not. Thus L is diagonalisable while L' is not. The minimal polynomials are

$$M_{\mathsf{L}} = (\xi - 1_{\mathsf{F}})(\xi - 2_{\mathsf{F}}), \qquad M_{\mathsf{L}'} = (\xi - 1_{\mathsf{F}})^2(\xi - 2_{\mathsf{F}}),$$

which also bears out the characterisation of diagonalisable endomorphisms in Proposition 5.8.50. •

5.8.12 Semisimple and absolutely semisimple endomorphisms

Since one can reasonably demand diagonalisability only when the field of scalars is algebraically closed, it is useful to have a comparable notion for fields that are not algebraically closed. This notion turns out to be slightly complicated as it is related to the character of the field F in a nontrivial way. This forces one to consider two flavours of ideas related to diagonalisation: semisimple and absolutely semisimple. As we shall see, the latter implies the former. Moreover, the two notions are equivalent for fields of characteristic zero, e.g., $\mathbb{R}$. Thus the reader looking to simplify their life can equate "semisimple" and "absolutely semisimple."

Let us first consider the notion of a semisimple endomorphism. Its characterisation has a rather geometric flavour.

5.8.52 Definition (Semisimple endomorphism) Let F be field and let V be an F-vector space. An endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is ***semisimple*** if, for every L-invariant subspace $\mathsf{U} \subseteq \mathsf{V}$, there exists an L-invariant complement to U in V. •

The following result explains the significance of semisimple endomorphisms in terms of the minimal polynomial.

5.8.53 Theorem (The minimal polynomial of a semisimple endomorphism) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Let $\mathrm{P}_1, \dots, \mathrm{P}_k \in \mathsf{F}[\xi]$ be distinct, monic, irreducible polynomials and let $\mathrm{l}_1, \dots, \mathrm{l}_k \in \mathbb{Z}_{>0}$ be such that*

$$\mathrm{M}_{\mathsf{L}} = \mathrm{P}_1^{\mathrm{l}_1} \cdots \mathrm{P}_k^{\mathrm{l}_k}.$$

Then L is semisimple if and only if $\mathrm{l}_1 = \cdots = \mathrm{l}_k = 1$.

Proof First note that, by Proposition 5.8.7, L is semisimple if and only if every submodule of V_{L} possesses a complementary submodule. We shall thus deal with the $\mathsf{F}[\xi]$-module V_{L}, and shall borrow the notation of Theorem 4.9.16.

Suppose that $l_1, \dots, l_k = 1$. By Theorem 5.8.29, to show that every submodule of V_{L} possesses a complement, it suffices to show that every submodule of a primary submodule M of V_{L} has a complement in M. That is to say, we can assume without loss of generality that $M_{\mathsf{L}} = P_1$ and so $\mathsf{V}_{\mathsf{L}} = \mathsf{V}_{\mathsf{L}}(P_1) = \mathsf{V}_{\mathsf{L}}[P_1]$. In this case, Lemma 3 of Theorem 4.9.16 gives V_{L} as a vector space over the field $\mathsf{F}[\xi]/(P_1)$. Submodule of V_{L} are then precisely subspaces of V_{L}, and since subspaces possess complements by Theorem 4.5.52, it follows that every submodule of V_{L} possesses a complementary submodule.

Now suppose that L is semisimple. Let $v \in \mathsf{V}_{\mathsf{L}}$ and let M_j be a complement to $\mathrm{span}_{\mathsf{F}[\xi]}(P_j \cdot v)$, $j \in \{1, \dots, k\}$. Then we have $v = (PP_j) \cdot v + u$ for some $P \in \mathsf{F}[\xi]$ and $u \in \mathsf{M}_j$. Thus $u = (1_{\mathsf{F}} - (PP_1)) \cdot v$, and so $P_j \cdot u = 0_{\mathsf{V}}$ since

$$P_j \cdot u = P_j(1_{\mathsf{F}} - (PP_1)) \cdot v \in \mathrm{span}_{\mathsf{F}[\xi]}(P_j \cdot v) \cap \mathsf{M}_j.$$

Moreover, suppose that $v \in \mathsf{V}_{\mathsf{L}}[P_j]$ for some $j \in \{1, \dots, k\}$. Then $P_j^{l_j} \cdot v = 0_{\mathsf{V}}$. Since $\mathsf{V}_{\mathsf{L}}(P_j)$ is a submodule, our above argument gives $(P_j(1 - PP_j)) \cdot v = 0_{\mathsf{V}}$ for some $P \in \mathsf{F}[\xi]$. Thus

$$P_j \cdot v = (PP_j^2) \cdot v = (P^2 P_j^3) \cdot v = \cdots = (P^m P_j^{m+1}) \cdot v = \cdots .$$

But this implies that $\mathsf{V}_{\mathsf{L}}(P_j) = \mathsf{V}_{\mathsf{L}}[P_j]$ for each $j \in \{1, \dots, k\}$. This in turn implies, by Theorem 5.8.29, that $M_{\mathsf{L}} = P_1 \cdots P_k$, as desired. ■

For endomorphisms of vector spaces over algebraically closed fields, or more generally, for endomorphisms whose minimal polynomial splits, we have the following result. This shows that the notion of semisimplicity generalises the notion of diagonalisability.

5.8.54 Proposition (Semisimplicity when the minimal polynomial splits) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ have a minimal polynomial that splits. Then L is semisimple if and only if diagonalisable.*

Proof First suppose that L is semisimple. By Theorem 5.8.53, since M_{L} splits we have

$$M_{\mathsf{L}} = (\xi - \lambda_1)\cdots(\xi - \lambda_k)$$

for distinct $\lambda_1, \dots, \lambda_k \in \mathsf{F}$. By Proposition 5.8.50 it follows that L is diagonalisable.

Conversely, if L is diagonalisable, by Proposition 5.8.50 it follows that M_{L} is a product of distinct degree one factors. By Theorem 5.8.53 it follows that L is semisimple. ■

One might now speculate on the relationship between semisimplicity of an endomorphism and the diagonalisability of the endomorphism after extending from a field to a field where the minimal polynomial splits (e.g., the algebraic closure). The natural conjecture is the following: An endomorphism over a field F is semisimple if and only if its minimal polynomial splits in some extension K of F. *This is actually not true, in general*. It is true much of the time, but not all of the time. For example, as we shall see, it is true when F has characteristic zero. The general story is rooted in the notion of separable polynomials and separable field extensions which we discussed in Section 4.6.7. Let us recall here the main ideas.

1. A polynomial $A \in \mathsf{F}[\xi]$ is separable when it splits into distinct degree one factors in the algebraic closure $\bar{\mathsf{F}}$ of F.
2. An irreducible polynomial is always separable if F has characteristic zero (Proposition 4.6.43).
3. An extension K of F is separable if every $a \in \mathsf{K}$ is the root of a separable polynomial $A \in \mathsf{F}[\xi]$.
4. Algebraic extensions of fields of characteristic zero are separable.

The reader wishing to simplify things to the interesting case where $\mathsf{F} = \mathbb{R}$ should keep points 2 and 4 in mind.

Now let us proceed with the development.

5.8.55 Definition (Absolutely semisimple endomorphism) Let F be a field and let V be a finite-dimensional F-vector space. An endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is ***absolutely semisimple*** if there exists an extension K of F such that L_{K} is diagonalisable. •

For absolutely semisimple endomorphisms we have the following characterisation.

5.8.56 Theorem (Characterisation of absolutely semisimple endomorphism) *Let F be a field with algebraic closure $\bar{\mathsf{F}}$ and let V be a finite-dimensional F-vector space. For an endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ the following statements are equivalent:*

(i) *L is absolutely semisimple;*

(ii) *$\mathsf{L}_{\bar{\mathsf{F}}}$ is diagonalisable;*

(iii) M_{L} *is separable.*

Proof (i) $\Longrightarrow$ (ii) Suppose that K is an extension of F for which L_{K} is diagonalisable. Then, by Proposition 5.8.50, the minimal polynomial M_{L} splits in K into a product of distinct degree 1 factors. Thus $\mathsf{L}_{\bar{\mathsf{F}}}$ splits into the same product of distinct degree 1 factors, and so is diagonalisable by Proposition 5.8.50.

(ii) $\Longrightarrow$ (iii) If $\mathsf{L}_{\bar{\mathsf{F}}}$ is diagonalisable then M_{L} splits in $\bar{\mathsf{F}}$ into a product of distinct degree 1 factors by Proposition 5.8.50. Thus M_{L} is separable by Proposition 4.6.42.

(iii) $\Longrightarrow$ (i) By Proposition 4.6.42, if M_{L} is separable then there exists an extension K in which M_{L} splits into a product of distinct degree 1 factors. By Proposition 5.8.50 it follows that L_{K} is diagonalisable. ■

The relationship between semisimple and absolutely semisimple endomorphisms is contained in the following result.

5.8.57 Corollary (Absolutely semisimple implies semisimple) *Let* F *be a field with* V *a finite-dimensional* F*-vector space. If* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *is absolutely semisimple then it is semisimple.*

Proof If L is absolutely semisimple then M_{L} is separable by Theorem 5.8.56. Write the decomposition of M_{L} as a product of powers of distinct irreducible factors:

$$M_{\mathsf{L}} = P_1^{l_1} \cdots P_k^{l_k}.$$

Separability of M_{L} implies that $l_1 = \cdots = l_k = 1$ since otherwise M_{L} will certainly have repeated roots in $\bar{\mathsf{F}}$. It follows from Theorem 5.8.53 that L is semisimple. ■

The following corollary simplifies much of the complicated business concerning semisimple and absolutely semisimple endomorphisms.

5.8.58 Corollary (In characteristic zero semisimple equals absolutely semisimple) *Let* F *be a field of characteristic zero with algebraic closure* $\bar{\mathsf{F}}$ *and let* V *be a finite-dimensional* F*-vector space. For an endomorphism* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *the following statements are equivalent:*

(i) L *is semisimple;*

(ii) L *is absolutely semisimple;*

(iii) if $\mathrm{P}_1, \dots, \mathrm{P}_k \in \mathsf{F}[\xi]$ *are distinct, monic, irreducible polynomials and if* $\mathrm{l}_1, \dots, \mathrm{l}_k \in \mathbb{Z}_{>0}$ *are such that*

$$\mathrm{M}_{\mathsf{L}} = \mathrm{P}_1^{\mathrm{l}_1} \cdots \mathrm{P}_{\mathrm{k}}^{\mathrm{l}_{\mathrm{k}}},$$

then $\mathrm{l}_1 = \cdots = \mathrm{l}_{\mathrm{k}} = 1$*;*

(iv) $\mathsf{L}_{\bar{\mathsf{F}}}$ *is diagonalisable.*

Proof The result will follow from Theorems 5.8.53 and 5.8.56 and Corollary 5.8.58 if we can show that M_{L} is separable. This in turn follows if we can show that $P_1, \dots, P_k$ are irreducible. But this follows from Proposition 4.6.43 since there it is shown that all irreducible polynomials over fields of characteristic zero are separable. ■

Let us give an example of an endomorphism that is semisimple, but not absolutely semisimple.

5.8.59 Example (Semisimple but not absolutely semisimple endomorphism) We use Example 4.6.48 as our starting point. Thus we consider the field $\mathbb{Z}_2$ with $\mathsf{F} = \mathbb{Z}_2(\eta)$ the field of rational functions in indeterminate η. We take $\mathsf{V} = \mathsf{F}^2$ and define $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by the following 2×2 matrix in F:

$$\mathsf{L} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ \eta & 0_{\mathsf{F}} \end{bmatrix}.$$

The characteristic polynomial is $C_{\mathsf{L}} = \xi^2 - \eta$. As we argue in Example 4.6.48, this polynomial is irreducible so L is semisimple by Theorem 5.8.53. We also argue in Example 4.6.48 that C_{L} is not separable, and so L is not absolutely semisimple by Theorem 5.8.56. •

5.8.13 Triangularisable and nilpotent endomorphisms

In Section 5.8.10 we showed that when the minimal polynomial splits, one can proceed from the rational canonical form to the Jordan canonical form. Note that a diagonal matrix is trivially in Jordan canonical form. When a matrix is not diagonalisable, but its minimal polynomial splits, then the Jordan canonical form shows us that the matrix can be put into a form where the matrix differs from a diagonal matrix only by 1_{F}'s in the entries above the diagonal. In this section we wish to examine this structure without making reference to the rational canonical form. In this way we can free the discussion from reliance on modules over principal ideal domains.

