The Nature and Origins of Modern Mathematics: an Elementary Introduction

Andrew McLennan

School of Economics
University of Queensland
Level 6 Colin Clark Building
St Lucia, Qld 4072
Australia

mclennan@socsci.umn.edu

July 27, 2009

Preface

Contemporary mathematics is a very different thing from the mathematics of 150 years ago. To a certain extent this is simply because we know a lot more, but the more radical changes are transformations of the most fundamental concepts of the subject.

At a certain point in the 19th century mathematicians realized that set theory could be used to give exact descriptions of all the objects they worked with. The most obvious and immediate benefit is increased clarity and rigor, but that is far from the end of the story. The methods used to give precise definitions of existing concepts can also be used to define novel structures, and in the 20th century this led to the emergence of many entirely new fields of research. A bit more subtly, the axiomatic method based on set theory can be used to take a concept apart, to break it down into more fundamental elements, to recombine these elements, and ultimately to reformulate the original concept in ways that discard inessential aspects inherited from particular applications while retaining a critical core. This is the process of abstraction.

This book describes some of the resulting concepts. Up to a point its trajectory is quite similar to the mathematical curriculum at the secondary school and early university level: fundamentals of mathematical reasoning, basic facts about real numbers, continuity and convergence, some algebra, and then the calculus. Every idea had some predecessor in the mathematical thought of Sir Isaac Newton. But instead of thinking of these as a collection of problem-solving methods or "skills," we will be entirely concerned with viewing them as a system of interrelated definitions that combine to create a mathematics that is more general, unified, and powerful than anything Newton could have imagined. The final chapters use these concepts to develop geometric structures that go far beyond geometry as it was understood in the 18th century, but which are now fundamental in mathematics and physics.

These concepts are, in themselves, quite simple. Whatever difficulties they entail arise in two ways. First, understanding one of them is primarily a matter of seeing that the definition is a “correct” response to some need, or a more general version of another concept of proven value. We will emphasize relationships between the concepts, their historical origins, and certain fundamental results that “validate” them, but this is only the beginning of an accumulation of experience that generates an ever-evolving sense and appreciation of each of them. The second source of difficulties is that *truly* understanding these concepts means understanding their implications and larger consequences for mathematics. At present mathematical knowledge is exploding, and the distant future is unknowable; my own guess is that for at least a few more generations our understanding of these ideas will become increasingly incomplete. But while these thoughts should make professional mathematicians feel humble, the reader certainly isn’t expected to grapple with such profound mysteries.

This book is different from other books about mathematics you may have seen. It aims at a broad audience, and assumes very little in the way of prior mathematical background. It is, I hope, particularly well suited for high school and college students who are interested in mathematics but have a hard time finding books that don’t assume background they lack. At the same time it is a book *of* mathematics, with formal definitions, theorems, and proofs, rather than a book *about* what mathematics “is like,” or what it aims to accomplish, or the biographies of mathematicians. Even though it treats some subjects that are usually thought to be advanced, it is in various ways easier reading than other math books, focusing on the simplest aspects of each topic, with thorough, detailed, and gentle explanations of the steps in each argument. There is very little in the way of “gotcha” cleverness. It rambles a bit, so that not that much needs to be carried forward from one chapter to the next. I would hope that it is accessible to anyone of normal intelligence who approaches it with patience, going slowly enough to really understand each new idea and argument, and with sincere interest. I have tried in various ways to make it a bit more entertaining, in the everyday sense, than scholarly math books which presume a readership of addicts like myself. But your mileage may vary.

In spite of everything, there will almost certainly be times when what you’re reading doesn’t seem to make sense. The first thing to do is to retrace your mental steps and try to puzzle things out; usually a confusion has its roots in an earlier misapprehension of some small detail. But sometimes that won’t work, at which point it would be best to ask someone. Contrary to what your teachers may have told you about the only stupid questions being those that are unasked, *almost all your questions will be dumb.* Really dumb. Everything here is simple, once you've seen it in the proper light, so most likely the answers to your questions will make you look, and feel, like a goofball. Try to have a thick skin about this; it happens to everyone.

There is also a lot of terminology. The first time a technical term appears it will be in **bold face**, either in a definition or (more commonly) with some precise, but less formal, explanation of its meaning. There will probably be many times when you encounter a term you don't recall, or whose meaning you don't recall, or recall vaguely. Even if you are just a little bit unsure, it's a good idea to review the definition. (You should be able to find it quickly by looking in the index.) This is largely a book about definitions, and the perspective on mathematics they embody and express; using the jargon freely once it's been introduced is an important strategy for reinforcing your familiarity and understanding. This may seem harsh, but it's a bit like learning a foreign language: a good instructor conducts the course entirely in that language from a very early stage.

Almost certainly you already know that mathematical notation makes heavy use of the Greek alphabet. But you might see a few Greek letters that are new to you, and it helps to know how they are pronounced, and a bit about how the Greek and Roman alphabets are related. If, instead of trying to memorize the Greek alphabet now, you keep a reference handy (perhaps bookmarked in your browser) and look up unfamiliar letters as you go along, you'll learn everything you need to know effortlessly.

Will reading this book help you get good grades in math courses?

I started writing it in response to frustrations I felt in teaching a course that is usually called Mathematics for Economists. In part because the amount of time in a one semester course is very short, in part because students primarily want to know how to solve the problems that will determine their grades, and in part because the problem solving techniques developed in this course are inputs to other courses in economics, I felt, and largely succumbed to, pressure to focus on the "how to do it" aspects of the subject, shortchanging the conceptual underpinnings. Some of my students were excellent, but many were woeful products of years of precisely this sort of instruction, with an almost transcendental inability to deal with the subject matter that went far beyond any lack of native intelligence. In actual fact many of them were at least as smart as most people, or significantly smarter. And no normal human being is truly so stupid as to be incapable of understanding elementary mathematics, which is much simpler than many commonplace aspects of everyday life.

Imagine a piano student who works for years with scales, triads, and exercises, but who never plays or hears any actual music. It would be a hopeless struggle to "remember" information that had been stripped of any meaning. Now suppose that one day this person attends a concert. The next day her "skills" would be no different than they had been before, but she could begin to practice in an entirely different manner, especially if she continued to listen to music, and began to play real music herself. My highest hope for this book is that in some readers it will trigger such a process, but it is a starting point, not a cure.

There is a much simpler answer to the question above: kittens that like to play inevitably turn into cats who know how to catch mice. Professional mathematicians are primarily motivated by intellectual stimulation and aesthetic pleasure; the "unreasonable effectiveness" of mathematics as a tool for dealing with the world is, for them, an unintended side effect. The last chapter recommends several other books that convey this sense of mathematics to less experienced readers, and there's no reason not to start exploring them now.

Contents

Chapter 1

What Mathematics Is

1.1 Why Read This Book?

> *You enter the first room of the mansion and it's completely dark. You stumble around bumping into the furniture, but gradually you learn where each piece of furniture is. Finally, after six months or so, you find the light switch, you turn it on, and suddenly it's all illuminated. You can see exactly where you were. Then you move into the next room and spend another six months in the dark.*
>
> –Andrew Wiles

Does this sound like you taking a math course? Maybe you're thinking "Yeah, except for the part about how 'suddenly it's all illuminated.' " First of all, you're in excellent company: Andrew Wiles (b. 1953) is describing his experience working out the proof of what was until then the most famous unresolved conjecture in mathematics. For everyone, especially the best, new mathematics is baffling until you understand it, but once you really understand it, it's simple.

Unlike most books, which focus on one nut or bolt at a time, this book proceeds with a lighter touch, aiming to first give you a sense of the overall structure of mathematics, its methods, and its larger agenda. One concrete purpose is to serve as a supplemental reading for students taking courses in advanced calculus and real analysis. A supplemental reading should do something different from the course's main text, and if I had to summarize the difference in just a few words, I would say that whereas textbooks conceive of the learning process as work, this book is meant to be entertaining.

Like a popular book about science, it takes a broader (and somewhat superficial) view of the material, emphasizing certain key concepts as they stand in relation to each other and the larger goals of the subject. Following a pedagogical method that is standard in physics, but rather uncommon in mathematics, the historical development of these ideas is used to portray them as creative responses to the problems and opportunities of the eras in which they were created, not just static facts devoid of drama.

For many people a course in real analysis is their first exposure to "higher" mathematics, with clear axiomatic foundations and rigorous justifications of all assertions. Prior to this point, mathematics may have seemed like a grabbag of algorithms for performing various calculations, but real analysis and subsequent courses are primarily concerned with theorems and proofs. Computations are still important, but if you really understand the logic of the material, you should be able to figure out how to compute when the need arises, and merely knowing how to compute is no substitute for real understanding. There is an undeniable sense in which this is a "harder" kind of mathematics—you are asked to perform (by your textbook, but not here!) at a higher level—but knowing *exactly* why things are the way they are is in many ways, in the end, the simplest and easiest approach.

An initial acquaintance with the concepts described here doesn't require a huge effort. In order to understand this book you have to be able to follow a logically compelling argument, like a juror at a trial, but there are only a few algebraic computations that could be described as complex. While some of the proofs have surprising aspects, they are mostly straightforward, and there is nothing terribly deep or complex here. This book is not meant to be "studied," and you are not expected to do lots of problems as you go through it. The problems at the end were added as an afterthought. Possibly you'll enjoy them—many introduce interesting concepts and results—and working a few for each chapter may help reinforce and consolidate your understanding, but it's up to you. The main text was written with the expectation that you'll simply be reading.

And it is all breathtakingly beautiful. The concepts described in this book are among the greatest contributions to science ever, as important as evolution and relativity in making the last couple centuries a watershed in human affairs. Results such as the fundamental theorem of algebra and the existence of non-Euclidean geometries, which stymied the most talented mathematicians for decades or centuries, are made accessible, not just to experts, but to beginners.

The rigorous, proof-oriented approach to mathematics requires exact axiomatic foundations, which means going back to the very beginnings of the subject and redeveloping everything from scratch. For some readers this will mean that the early parts of the book, especially this chapter and the next, are largely review, but the point of view will probably be somewhat unfamiliar, and there are a few advanced ideas and intriguing tangents to spice things up.