Let us begin with the definition of a what seems to be a generalisation of an endomorphism that can be put into Jordan canonical form.

5.8.60 Definition (Triangularisable endomorphism) Let F be a field and let V be a finite-dimensional F-vector space. An endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is ***triangularisable*** if there exists a basis $\mathscr{B}$ for V such that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is upper triangular. •

The condition in the definition that the matrix representative be upper triangular can be replaced with the condition that the matrix representative be lower triangular; see Exercise 5.8.11.

There is a simple characterisation of triangularisable endomorphisms.

5.8.61 Theorem (Characterisation of triangularisable endomorphism) *Let F be a field and let V be a finite-dimensional F-vector space. An endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is triangularisable if and only if M_{L} splits in V.*

Proof Suppose that L is triangularisable and let $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ be upper diagonal. By Exercise 5.3.7 the eigenvalues of L are the diagonal entries of $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$, and so are in F. Thus, since the roots of M_{L} are the eigenvalues of L, it follows that M_{L} splits.

Now suppose that M_{L} splits. By Theorem 5.8.44 it follows that there exists a basis $\mathscr{B}$ in which $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is block diagonal with the diagonal blocks being Jordan blocks. In particular, $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is upper triangular. However, let us provide an independent proof of this.

Since M_{L} splits in F there exists an eigenvalue of L and so there is a one-dimensional L-invariant subspace V_1 of V. By Exercise 5.7.2 the $(n-1)$-dimensional (with $n = \dim(\mathsf{V})$) subspace $\operatorname{ann}(\mathsf{V}_1)$ is L'-invariant. By Exercise 5.8.5, $M_{\mathsf{L}} = M_{\mathsf{L}'}$. Thus $\mathsf{L}'|\operatorname{ann}(\mathsf{V}_1)$ splits in F and so $\mathsf{L}'|\operatorname{ann}(\mathsf{V}_1)$ possesses a one-dimensional L'-invariant subspace which then gives an $(n-2)$-dimensional L'-invariant subspace containing $\operatorname{ann}(\mathsf{V}_1)$. Since V is finite-dimensional, corresponding to this subspace is a two-dimensional subspace V_2 of V, invariant under L and containing V_1. Continuing in this way we arrive at a sequence

$$\mathsf{V}_1 \subseteq \mathsf{V}_2 \subseteq \cdots \subseteq \mathsf{V}_n = \mathsf{V}$$

of L-invariant subspaces such that $\dim(\mathsf{V}_j) = j$, $j \in \{1,\dots,n\}$. Now let $\mathscr{B} = \{e_1,\dots,e_n\}$ be a basis for V for which $\{e_1,\dots,e_j\}$ is a basis for V_j. By definition of the subspaces $\mathsf{V}_1,\dots,\mathsf{V}_n$ we have

$$\begin{aligned}\mathsf{L}(e_1) &= a_{11}e_1,\\ \mathsf{L}(e_2) &= a_{12}e_1 + a_{22}e_2,\\ &\vdots\\ \mathsf{L}(e_n) &= a_{1n}e_1 + \cdots + a_{nn}e_n,\end{aligned}$$

where the a_{ij}'s are in F. By definition of the matrix representative this means that $[\mathsf{L}]^{\mathscr{B}}_{\mathscr{B}}$. ■

This theorem can be used to give another proof of the Cayley–Hamilton Theorem.

5.8.62 Corollary (Cayley–Hamilton Theorem yet again) *If F is a field, if V is a finite-dimensional F-vector space, and if $\mathsf{L} \in \operatorname{End}_{\mathsf{F}}(\mathsf{V})$, then $\operatorname{Ev}_{\mathsf{F}}(\mathsf{C}_{\mathsf{L}})(\mathsf{L}) = 0_{\operatorname{End}_{\mathsf{F}}(\mathsf{V})}$, i.e., L satisfies its own characteristic polynomial.*

Proof Let us first prove a lemma.

1 Lemma *If $\mathbf{A} \in \mathrm{Mat}_{\mathrm{n}\times\mathrm{n}}(\mathsf{F})$ is an upper triangular matrix with diagonal entries $(\mathrm{d}_1,\dots,\mathrm{d}_\mathrm{n})$, then*

$$(\mathbf{A} - \mathrm{d}_1\mathbf{I}_\mathrm{n})\cdots(\mathbf{A} - \mathrm{d}_\mathrm{n}\mathbf{I}_\mathrm{n}) = \mathbf{0}_{\mathrm{n}\times\mathrm{n}}.$$

Proof We prove this by induction on n, the result being obviously true when $n = 1$. Let $\{\boldsymbol{e}_1,\dots,\boldsymbol{e}_n\}$ be the standard basis for F^n. We then clearly have $(\boldsymbol{A} - d_n\boldsymbol{I}_n)\boldsymbol{e}_n = \boldsymbol{0}_{\mathsf{F}^n}$, simply using the fact that $\boldsymbol{A}$ is upper triangular with $\boldsymbol{A}(n,n) = d_n$. Therefore,

$$(\boldsymbol{A} - d_1\boldsymbol{I}_n)\cdots(\boldsymbol{A} - d_n\boldsymbol{I}_n)\boldsymbol{e}_n = \boldsymbol{0}_{\mathsf{F}^n}.$$

Note that the matrices $\boldsymbol{A} - d_1\boldsymbol{I}_n,\dots,\boldsymbol{A} - d_n\boldsymbol{I}_n$ commute so that

$$(\boldsymbol{A} - d_1\boldsymbol{I}_n)\cdots(\boldsymbol{A} - d_n\boldsymbol{I}_n) = (\boldsymbol{A} - d_n\boldsymbol{I}_n)(\boldsymbol{A} - d_1\boldsymbol{I}_n)\cdots(\boldsymbol{A} - d_{n-1}\boldsymbol{I}_n).$$

By the induction hypothesis,

$$(\boldsymbol{A} - d_1\boldsymbol{I}_n)\cdots(\boldsymbol{A} - d_{n-1}\boldsymbol{I}_n)\boldsymbol{e}_j = \boldsymbol{0}_{\mathsf{F}^n}$$

for $j \in \{1, \dots, n-1\}$ (why?). Therefore,

$$(\boldsymbol{A} - d_1\boldsymbol{I}_n)\cdots(\boldsymbol{A} - d_n\boldsymbol{I}_n)\boldsymbol{e}_j = \boldsymbol{0}_{\mathsf{F}^n}$$

for all $j \in \{1, \dots, n\}$, which is the result. ▼

Now let $\bar{\mathsf{F}}$ be the algebraic closure of F so that M_{L} and C_{L} split in $\bar{\mathsf{F}}$. Then let $\{e_1, \dots, e_n\}$ be a basis for $\bar{\mathsf{F}}^n$ for which $\boldsymbol{A} \triangleq [\mathsf{L}_{\bar{\mathsf{F}}}]^{\mathscr{B}}_{\mathscr{B}}$ is upper triangular. Then, by the lemma,

$$(\boldsymbol{A} - \lambda_1\boldsymbol{I}_n)\cdots(\boldsymbol{A} - \lambda_n\boldsymbol{I}_n) = \boldsymbol{0}_{n\times n},$$

where $\lambda_1, \dots, \lambda_n \in \bar{\mathsf{F}}$ are the eigenvalue of $\mathsf{L}_{\bar{\mathsf{F}}}$. But we also have

$$\mathsf{C}_{\mathsf{L}} = (\xi - \lambda_1)\cdots(\xi - \lambda_n)$$

since the roots up the characteristic polynomial are the eigenvalues. Thus

$$\mathrm{Ev}_{\bar{\mathsf{F}}}(\mathsf{C}_{\mathsf{L}})(\mathsf{L}_{\bar{\mathsf{F}}}) = 0_{\mathrm{End}_{\bar{\mathsf{F}}}(\mathsf{V}_{\bar{\mathsf{F}}})}.$$

Since $\mathrm{Ev}_{\bar{\mathsf{F}}}(\mathsf{C}_{\mathsf{L}})(\mathsf{L}_{\bar{\mathsf{F}}}) = (\mathrm{Ev}_{\mathsf{F}}(\mathsf{C}_{\mathsf{L}})(\mathsf{L}))_{\bar{\mathsf{F}}}$ the result follows. ■

As we know from Theorem 5.8.44, the upper triangular form is too coarse a canonical form for a matrix with a minimal polynomial that splits. The Jordan canonical form is a far more structured canonical form. In order to work our way towards the Jordan canonical form without making use of the rational canonical form, we introduce a particular class of triangular endomorphisms.

5.8.63 Definition (Nilpotent endomorphism) Let F be a field and let V be a finite-dimensional F-vector space. An endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is ***nilpotent*** if there exists $k \in \mathbb{Z}_{\geq 0}$ such that $\mathsf{L}^k = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. The least integer k for which this holds is called the ***index of nilpotency*** of L. •

Let us give some characterisations and properties of nilpotent endomorphisms.

5.8.64 Proposition (Properties of nilpotent endomorphisms) *Let F be a field and let V be a finite-dimensional F-vector space. For an endomorphism $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ the following statements are equivalent:*

(i) L *is nilpotent with index of nilpotency* k*;*

(ii) $\mathrm{M}_{\mathsf{L}} = \xi^{\mathrm{k}}$.

If particular, nilpotent endomorphisms are triangularisable.

Proof Suppose that L is nilpotent with index of nilpotency k. Then $\mathrm{Ev}_{\mathsf{F}}(\xi^k)(\mathsf{L}) = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$ from which we conclude that $M_{\mathsf{L}}|\xi^k$. Moreover, since $\mathrm{Ev}_{\mathsf{F}}(\xi^{k-1})(\mathsf{L}) \neq 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$ it follows that $M_{\mathsf{L}} = \xi^k$.

If $M_{\mathsf{L}} = \xi^k$ then L is nilpotent with index of nilpotency at most k. Since M_{L} is the minimal polynomial, it follows that $\mathrm{Ev}_{\mathsf{F}}(\xi^{k-1})(\mathsf{L}) \neq 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$ so that the index of nilpotency is exactly k.

If L is nilpotent the first part of the proof shows that M_{L} splits in F. Thus L is triangularisable by Theorem 5.8.61. ■

Now we provide a very structured canonical form for nilpotent endomorphisms. In order to do this it is convenient to introduce a piece of notation.

5.8.65 Definition (Nilpotent block) Let F be a field and let $k \in \mathbb{Z}_{>0}$. The matrix

$$N(k) \triangleq \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 1_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ 0_{\mathsf{F}} & 0_{\mathsf{F}} & 0_{\mathsf{F}} & \cdots & 0_{\mathsf{F}} & 0_{\mathsf{F}} \end{bmatrix}$$

is the ***nilpotent block*** associated with k. •

Obviously the nilpotent block is related to the Jordan block by $N(k) = J(\lambda, k)$. Now we can give the canonical form for a nilpotent endomorphism.

5.8.66 Theorem (Nilpotent canonical form) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ be nilpotent. Then there exists $\mathrm{k}_1, \dots, \mathrm{k}_\mathrm{r} \in \mathbb{Z}_{>0}$ and a basis $\mathscr{B}$ for V such that*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \mathrm{diag}(\mathbf{N}(\mathrm{k}_1), \dots, \mathbf{N}(\mathrm{k}_\mathrm{r})).$$

Moreover, if there exists a basis for V such that the matrix representative of L in this basis is

$$\mathrm{diag}(\mathbf{N}(\mathrm{l}_1), \dots, \mathbf{N}(\mathrm{l}_\mathrm{s})),$$

then $\mathrm{r} = \mathrm{s}$ and there exists $\sigma \in \mathfrak{S}_\mathrm{r}$ such that $\mathrm{l}_\mathrm{j} = \mathrm{k}_{\sigma(\mathrm{j})}$ for $\mathrm{j} \in \{1, \dots, \mathrm{r}\}$.

Proof Some of the ideas in the proof, and indeed the theorem itself, follows from our discussion of the rational canonical form, particularly Theorem 5.8.44. However, we give an independent, self-contained proof.

For $v_0 \in \mathsf{V}$ let $\mathrm{nil}(v_0, \mathsf{L})$ be the least integer k for which $\mathsf{L}^k(v_0) = 0_{\mathsf{V}}$. Also let

$$\mathsf{C}(v_0, \mathsf{L}) = \mathrm{span}_{\mathsf{F}}(\mathsf{L}^j(v_0)| \ j \in \mathbb{Z}_{\geq 0}).$$

Let us prove a lemma.

1 Lemma *The set $\{\mathsf{L}^{\mathrm{nil}(\mathrm{v}_0,\mathsf{L})-1}(\mathrm{v}_0), \dots, \mathsf{L}(\mathrm{v}_0), \mathrm{v}_0\}$ is a basis for $\mathsf{C}(\mathrm{v}_0, \mathsf{L})$. Moreover, the matrix representative for L in this basis is $\mathbf{N}(\mathrm{nil}(\mathrm{v}_0, \mathsf{L}))$.*

Proof Let us first work with the reordered set $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}\}$, as it is convenient to do this.

Clearly

$$\mathrm{span}_{\mathsf{F}}(v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0)) \subseteq \mathsf{C}(v_0, \mathsf{L}).$$

If $v \in \mathsf{C}(v_0, \mathsf{L})$ then we can write

$$v = c_0 v_0 + c_1 \mathsf{L}(v_0) + \dots + c_k \mathsf{L}^k(v_0)$$

for some $k \in \mathbb{Z}_{\geq 0}$ and $c_0, c_1, \dots, c_k \in \mathsf{F}$. Since $\mathsf{L}^j(v_0) = 0_{\mathsf{V}}$ for $j \geq \mathrm{nil}(v_0, \mathsf{L})$ it follows that

$$v_0 \in \mathrm{span}_{\mathsf{F}}(v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0))$$

and so $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}\}$ spans $\mathsf{C}(v_0, \mathsf{L})$.

Now suppose that

$$c_0 v_0 + c_1 \mathsf{L}(v_0) + \cdots + c_{\mathrm{nil}(v_0,\mathsf{L})-1}\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0) = 0_{\mathsf{V}}. \tag{5.42}$$

Applying $\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}$ to this equation gives $c_0\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0) = 0_{\mathsf{V}}$ which gives $c_0 = 0_{\mathsf{F}}$ by Proposition 4.5.3(vi). Now, with $c_0 = 0_{\mathsf{F}}$, apply $\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-2}$ to (5.42) to get $c_1\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0) = 0_{\mathsf{V}}$. This gives $c_1 = 0_{\mathsf{F}}$. Carrying on in this way we deduce that

$$c_0 = c_1 = \cdots = c_{\mathrm{nil}(v_0,\mathsf{L})-1} = 0_{\mathsf{F}},$$

giving linear independence of $\{v_0, \mathsf{L}(v_0), \dots, \mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}\}$.

Now we compute

$$\begin{aligned}\mathsf{L}(\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0)) &= 0_{\mathsf{V}},\\ \mathsf{L}(\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-2}(v_0)) &= \mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0),\\ &\vdots\\ \mathsf{L}(\mathsf{L}(v_0)) &= \mathsf{L}^2(v_0),\\ \mathsf{L}(v_0) &= \mathsf{L}(v_0),\end{aligned}$$

from which we immediately deduce that the matrix representative of L in the basis $\{\mathsf{L}^{\mathrm{nil}(v_0,\mathsf{L})-1}(v_0), \dots, \mathsf{L}(v_0), v_0\}$ is indeed $\boldsymbol{N}(\mathrm{nil}(v_0,\mathsf{L}))$. ▼

The following simple lemma will be useful.

2 Lemma *With the above notation, the ideal of* $\mathsf{F}[\xi]$ *given by*

$$\{\mathrm{P} \in \mathsf{F}[\xi] \mid \mathrm{Ev}_{\mathsf{F}}(\mathrm{P})(\mathsf{L})(\mathrm{v}_0) = 0_{\mathsf{V}}\}$$

is generated by $\xi^{\mathrm{nil}(\mathrm{v}_0,\mathsf{L})}$.

Proof Let

$$\mathsf{I}(v_0,\mathsf{L}) = \{P \in \mathsf{F}[\xi] \mid \mathrm{Ev}_{\mathsf{F}}(P)(\mathsf{L})(v_0) = 0_{\mathsf{V}}\}.$$

Clearly $(\xi^{\mathrm{nil}(v_0,\mathsf{L})}) \subseteq \mathsf{I}(v_0,\mathsf{L})$. Let $P \in \mathsf{I}(v_0,\mathsf{L})$ be given by

$$P = a_k\xi^k + \cdots + a_1\xi + a_0.$$

Thus

$$a_0 v_0 + a_1\mathsf{L}(v_0) + \cdots + a_k\mathsf{L}^k(v_0) = 0_{\mathsf{V}}.$$

Arguing as in the linear independence part of the proof of Lemma 1 we can show that $a_0 = a_1 = \cdots = a_{\mathrm{nil}(v_0,\mathsf{L})-1} = 0_{\mathsf{F}}$. Thus $\xi^{\mathrm{nil}(v_0,\mathsf{L})-1}|P$ and so $P \in (\xi^{\mathrm{nil}(v_0,\mathsf{L})-1})$ by Proposition 4.2.61. ▼

Now we prove another lemma.

3 Lemma *There exists* $\mathrm{v}_1, \dots, \mathrm{v}_\mathrm{r} \in \mathsf{V}$ *such that*

$$\mathsf{V} = \mathsf{C}(\mathrm{v}_1, \mathsf{L}) \oplus \cdots \oplus \mathsf{C}(\mathrm{v}_\mathrm{r}, \mathsf{L}).$$

Proof The proof is by induction on $\dim(\mathsf{V})$. When $\dim(\mathsf{V}) = 0$ the result is trivial. Now suppose that the result is true for vector spaces of dimension $1, \dots, n-1$ and suppose that $\dim(\mathsf{V}) = n$. Since L is nilpotent, $(\det \mathsf{L})^k = \det(\mathsf{L}^k) = 0_\mathsf{F}$ for some $k \in \mathbb{Z}_{\geq 0}$. Thus $\det \mathsf{L} = 0_\mathsf{F}$ and so $\mathrm{rank}(\mathsf{L}) < n$ by Theorem 5.4.35. Therefore, $\dim(\mathsf{L}(\mathsf{V})) < n$ and by the induction hypothesis there exists nonzero $u_1, \dots, u_s \in \mathsf{L}(\mathsf{V})$ such that

$$\mathsf{L}(\mathsf{V}) = \mathsf{C}(u_1, \mathsf{L}|\mathsf{L}(\mathsf{V})) \oplus \cdots \oplus \mathsf{C}(u_s, \mathsf{L}|\mathsf{L}(\mathsf{V})).$$

Since $u_1, \dots, u_s \in \mathrm{image}(\mathsf{L})$ there exists $v_1, \dots, v_s \in \mathsf{V}$ such that $\mathsf{L}(v_j) = u_j$, $j \in \{1, \dots, s\}$. Clearly $\mathrm{nil}(v_j, \mathsf{L}) \geq 2$, $j \in \{1, \dots, s\}$, since $u_1, \dots, u_s$ are nonzero. We claim that

$$\mathsf{C}(v_1, \mathsf{L}) \cap \cdots \cap \mathsf{C}(v_s, \mathsf{L}) = \{0_\mathsf{V}\}.$$

Suppose that $w_j \in \mathsf{C}(v_j, \mathsf{L})$, $j \in \{1, \dots, s\}$, are such that $w_1 + \cdots + w_s = 0_\mathsf{V}$. Then

$$\mathsf{L}(w_1) + \cdots + \mathsf{L}(w_s) = 0_\mathsf{V}.$$

Since

$$\begin{aligned}\mathsf{L}(w_1) + \cdots + \mathsf{L}(w_s) &\in \mathsf{L}(\mathsf{C}(v_1, \mathsf{L}) \cap \cdots \cap \mathsf{C}(v_s, \mathsf{L})) \\ &= \mathsf{C}(u_1, \mathsf{L}|\mathsf{L}(\mathsf{V})) \cap \cdots \cap \mathsf{C}(u_s, \mathsf{L}|\mathsf{L}(\mathsf{V})) = \{0_\mathsf{V}\},\end{aligned}$$

it follows that $\mathsf{L}(w_j) = 0_\mathsf{V}$, $j \in \{1, \dots, s\}$. Since $w_j \in \mathsf{C}(v_j, \mathsf{L})$, by Lemma 1 it follows that

$$w_j = \sum_{l=0}^{\mathrm{nil}(v_j, \mathsf{L})} c_{jl} \mathsf{L}^l(v_j)$$

for some $c_{jl} \in \mathsf{F}$, $j \in \{1, \dots, s\}$, $l \in \{0, 1, \dots, \mathrm{nil}(v_j, \mathsf{L})\}$. Define $P_j \in \mathsf{F}[\xi]$ by

$$P_j = \sum_{l=0}^{\mathrm{nil}(v_j, \mathsf{L})} c_{jl} \xi^l$$

so that

$$0_\mathsf{V} = \mathsf{L}(w_j) = \sum_{l=0}^{\mathrm{nil}(v_j, \mathsf{L})} c_{jl} \mathsf{L}^l(\mathsf{L}(v_j)) = \mathrm{Ev}_\mathsf{F}(P_j)(\mathsf{L}) \cdot u_j$$

By Lemma 1 it follows that $\xi^{\mathrm{nil}(u_j, \mathsf{L})} | P_j$. Thus $P_j = \xi Q_j$ for some $Q_j \in \mathsf{F}[\xi]$. Therefore,

$$w_j = \mathrm{Ev}_\mathsf{F}(P_j)(\mathsf{L}) \cdot v_j = \mathrm{Ev}_\mathsf{F}(Q_j)(\mathsf{L}) \cdot u_j \in \mathsf{C}(u_j, \mathsf{L}|\mathsf{L}(\mathsf{V})).$$

Therefore,

$$w_1 + \cdots + w_s \in \mathsf{C}(u_1, \mathsf{L}|\mathsf{L}(\mathsf{V})) \cap \cdots \cap \mathsf{C}(u_s, \mathsf{L}|\mathsf{L}(\mathsf{V})) = \{0_\mathsf{V}\},$$

and so $w_j = 0_{\mathsf{V}}$, $j \in \{1, \dots, s\}$. This gives

$$\mathsf{C}(v_1, \mathsf{L}) \cap \dots \cap \mathsf{C}(v_s, \mathsf{L}) = \{0_{\mathsf{V}}\}$$

as claimed.