Going back to the very beginning also means that this book is *in principle* accessible to readers who have only advanced as far as high school algebra. Whether this is true *in practice* depends on the reader's ability and motivation. If you're a secondary school student (or a curious layperson) who enjoys mathematics, and you think you're reasonably good at it, by all means give this book a try! When I was young I was frustrated by the paucity of books that were accessible, and aimed at letting me advance in the subject by reading on my own. If you feel the same way, you're the sort of reader I have had in my mind's eye.

Especially if you have less preparation, you'll probably find that this book is much more tiring than other kinds of reading. With a book of crossword puzzles the usual pace would be one or two a day, or maybe three if you felt exceptionally enthusiastic. Thinking about each section here as a puzzle, to be solved and savored, will put you on a good pace. Learning mathematics reprograms the mind at a deep level, and this is a process that takes time, and sleep. So read slowly, trying to fully understand each step, and take some time out to digest before going on to the next section or chapter. As you reflect on what you've read, you'll often notice some new connection or unexpected perspective.

Reading a book about rigorous mathematics is a bit like walking a tightrope—if you don't correctly understand something, subsequent material quickly becomes confusing or impenetrable—and perhaps most readers won't make it to the end. This happens to all readers of math books at all levels, and you shouldn't feel bad about it. Possibly you'll pick up where you left off a couple weeks, or a couple years, later, with your prior confusion clarified. But even if you don't, by going as far as you could you will have succeeded in pushing your understanding to a new level.

1.2 What Is "Doing" Mathematics?

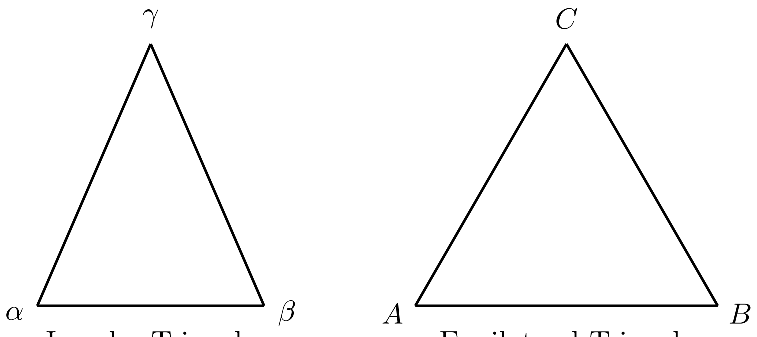
I recently read some autobiographical comments by a mathematician who said that when he was a high school student, the idea that new mathematics was being created in the present day would have seemed as bizarre to him as imagining that professors of English sat around making up new words.

One can hardly imagine learning about music without quickly realizing that it is something that people compose and perform, but in elementary and secondary school mathematics is often presented as an entirely impersonal collection of true facts and computational methods. Computational questions are used in math courses to test the student's facility, and too often students (and their instructors!) mistakenly believe that the ability to solve such problems is the *goal* of the course. If you just practice the methods of doing standard problems you might scrape by, but the real point of learning mathematics is to develop the ability to understand logical and quantitative arguments, and to create original, valid arguments yourself. Before anything else, one should have a sense of what people are trying to do when they "do" mathematics.

It seems that elementary mathematics emerged in the ancient world as a response to practical problems such as keeping accounts or measuring land. As knowledge accumulated and became more voluminous, presumably there arose a desire to make it more systematic. All this is pretty murky and speculative, but what we do know for sure is that this process led eventually to the discovery of the **axiomatic method**, as embodied in Euclid's (325-265 BC) *Elements*.

In the axiomatic method a substantial body of knowledge is organized as a combination of a small number of fundamental propositions, called **axioms**, that are taken as given, and a large number of logical consequences. It is a fundamental method of all of science (not just mathematics) for at least three reasons. First, it simplifies, clarifies, and organizes everything. In your own study of mathematics you should aim at capturing the psychological benefits of the axiomatic approach by organizing your own knowledge around first principles and deductive methods. Second, the distinction between assumptions and logical inferences is critical in science: assumptions are open to doubt, but logic is not, so if a conclusion seems dubious or downright counterfactual, some assumption has to be modified.

Most important, though, is that axiomatic organization of scientific knowledge almost always suggests a host of specific questions and general directions for further research. The explosive growth of scientific knowledge during the last few centuries is, in large part, the natural consequence of a relentless pursuit of what seem, after logical organization of existing knowledge, to be the most fundamental unresolved issues.



Isoceles Triangle Equilateral Triangle

Figure 1.1

For a concrete illustration of how this might happen (the actual historical process was much more complicated) we’ll consider the notion of **symmetry**. We’ll begin with the observation that, in some vague sense, an equilateral triangle is “more symmetric” than an isoceles triangle. Just what is this “symmetry” thing, of which there might be more or less in any particular instance? Well, the general idea of symmetry has a huge number of applications, and one thing they all have in common is interchange of various elements of the object under consideration. Concretely, one can place a copy of an equilateral triangle on top of the original triangle in six different ways that preserve the distances between all vertices, but for an isoceles triangle there are only two ways to do this.

So, it seems that a symmetry of an equilateral triangle can be represented by a function[1] mapping vertices to vertices, say $\sigma(A) = B$, $\sigma(B) = C$, and $\sigma(C) = A$. It seems natural to ask about the properties of such mappings, and it seems evident that a mapping representing a symmetry should be a **bijection**: exactly one element of its domain is mapped to each element of its range[2].

[1] The discussion below assumes you already know what sets and functions are. If you don’t, you should first read the description of these concepts at the beginning of Section 1.4.

[2] As you may already know, this property is usually broken down into two parts. A function $f : X \to Y$ is **one-to-one**, or **injective**, or an **injection**, if, for each element y of the range Y, there is at most one x in X such that $f(x) = y$. It is **onto**, or **surjective**, or a **surjection**, if, for each $y \in Y$ there is at least one $x \in X$ such that $f(x) = y$. For a function between two finite sets with the same number of elements, or from a finite set to itself, these two conditions amount to the same thing.

Actually, a mathematician is inclined to ask a somewhat different question: what are the properties of the *collection* of all mappings representing symmetries of a given object? It turns out that there are three key properties. First, the **identity function** should always be a symmetry. For any set X, Id_X will denote the mapping that takes each element of X to itself. Thus:

$$\mathrm{Id}_{\{A,B,C\}}(A) = A, \quad \mathrm{Id}_{\{A,B,C\}}(B) = B, \quad \mathrm{Id}_{\{A,B,C\}}(C) = C.$$

Second, the composition of any two mappings representing symmetries should, in turn, be a mapping representing a symmetry. In general, if f maps the set X into the set Y, and g maps the set Y into the set Z, then the **composition** of f and g is the mapping $g \circ f$ of X into Z that takes each $x \in X$ to $g(f(x))$. Suppose $\tau(A) = A$, $\tau(B) = C$, and $\tau(C) = B$. Then (as you should verify for yourself) $\tau \circ \sigma$ takes A to C, B to itself, and C to A. Third, the **inverse** of any mapping representing a symmetry should represent a symmetry. If f is a bijection mapping X into Y, then f^{-1} is the unique mapping of Y into X satisfying $f^{-1} \circ f = \mathrm{Id}_X$. So, σ^{-1}, which takes A to C, B to A, and C to B, should represent a symmetry.

Now let's consider a different situation exhibiting symmetries. Alice, Bob, and Carol play the following game. Each takes a black pawn and a white pawn behind his or her back, then, without letting the other players see, brings forward a hand containing a single pawn. If all three players chose the black pawn, or they all chose the white pawn, then no money changes hands. If two chose black and one chose white, or if two chose white and one chose black, then the two who chose the same color each pay $1 to the third player. Evidently this game is symmetric insofar as the rules are "invariant" under any bijective mapping of the set $\{\mathrm{Alice}, \mathrm{Bob}, \mathrm{Carol}\}$ to itself.

The point of this example is that the relevant symmetries on the set $\{A, B, C\}$ of vertices of an equilateral triangle are, in some obvious but as yet unexpressed sense, "the same" as the relevant symmetries on the set $\{\mathrm{Alice}, \mathrm{Bob}, \mathrm{Carol}\}$ of players of this game. The technique modern mathematics uses to capture such notions is *abstraction*: we define a new type of object that embodies the common features of these two symmetric situations while discarding the aspects that are particular to one or the other of the two examples.

Sometimes this process goes further than one might expect, arriving at definitions that can seem completely baffling if one has not already traced through the process that led to them. The definition below might seem mystifying if you didn't know that it is based on two further observations about composition of functions. First, if $f : X \to Y$ is a function, then

$$f \circ \mathrm{Id}_X = f = \mathrm{Id}_Y \circ f.$$

Second, composition is **associative**: if $f : X \to Y$, $g : Y \to Z$, and $h : Z \to W$ are functions, then for any x the three steps in figuring out what $h(g(f(x)))$ is can be thought of as the result of combining pairwise compositions in two different ways, but the grouping doesn't affect the result:

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

Here is our first main concept, which will come up again and again.

Definition 1.1. *A* **group** *is a set G with a binary operation (that is, a function taking ordered pairs of elements of G to elements of G, that we will write using the notational conventions of multiplication) satisfying the following conditions:*

(a) The operation is associative: $g(g'g'') = (gg')g''$ for all $g, g', g'' \in G$.

(b) There is an $e_G \in G$, called the ***identity element****, such that*

$$e_G g = g = g e_G$$

for all $g \in G$.

(c) For each $g \in G$ there is an ***inverse*** *$g^{-1} \in G$ such that*

$$gg^{-1} = e_G = g^{-1}g.$$

The set of all bijections from $\{A, B, C\}$ to itself is a group, as is the set of all bijections from $\{\text{Alice}, \text{Bob}, \text{Carol}\}$ to itself. If we want to emphasize that these are really the same group we can proceed as follows. First, for any positive integer n we define S_n to be the set of all bijections from $\{1, \ldots, n\}$ to itself. This is called the **symmetric group** on $\{1, \ldots, n\}$; since compositions and inverses of bijections are bijections, it clearly satisfies the three conditions above. An **action** of a group G on a set X is a function taking each pair (g, x) in which $g \in G$ and $x \in X$ to an element of X, which we denote by gx. This function must have the following properties:

(i) $e_G x = x$ for all $x \in X$;

(ii) $g(g'x) = (gg')x$ for all $g, g' \in G$ and $x \in X$.