Let U be a complement to $\ker(\mathsf{L}) \cap \operatorname{image}(\mathsf{L})$ in $\ker(\mathsf{L})$. That is, suppose that

$$\ker(\mathsf{L}) = (\ker(\mathsf{L}) \cap \operatorname{image}(\mathsf{L})) \oplus \mathsf{U}.$$

We claim that

$$\mathsf{V} = \mathsf{C}(v_1, \mathsf{L}) \oplus \dots \oplus \mathsf{C}(v_s, \mathsf{L}) \oplus \mathsf{U}.$$

First let

$$v \in (\mathsf{C}(v_1, \mathsf{L}) \oplus \dots \oplus \mathsf{C}(v_s, \mathsf{L})) \cap \mathsf{U}.$$

Then

$$v \in (\mathsf{C}(v_1, \mathsf{L}) \oplus \dots \oplus \mathsf{C}(v_s, \mathsf{L})) \cap \ker(\mathsf{L}).$$

Therefore

$$v = w_1 + \dots + w_s, \qquad w_j \in \mathsf{C}(v_j, \mathsf{L}),\ j \in \{1, \dots, s\},$$

and

$$\mathsf{L}(w_1 + \dots + w_s) = 0_{\mathsf{V}}.$$

Since $\mathsf{L}(w_j) \in \mathsf{C}(v_j, \mathsf{L})$ we must have $\mathsf{L}(w_j) = 0_{\mathsf{V}}$, $j \in \{1, \dots, s\}$. Now a duplication of the above argument for $w_1, \dots, w_s$ gives $w_1 = \dots = w_s = 0_{\mathsf{V}}$. Thus

$$(\mathsf{C}(v_1, \mathsf{L}) \oplus \dots \oplus \mathsf{C}(v_s, \mathsf{L})) \cap \mathsf{U} = \{0_{\mathsf{V}}\}.$$

Since $\mathsf{U} \subseteq \ker(\mathsf{L})$ it follows that there exists a basis $\{u_1, \dots, u_m\}$ for U such that $\mathsf{L}(u_j) = 0_{\mathsf{V}}$, $j \in \{1, \dots, m\}$. That is to say, $\mathsf{C}(u_j, \mathsf{L}) = \operatorname{span}_{\mathsf{F}}(u_j)$, $j \in \{1, \dots, m\}$. Thus we have

$$\mathsf{V} = \mathsf{C}(v_1, \mathsf{L}) \oplus \dots \oplus \mathsf{C}(v_s, \mathsf{L}) \oplus \mathsf{C}(u_1, \mathsf{L}) \oplus \dots \oplus \mathsf{C}(u_m, \mathsf{L}),$$

giving the lemma. ▼

The above lemma, combined with Lemma 1, gives the existence of a basis as in the theorem statement. To show uniqueness (up to permutation) we again use induction on $\dim(\mathsf{V})$. Clearly there is only one matrix representation in the form of the theorem statement when $\dim(\mathsf{V}) = 0$. Suppose that this is so for all vector spaces of dimension $1, \dots, n-1$ and let $n = \dim(\mathsf{V})$. Let

$$\mathsf{V} = \mathsf{C}(v_1, \mathsf{L}) \oplus \mathsf{C}(v_s, \mathsf{L})$$

be a decomposition of V corresponding to a basis in which the matrix representative of L is block diagonal with diagonal blocks being nilpotent blocks. Thus v_j is such that $\{\mathsf{L}^{\operatorname{nil}(v_j, \mathsf{L})-1}(v_j), \dots, \mathsf{L}(v_j), v_j\}$ is a basis for $\mathsf{C}(v_j, \mathsf{L})$. Some of these blocks may be one-dimensional; let us suppose that these are collected such that

$$\begin{aligned} \dim(\mathsf{C}(v_j, \mathsf{L})) &> 1, \qquad j \in \{1, \dots, m\}, \\ \dim(\mathsf{C}(v_j, \mathsf{L})) &= 1, \qquad j \in \{m+1, \dots, s\}. \end{aligned}$$

Thus

$$\mathsf{C}(v_{m+1}, \mathsf{L}) \oplus \cdots \oplus \mathsf{C}(v_s, \mathsf{L}) \subseteq \ker(\mathsf{L}).$$

Moreover, it is easy to see from the form of the matrix representative for L that

$$\mathsf{L}(\mathsf{V}) = \mathsf{C}(\mathsf{L}(v_1), \mathsf{L}|\mathsf{L}(\mathsf{V})) \oplus \cdots \oplus \mathsf{C}(\mathsf{L}(v_m), \mathsf{L}|\mathsf{L}(\mathsf{V}))$$

and that

$$\dim(\mathsf{C}(\mathsf{L}(v_j), \mathsf{L}|\mathsf{L}(\mathsf{V}))) = \dim(\mathsf{C}(v_j, \mathsf{L})) - 1.$$

By the induction hypothesis the number m and the dimensions of $\mathsf{C}(\mathsf{L}(v_j), \mathsf{L}|\mathsf{L}(\mathsf{V}))$, $j \in \{1, \dots, m\}$, are determined uniquely by $\mathsf{L}|\mathsf{V}$. Thus the number m of blocks of size greater than one and the dimensions of $\mathsf{C}(v_j, \mathsf{L})$, $j \in \{1, \dots, m\}$, are determined uniquely by L. Thus the size of all nilpotent blocks in the matrix representative for L are uniquely determined by L. ■

Note that the given proof of the theorem is independent of the rational canonical form, and so free from all the business about modules over principal ideal domains. The price one pays for this is a fairly complicated proof, since one essentially has to generate "by hand" all of the ideas one gets from the theory of modules over principal ideal domains.

Let us give a name to the canonical form of the preceding theorem.

5.8.67 Definition (Nilpotent canonical form) Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ be nilpotent. The matrix

$$\mathrm{diag}(\boldsymbol{N}(k_1), \dots, \boldsymbol{N}(k_r))$$

associated with L by Theorem 5.8.66, with $k_1 \geq \cdots \geq k_r$, is the ***nilpotent canonical form*** for L. •

5.8.14 The Jordan decomposition

In this section we shall give what amounts to a generalisation of the Jordan canonical form. Since the Jordan canonical form is valid only for endomorphisms whose minimal polynomial splits, one might inquire whether there is an analogue to this in the case where the minimal polynomial does not split. One might imagine that this is related to whether the endomorphism has a Jordan canonical form in some extension of the field in which one is working. This is indeed the case as the following development shows.

The reader who wants to keep things simple and consider only $\mathbb{R}$-vector spaces may wish to recall that "absolutely semisimple" equals "semisimple" in this case.

5.8.68 Definition (Jordan decomposition of an endomorphism) Let F be a field and let V be a finite-dimensional F-vector space. A ***Jordan decomposition*** for $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is an additive decomposition $\mathsf{L} = \mathsf{S}(\mathsf{L}) + \mathsf{N}(\mathsf{L})$ of L with the following properties:

(i) $\mathsf{S}(\mathsf{L})$ is absolutely semisimple;

(ii) $\mathsf{N}(\mathsf{L})$ is nilpotent;

(iii) $\mathsf{S}(\mathsf{L})$ and $\mathsf{N}(\mathsf{L})$ commute. •

A Jordan decomposition does not always exist. However, it does in all cases of interest to us. But let us state the general theorem first. The proof of the theorem makes free use of some ideas from Section 4.6. Readers wanting to simplify life by thinking of $\mathbb{R}$-vector spaces may think of the extension $\mathbb{C}$ of $\mathbb{R}$ where the automorphisms of the extension are either the identify map or complex conjugation.

5.8.69 Theorem (Existence of Jordan decomposition) *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Then the following statements are equivalent:*

(i) there exists a separable extension K of L containing the eigenvalues of L;

(ii) there exists a unique Jordan decomposition for L.

Moreover, $\mathsf{C}_{\mathsf{L}} = \mathsf{C}_{\mathsf{S}(\mathsf{L})}$.

Proof We first prove the theorem when the minimal polynomial splits.

1 Lemma *Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$. If M_{L} splits in F then there exists a unique Jordan decomposition for L. Moreover, the characteristic polynomial of L agrees with that of its semisimple part.*

Proof By Corollary 5.8.32 let us write

$$\mathsf{V} = \overline{\mathsf{W}}(\lambda_1, \mathsf{L}) \oplus \cdots \oplus \overline{\mathsf{W}}(\lambda_k, \mathsf{L}),$$

where $\lambda_1, \dots, \lambda_k$ are the distinct eigenvalues of L. For brevity let us denote $\mathsf{W}_j = \overline{\mathsf{W}}(\lambda_j, \mathsf{L})$, $j \in \{1, \dots, k\}$. Let us denote $\mathsf{L}_j = \mathsf{L}|\mathsf{W}_j$, $j \in \{1, \dots, k\}$, this making sense by Proposition 5.4.59. Let us define $\mathsf{S}_j = \lambda_j \,\mathrm{id}_{\mathsf{W}_j}$ and $\mathsf{N}_j = \mathsf{L}_j - \mathsf{S}_j$. Clearly S_j is diagonalisable. Also, since

$$\mathsf{W}_j = \ker((\lambda_j \,\mathrm{id}_{\mathsf{W}_j} - \mathsf{L}_j)^{k_j}) = \ker(\mathsf{N}_j^{k_j})$$

for some $k_j \in \mathbb{Z}_{\geq 0}$, it follows that N_j is nilpotent. One can check by direct computation that $\mathsf{S}_j\mathsf{N}_j = \mathsf{N}_j\mathsf{S}_j$.

Now define $\mathsf{S}(\mathsf{L})$ by asking that $\mathsf{S}(\mathsf{L})|\mathsf{W}_j = \mathsf{S}_j$ and define $\mathsf{N}(\mathsf{L})$ by asking that $\mathsf{N}(\mathsf{L})|\mathsf{W}_j = \mathsf{N}_j$. It is clear that $\mathsf{S}(\mathsf{L})$ is diagonalisable and so absolutely semisimple. It is also clear that $\mathsf{S}(\mathsf{L})\mathsf{N}(\mathsf{L}) = \mathsf{N}(\mathsf{L})\mathsf{S}(\mathsf{L})$. That $\mathsf{N}(\mathsf{L})$ is nilpotent follows since $\mathsf{N}(\mathsf{L})^k|\mathsf{W}_j = 0_{\mathrm{End}_{\mathsf{F}}(\mathsf{W}_j)}$ if $k = \max\{k_1, \dots, k_r\}$. This gives the existence of a Jordan decomposition for L. It is evident that $C_{\mathsf{L}} = C_{\mathsf{S}(\mathsf{L})}$.

To see that this decomposition is unique, suppose that $\mathsf{L} = \mathsf{S} + \mathsf{N}$ where S is absolutely semisimple, N is nilpotent, and S and N commute. Using the fact that $\mathsf{L} = \mathsf{S} + \mathsf{N}$ one can directly verify that L commutes with both S and N. Thus W_j, $j \in \{1, \dots, k\}$, is invariant under both S and N by Exercise 5.4.11. We claim that $\mathsf{S}|\mathsf{W}_j = \mathsf{S}_j = \lambda_j \,\mathrm{id}_{\mathsf{W}_j}$. Since S is absolutely semisimple let K be an extension of F such that S_{K} is diagonalisable. Then $\mathsf{S}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}}$ is diagonalisable (why?) and so $\mathsf{S}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}} - \mathsf{S}_{j,\mathsf{K}}$ is diagonalisable since $\mathsf{S}_{j,\mathsf{K}}$ has a diagonal matrix representative in every basis. Thus

$$\mathsf{N}_{j,\mathsf{K}} - \mathsf{N}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}} = \mathsf{S}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}} - \mathsf{S}_{j,\mathsf{K}}$$

is diagonalisable. Note that $\mathsf{N}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}}$ commutes with both $\mathsf{S}_{j,\mathsf{K}}$ and $\mathsf{L}_{j,\mathsf{K}}$. Therefore, we have

$$(\mathsf{N}_{j,\mathsf{K}} - \mathsf{N}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}})^k = \sum_{l=0}^{k} B_{k,l}(\mathsf{N}_{j,\mathsf{K}})^l(-\mathsf{N}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}})^{k-l},$$

using the Binomial Theorem. Therefore, for k sufficiently large we have

$$(\mathsf{N}_{j,\mathsf{K}} - \mathsf{N}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}})^k = 0_{\mathrm{End}_{\mathsf{K}}(\mathsf{W}_{j,\mathsf{K}})},$$

and so $\mathsf{N}_{j,\mathsf{K}} - \mathsf{N}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}}$, and therefore, $\mathsf{S}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}} - \mathsf{S}_{j,\mathsf{K}}$, is nilpotent. Thus $\mathsf{S}_{\mathsf{K}}|\mathsf{W}_{j,\mathsf{K}} - \mathsf{S}_{j,\mathsf{K}}$ is nilpotent and diagonalisable, and therefore must be zero. This gives $\mathsf{S}|\mathsf{W}_j = \mathsf{S}_j$ and so $\mathsf{S} = \mathsf{S}(\mathsf{L})$. Clearly then $\mathsf{N} = \mathsf{N}(\mathsf{L})$. ▼

Now let us proceed with the proof of the theorem. First suppose that L admits a Jordan decomposition, which we denote by $\mathsf{L} = \mathsf{S} + \mathsf{N}$. Then $\mathsf{L}_{\bar{\mathsf{F}}} = \mathsf{S}_{\bar{\mathsf{F}}} + \mathsf{N}_{\bar{\mathsf{F}}}$ with $\mathsf{S}_{\bar{\mathsf{F}}}$ diagonalisable. By the lemma, $C_{\mathsf{L}} = C_{\mathsf{S}}$. By Theorem 5.8.56 it follows that the eigenvalues of L lie in a separable extension of F.

Now we prove the converse. For simplicity let us suppose that we have chosen a basis, thus identifying V with F^n and endomorphisms with matrices. Thus we consider $\mathsf{L} \in \mathrm{Mat}_{n\times n}(\mathsf{F})$. Suppose that the eigenvalues of L lie in a separable extension K of F. By Corollary 4.6.51 we suppose K to be Galois. Thus M_{L} splits in K by Proposition 5.8.13. By the lemma we write $\mathsf{L}_{\mathsf{K}} = \mathsf{S} + \mathsf{N}$ as the Jordan decomposition of L_{K}. Thus $\mathsf{S} \in \mathrm{Mat}_{n\times n}(\mathsf{K})$ is diagonalisable, $\mathsf{N} \in \mathrm{Mat}_{n\times n}(\mathsf{K})$ is nilpotent, and S and N commute. If $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$ and if $\boldsymbol{A} \in \mathrm{Mat}_{n\times n}(\mathsf{K})$ denote by $\boldsymbol{A}^{\phi} \in \mathrm{Mat}_{n\times n}(\mathsf{K})$ the matrix obtained by applying ϕ to the entries in $\boldsymbol{A}$. We have

$$\mathsf{L}_{\mathsf{K}} = \mathsf{L}_{\mathsf{K}}^{\phi} = (\mathsf{S} + \mathsf{N})^{\phi} = \mathsf{S}^{\phi} + \mathsf{N}^{\phi}$$

and

$$\mathsf{S}^{\phi}\mathsf{N}^{\phi} = (\mathsf{S}\mathsf{N})^{\phi} = (\mathsf{N}\mathsf{S})^{\phi} = \mathsf{N}^{\phi}\mathsf{S}^{\phi}$$

for every $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$, where we use the fact in the first equation that the entries in L are in F, and so are fixed by ϕ. We claim that S^{ϕ} is diagonalisable. Indeed, since S is diagonalisable we have

$$\boldsymbol{P}\mathsf{S}\boldsymbol{P}^{-1} = \boldsymbol{D}$$

for an invertible matrix $\boldsymbol{P}$ and a diagonal matrix $\boldsymbol{D}$. Thus

$$\boldsymbol{P}^{\phi}\mathsf{S}^{\phi}(\boldsymbol{P}^{-1})^{\phi} = \boldsymbol{D}^{\phi},$$

and so S^{ϕ} is indeed diagonalisable. This shows that $\mathsf{L} = \mathsf{S}^{\phi} + \mathsf{N}^{\phi}$ is a Jordan decomposition for L. By the uniqueness part of the lemma above this implies that $\mathsf{S}^{\phi} = \mathsf{S}$ and $\mathsf{N}^{\phi} = \mathsf{N}$ for every $\phi \in \mathrm{Aut}_{\mathsf{F}}(\mathsf{K})$. By Proposition 4.6.50 this implies that S and N have entries in F. Thus $\mathsf{L} = \mathsf{S} + \mathsf{N}$ is the unique Jordan decomposition for L, as desired.

The final assertion of the theorem follows from the corresponding assertion from the lemma above. ■

For us the following result is the most useful.

5.8.70 Corollary (Jordan decomposition for fields of characteristic zero) *Let* F *be a field of characteristic zero and let* V *be a finite-dimensional* F*-vector space. If* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ *then* L *possesses a unique Jordan decomposition.*

Proof By Proposition 4.6.47 any splitting field for the minimal polynomial is separable, so the eigenvalues of L are always contained in a separable field extension of F. ■

Since a Jordan decomposition generally exists, it is most compelling to consider an example when it does not.

5.8.71 Example (An endomorphism not having a Jordan decomposition) We continue with Example 5.8.59. Thus we let $\mathsf{F} = \mathbb{Z}_2(\eta)$ be the field of rational functions with coefficients in $\mathbb{Z}_2$. We take

$$\mathsf{L} = \begin{bmatrix} 0_{\mathsf{F}} & 1_{\mathsf{F}} \\ \eta & 0_{\mathsf{F}} \end{bmatrix}.$$

The characteristic polynomial $\xi^2 - \eta$ is not separable, and so the eigenvalues cannot be contained in a separable extension. Therefore, L does not have a Jordan decomposition. However, the matrix form for L is in rational canonical form. This illustrates one of the advantages of the rational canonical form: it always exists. •

From the Theorem 5.8.69 we arrive at another proof of the Jordan canonical form, a proof not depending on the rational canonical form with its dependence on the theory of modules over principal ideal domains.

5.8.72 Corollary (Jordan canonical form again) *Let* F *be a field, let* V *be a finite-dimensional* F*-vector space, and let* $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$*. The following statements are equivalent:*

(i) M_{L} *(or, equivalently,* C_{L}*) splits in* F*;*

(ii) there exists

(a) $k \in \mathbb{Z}_{>0}$*,*

(b) distinct $\lambda_j \in \mathsf{F}$*,* $j \in \{1, \dots, k\}$*,*

(c) $p_j \in \mathbb{Z}_{>0}$*,* $j \in \{1, \dots, k\}$*, and*

(d) $\mathbf{l}_j \in \mathbb{Z}_{>0}^{p_j}$*,* $j \in \{1, \dots, k\}$*,*

such that the elementary divisors of L *are the multiset*

$$\{(\xi - \lambda_1)^{l_{11}}, \dots, (\xi - \lambda_1)^{l_{1m_1}}, \dots, (\xi - \lambda_k)^{l_{k1}}, \dots, (\xi - \lambda_k)^{l_{km_k}}\}.$$

Moreover, if either statement holds then there exists a basis $\mathscr{B}$ *for* V *such that*

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \mathrm{diag}(\mathbf{J}(\lambda_1, \mathbf{l}_1), \dots, \mathbf{J}(\lambda_k, \mathbf{l}_k), \dots, \mathbf{J}(\lambda_k, l_{km_k})).$$

Finally, if

$$[\mathsf{L}]_{\mathscr{B}'}^{\mathscr{B}'} = \mathrm{diag}(\mathbf{J}(\mu_1, r_{11}), \dots, \mathbf{J}(\mu_1, r_{1n_1}), \dots, \mathbf{J}(\mu_s, r_{p1}), \dots, \mathbf{J}(\mu_p, r_{pn_p}))$$

is a matrix representative of L *in another basis, then* $p = k$ *and there exists a permutation* $\sigma \in \mathfrak{S}_k$ *such that* $\mu_j = \lambda_\sigma(j)$*,* $n_j = m_{\sigma(j)}$*, and* $r_{ja} = l_{\sigma(j)a}$ *for* $j \in \{1, \dots, k\}$ *and* $a \in \{1, \dots, m_j\}$*.*

Proof As with Theorem 5.8.44, it is immediate that (ii) implies (i). The converse follows from the lemma in the proof of Theorem 5.8.69, along with the nilpotent canonical form from Theorem 5.8.66. We leave to the reader the straightforward task of putting the pieces together. ■

5.8.15 The $\mathbb{R}$-Jordan canonical form

In this section we focus on endomorphisms of $\mathbb{R}$-vector spaces. If the eigenvalues of an endomorphism are real, then we can apply Corollary 5.8.72 to get a basis where the matrix representative is in Jordan canonical form. However, if there are complex eigenvalues, then the Jordan canonical will necessarily be complex. One can ask whether it is possible to obtain a *real* canonical form. Indeed it is, and we give this here.

The key idea to organise the discussion is the following.

5.8.73 Definition ($\mathbb{R}$-Jordan block) Let $\sigma, \omega \in \mathbb{R}$ with $\omega \neq 0$ and denote

$$B(\sigma, \omega) = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}.$$

(i) For $k \in \mathbb{Z}_{>0}$, the $2k \times 2k$-matrix

$$J(\sigma, \omega, k) \triangleq \begin{bmatrix} B(\sigma,\omega) & I_2 & 0_{2\times 2} & \cdots & 0_{2\times 2} & 0_{2\times 2} \\ 0_{2\times 2} & B(\sigma,\omega) & I_2 & \cdots & 0_{2\times 2} & 0_{2\times 2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \cdots & B(\sigma,\omega) & I_2 \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & \cdots & 0_{2\times 2} & B(\sigma,\omega) \end{bmatrix}$$

is the ***$\mathbb{R}$-Jordan block*** associated with k and $\sigma + i\omega \in \mathbb{C}$.

(ii) For $r \in \mathbb{Z}_{>0}$ and $\boldsymbol{k} = (k_1, \dots, k_r) \in \mathbb{Z}^r_{>0}$, the matrix

$$J(\sigma, \omega, \boldsymbol{k}) = \begin{bmatrix} J(\sigma,\omega,k_1) & 0 & \cdots & 0 \\ 0 & J(\sigma,\omega,k_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J(\sigma,\omega,k_r) \end{bmatrix}$$

is the ***Jordan arrangement*** associated with $\boldsymbol{k}$, and σ and ω. •

With this notation we can state the following canonical form for $\mathbb{R}$-endomorphisms.

5.8.74 Theorem ($\mathbb{R}$-Jordan canonical form) *Let* V *be a finite-dimensional* $\mathbb{R}$*-vector space. For* $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$ *suppose that* $\lambda_j \in \mathbb{R}$, $j \in \{1, \dots, m\}$, *and* $\sigma_j, \omega_j \in \mathbb{R}$, $\omega_j \neq 0$, $j \in \{1, \dots, r\}$, *are such that*

$$\lambda_1, \dots, \lambda_m, \sigma_1 + i\omega_1, \dots, \sigma_k + i\omega_k$$

are the distinct eigenvalues of L. *Then there exists* *(i)* $\mathrm{p_j} \in \mathbb{Z}_{>0}$, $\mathrm{j} \in \{1, \dots, \mathrm{m}\}$,

(ii) $\mathbf{k}_\mathrm{j} \in \mathbb{Z}^{\mathrm{p_j}}_{>0}$, $\mathrm{j} \in \{1, \dots, \mathrm{m}\}$,

(iii) $\mathrm{q_j} \in \mathbb{Z}_{>0}$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$,

(iv) $\mathbf{l}_\mathrm{j} \in \mathbb{Z}^{\mathrm{q_j}}_{>0}$, $\mathrm{j} \in \{1, \dots, \mathrm{k}\}$,

and a basis $\mathscr{B}$ *for* V *such that*

$$[\mathsf{L}]^{\mathscr{B}}_{\mathscr{B}} = \mathrm{diag}(\mathbf{J}(\lambda_1, \mathbf{k}_1), \dots, \mathbf{J}(\lambda_\mathrm{m}, \mathbf{k}_\mathrm{m}), \mathbf{J}(\sigma_1, \omega_1, \mathbf{l}_1), \dots, \mathbf{J}(\sigma_\mathrm{r}, \omega_\mathrm{r}, \mathbf{l}_\mathrm{k})).$$

Moreover, this form of the matrix representative is unique up to reordering of the diagonal blocks.