In the first action we have in mind an element of S_n interchanges the elements of $\{A, B, C\}$ in "the same way" that it interchanges the elements of $\{1, 2, 3\}$. For example, if $\sigma(1) = 1$, $\sigma(2) = 3$, and $\sigma(3) = 2$, then $\sigma A = A$, $\sigma B = C$, and $\sigma C = B$. The action of S_3 on $\{\mathrm{Alice}, \mathrm{Bob}, \mathrm{Carol}\}$ is defined analogously.

The action is said to be **faithful** if e_G is the only element of G that induces the identity function on X. That is, if $gx = x$ for all $x \in X$, then $g = e_G$. We can now sum it all up quite succinctly by saying that in the two situations described above S_3 acts faithfully on $\{A, B, C\}$ and $\{\mathrm{Alice}, \mathrm{Bob}, \mathrm{Carol}\}$ respectively.

This all seems pretty simple, and the ancient Greeks clearly knew about symmetry, and were interested in it, so you might guess that these definitions have been around for at least 2500 years, but you would be wrong. The concept of a group is only about 200 years old. It's not so easy to psychoanalyze the failings of our forebears, but one can suggest three factors to account for this.

First, until recently people have generally believed that mathematical objects are, in some sense, already "out there." In ancient Greece the **Pythagorean school** of philosophy held that all numbers are rational, i.e., ratios of integers. Note the denigration suggested by this terminology: even if other numbers do exist, they're kind of nutty. Possibly you were taught, at a certain age, that -1 doesn't have a square root, then later you learned that it does, sort of, except that $\sqrt{-1}$ is "imaginary" and not "real."

Probably more important than such prejudices, which mathematicians would have been happy to overcome, even if the solid citizens were dubious, is the fact that the technology for constructing new mathematical objects using set theory is a recent development. We'll say much more about this in a little bit.

Finally, even though the central definitions of mathematics are, in the end, simple, the historical process that created them wasn't. Nowadays there are thousands of new definitions proposed every year, in the course of mathematicians doing their work. For specific, well defined projects this can be straightforward, but the biggest concepts in mathematics, such as symmetry, or number, or geometry, are studied continually for centuries on end, and the formulations of them that are most popular in any period are obtained by recrafting earlier approaches to fit the applications of greatest current interest. Although symmetries have been around in various forms for a long time, certain aspects of their importance became apparent during the first part of the 19th century. The definitions above, and the overall way we think about groups, emerged gradually during the next fifty or one hundred years. As you probably noticed, by themselves these definitions don't really say anything about symmetry that you didn’t know beforehand, and while they might seem simple and natural, we haven’t yet presented any particular reason to think they’ll be useful or interesting.

Much of the recent growth of mathematics is the result of a more or less automatic consequence of abstraction: a new concept, say the notion of a group, is introduced because it is relevant and illuminating in its application to preexisting mathematical phenomena, but it then becomes an object of study in itself. To give some flavor of this, and to introduce ideas that will recur in different contexts later, I am going to quickly state some of the basic definitions and results of group theory. The material below is written in the terse, “just the facts,” style of modern mathematics, but please don’t be intimidated. At this point you can’t possibly “truly understand” either the importance or the ramifications of what follows, and you shouldn’t expect to comprehend it “fully.” (Realistically, you’ll probably need to review it, possibly more than once, as the concepts are applied later.) Just read it slowly and carefully, trying to see that it makes sense on its own terms.

Let G be a group. One important fact about groups is that if $g, \varepsilon \in G$ satisfy $g = g\varepsilon$, then

$$\varepsilon = e_G\varepsilon = (g^{-1}g)\varepsilon = g^{-1}(g\varepsilon) = g^{-1}g = e_G.$$

A symmetric argument shows that $\varepsilon = e_G$ whenever $g = \varepsilon g$. That is, there is only one element of G that acts like e_G, in connection with any element of G, from either side. Another important fact is that $g' = g^{-1}$ whenever $gg' = e_G$ because

$$g^{-1} = g^{-1}e_G = g^{-1}(gg') = (g^{-1}g)g' = e_Gg' = g'.$$

Again, a symmetric argument implies that $g' = g^{-1}$ whenever $g'g = e_G$. Thus, for each $g \in G$, there is only one element of G that acts, from either side, like g^{-1}. In particular, each $g \in G$ is the inverse of its inverse:

$$(g^{-1})^{-1} = g.$$

Now let H be a second group. A **homomorphism** from G to H is a function $\varphi : G \to H$ that “respects” or “commutes with” the group operations:

$$\varphi(gg') = \varphi(g)\varphi(g')$$

for all $g, g' \in G$. It is always the case that $\varphi(e_G) = e_H$ because

$$\varphi(e_G) = \varphi(e_G)\varphi(e_G)\varphi(e_G)^{-1} = \varphi(e_Ge_G)\varphi(e_G)^{-1} = \varphi(e_G)\varphi(e_G)^{-1} = e_H.$$

For any $g \in G$ we have $\varphi(g^{-1}) = \varphi(g)^{-1}$ because

$$\varphi(g)\varphi(g^{-1}) = \varphi(gg^{-1}) = \varphi(e_G) = e_H.$$

A homomorphism φ is said to be an **isomorphism** if it is bijective, in which case we say that G and H are **isomorphic**. The most basic and important fact about isomorphisms is that the inverse of an isomorphism is also an isomorphism. To prove this we first note that the inverse of any bijection is a bijection, so the key point is that the inverse of an isomorphism is a homomorphism. Take two elements h and h' of H, set $g := \varphi^{-1}(h)$ and $g' := \varphi^{-1}(h')$, and compute that

$$\varphi^{-1}(hh') = \varphi^{-1}(\varphi(g)\varphi(g')) = \varphi^{-1}(\varphi(gg')) = gg' = \varphi^{-1}(h)\varphi^{-1}(h').$$

By the way, the symbol ':=' is the **assignment operator**. In most mathematics books it's written as '=', leaving the reader to determine from the context whether the sentence in question is an assertion that two things are equal or a definition of the thing on the left as a symbol whose meaning is the thing on the right.

If $f : X \to Y$ is a function and $f(x) = y$, then we say that y is the **image** of x under f, and that x is a **preimage** of y. Another important point about notation is that whenever $f : X \to Y$ is a function and $B \subset Y$, $f^{-1}(B)$ denotes the set $\{ x \in X : f(x) \in B \}$ of preimages of elements of B. This makes sense regardless of whether f is invertible, in the sense of being one-to-one and onto. Usually we'll write $f^{-1}(y)$ in place of the more cumbersome $f^{-1}(\{y\})$ when $y \in Y$, but now you have to be careful: if f is invertible, $f^{-1}(y)$ will typically denote the *element* of X that is mapped to y by f, and otherwise it denotes the *set* of preimages of y.

An isomorphism from a group to itself is called an **automorphism**. That Id_G is an automorphism is a simple and obvious, but crucially important, fact. There are automorphisms that are called **inner automorphisms** because they come from the group itself: for any $\gamma \in G$ let $C_\gamma : G \to G$ be the function that takes $g \in G$ to $C_\gamma(g) = \gamma g \gamma^{-1}$. This is a homomorphism because

$$C_\gamma(gg') = \gamma gg'\gamma^{-1} = \gamma g e_G g'\gamma^{-1} = \gamma g\gamma^{-1}\gamma g'\gamma^{-1} = C_\gamma(g)C_\gamma(g')$$

for all $g, g' \in G$, and $C_{\gamma^{-1}}$ is the inverse of C_γ (please convince yourself that this is so) so C_γ is an automorphism. Note that $\mathrm{Id}_G = C_{e_G}$ is an inner automorphism. An automorphism that isn't inner is called an **outer automorphism**.

A **subgroup** of G is a subset $G' \subset G$ such that:

(i) $e_G \in G'$;

(ii) $g^{-1} \in G'$ whenever $g \in G'$;

(iii) $gg' \in G'$ whenever $g, g' \in G'$.

That is, a subgroup of G is a subset containing e_G that is "closed" under inversion and the group operation, in the sense that $\{\, g^{-1} : g \in G' \,\}$ and $\{\, gg' : g, g' \in G' \,\}$ are both contained in G'. (Since $(g^{-1})^{-1} = g$ for all g, the first set is actually equal to G', and the second set is equal to G' because G' contains e_G.) Observe that G itself and $\{e_G\}$ are always subgroups. (That is, please check (i)-(iii) in your head.) To a large extent group theory regards the "structure" of a group as synonymous with its system of subgroups.

Let $\varphi : G \to H$ be a homomorphism. The following argument shows that if H' is a subgroup of H, then

$$\varphi^{-1}(H') := \{\, g \in G : \varphi(g) \in H' \,\}$$

is a subgroup of G. Above we showed that $\varphi(e_G) = e_H$, so $e_G \in \varphi^{-1}(H')$. If $g \in \varphi^{-1}(H')$, then $g^{-1} \in \varphi^{-1}(H')$ because

$$\varphi(g^{-1}) = \varphi(g)^{-1} \in \{\, h^{-1} : h \in H' \,\} = H'.$$

If $g, g' \in \varphi^{-1}(H')$, then $gg' \in \varphi^{-1}(H')$ because

$$\varphi(gg') = \varphi(g)\varphi(g') \in \{\, hh' : h, h' \in H' \,\} = H'.$$

The **kernel** of φ is

$$\ker(\varphi) := \varphi^{-1}(e_H).$$

Since $\{e_H\}$ is a subgroup of H, $\ker(\varphi)$ is a subgroup of G, but it turns out that not every subgroup can be the kernel of a homomorphism. A **normal subgroup** of G is a subgroup N such that $C_\gamma(g) \in N$ whenever $g \in N$ and $\gamma \in G$. If $g \in \ker(\varphi)$ and γ is any element of G, then $C_\gamma(g) \in \ker(\varphi)$ by virtue of the calculation

$$\begin{aligned}\varphi(C_\gamma(g)) = \varphi(\gamma g \gamma^{-1}) &= \varphi(\gamma)\varphi(g)\varphi(\gamma^{-1}) = \varphi(\gamma)e_H\varphi(\gamma^{-1}) \\ &= \varphi(\gamma)\varphi(\gamma^{-1}) = \varphi(\gamma)\varphi(\gamma)^{-1} = e_H.\end{aligned}$$

Thus the kernel of φ is a normal subgroup of G.

Here's an example of a subgroup that is not normal. Let $\sigma \in S_3$ be the function that takes 1 to 2, 2 to 1, and 3 to itself. Then $\sigma^{-1} = \sigma$, so $G' := \{\, \mathrm{Id}_{\{1,2,3\}}, \sigma \,\}$ obviously contains all products and inverses of its elements and is consequently a subgroup of S_3. If $\gamma \in S_3$ takes 1 to itself, 2 to 3, and 3 to 2, then

$$\gamma = \begin{cases} 1 \to 1 \\ 2 \to 3 \\ 3 \to 2, \end{cases} \quad \sigma = \begin{cases} 1 \to 2 \\ 2 \to 1 \\ 3 \to 3, \end{cases} \quad \text{and } \gamma^{-1} = \begin{cases} 1 \to 1 \\ 2 \to 3 \\ 3 \to 2, \end{cases} \quad \text{so} \quad C_\gamma(\sigma) = \begin{cases} 1 \to 3 \\ 2 \to 2 \\ 3 \to 1. \end{cases}$$

(It doesn't matter in this particular instance because $\gamma^{-1} = \gamma$, but our notation for compositions lead to compositions like the one above being computed by reading from right to left, first finding the effect of γ^{-1}, then the subsequent effect of σ, and finally the effect of γ.) Since it does not contain $C_\gamma(\sigma)$, G' is *not* a normal subgroup of S_3.

Everything so far is quite elementary and basic. If you feel a bit overwhelmed, rest assured that that is natural: for many nonmathematical subjects the learning process can be primarily a matter of remembering the sorts of things that the human mind finds easy to remember, in part because they are easily related to other things you already know. The concepts above are second nature for any mathematician, but only as a result of seeing them applied again and again over the years. There are many groups in the rest of the book, so the concepts will probably seem quite familiar by the time you reach the end.

The next definition is quite a bit less elementary. The group G is said to be **simple** if its only normal subgroups are $\{e_G\}$ and G itself. (One thing you should know about mathematical terminology is that "simple" objects are usually not so simple. Truly simple things are typically said to be "trivial.") As it happens, one of the most celebrated recent advances of mathematics is the completion of the **classification of finite simple groups**. That is, there is now a list of exactly described finite simple groups, and any finite simple group is isomorphic to some element of the list. The proof of the theorem stating that this is so is scattered in about 500 journal articles comprising over ten thousand pages, almost all of which are written in the dense style we saw above. Currently a group of mathematicians is working to boil this down to a simplified and unified presentation that is expected to occupy "only" about 5000 pages.

Pretty much everyone can directly experience the wonderful flowering of music, film, and other arts, echoing around the world these days. All educated people know that we are living in an era of profound and rapid scientific advances, even if each of us is limited in our ability to understand the specifics. Unfortunately, only a small fraction of the population knows that this is a period of equally wonderful progress in mathematics, and only experts can fully appreciate the beauty and magnificence of these contributions to world civilization.

1.3 Proofs

In mathematics a proposition is “known” if it has been proven. In this sense, mathematics is all about proofs, but many students arrive at college level mathematics without having seen any proofs, and among those who have seen some, many have never had to write their own proofs. Here we address some basic questions. What is a proof? Why are they the standard of truth and knowledge in mathematics? How are proofs conceived and constructed? What should you be trying to do when you read one? How can you learn to write proofs yourself?

The fundamental idea is quite simple: a proof is a logically compelling argument showing that certain premises imply a desired conclusion. That is, we wish to show that a proposition P implies another proposition Q. We do this by constructing a sequence of intermediate propositions $R_1, \ldots, R_n$, where $R_1 = P$ and $R_n = Q$, such that for each $i = 2, \ldots, n$, R_i is an “obvious” or “elementary” consequence of $R_1, \ldots, R_{i-1}$ and other facts of mathematics that are already known. Everybody understands what a prosecutor is trying to do in the courtroom, and at a first cut a mathematical proof is the same sort of thing: an argument that is airtight.

Things get a little bit complicated, both theoretically and practically, when one delves into the details. What constitutes a “valid inference?” This has been an important issue in philosophy from ancient Greece to the present, but, practically speaking, it isn’t a serious problem at the beginning level, since everyone knows what simple logical inferences look like. The inferences in the vast majority of proofs, including all the proofs in this book, are simple in this sense. We won’t worry about it.

Other mathematical sciences use proofs, but by and large their ethos concerning what is known is more permissive, including empirical regularities or propositions that seem overwhelmingly likely, but for which no proof has yet been found. Why are mathematicians such purists? Actually, conjectures and open problems play an important role in mathematical research, so it is not quite correct to say that proof is the only accepted form of “knowledge.” In this sense the bright red line between theorems and “conjectures for which we have compelling evidence” is a social phenomenon, and to describe it carefully would take many pages. But at the heart of any detailed explanation is the idea that in mathematics “knowing” which things are true, or very likely to be true, is much less important than having an exact understanding of why they are true.

An example illustrates why this is so. In 1742 a Prussian mathematician named Christian Goldbach (1690-1764) wrote a letter to Leonhard Euler (1707-1783) proposing a conjecture which, after slight reworking, is that every even number is the sum of two prime numbers. At present this has been verified by computers for all even numbers less than 3×10^{17}, and there are probabilistic arguments that suggest that it is, in a certain sense, overwhelmingly likely to be true. Prime numbers are distributed in an irregular and "ragged" way, and the whole thing feels rather hopeless, so it might seem reasonable to just accept Goldbach's conjecture and go on to other things. There are at least three reasons mathematicians have a different attitude.

First off, whether or not Goldbach's conjecture is actually true is not very important in itself. Nothing anybody wants to do in the world would be affected by a very large even number that happened to not be the sum of two primes. In other sciences we agree to "know" certain things that haven't been established with complete rigor because there are bridges that need to be built and diseases we would like to treat.

Second, the absolute certainty provided by proof has allowed mathematics to reach incredible heights. Proofs that go on for hundreds of pages are possible precisely because no doubt is allowed to creep in at any stage. Think of building a machine with thousands of parts. If each part has a failure rate of one in one thousand, the machine probably won't work. Mathematicians do have some tolerance for research showing that unresolved conjectures would have interesting consequences—among other things, such work plays an important part in establishing that some conjectures are more important than others—but there are important practical reasons for not letting this sort of thing get out of hand.

Finally, in mathematics the quest is more important than the destination. Hard open problems stimulate the development of new techniques that broaden and deepen our overall understanding of mathematics. Building on earlier work by G. H. Hardy (1877-1947) and J. E. Littlewood (1885-1977), in 1937 Ivan Vinogradov (1891-1983) showed that every sufficiently large odd number can be written as a sum of three primes. That is, there is an integer N such that for every integer $n > N$ there are primes p, q, r such that $2n + 1 = p + q + r$. The ideas involved in that work are far beyond the scope of this book, but hopefully the reader can imagine how these developments might be much more interesting than simply knowing whether Goldbach's conjecture happens to be true.

If you glance at a book about "higher" mathematics, you'll quickly see that it consists largely of definitions, theorems, and proofs, with a bit of less formal explanation thrown in, but not much. If it is a textbook, the problems mostly ask the students to supply proofs. The view of mathematics embodied in this approach is that there is an established, logically structured, body of material that is generally accepted, and that the task of the author, and any student, is to first forge a secure connection with this larger structure, then develop the book's specific topic by extending that structure's logic, paying meticulous attention to getting each detail right.

This approach also embodies certain beliefs about the psychology of learning mathematics, and the role of mathematical writing in that process. The student's primary focus should always be on *why* things are the way they are. It is generally thought that the most effective way to communicate that in writing is to say *as little* as the logical structure of the material allows, precisely in order to highlight that structure.

If you are accustomed to focusing on how to do calculations, you won't already know the proper way to learn this sort of mathematics. It will take some getting used to, you'll have to make several adjustments, and there are some pitfalls you'll need to avoid. And to be frank, even if you succeed in all this, learning new mathematics will still be hard. To an important extent mathematicians enjoy the subject precisely because it is something they can really sink their teeth into.

The most important thing to get used to is that you have to go one step at a time, paying attention to each aspect of the definition, theorem, or proof at hand, before going on to the next item. If you skip over even a few fine points, pretty soon what you are trying to read will stop making sense, and you will have to go back. Actually, you will need to review material you have already read fairly often, either because you don't remember things or you get confused, even if you do sincerely try to apprehend each element. Measured in pages per hour, reading mathematics is a very slow process. In part this is because the content per page is quite high—with everything stripped down to the minimum, a 200 page math book is actually *much* longer than a 700 page book about, say, history—but it is also the case that learning mathematics is just inherently slow.

What you should always be trying to do is to turn all the little details into a larger, simpler picture of the key ideas underlying the main results, and more generally why the topic is interesting, important, and structured the way it is. Ultimately, understanding a mathematical topic is a matter of achieving a mental state in which you could reverse engineer the particular details by starting with the big picture and applying standard methods of proof and the general background obtained from prior study, the things that everyone knows. (The meaning of "everyone" varies according to your level!) Your ability to do this is very much a function of your command of the basics, and an important technique for strengthening them is to continually ask yourself whether you really understand the calculations and earlier results that are being applied in the proof you are reading right now, recreating them mentally, or even looking things up, if you are even a bit unsure. This means going even slower, which will try your patience, but in a sense it is merely a matter of supplying the sort of repetition and reinforcement that occur naturally in other sorts of writing, but which are lost in the process of stripping mathematical exposition down to the minimum. If this attitude toward the material is habitual, in the long run it will speed things up because you will be able to read with greater assurance and confidence.

A very different set of issues arise when a student starts writing proofs. It often happens that the first assignment asking for a proof leaves the student paralyzed, feeling that it must be simple, but not knowing where to begin. It's actually a lot like learning a computer programming language, in that the first step, writing a program that simply prints "Hello, world!", is hard because it applies several aspects of the language, whereas everything that comes later can be assimilated one step at a time. To see what's involved, let's look at a simple, but very famous, proof.