Proof Let us first consider the case where L has non-real eigenvalues $\lambda = \sigma + \mathrm{i}\omega$ and $\bar{\lambda} = \sigma - \mathrm{i}\omega$, possibly with multiplicity greater than 1. Thus

$$\mathsf{V}_\mathbb{C} = \overline{\mathsf{W}}(\lambda, \mathsf{L}_\mathbb{C}) \oplus \overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_\mathbb{C}).$$

Let $\mathscr{B} = \{u_1 + \mathrm{i}v_1, \dots, u_n + \mathrm{i}v_n\}$ be a basis for $\overline{\mathsf{W}}(\lambda, \mathsf{L}_\mathbb{C})$ for which the matrix representative of $\mathsf{L}_\mathbb{C}|\overline{\mathsf{W}}(\lambda, \mathsf{L}_\mathbb{C})$ is in Jordan canonical form. An application of the definition of matrix representative then shows that $\overline{\mathscr{B}} = \{u_1 - \mathrm{i}v_1, \dots, u_n - \mathrm{i}v_n\}$ is a basis $\overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_\mathbb{C})$ for which the matrix representative of $\mathsf{L}_\mathbb{C}|\overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_\mathbb{C})$ is in Jordan canonical form. Suppose that the first k basis vectors in $\mathscr{B}$ correspond to a $k \times k$ Jordan block in the matrix representative for $\mathsf{L}_\mathbb{C}|\overline{\mathsf{W}}(\lambda, \mathsf{L}_\mathbb{C})$ and that the first k basis vectors in $\overline{\mathscr{B}}$ correspond to a $k \times k$ Jordan block in the matrix representative for $\mathsf{L}_\mathbb{C}|\overline{\mathsf{W}}(\bar{\lambda}, \mathsf{L}_\mathbb{C})$. The subset

$$\{u_1 + \mathrm{i}v_1, u_1 - \mathrm{i}v_1, \dots, u_k + \mathrm{i}v_k, u_k - \mathrm{i}v_k\}$$

of basis vectors then spans a $2k$-dimensional $\mathsf{L}_\mathbb{C}$-invariant subspace. As was shown in Lemma 1 in the proof of Theorem 5.4.68, this subspace is the complexification of a L-invariant $\mathbb{R}$-subspace of V, and a basis for this $\mathbb{R}$-subspace is $\{u_1, v_1, \dots, u_k, v_k\}$. Let us determine the matrix representative of L restricted to this subspace. We have

$$\begin{aligned}\mathsf{L}_\mathbb{C}(u_1 + \mathrm{i}v_1) &= \mathsf{L}(u_1) + \mathrm{i}\mathsf{L}(v_1) = \lambda(u_1 + \mathrm{i}v_1) \\ &= \sigma u_1 - \omega v_1 + \mathrm{i}(-\omega u_1 + \sigma v_1)\end{aligned}$$

which gives

$$\mathsf{L}(u_1) = \sigma u_1 - \omega v_1, \quad \mathsf{L}(v_1) = -\omega u_1 + \sigma v_1.$$

For $j \in \{2, \dots, k\}$ we have

$$\begin{aligned}\mathsf{L}_\mathbb{C}(u_j + \mathrm{i}v_j) &= \mathsf{L}(u_j) + \mathrm{i}\mathsf{L}(v_j) = u_{j-1} + \mathrm{i}v_{j-1} + \lambda(u_j + \mathrm{i}v_j) \\ &= u_{j-1} + \sigma u_j - \omega v_j + \mathrm{i}(v_{j-1} - \omega u_j + \sigma v_j)\end{aligned}$$

which gives

$$\mathsf{L}(u_j) = u_{j-1} + \sigma u_j - \omega v_j, \quad \mathsf{L}(v_j) = v_{j-1} - \omega u_j + \sigma v_j.$$

Thus the matrix representative of L restricted to this subspace is $\boldsymbol{J}(\sigma, \omega, k)$. Doing this for all of the nilpotent blocks in the basis for $\mathsf{V}_\mathbb{C}$ gives the theorem in the special case.

The general case is proved by carrying out the above procedure for each of the complex generalised eigenspaces, and simply applying Corollary 5.8.72 for the real generalised eigenspaces. ■

We shall see in Section V-5.2.2 that this theorem is of great use in determining the nature of the solutions to linear ordinary differential equations.

Let us give a name to the canonical form of the theorem.

5.8.75 **Definition ($\mathbb{R}$-Jordan canonical form)** For a $\mathbb{R}$-vector space V and an endomorphism $\mathsf{L} \in \mathrm{End}_{\mathbb{R}}(\mathsf{V})$, the matrix representative

$$[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}} = \mathrm{diag}(\boldsymbol{J}(\lambda_1, k_1), \dots, \boldsymbol{J}(\lambda_m, k_m), \boldsymbol{J}(\sigma_1, \omega_1, l_1), \dots, \boldsymbol{J}(\sigma_r, \omega_r, l_r)).$$

of Theorem 5.8.74 is the $\mathbb{R}$-***Jordan canonical form***. •

5.8.16 A worked example

In this section we work out an example, computing the rational canonical form and the real Jordan canonical form. The intention is not so much to illustrate how one computes the canonical form; this is generally impossible and if it needs to be done can be done effectively by any one of the computational packages available. The intent is to show how some of the concepts in this section fit together. We shall omit all elementary linear algebra computations involving solving of linear equations. These are easily done using computer packages. Only insane people do such computations by hand anymore.

We take $\mathsf{V} = \mathbb{R}^6$ and let L be represented by the 6×6-matrix

$$\mathsf{L} = \begin{bmatrix} -1 & 0 & 0 & 0 & -2 & 0 \\ -2 & 2 & 0 & 1 & 0 & 0 \\ -2 & 1 & 2 & 1 & 0 & 0 \\ -6 & 0 & 0 & 2 & -4 & 0 \\ 2 & 0 & 0 & 0 & -1 & 0 \\ -2 & 0 & 0 & 0 & 3 & 2 \end{bmatrix}.$$

Characteristic and minimal polynomials

The characteristic polynomial is readily computed to be

$$\boxed{C_{\mathsf{L}} = \det(\xi\,\mathrm{id}_{\mathsf{V}} - \mathsf{L}) = \xi^6 - 6\xi^5 + 13\xi^4 - 24\xi^3 + 72\xi^2 - 128\xi + 80}.$$

Typically this is where the computation will stop, since one cannot effectively factor polynomials of degree greater than 4 (cf. Theorem 4.7.12). However, we have cooked this example to have nice eigenvalues, and these are the elements of the multiset

$$\{2, 2, 2, 2, -1 + 2\mathrm{i}, -1 - 2\mathrm{i}\}.$$

Thus

$$C_{\mathsf{L}} = (\xi - 2)^4(\xi^2 + 2\xi + 5)$$

since the roots of $\xi^2 + 2\xi + 5$ are $-1 \pm 2\mathrm{i}$. Now let us determine the minimal polynomial. We know that since the eigenvalues are roots of the characteristic polynomial and vice versa (Proposition 5.8.13), it must be the case that $M_{\mathsf{L}} = (\xi-2)^k(\xi^2+2\xi+5)$. All one can do is compute $\mathrm{Ev}_{\mathsf{F}}((\xi-2)^k(\xi^2+2\xi+5))(\mathsf{L})$ for various k and find the smallest k for which this expression is $0_{\mathrm{End}_{\mathsf{F}}(\mathsf{V})}$. We determine that this minimal k is 3 and so

$$\boxed{M_{\mathsf{L}} = (\xi-2)^3(\xi^2+2\xi+5)}.$$

Invariant subspaces

The key to most of the computations is knowing bases for the eigenspaces and generalised eigenspaces. These can be computed for L and for its complexification. The real invariant subspaces corresponding to the eigenvectors will be

$$\mathsf{W}_1 = \ker(\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}}), \quad \mathsf{W}_2 = \ker(\mathsf{L}^2+2\mathsf{L}+5\,\mathrm{id}_{\mathsf{V}}).$$

We determine these to be

$$\boxed{\begin{aligned}\mathsf{W}_1 &= \mathrm{span}_{\mathbb{R}}((0,0,1,0,0,0),(0,0,0,0,0,1)),\\ \mathsf{W}_2 &= \mathrm{span}_{\mathbb{R}}((1,0,0,2,0,0),(0,0,0,0,-1,1)).\end{aligned}}$$

The real invariant subspaces corresponding to the generalised eigenvectors will be

$$\overline{\mathsf{W}}_1 = \ker((\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}})^3), \quad \overline{\mathsf{W}}_2 = \ker(\mathsf{L}^2+2\mathsf{L}+5\,\mathrm{id}_{\mathsf{V}}).$$

We compute these to be

$$\boxed{\begin{aligned}\overline{\mathsf{W}}_1 &= \mathrm{span}_{\mathbb{R}}((0,1,0,0,0,0),(0,0,1,0,0,0),(0,0,0,1,0,0),(0,0,0,0,0,1)),\\ \overline{\mathsf{W}}_2 &= \mathrm{span}_{\mathbb{R}}((1,0,0,2,0,0),(0,0,0,0,-1,1)).\end{aligned}}$$

Let us determine the complex invariant subspaces. For the eigenvectors these are

$$\begin{aligned}\mathsf{W}_{1,\mathbb{C}} &= \ker(\mathsf{L}_{\mathbb{C}}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}}),\\ \mathsf{W}_{2,\mathbb{C}} &= \ker(\mathsf{L}_{\mathbb{C}}-(-1+2\mathrm{i})\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}}),\\ \mathsf{W}_{3,\mathbb{C}} &= \ker(\mathsf{L}_{\mathbb{C}}-(-1-2\mathrm{i})\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}}).\end{aligned}$$

Doing the computations gives

$$\boxed{\begin{aligned}\mathsf{W}_{1,\mathbb{C}} &= \mathrm{span}_{\mathbb{R}}((0,0,1,0,0,0),(0,0,0,0,0,1)),\\ \mathsf{W}_{2,\mathbb{C}} &= \mathrm{span}_{\mathbb{R}}((-\mathrm{i},0,0,-2\mathrm{i},-1,1)),\\ \mathsf{W}_{3,\mathbb{C}} &= \mathrm{span}_{\mathbb{R}}((\mathrm{i},0,0,2\mathrm{i},-1,1)).\end{aligned}}$$

For the generalised eigenspaces we have

$$\begin{aligned}\overline{\mathsf{W}}_{1,\mathbb{C}} &= \ker((\mathsf{L}_{\mathbb{C}}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})^3),\\ \overline{\mathsf{W}}_{2,\mathbb{C}} &= \ker(\mathsf{L}_{\mathbb{C}}-(-1+2\mathrm{i})\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}}),\\ \overline{\mathsf{W}}_{3,\mathbb{C}} &= \ker(\mathsf{L}_{\mathbb{C}}-(-1-2\mathrm{i})\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}}).\end{aligned}$$

For these we compute

$$\boxed{\begin{aligned}\overline{\mathsf{W}}_{1,\mathbb{C}} &= \operatorname{span}_{\mathbb{R}}((0,1,0,0,0,0),(0,0,1,0,0,0),(0,0,0,1,0,0),(0,0,0,0,0,1)),\\ \overline{\mathsf{W}}_{2,\mathbb{C}} &= \operatorname{span}_{\mathbb{R}}((-\mathrm{i},0,0,-2\mathrm{i},-1,1)),\\ \overline{\mathsf{W}}_{3,\mathbb{C}} &= \operatorname{span}_{\mathbb{R}}((\mathrm{i},0,0,2\mathrm{i},-1,1)).\end{aligned}}$$

Note that the basis for W_2 is obtained by taking the real and imaginary parts of the basis for $\mathsf{W}_{2,\mathbb{C}}$, just as in Theorem 5.4.68.

Elementary divisors and invariant factors

To determine L up to similarity, we determine its elementary divisors. As we have seen (for example, in Examples 5.8.16 and 5.8.37), the minimal polynomial does not necessarily determine the elementary divisors, and so does not necessarily determine L up to similarity. However, we do know that M_{L} is the least common multiple of its elementary divisors (this is Exercise 5.8.4). We also know that the product of the elementary divisors must be the characteristic polynomial (Proposition 5.8.35). In this case this allows us to determine the elementary divisors uniquely.

First of all, it must be the case that one of the elementary divisors must be $\xi^2 + 2\xi + 5$. The others must then be of the form $(\xi - 2)^l$ where the least common multiple of the l's is 3 and the l's sum to four. This means that we must have $l_1 = 1$ and $l_2 = 3$. Therefore, the elementary divisors are

$$\boxed{E_1 = \xi - 2, \quad E_2 = (\xi - 2)^3, \quad E_3 = \xi^2 + 2\xi + 5}.$$

The invariant factors are easily computed from these (cf. the proof of Theorem 4.9.21) to be

$$\boxed{D_1 = \xi - 2, \quad D_2 = (\xi - 2)^3(\xi + 2\xi + 5)}.$$

We shall not have any use for the invariant factors in what we do here, but give them just for fun.