Theorem 1.2 (Euclid)**.** *There are infinitely many prime numbers.*

Proof. Suppose there are only finitely many prime number $p_1, \dots, p_n$. Let $r = p_1 \cdot \ldots \cdot p_n + 1$. Then r has a factorization as a product of prime numbers, but no p_i can divide r, so there must be primes that are not included in the list $p_1, \dots, p_n$. This contradiction completes the proof. □

Before talking about the content of this argument, lets look at the purely mechanical features. We see a heading consisting of "Theorem" and an identifying number, both in bold face. In this case, but not always, there is an attribution consisting of a name in parentheses. Usually this is either the name of the theorem, if it has one, or the name of the person who first proved the theorem, but in this particular case we know that the theorem appeared in Euclid's *Elements*, but we don't know very much about its prior history. Following this there is a statement in italics called the **assertion**. Some space is skipped, and the actual proof is bracketed by the word "Proof" (unindented and in italics) and a square box.

The square box is a replacement in modern texts for the more traditional symbol 'Q.E.D.' This is an abbreviation for the the Latin phrase *quod erat* *demonstratum*, which means “thus it is demonstrated.” Possibly the symbol ‘Q.E.D.’ was truly useful back in the days when Latin had some real status as a universal language, but now it’s just a piece of trivia. More important is the fact that the reader often needs to be told that a proof is over. Even though this is the meaning of the square box, it can still be helpful to say this in words, as we have done here.

This format embodies and enforces an important principle called **information hiding**: all the information required to understand the proof is contained in the assertion, except to the extent that a proof invokes theorems that were proved earlier, and the assertion can be applied in later arguments without knowing how the proof works. A similar idea has been found to be very useful in organizing the code of large computer programming projects. The declaration of a subroutine makes a promise about what will happen when the subroutine is invoked. In order to use the subroutine you do not need to know the details of how this promise is fulfilled. The declaration will typically also specify the resources that are utilized by the implementation of the subroutine. This sets an outer bound on what you need to know in order to understand the implementation. Information hiding seems to be an indispensable principle for organizing large bodies of precise interrelated technical information in a way that can be understood and manipulated by people.

Turning to the actual content of the proof, we see a very common and useful idea called **proof by contradiction**. Logically, it is expressed by the following formula:

$$[(P \wedge \neg Q) \Rightarrow Q] \Rightarrow [P \Rightarrow Q].$$

Here P and Q are variables that represent “elementary” propositions, and $\wedge$, $\neg$, and $\Rightarrow$ mean “and,” “not,” and “implies” respectively. In words this formula says that if we can prove that Q is true whenever both P and $\neg Q$ hold, then Q must be true whenever P holds. If our goal is to prove that P implies Q, then, in the proof, we can add the assumption that Q is false to the other assumptions embodied in P. A more general version of this idea is expressed by the formula

$$[[(P \wedge \neg Q) \Rightarrow R] \wedge \neg R] \Rightarrow [P \Rightarrow Q].$$

If P and $\neg Q$ together imply some proposition R (e.g., $1 = 0$) that we know to be false, then Q must be true whenever P is true.

Here’s another famous example of this sort of argument.

Theorem 1.3. *There do not exist nonzero integers a and b such that $a^2 = 2b^2$, so $\sqrt{2}$ is an irrational number.*

Proof. Suppose that, contrary to the assertion, there do exist such a and b. The prime factorization of a^2 is unique, so it must be obtained by squaring the prime factorization of a, and consequently it has 2 raised to an even power. But the same reasoning shows that the prime factorization of b^2 has 2 raised to an even power, so that the prime factorization of $2b^2$ has 2 raised to an odd power. This is a contradiction of the uniqueness of prime factorization. □

Another important proof technique is called **induction**. It has the following logical pattern.

$$[P_0 \wedge (\forall n > 0)[P_{n-1} \Rightarrow P_n]] \Rightarrow (\forall n \geq 0)P_n.$$

The symbol '$\forall$' is read "for all." The idea is that there is a sequence of propositions $P_0, P_1, P_2, \ldots$ that we want to prove. If we can prove that P_0 is true, then it is enough to prove P_n, for general n, with the additional hypothesis that P_{n-1} is true. (To be a bit more precise, it is actually enough to prove P_n with the additional hypothesis that $P_0, P_1, \ldots, P_{n-1}$ are *all* true.)

We will use induction to prove the binomial theorem, which is a famous and very useful result that is used to expand expressions of the form $(x+y)^n$. First of all you have to know that for any positive integer n,

$$n! := 1 \cdot 2 \cdots (n-1) \cdot n$$

(pronounced "n factorial") is the product of all the integers between 1 and n. For example $4! = 1 \cdot 2 \cdot 3 \cdot 4 = 24$. We also set $0! := 1$; there are deep explanations of why this is the "right" definition of 0!, but we'll just accept it as a convention. For a positive integer n and an integer k with $0 \leq k \leq n$ we define

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}.$$

Lemma 1.4. *For any integers n and k with $n \geq 1$ and $1 \leq k \leq n$,*

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof. We compute that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n\cdot(n-1)!}{k!(n-k)!} = \frac{(k+(n-k))\cdot(n-1)!}{k!(n-k)!}$$
$$= \frac{(n-1)!}{(k-1)!(n-k)!} + \frac{(n-1)!}{k!(n-k-1)!} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

□

This time the result was called a **lemma** because its primary role to serve as an intermediate step in the proof of another theorem. **Propositions** are usually results that are less important than theorems, but which nonetheless have some conceptual interest. A **corollary** is a simple consequence of some result, usually the one that comes right before it.

Here is another proof using induction:

Corollary 1.5. *For any integers n and k with $n \ge 1$ and $0 \le k \le n$, $\binom{n}{k}$ is an integer.*

Proof. If $k = 0$ or $k = n$, then $\binom{n}{k} = \frac{n!}{n!\cdot 0!} = 1$, obviously. In particular, we have $\binom{1}{0} = 1 = \binom{1}{1}$, so the result is true for $n = 1$. Suppose, for some $n \ge 2$, that we have already shown that $\binom{n-1}{0}, \binom{n-1}{1}, \ldots, \binom{n-1}{n-2}, \binom{n-1}{n-1}$ are integers. Then $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n-1}, \binom{n}{n}$ are integers: if $k = 0$ or $k = n$, then $\binom{n}{k} = 1$, as we have already noted, and if $0 < k < n$, then $\binom{n}{k}$ is an integer by virtue of the lemma above.

□

The symbol $\binom{n}{k}$ is called a **binomial coefficient** and pronounced "n choose k" because it is the number of distinct k-element subsets of $\{1, \ldots, n\}$, or any set with n elements. To show this we argue by induction on n, using the formula in the proof above. Arbitrarily, choose some element of $\{1, \ldots, n\}$ that we'll call "the last" element. For $k = 0$ or $k = n$ the claim is obvious, there is one null set and one subset containing all n elements. For $0 < k < n$, any k-element subset is either the last element together with some $(k-1)$-element subset of the remainder or a k element subset of the remainder. Assuming that the claim has already been established with n replaced by $n-1$, there are $\binom{n-1}{k-1}$ subsets of the first type and $\binom{n-1}{k}$ subsets of the second type, so the claim follows from the hypothesis that the claim is true with n replaced by $n-1$, and the lemma above.

Here is the binomial theorem, with an inductive proof.

Theorem 1.6 (Binomial Theorem). *For any numbers x and y, and any integer $n \geq 1$,*

$$(x+y)^n = \sum_{i=0}^{n} \tbinom{n}{i} x^{n-i} y^i.$$

Proof. This is obviously true when $n = 1$. Suppose it has already been established with $n-1$ in place of n. We compute that

$$\begin{aligned}(x+y)^n = (x+y)(x+y)^{n-1} &= (x+y)\sum_{i=0}^{n-1} \tbinom{n-1}{i} x^{n-i-1} y^i \\ &= \sum_{i=0}^{n-1} \tbinom{n-1}{i} x^{n-i} y^i + \sum_{i=0}^{n-1} \tbinom{n-1}{i} x^{n-i-1} y^{i+1}.\end{aligned}$$

Changing the index in the second sum by one gives

$$\begin{aligned}(x+y)^n &= \sum_{i=0}^{n-1} \tbinom{n-1}{i} x^{n-i} y^i + \sum_{i=1}^{n} \tbinom{n-1}{i-1} x^{n-i} y^i \\ &= \tbinom{n-1}{0} x^n + \sum_{i=1}^{n-1} \Big(\tbinom{n-1}{i} + \tbinom{n-1}{i-1} \Big) x^{n-i} y^i + \tbinom{n-1}{n-1} y^n \\ &= \tbinom{n}{0} x^n + \sum_{i=1}^{n-1} \tbinom{n}{i} x^{n-i} y^i + \tbinom{n}{n} y^n\end{aligned}$$

where the last equality applies Lemma 1.4 and the fact that $\tbinom{n-1}{0} = 1 = \tbinom{n}{0}$ and $\tbinom{n-1}{n-1} = 1 = \tbinom{n}{n}$. □

In each of the proofs we have seen so far we invoked certain mathematical truths without proving them first. Specifically, the first proof uses the fact that any integer can be written as a product of prime numbers, and that if an integer r is divisible by another integer $p \geq 2$, then $r+1$ is *not* divisible by p. The second proof uses the fact that the factorization of an integer into a product of primes is unique, up to the ordering of the factors. The inductive proofs use things like the commutative, associate, and distributive properties of addition and multiplication. For those new to writing proofs, these facts might all be very well known, and at the same time it is not clear what it is, precisely, that justifies the assumption that the *reader* also knows them. There is also an important related question: how much detail do you need to include? These questions have a practical and a theoretical aspect.

From a practical point of view, a proof occurs in some larger context which creates certain assumptions about what is known, it is directed at some actual or imagined audience, and it is intended to create certain effects that go beyond its purely logical mission. The proofs you read will usually be in books that (if they are logically organized) begin with some declaration of expectations concerning the reader's prior study in mathematics, and some statement of the fundamental assumptions of the work. Many books, including this one, begin at a lower level than the central topic, reviewing material that should be familiar to any plausible reader, precisely in order to create a commonly understood framework.

The proofs that you'll write while taking courses are a bit different, since your goal is to demonstrate that you understand the material, and that you are a smart person who can present arguments clearly and precisely. The question of what you can assume will be answered, to some extent, by the material already presented in the course, but really you should think in terms of a proof having some central idea. Your goal should be to present that central idea clearly, with enough detail to convince the instructor that you are aware of and know how to handle any nuances in the argument.