Rational canonical form

From the elementary divisors and invariant factors above we can immediately write down the rational canonical form and the invariant factor canonical form. However, we also want the bases in which L have these canonical forms as matrix representative.

Let us first work with the rational canonical form. We will have one invariant subspace V_j for each elementary divisor E_j, $j \in \{1,2,3\}$, and the characteristic polynomial of L on V_j subspace will be E_j. Let us first work with the two elementary divisors $E_1 = \xi - 2$ and $E_2 = (\xi - 2)^3$. The invariant subspace corresponding to E_1 will be a one-dimensional subspace corresponding to some eigenvector for the eigenvalue 2. This one-dimensional subspace should be chosen from the two-dimensional subspace W_1. We know from the general theory (think about the Jordan canonical form) that this will be the unique one-dimensional subspace of W_1 that is complementary to image(L). One can verify that

$$\mathsf{V}_1 = \operatorname{span}_{\mathbb{R}}((0,0,0,0,0,1))$$

does the trick. For the elementary divisor E_2 we first take a complement in $\overline{\mathsf{W}}_1$ to V_1. A convenient such complement is

$$\mathsf{V}_2 = \operatorname{span}_{\mathbb{R}}((0,1,0,0,0,0),(0,0,1,0,0,0),(0,0,0,1,0,0)).$$

We know that L is cyclic restricted to this subspace. Following Theorem 5.8.20, in this subspace we seek a vector v_0 such that $\{v_0, \mathsf{L}(v_0), \mathsf{L}^2(v_0)\}$ is a basis for V_2. By Proposition 5.8.22, one should choose $v_0 \in \mathsf{V}_2$ that satisfies $(\mathsf{L} - 2\,\mathrm{id}_{\mathsf{B}})^2(v_0) \neq 0_{\mathsf{V}}$. The fact is that almost any $v_0 \in \mathsf{V}_2$ will do. However, it is also easy to choose vectors for which this will not work. In fact, of the three basis vectors we have written for V_2, the only one that will work is

$$v_0 = (0,0,0,1,0,0).$$

We compute

$$\mathsf{L}(v_0) = (0,1,1,2,0,0), \quad \mathsf{L}^2(v_0) = (0,4,5,4,0,0).$$

We now follow the proof of Proposition 5.8.24 to deduce a basis for V_2 for which $\mathsf{L}|\mathsf{V}_2$ is in companion form. We first denote

$$e_1 = \mathsf{L}^2(v_0), \quad e_2 = \mathsf{L}(v_0), \quad e_3 = v_0.$$

Note that

$$(\xi - 2)^3 = \xi^3 - 6\xi^2 + 12\xi - 8.$$

Then we define

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & 0 \\ -6 & 1 & 0 \\ 12 & -6 & 1 \end{bmatrix}$$

and

$$f_k = \sum_{j=1}^{3} \boldsymbol{T}(j,k) e_j, \qquad k \in \{1,2,3\},$$

so that

$$f_1 = (0,-2,-1,4,0,0), \quad f_2 = (0,1,1,-4,0,0), \quad f_3 = (0,0,0,1,0,0).$$

Now we need to choose a basis such that $\mathsf{L}|\mathsf{W}_2$ is in companion form. We need to choose $v_0 \in \mathsf{W}_2$ such that $\{v_0, \mathsf{L}(v_0)\}$ is a basis for W_2. In this case, because W_2 has no L-invariant subspaces, we can choose any $v_0 \in \mathsf{W}_2$. Let us take

$$v_0 = (1, 0, 0, 2, 0, 0).$$

Now we apply the procedure in the proof of Proposition 5.8.24 to arrive at a basis where $\mathsf{L}|\mathsf{W}_2$ is in companion form. We first define $e_1 = \mathsf{L}(v_0)$ and $e_2 = v_0$. We then take

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

and

$$f_k = \sum_{j=1}^{2} \boldsymbol{T}(k, j) e_j, \qquad k \in \{1, 2\}.$$

We compute

$$f_1 = (1, 0, 0, 2, 2, -2), \quad f_2 = (1, 0, 0, 2, 0, 0).$$

Now we can determine the rational canonical form by computing the matrix representative of L in the basis

$$\{(0, 0, 0, 0, 0, 1), (0, -2, -1, 4, 0, 0), (0, 1, 1, -4, 0, 0), \\ (0, 0, 0, 1, 0, 0), (1, 0, 0, 2, 2, -2), (1, 0, 0, 2, 0, 0)\}$$

This is done by defining $\boldsymbol{P} \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ by

$$\boldsymbol{P}^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 4 & -4 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & -2 & 0 \end{bmatrix},$$

i.e., by putting the basis vectors in the columns of a matrix, and taking the inverse. We then have the rational canonical form of L as

$$\boxed{\boldsymbol{P}\mathsf{L}\boldsymbol{P}^{-1} = \left[\begin{array}{c|ccc|cc} 2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 8 & -12 & 6 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -5 & -2 \end{array}\right]},$$

noting that L is the matrix representative of L in the standard basis.

Invariant factor canonical form

Next we determine the invariant factor canonical form, just for completeness. Here we will have two invariant subspaces, which we denote by V_1 and V_2, corresponding to the two invariant factors D_1 and D_2. We already determined a basis for V_1 above:

$$\mathsf{V}_1 = \operatorname{span}_{\mathbb{R}}((0,0,0,0,0,1)).$$

We also have

$$\mathsf{V}_2 = \operatorname{span}_{\mathbb{R}}\{(0,1,0,0,0,0),(0,0,1,0,0,0),(0,0,0,1,0,0),\\ (1,0,0,2,0,0),(0,0,0,0,-1,1)\}.$$

To determine a basis for which $\mathsf{L}|\mathsf{V}_2$ is in companion form, we first need to find a vector v_0 such that $\{v_0,\mathsf{L}(v_0),\dots,\mathsf{L}^4(v_0)\}$ is a basis for V_2. We take v_0 to be the sum of the vectors chosen above when we determined the companion for L associated to the elementary divisors E_2 and E_3:

$$v_0 = (0,0,0,1,0,0) + (1,0,0,2,0,0) = (1,0,0,3,0,0).$$

One can verify that

$$\{e_1 = \mathsf{L}^4(v_0),\dots,e_4 = \mathsf{L}(v_0), e_5 = v_0\}$$

is a basis for V_2. We then note that

$$D_2 = (\xi-2)^3(\xi^2+2\xi+5) = \xi^5 - 4\xi^4 + 5\xi^3 - 14\xi^2 + 44\xi - 40.$$

Following the proof of Proposition 5.8.24 we define

$$\boldsymbol{T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0\\ -4 & 1 & 0 & 0 & 0\\ 5 & -4 & 1 & 0 & 0\\ -14 & 5 & -4 & 1 & 0\\ 44 & -14 & 5 & -4 & 1 \end{bmatrix}$$

and

$$f_k = \sum_{j=1}^{5} \boldsymbol{T}(j,k)e_j.$$

This gives

$$\begin{aligned} f_1 &= (-8,-10,-5,4,-16,16),\\ f_2 &= (4,1,3,-4,24,-24),\\ f_3 &= (6,0,1,13,-12,12),\\ f_4 &= (-5,1,1,-12,2,-2),\\ f_5 &= (1,0,0,3,0,0). \end{aligned}$$

Now we compute the invariant factor canonical form by determining the matrix representative of L in the basis

$$\{(0,0,0,0,0,1),(-8,-10,-5,4,-16,16),(4,1,3,-4,24,-24),\\ (6,0,1,13,-12,12),(-5,1,1,-12,2,-2),(1,0,0,3,0,0)\}.$$

Thus we take

$$\boldsymbol{P}^{-1}=\begin{bmatrix}0&-8&4&6&-5&1\\0&-10&1&0&1&0\\0&-5&3&1&1&0\\0&4&-4&13&-12&3\\0&-16&24&-12&2&0\\1&16&-24&12&-2&0\end{bmatrix}$$

and compute the invariant factor canonical form to be

$$\boxed{\boldsymbol{P}\mathsf{L}\boldsymbol{P}^{-1}=\left[\begin{array}{c|ccccc}2&0&0&0&0&0\\ \hline 0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\\0&40&-44&14&-5&4\end{array}\right]}.$$

Complex Jordan canonical form

The computation of the Jordan canonical form is relatively straightforward. The key is determining basis where the nilpotent part of the transformation is in nilpotent canonical form. But this is straightforward, given Lemma 1 of Theorem 5.8.66.

We shall have four Jordan blocks in our canonical form corresponding to the elementary divisors E_1, E_2, and E_3; there are two Jordan blocks for E_3 since there are two distinct complex conjugate eigenvalues associated with this factor. This gives us four invariant subspaces of $\mathbb{C}^6$ whose bases we have already computed:

$$\begin{aligned}\mathsf{V}_1 &= \operatorname{span}_{\mathbb{C}}((0,0,0,0,0,1)),\\ \mathsf{V}_2 &= \operatorname{span}_{\mathbb{C}}((0,1,0,0,0,0),(0,0,1,0,0,0),(0,0,0,1,0,0)),\\ \mathsf{V}_3 &= \operatorname{span}_{\mathbb{C}}((-\mathrm{i},0,0,-2\mathrm{i},-1,1)),\\ \mathsf{V}_4 &= \operatorname{span}_{\mathbb{C}}((\mathrm{i},0,0,2\mathrm{i},-1,1)).\end{aligned}$$

The only one of these subspaces where we have to fuss about the nilpotent canonical form is V_2. According to Lemma 1 of Theorem 5.8.66 we should find a vector $v_0 \in \mathsf{V}_2$ such that $(\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})^2(v_0) \neq 0_{\mathsf{V}_{\mathbb{C}}}$. Such a vector is easily determined to be

$$v_0 = (0,0,0,1,0,0).$$

Then we are guaranteed that $(\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})|\mathsf{V}_2$ will be in nilpotent canonical form with respect to the basis $\{(\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})^2(v_0), (\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})(v_0), v_0\}$. We compute

$$(\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})(v_0) = (0,1,1,2,0,0), \quad (\mathsf{L}-2\,\mathrm{id}_{\mathsf{V}_{\mathbb{C}}})^2(v_0) = (0,4,5,4,0,0).$$

Now define

$$\boldsymbol{P}^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & -i & i \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2i & 2i \\ 0 & 0 & 0 & 0 & -1 & -1 \\ 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

We then have the complex Jordan canonical form as

$$\boxed{\boldsymbol{P}\mathsf{L}\boldsymbol{P}^{-1} = \left[\begin{array}{c|ccc|cc} 2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1+2i & 0 \\ 0 & 0 & 0 & 0 & 0 & -1-2i \end{array}\right]}.$$

Real Jordan canonical form

It is now straightforward to compute the real Jordan canonical form. We need only consider the basis formed by taking the real and imaginary parts of the basis vector for V_3. Thus we define

$$\boldsymbol{P}^{-1} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and compute

$$\boxed{\boldsymbol{P}\mathsf{L}\boldsymbol{P}^{-1} = \left[\begin{array}{c|ccc|cc} 2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 0 & -2 & -1 \end{array}\right]}.$$

5.8.17 Notes

Our proof of the Cayley–Hamilton Theorem in Theorem 5.8.19 follows Chisala [1998].

In [Bourbaki 1990, A.VII] the reader can find a discussion of semisimple endomorphisms for arbitrary vector spaces. Along with this discussion there is also included a discussion of the general notion of the Jordan canonical form. To properly allow for fields of nonzero characteristic, one must enter lightly into Galois theory.

Exercises

5.8.1 Let F be a field, let I be an index set, and let V be an F-vector space.

(a) Show that the relation of similarity in $\mathrm{Mat}_{I\times I}(\mathsf{F})$ is an equivalence relation.