Proofs are written in English: you should use proper grammar (or at least try if you're not a native speaker) and write in complete, correctly punctuated sentences. When you can use words in place of symbols without loss of precision, do so. In particular, the logical symbols $\wedge$, $\vee$ ("or"), $\neg$, $\Rightarrow$, $\forall$, and $\exists$ ("there exists") should generally be avoided except when one wishes to emphasize the logical structure. Creating a well written proof usually involves extensive rewriting. Often this is a matter of aiming for greater brevity without loss of content, but more generally good mathematical writing results from a process in which the author just keeps asking if there is some way to make things even a little bit easier for the reader until she really can't think of anything. This is hard work, but it can give a surprising amount of aesthetic satisfaction. Whereas the central definitions of mathematics are embodiments of profound thought and centuries of experience, proofs can be surprising, clever, and charming.

Both for the proofs you read and those you write, there is a difference between a "proof in logic" and a proof aimed at a human audience. For all but the simplest theorems, a complete and exact proof in which every step was spelled out explicitly would be long and virtually unreadable, at least by humans. (There is an active research program developing languages that express proofs exactly, so that computers can verify them. Converting one page of mathematics written for people into such a language currently takes roughly one week.) A proof for humans is really a compelling argument to the effect that a proof in logic could be constructed. There is an expectation that the reader will be able to fill in "obvious" details, and the appropriate style depends very much on the level of the intended audience.

An additional problem, both for reading proofs and writing them, is that although they must (with minor exceptions) be presented in a logically linear fashion, to guard against circular reasoning, they are best thought of as the result of a "top-down" way of thinking. That is, the proposition we want to prove is first understood as a consequence of a few "big" intermediate steps, then we look for proofs of these steps, perhaps breaking some of these into smaller pieces, and so forth. When reading a proof, you should not be content with merely seeing how each step is a consequence of what came before. In addition, you should try to understand the larger architecture of the argument, and you should try to imagine the process by which the author passed from the main ideas to the details.

After you've had a little practice, the problem of what you can legitimately assume in a proof will not seem so hard. But that does not mean that we have resolved the issue from a theoretical point of view. In fact this question—"Where does mathematics begin?"—is a very hard one that is still not completely settled. The next section describes an overall approach to it that is at least very effective in a practical sense.

1.4 Foundations: Sets, Relations, and Functions

> Whence it is manifest that if we could find characters or signs appropriate for expressing all our thoughts as definitely and as exactly as arithmetic expresses numbers or geometric analysis expresses lines, we could in all subjects in so far as they are amenable to reasoning accomplish what is done in Arithmetic and Geometry.
>
> For all inquiries which depend on reasoning would be performed by the transposition of characters and by a kind of calculus, which would immediately facilitate the discovery of beautiful results. For we should not have to break our heads as much as is necessary today, and yet we should be sure of accomplishing everything the given facts allow.
>
> Moreover, we should be able to convince the world what we should have found or concluded, since it would be easy to verify the calculation either by doing it over or by trying tests similar to that of casting out nines in arithmetic. And if someone would doubt my results, I should say to him: "Let us calculate, Sir," and thus by taking to pen and ink, we should soon settle the question.

Gottfried Wilhelm Leibniz (1646-1716) *The Method of Mathematics*

Optimism about Leibniz' dream peaked around the year 1900, as a result of the development of set theory. Here we'll first explain how set theory is useful to mathematics, then say a bit about why things didn't work out as well as some had hoped.

You probably already know that a set is a collection of things called **elements**. For instance $\{\mathrm{you}, \mathrm{me}\}$ denotes a set whose two elements are simply listed. A set is determined by its elements: if two sets have the same elements, they are the same set. A set can contain a single element, if which case it is called a **singleton**. Don't confuse a singleton with its unique element: a and $\{a\}$ are not the same thing! The set that has no elements is called the **null set** or **empty set** and is denoted by $\emptyset$. We say that A is a **subset** of B, and write $A \subset B$, if every element of A is also an element of B. It is a **proper subset** of B if, in addition, there is at least one element of B that is not an element of A.

The basic operations for constructing sets include union, intersection, and set difference: if A and B are sets, then their **union** $A \cup B$ is the set containing all the elements of A and all the elements of B, their **intersection** $A \cap B$ is the set containing all the elements of A that are also elements of B, and the **set difference** $A \setminus B$ is the set containing all the elements of A that are *not* elements of B. In addition, one may define a subset of a given set by selecting out those elements that have a certain property. If $P(b)$ means "b is red," and B is the set of balloons, then $\{\, b \in B : P(b) \,\}$ is the set of red balloons. More generally, almost any method of constructing mathematical objects can be used to define sets; we'll see many examples as we go along.

Set theory provides an all-purpose toolkit for precisely describing mathematical objects and concepts we already know about, and for defining new ones. For example, everyone knows that ordered pairs like (x, y) are important in mathematics, but just what *is* an ordered pair? Using set theory, we can *define* (x, y) to be $\{\{x\}, y\}$. There are other ways to define an ordered pair, and nobody really wants to work with this definition, so it probably all seems pretty boring. But that's the whole point! By agreeing on such a definition, all the ambiguity and potential for controversy is eliminated, just as Leibniz had hoped.

Continuing, the **cartesian product** of two sets A and B is *defined* to be

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

Ordered triples, quadruples, etc., and cartesian products of three sets, four sets, etc., can be defined in many analogous ways, at least some of which should be obvious, and which are much too tedious to describe here.

A **binary relation** is *defined* to be an ordered triple $r = (A, B, R)$ in which A and B are sets and $R \subset A \times B$. For example, the relation 'is taller than' could be the triple (H, H, S) in which H is the set of people and S is the set of pairs (p, q) is which p and q are people and p is taller than q. As with our formal definition of ordered pairs, this definition of a relation is useful because it is precise and fully general, but in almost all cases we will think of a relation, say 'less than,' as a symbol like '$<$' such that '$a < b$' is a proposition such that $a < b$ is true for some $(a, b) \in A \times B$ and false for others.

A **function** is *defined* to be binary relation $f = (A, B, F)$ with the additional property that for each $a \in A$ there is exactly one $b \in B$ (usually written $f(a)$) such that $(a, b) \in F$. In this circumstance we say that A is the **domain** of f, B is the **range** of f, and F is the **graph** of f. The symbol '$f : A \to B$' is treated grammatically as a noun (any sentence with this symbol should be grammatical if the symbol is replaced by 'f') and indicates that f is a function with domain A and range B.

The **image** of f is

$$f(A) := \{\, f(a) : a \in A \,\}.$$

Note that $(A, f(A), F)$ is a function that is different from f if the image of f is a proper subset of B. This is why it is not quite correct to identify the function f with F. More generally, if $A' \subset A$, then $f(A') := \{\, f(a) : a \in A' \,\}$. Then $f(a)$ is an element of A and (as we have defined things) $f(\{a\})$ is a singleton subset of A, but standard practice is to write $\{f(a)\}$, so that the symbol $f(\{a\})$ never occurs.

If $B' \subset B$, the **preimage** of B' is

$$f^{-1}(B') := \{\, a \in A : f(a) \in B' \,\}.$$

Usually we will write $f^{-1}(b)$ in place of $f^{-1}(\{b\})$, but as we mentioned in our discussion of group theory, when the function f is invertible, $f^{-1}(b)$ usually denotes the element that is mapped to b. In practice, this is less confusing than it sounds; a little common sense will usually guide you to the right interpretation.

If $A' \subset A$, the **restriction** of f to A' is the function $f|_{A'} : A' \to B$ with the definition

$$f|_{A'} := (A', B, F \cap (A' \times B)).$$

This is one of the two most important ways of creating a new function from given functions (composition is the other one) and the properties of restrictions are often so simple and straightforward that they are regarded as too obvious to mention. But in order to develop a secure grasp of the foundations you have to pay careful attention to the nuts and bolts, so we will lean in the direction of discussing restrictions explicitly.

The importance of functions had been recognized long before set theory was developed, but there was no single definition, and there was a tendency to think of a function as synonymous with the formula that defined it. This has all the usual disadvantages of lack of standardization, and in addition it created a tendency to overlook the fact that a single formula can define more than one function. For example, the formula $f(x) = x^2$ defines a function from the integers to the integers, another function from the rational numbers to the rational numbers, and a third function from the real numbers to the real numbers. Set theory gave mathematicians a language that allowed them to settle on one general definition of the term "function," with obvious and enormous benefits for mathematical communication.

By the way, the function concept is another one of those "profound" ideas that seems to embody some very deep wisdom about how mathematical information should be organized, even though it has a very simple definition. It emerged gradually out of the experience of mathematicians, and it's not so easy to say why it works so well. (It's easy to see that in some sense sets are like nouns and functions are like verbs, but do we really understand why human languages have this organization?) In somewhat the same way, academics have found that if they save copies of journal articles in file folders, as most do, the only method that allows you to find what you are looking for is to label the folders with the names of the authors. (Filing by topic works *very* poorly.) There is no obvious reason why no other method is effective, but it is perhaps not at all coincidental that there is a well defined (and easily computed!) function from the set of journal articles to the set of authors that maps each article to its first author.

It is a digression, and a more advanced concept than one would typically expect at this level, but I would like to explain another organizational principle that mathematicians have noticed, and found very useful, during the last sixty or so years. A **category** $\mathcal{C}$ consists of:

(a) a class[3] $\mathrm{Ob}(\mathcal{C})$ of things called **objects**;

(b) for each pair of objects $A, B \in \mathrm{Ob}(\mathcal{C})$, a set $\mathcal{C}(A, B)$ of **morphisms** from A to B;

[3] The concept of a **class** is a variant of the set concept that will be explained below.

(c) for each triple of objects $A, B, C \in \mathrm{Ob}(\mathcal{C})$ a function from $\mathcal{C}(A, B) \times \mathcal{C}(B, C)$ to $\mathcal{C}(A, C)$ called **composition**. The image of (f, g) under this mapping is denoted by $g \circ f$.