(b) Show that the relation of similarity in $\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is an equivalence relation.

5.8.2 Prove Proposition 5.8.3.

5.8.3 Let F be a field with V an F-vector space. Show that there exists a unique isomorphism ι_{V} of $\mathsf{F}[\xi]\otimes\mathsf{V}$ with $\mathsf{V}[\xi]$ satisfying $\iota_{\mathsf{V}}(P\otimes v) = P\cdot v$, where

$$P\cdot v = a_k\xi^k\cdot v + \cdots + a_1\xi\cdot v + a_0 v,$$

if

$$P = a_k\xi^k + \cdots + a_1\xi + a_0.$$

5.8.4 Let F be a field and let V be a finite-dimensional F-vector space. Show that the minimal polynomial of L is the least common multiple of the elementary divisors of L.

5.8.5 Let F be a field and let V be a finite-dimensional F-vector space. For $\mathsf{L}\in\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ consider the dual endomorphism $\mathsf{L}'\in\mathrm{End}_{\mathsf{F}}(\mathsf{V}')$.

(a) Show that L and L' have the same minimal polynomial.

(b) Show that L and L' have the same characteristic polynomial.

5.8.6 Let F be a field, let V be a finite-dimensional F-vector space, and let $\mathsf{L}\in\mathrm{End}_{\mathsf{F}}(\mathsf{V})$. Recall from Definition 5.4.11 that $\langle\mathsf{L},\{v_0\}\rangle$ denotes the smallest L-invariant subspace containing v_0. Show that $\langle\mathsf{L},\{v_0\}\rangle$ is L-cyclic.

5.8.7 Let F be a field and let V be a finite-dimensional F-vector space. For $\mathsf{L}\in\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ consider the dual endomorphism $\mathsf{L}'\in\mathrm{End}_{\mathsf{F}}(\mathsf{V}')$. Show that L and L' have the same elementary divisors and invariant factors.

5.8.8 Let V be a two-dimensional $\mathbb{R}$-vector space and let $\mathsf{L}\in\mathrm{End}_{\mathbb{R}}(\mathsf{V})$.

(a) Write the characteristic polynomial of L in terms of $\det\mathsf{L}$ and $\operatorname{tr}\mathsf{L}$.

(b) Give the condition, expressed in terms of $\det\mathsf{L}$ and $\operatorname{tr}\mathsf{L}$, for L to have real eigenvalues.

5.8.9 Let F be a field and let $\mathsf{V} = \mathsf{F}_0^\infty$. Define $\mathsf{L}\in\mathrm{End}_{\mathsf{F}}(\mathsf{V})$ by $\mathsf{L}(e_j) = j_{\mathsf{F}}$ (recall that $j_{\mathsf{F}} = 1_{\mathsf{F}}+\cdots+1_{\mathsf{F}}$ denotes the j-fold sum of 1_{F}) , where $\{e_j\}_{j\in\mathbb{Z}_{>0}}$ is the standard basis.

(a) Give the primary decomposition of V_{L}.

(b) What is the “minimal polynomial” of L (using the definition that the minimal polynomial generates the ideal $\mathrm{ann}(\mathsf{V}_{\mathsf{L}})$)?

5.8.10 Let V be a finite-dimensional $\mathbb{R}$-vector space and let $\mathsf{L} \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{V};\mathsf{V})$ have the property that its minimal polynomial is irreducible over $\mathbb{R}$. Show that either $\mathsf{L} = a\,\mathrm{id}_{\mathsf{V}}$ for $a \in \mathbb{R}$ or that $\mathsf{L} = a\,\mathrm{id}_{\mathsf{V}} + b\mathsf{J}$, where $a, b \in \mathbb{R}$ with $b \neq 0$, and where $\mathsf{J} \in \mathrm{Hom}_{\mathbb{R}}(\mathsf{V};\mathsf{V})$ has the property that $\mathsf{J}^2 = -\,\mathrm{id}_{\mathsf{V}}$.

Hint: *A polynomial that is irreducible over* $\mathbb{R}$ *has the form* $\xi - \mathsf{a}$ *for* $\mathsf{a} \in \mathbb{R}$ *or the form* $(\xi - \mathsf{a})^2 + \mathsf{b}^2$ *for* $\mathsf{a}, \mathsf{b} \in \mathbb{R}$ *with* $\mathsf{b} \neq 0$ *(see Exercise 4.7.4).*

5.8.11 Let F be a field and let V be a finite-dimensional F-vector space. Show that $\mathsf{L} \in \mathrm{End}_{\mathsf{F}}(\mathsf{V})$ is triangularisable if and only if there exists a basis $\mathscr{B}$ for V such that $[\mathsf{L}]_{\mathscr{B}}^{\mathscr{B}}$ is lower triangular.

Bibliography

Bernstein, S. N. [1912] *Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités*, Communication de la Société Mathématique de Kharkov, **13**, pages 1–2.

Bliss, G. A. [1917] *A necessary and sufficient condition for the existence of a Stieltjes integral*, Proceedings of the National Academy of Sciences of the United States of America, **3**(11), pages 633–637, ISSN: 1091-6490, URL: http://www.jstor.org/stable/83517 (visited on 07/23/2014).

Bourbaki, N. [1990] *Algebra II*, Elements of Mathematics, Springer-Verlag: New York/Heidelberg/Berlin, ISBN: 978-3-540-00706-7.

Bridges, D. S. and Richman, F. [1987] *Varieties of Constructive Mathematics*, number 97 in London Mathematical Society Lecture Note Series, Cambridge University Press: New York/Port Chester/Melbourne/Sydney, ISBN: 978-0-521-31802-0.

Campoli, O. A. [1988] *A principal ideal domain that is not a Euclidean domain*, The American Mathematical Monthly, **95**(9), pages 868–871, ISSN: 0002-9890, DOI: 10.2307/2322908.

Chisala, B. P. [1998] *A quick Cayley–Hamilton*, The American Mathematical Monthly, **105**(9), pages 842–844, ISSN: 0002-9890, DOI: 10.2307/2589214.

Cohen, P. J. [1963] *A minimal model for set theory*, American Mathematical Society. Bulletin. New Series, **69**, pages 537–540, ISSN: 0273-0979, DOI: 10.1090/S0002-9904-1963-10989-1.

Dirichlet, J. P. G. L. [1842] *Verallgemeinerung eines Satzes aus der Lehre von den Kettenbrüchen nebst einigen Anwendungen auf die Theorie der Zahlen*, Bericht über die Verhandlungen der Königlich Preussischen Akademie der Wissenschaften, pages 93–95.

Gilmer, R. [1968] *A note on the algebraic closure of a field*, The American Mathematical Monthly, **75**(10), pages 1101–1102, ISSN: 0002-9890, DOI: 10.2307/2315743.

Gödel, K. [1931] *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme*, Monatshefte für Mathematik, **38**(1), pages 173–189, ISSN: 0026-9255, DOI: 10.1007/s00605-006-0423-7.

Gordon, R. A. [1998] *The use of tagged partitions in elementary real analysis*, The American Mathematical Monthly, **105**(2), pages 107–147, ISSN: 0002-9890, DOI: 10.2307/2589642.

Hamming, R. W. [1980] *The unreasonable effectiveness of mathematics*, The American Mathematical Monthly, **87**(2), pages 81–90, ISSN: 0002-9890, DOI: 10.2307/2321982.

Hobson, E. W. [1957] *The Theory of Functions of a Real Variable and the Theory of Fourier's Series*, Dover Publications, Inc.: New York, NY.

Hörmander, L. [1966] *An Introduction to Complex Analysis in Several Variables*, Van Nostrand Reinhold Co.: London, Reprint: [Hörmander 1990].

— [1990] *An Introduction to Complex Analysis in Several Variables*, 3rd edition, number 7 in North Holland Mathematical Library, North-Holland: Amsterdam/New York, ISBN: 978-0-444-88446-6, Original: [Hörmander 1966].

Krantz, S. G. and Parks, H. R. [2002] *A Primer of Real Analytic Functions*, 2nd edition, Birkhäuser Advanced Texts, Birkhäuser: Boston/Basel/Stuttgart, ISBN: 978-0-8176-4264-8.

Kronecker, L. [1899] *Werke*, volume 3, Teubner: Leipzig.

Kueh, K.-L. [1986] *A note on Kronecker's approximation theorem*, The American Mathematical Monthly, **93**(7), pages 555–556, ISSN: 0002-9890, DOI: 10.2307/2323034.

Lang, S. [2005] *Algebra*, 3rd edition, number 211 in Graduate Texts in Mathematics, Springer-Verlag: New York/Heidelberg/Berlin, ISBN: 978-0-387-95385-4.

Lewin, J. [1991] *A simple proof of Zorn's lemma*, The American Mathematical Monthly, **98**(4), pages 353–354, ISSN: 0002-9890, DOI: 10.2307/2323807.

McCarthy, J. [1953] *An everywhere continuous nowhere differentiable function*, The American Mathematical Monthly, **60**(10), page 709, ISSN: 0002-9890, DOI: 10.2307/2307157.

McDonald, B. R. [1984] *Linear Algebra over Commutative Rings*, Monographs and Textbooks in Pure and Applied Mathematics, Dekker Marcel Dekker: New York, NY, ISBN: 978-0-8247-7122-5.

Moore, G. H. [1982] *Zermelo's Axiom of Choice: Its Origins, Development, and Influence*, Springer-Verlag: New York/Heidelberg/Berlin, ISBN: 0-387-90670-3, Reprint: [Moore 2013].

— [2013] *Zermelo's Axiom of Choice: Its Origins, Development, and Influence*, Dover Publications, Inc.: New York, NY, ISBN: 978-0-486-48841-7, Original: [Moore 1982].

Motzkin, T. S. [1949] *The Euclidean algorithm*, American Mathematical Society. Bulletin. New Series, **55**, pages 1142–1146, ISSN: 0273-0979, DOI: 10.1090/S0002-9904-1949-09344-8.

Niven, I. [1947] *A simple proof that π is irrational*, American Mathematical Society. Bulletin. New Series, **53**, page 509, ISSN: 0273-0979, DOI: 10.1090/S0002-9904-1947-08821-2.

Pollard, S. [1920] *The Stieltjes integral and its generalizations*, The Quarterly Journal of Mathematics. Oxford. Second Series, **49**(10), pages 73–138, ISSN: 1464-3847.

Robinson, A. [1974] *Non-Standard Analysis*, Princeton Mathematical Series, Princeton University Press: Princeton, NJ, Reprint: [Robinson 1996].

— [1996] *Non-Standard Analysis*, Princeton Landmarks in Mathematics, Princeton University Press: Princeton, NJ, ISBN: 978-0-691-04490-3, Original: [Robinson 1974].

Siksek, S. and El-Sedy, E. [2004] *Points of non-differentiability of convex functions*, Applied Mathematics and Computation, **148**(3), pages 725–728, ISSN: 0096-3003, DOI: 10.1016/S0096-3003(02)00932-3.

Stillwell, J. [1994] *Galois theory for beginners*, The American Mathematical Monthly, **101**(1), pages 22–27, ISSN: 0002-9890, DOI: 10.2307/2325119.

Suppes, P. [1960] *Axiomatic Set Theory*, The University Series in Undergraduate Mathematics, Van Nostrand Reinhold Co.: London, Reprint: [Suppes 1972].

— [1972] *Axiomatic Set Theory*, Dover Publications, Inc.: New York, NY, ISBN: 978-0-486-61630-8, Original: [Suppes 1960].

Theory, A. N. [1987] *Ian N. Stewart and David O. Tall*, Chapman & Hall: New York/-London, ISBN: 978-0-412-29870-7.

Wigner, E. P. [1960] *The unreasonable effectiveness of mathematics in the natural sciences*, Communications on Pure and Applied Mathematics, **13**(1), pages 1–14, ISSN: 0010-3640, DOI: 10.1002/cpa.3160130102.

Young, W. H. [1913] *On integration with respect to a function of bounded variation*, Proceedings of the London Mathematical Society. Third Series, **13**(2), pages 109–150, ISSN: 0024-6115.