Usually, but not always, the objects are sets, the morphisms are functions, and "composition" is composition of functions. This structure is required to have the following properties:

(i) Composition is **associative**: if $A, B, C, D \in \mathrm{Ob}(\mathcal{C})$, $f \in \mathcal{C}(A, B)$, $g \in \mathcal{C}(B, C)$, and $h \in \mathcal{C}(C, D)$, then

$$h \circ (g \circ f) = (h \circ g) \circ f.$$

(ii) For each object $A \in \mathrm{Ob}(\mathcal{C})$ there is an **identity** $\mathrm{Id}_A \in \mathcal{C}(A, A)$ such that $\mathrm{Id}_B \circ f = f$ and $f \circ \mathrm{Id}_A = f$ whenever $A, B \in \mathrm{Ob}(\mathcal{C})$ and $f \in \mathcal{C}(A, B)$.

For an initial example, we can point out that sets and functions constitute a category. Groups and homomorphisms give a slightly more interesting example. Suppose that G, G', and G'' are groups and $\varphi : G \to G'$ and $\varphi' : G' \to G''$ are homomorphisms. Then

$$\varphi'(\varphi(gh)) = \varphi'(\varphi(g)\varphi(h)) = \varphi'(\varphi(g))\varphi'(\varphi(h))$$

for all $g, h \in G$, so $\varphi' \circ \varphi : G \to G''$ is a homorphism. Clearly Id_G is a homomorphism, and properties (i) and (ii) are satisfied by homomorphisms because they are satisfied by functions.

You have probably noticed that the definition of a category is extremely long-winded, and at the same time everything it says is (in the applications we have seen, and many others) quite trivial. In fact one very useful aspect of this concept is that saying, for example, that "groups and homomorphisms constitute a category" compresses a lot of easily understood, easily verified information into a neat little package. Of course this alone wouldn't make the concept useful if the situation didn't arise frequently, but it does. Actually, categories are so ubiquitous in mathematics that there is not much interesting mathematics associated with the concept itself, at least until one studies very advanced topics, in more or less the same way that there is not much of interest to say about functions in general, even though there are many interesting types of functions. For most of our applications of the concept there is no useful "theory of categories," and in this sense we don't really *need* the concept. But the experience of mathematicians has been that it is good to organize mathematics as the study of this or that category, and much of the material in this book conforms to this principle, so I think it will be illuminating to keep the concept in mind.

The language of categories can already be used to give a very general explanation of one reason groups are so important. Some of the terminology we introduced earlier in connection with groups is actually applicable to any category $\mathcal{C}$. A morphism $f \in \mathcal{C}(X, Y)$ is an **isomorphism** if it has an **inverse**, which is a morphism $g \in \mathcal{C}(Y, X)$ such that $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$. An **endomorphism** of $X \in \mathrm{Ob}(\mathcal{C})$ is a morphism $f \in \mathcal{C}(X, X)$ whose domain and range are both X. An **automorphism** of X is an endomorphism of X that is also an isomorphism.

Theorem 1.7. *For any category $\mathcal{C}$ and any $X \in \mathrm{Ob}(\mathcal{C})$, the set of automorphisms of X is a group.*

Proof. We must first of all show that a composition of automorphisms is an automorphism, so suppose that f and f' are automorphism with inverses g and g'. Since composition is associative we have

$$(f' \circ f) \circ (g \circ g') = f' \circ (f \circ g) \circ g' = f' \circ \mathrm{Id}_X \circ g' = f' \circ g' = \mathrm{Id}_X$$

and

$$(g \circ g') \circ (f' \circ f) = g \circ (g' \circ f') \circ f = g \circ \mathrm{Id}_X \circ f = g \circ f = \mathrm{Id}_X,$$

so $f' \circ f$ is indeed an automorphism because $g \circ g'$ is its inverse. It is now easy to see that (a)-(c) of Definition 1.1 are satisfied: composition of automorphisms is associative because composition of morphisms is associative, Id_X is an identity element, and the category theoretic inverse of an automorphism is an automorphism, and an inverse in the group theoretic sense. □

In addition to providing a sort of universal language and toolbox for mathematics, set theory made some very important substantive contributions to mathematical understanding. The theory of cardinality for infinite sets is particularly important and useful. For finite sets it clearly makes sense to say that two sets A and B have the same **cardinality** if there if a bijection $b : A \to B$. Georg Cantor (1845-1918) took the step of applying this notion to infinite sets. In particular, we say that a set A is **countable** if there is a bijection $b : \mathbb{N} \to A$ where $\mathbb{N} := \{1, 2, 3, \ldots\}$ is the set of **natural numbers**. If A' is an infinite subset of A, then $b^{-1}(A') = \{n_1, n_2, \ldots\}$ is an infinite subset of $\mathbb{N}$, and we can define a bijection $b' : \mathbb{N} \to A'$ by setting $b'(i) := b(n_i)$. Thus any infinite subset of a countable set is countable, so countability is the smallest infinite cardinality.

We now explain Cantor's remarkable proof that the set of real numbers, which is denoted by $\mathbb{R}$, is uncountable. We will use the fact that every real number has an infinite decimal expansion, but there is an obnoxious detail arising out of the fact that some numbers have more than one such expansion, e.g., $3.0000\ldots = 2.9999\ldots$. So, let S be the set of numbers between 0 and 1 whose decimal expansion involves only the digits '4', '5', and '6'. We argue by contradiction: suppose that $\mathbb{R}$ is countable, so that there is a bijection $b : \mathbb{N} \to \mathbb{R}$. Then $b^{-1}(S) = \{n_1, n_2, \ldots\}$ is an infinite subset of $\mathbb{N}$, and we can define a bijection $c : \mathbb{N} \to S$ by specifying that $c(i) := b(n_i)$. Think of c as a list:

$$\begin{aligned} c(1) &= 0.4656445\ldots \\ c(2) &= 0.5644546\ldots \\ c(3) &= 0.6454445\ldots \\ &\vdots \end{aligned}$$

Now construct a new number s with decimal expansion $0.d_1d_2d_3\ldots$ where $d_1 \in \{4,5,6\}$ is different from the first digit of $c(1)$, $d_2 \in \{4,5,6\}$ is different from the second digit of $c(2)$, $d_3 \in \{4,5,6\}$ is different from the third digit of $c(3)$, and so forth. Clearly the s constructed in this way is an element of S, but its construction guarantees that it is different from $c(1)$, different from $c(2)$, different from $c(3)$, etc. This contradicts the assumption that every element of S is in the list $c(1), c(2), c(3), \ldots$. Very pretty!

Here is another argument along seemingly similar lines. Suppose that $S_1, S_2, \ldots$ is a countable collection of sets, and that each S_i is countable. For the sake of simplicity let's assume that $S_i \cap S_j = \emptyset$ whenever $i \neq j$. We would like to show that the union $S_1 \cup S_2 \cup \ldots$ of all these sets is countable. For each i let $c_i : \mathbb{N} \to S_i$ be a bijection. We construct a bijection $\mathbf{c} : \mathbb{N} \to S_1 \cup S_2 \cup \ldots$ by sweeping out the diagonals of the diagram below one after the other, specifying that

$$\begin{array}{ccccc} \mathbf{c}(1) = c_1(1) & \mathbf{c}(3) = c_1(2) & \mathbf{c}(6) = c_1(3) & \mathbf{c}(10) = c_1(4) & \mathbf{c}(15) = c_1(5) \\ \mathbf{c}(2) = c_2(1) & \mathbf{c}(5) = c_2(2) & \mathbf{c}(9) = c_2(3) & \mathbf{c}(14) = c_2(4) & \\ \mathbf{c}(4) = c_3(1) & \mathbf{c}(8) = c_3(2) & \mathbf{c}(13) = c_3(3) & & \\ \mathbf{c}(7) = c_4(1) & \mathbf{c}(12) = c_4(2) & & & \\ \mathbf{c}(11) = c_5(1) & & & & \end{array}$$

and so forth. Again, a bit surprising, but simple and completely convincing once you've seen it. Or is it? Although just about any normal person (and most 19^{th} century mathematicians) would accept the argument above without any question, it actually applies an advanced and surprising idea.

The Axiom of Choice: If $r = (A, B, R)$ is a binary relation such that for each $a \in A$ there is at least one $b \in B$ such that $(a, b) \in R$, then there is a function $f : A \to B$ with $(a, f(a)) \in R$ for all $a \in A$.

In the discussion above it was assumed that for each i there is a nonempty set of functions like c_i. But the argument actually assumes that *there is a function $i \mapsto c_i$ that simultaneously specifies such a bijection for every i.* In the preceeding proof we could explicitly define a function $D : \mathbb{N} \to \{4, 5, 6\}$ such that $d_n = D(n)$ is different from the n^{th} digit of s_n for all $n \in \mathbb{N}$ by, for instance, letting $D(n)$ be the smallest element of $\{4, 5, 6\}$ different from the n^{th} digit of s_n. The set of all bijections between $\mathbb{N}$ and S_i isn't endowed with a structure that allows us to construct the desired function $i \mapsto c_i$ by specifying a "canonical" choice of c_i, and (although it is far from obvious at this point) there is simply no way to get the desired function without invoking something like the axiom of choice.

More generally, the axiom of choice is not a consequence of other "standard" assumptions of set theory, and it was a source of considerable controversy for many years. Nonconstructive reasoning of the sort employed by Cantor was sharply criticized by Leopold Kronecker (1823-1891) which resulted in Cantor being embattled for much of his career. Ernst Zermelo (1871-1953) gave a precise formulation of the axiom of choice in 1904, and over the next few decades it became clear that attempting to live without it would result in severe constraints on the sorts of mathematics that could be done. Nowadays the type of mathematics advocated by Kronecker, which is called **constructivism**, is a minor specialization that is of some interest more broadly because constructivist mathematics is, to some extent, a useful model of what computers can do. Although some logicians study axioms that might be thought of as possible replacements for the axiom of choice, all other mathematicians utilize it freely.

Possibly you're wondering whether there is any cardinality between countability and the cardinality of the real numbers, which is sometimes called the **cardinality of the continuum**. That's a *damn* good question. In 1900 David Hilbert (1862-1943) gave a lecture in which he laid out a list of unsolved problems that he thought were very important, and which he hoped might prove useful as targets to guide the development of mathematics during the coming century. Hilbert was then already the leading mathematician in the world, and he would go on to make many other important contributions, but nothing else he did is as famous as the **Hilbert Problems**. The **continuum hypothesis**—the conjecture that there is no cardinality between countability and the cardinality of the continuum—was the first problem on his list! As we'll explain in a little bit, the question was resolved a few decades later. Would you care to guess what the answer is?

So, as of 1900 mathematicians could see that sets could be used to represent just about any mathematical object. The next step in fulfilling Leibniz' vision was to develop a symbolic calculus that gave a precise formal language for defining sets, creating new sets from given sets, and more generally representing any valid mathematical argument as a sequence of allowed inferences within an exact system of symbolic logic. This project was attempted by Bertrand Russell (1872-1970) and Alfred North Whitehead (1861-1947) but Russell found an unexpected and extremely painful problem. Let S be the set of sets that are not elements of themselves. Is S an element of itself? Working through the two cases, we find that if it is, then it isn't, and if it isn't, then it is. Ouch!

Russell and Whitehead managed to salvage their project by developing something called the "theory of types" which gave a very finely described hierarchy of sets, carefully designed to prohibit the sorts of constructions that led to the paradox. Around the same time, Zermelo and Abraham Fraenkel (1891-1965) gave a different system of axioms describing allowed constructions of new sets from given sets that they hoped would provide a satisfactory foundation. These works, and the huge amount of research descended from them, are very complicated, and even professional mathematicians don't need to know that much about it, nor do many of them have the time to study very deeply in this area. Mostly they take what is generally called a "naive" approach to the subject, using the simplest constructions freely and knowing a few additional things like the theory of infinite cardinals that appear frequently in other areas of research. One idea that is useful is the notion of a **class**. I must confess that I really have no precise knowledge concerning how classes are described formally. The general intuition is that Russell's paradox arises because we falsely assume that the operations that are allowed for sets are also allowed for these more diffuse collections. By describing such collections as "classes," while sharply circumscribing the operations that classes allow, we are able to talk meaningfully about the "class of all sets" or "the class of all groups," as we did in our discussion of categories, even though there is no "set of all sets" or "set of all groups."

Leibniz was hoping not only for a language that could represent all mathematical objects (and all concepts of science, apparently) but also for computational procedures, analogous to the algorithms for addition and multiplication, that would allow any well posed problem to be answered in a mechanical and uncontroversial fashion. During the last one hundred years, along with the fantastic growth in computational technology, researchers have developed a detailed and precise understanding of what turn out to be rather severe limits to what computation can possibly accomplish. Barriers occur at several levels. For certain types of problems computational solution is possible in principle, but the fastest possible algorithms would consume a vast amount of computer time if applied to any problem instance outside of a few "toy" examples. In certain areas of mathematics there are types of problems for which there can never be a general algorithm, even though any instance of the problem has a definite answer.

Finally, there are questions that simply have no answers. In 1931 Kurt Gödel (1906-1978) showed that any sufficiently rich system of symbolic logic must include propositions that are **undecidable**, which means that neither the proposition nor its negation can be proved using the logic's formal rules of deduction. (Since we can always expand our axiom system by appending an undecidable proposition, or its negation, his argument actually shows that there are infinitely many undecidable propositions.) In 1940 he showed that the negation of the continuum hypothesis cannot be derived from the Zermelo-Fraenkel axiom system. In 1963 Paul Cohen (b. 1934) showed that the continuum hypothesis cannot be derived from the Zermelo-Fraenkel axioms, so it is undecidable.

As was the case with the axiom of choice, after this the continuum hypothesis can only be judged in terms of whether its consequences are more in accord with our intuitions, or more useful in applications, than the consequences of its negation. Many mathematicians talk about this as a matter of determining whether the continuum hypothesis is "true" or not, but to me it seems that such ways of speaking further compound the problem that the word 'true' is already overburdened with multiple meanings. If, on some basis, we decided that the continuum hypothesis was true, and then some mathematician showed that its negation implied wondrously beautiful theorems, would we really want to say that that person was doing "false mathematics?"

In any event, although these ideas are quite important in the history of mathematics, for the more mundane work of the rest of the book they are a distant background where the horizon meets the sky. The most important point for us is that even if we lack a completely precise formal apparatus of logical deduction, the language of set theory will allow us to proceed in a manner that is, in every practical sense, exact and rigorous.

Chapter 2

The Real Numbers

One of the things that makes math "hard" is that the practical side of the subject, which is what most people spend most of their time studying, is bound up with the real numbers. Because the set of real numbers is so familiar, it's easy to lose sight of the fact that it is actually an extremely complex mathematical structure. When one starts to approach the subject from the point of view of proofs it is important to develop clear understandings, or conventions, concerning the properties of the reals that are taken as given. This chapter lays out an axiom system for the real numbers. The axioms are numerous, but, with one possible exception, each of them expresses a property of the real numbers that has been familiar since you first learned about fractions and decimals and negative numbers and such, back in elementary school.

We'll also look at a number of structures that share some of the properties of the real numbers. Strictly speaking, this material is really not part of the standard curriculum in courses on calculus and linear algebra, and it is included mainly in the hope that you'll find it interesting. But the ideas we'll talk about are starting points of a great deal of mathematics that is central to the discipline, not to mention rich and deeply beautiful. And these seemingly unnecessary concepts and terminology will actually come up frequently throughout the rest of the book.

2.1 Fields

To start off with, here's a big, Big, *BIG* definition.

Definition 2.1. *A* **field** *is a triple* $(F, +, \cdot)$ *in which* F *is a set and*

$$+ : F \times F \to F \quad \textit{and} \quad \cdot : F \times F \to F$$

are binary operations (written using the usual conventions of addition and multiplication, i.e., the '$\cdot$' is usually omitted) with the following properties:

(F1) $x + (y + z) = (x + y) + z$ *for all* $x, y, z \in F$.

(F2) There is $0 \in F$ *such that* $x + 0 = x$ *for all* $x \in F$.

(F3) For each $x \in F$ *there is* $-x \in F$ *such that* $x + (-x) = 0$.

(F4) $x + y = y + x$ *for all* $x, y \in F$.

(F5) $x(yz) = (xy)z$ *for all* $x, y, z \in F$.

(F6) There is $1 \in F \setminus \{0\}$ *such that* $x \cdot 1 = x$ *for all* $x \in F$.

(F7) For each $x \in F \setminus \{0\}$ *there is* x^{-1} *such that* $x \cdot x^{-1} = 1$.

(F8) $xy = yx$ *for all* $x, y \in F$.

(F9) $x(y + z) = xy + xz$ *for all* $x, y, z \in F$.

There is a lot to digest here. Let's start off with some terminology. The elements 0 and 1 are called the **additive identity** and the **multiplicative identity** respectively, or just "zero" and "one," even in connection with the most complicated or abstract field. For a field element x, $-x$ and x^{-1} are its **additive inverse** and **multiplicative inverse** (or "negative x" and "x inverse") respectively. In words, axioms (F1) and (F5) say that addition and multiplication are **associative** while (F4) and (F8) say that these operations are **commutative**. Axiom (F9) is the **distributive law**.

Possibly you have already noticed that $(F, +)$ is a group. If a group operation is commutative, as is the case with $(F, +)$ by (F4), then the group is said to be **commutative** or **abelian**, in honor of Niels Henrik Abel (1802-1829). (Note that 'abelian' is not capitalized, unlike almost all other adjectives derived from mathematicians' names. I have no idea what motivation or historical accident led to this exception.) If we set

$$F^* := F \setminus \{0\},$$

then $(F^*, \cdot)$ is also an abelian group. In Chapter 1 we showed that the identity element of a group is unique (if you don't recall how to prove this, figure it out again) so 0 and 1 are the unique additive and multiplicative identities. We also showed that the identity element of a group is the only element of the group that acts like the identity in connection with any element of the group, and that a general group element has a unique inverse. In particular, additive and multiplicative inverses are unique.

There are many interesting and important fields. The set of **rational numbers** is denoted by $\mathbb{Q}$, earlier we mentioned that the set of **real numbers** is denoted by $\mathbb{R}$, and the set of **complex numbers** is denoted by $\mathbb{C}$. There are two other members of this group of standard symbols: the set of **natural numbers** is $\mathbb{N} := \{ 1, 2, 3, \dots \}$, which we met in the last chapter, and the set of **integers** is $\mathbb{Z} := \{ \dots, -2, -1, 0, 1, 2, \dots \}$. Presumably you are already familiar with natural numbers, integers, and rational numbers, even if the symbols are new, and their basic properties will be taken for granted in our work. No doubt you also know a lot about the real numbers, even if *really* knowing what the real numbers are is the subject of this chapter. Complex numbers come up when one studies quadratic equations, so probably almost all readers have seen them, but we'll also describe them explicitly in a bit, and learn a lot more about them later in the book.

Everyone knows that $\mathbb{Q}$ satisfies (F1)-(F9) and is consequently a field[1], but let's think a bit about what is involved in actually proving this. One approach is to start with basic properties of the integers and use them to prove the field axioms for $\mathbb{Q}$. A different approach, with wider applicability, supposes that we already know that $\mathbb{R}$ is a field. The general idea is so important that we'll state it as a theorem, even though the proof is very simple.

Theorem 2.2. *If F is a field and $K \subset F$, then K (endowed with the restrictions of addition and multiplication to $K \times K$) is a field if and only if $0, 1 \in K$ and K contains all sums, products, additive inverses, and multiplicative inverses of its elements.*

Proof. If K is a field, then it must be "closed" under addition and multiplication, i.e., it contains all sums and products of its elements, because this is part of the definition of a field. Since the additive identity of K, say $\tilde{0}$, satisfies $\tilde{0} + \tilde{0} = \tilde{0}$, and 0 is the only element of F that acts like an additive identity in connection with any element of F, we must have $\tilde{0} = 0$. Similar arguments show that the multiplicative identity of K must be 1, and that for any $x \in K$, the additive or multiplicative inverse of x, as an element of K, must agree with its inverse in F.

[1] Almost always F is written in place of $(F, +, \cdot)$, since it almost never happens that we wish to consider the set F endowed with some different structure.

Conversely, if K contains all sums and products of its elements, then the restricted field operations are functions from $K \times K$ to K, and (F1), (F4), (F5), (F8), and (F9) hold in K because they hold in F. If K contains 0 and 1, then (F2) and (F6) hold, and if K contains the inverses of its elements, then (F3) and (F7) hold. □

When the conditions described in this result hold we say that K is a **subfield** of F, and that F is an **extension** of K. Any subfield of $\mathbb{R}$ contains $\mathbb{N}$ because it contains 1 and all sums of its element, it contains $\mathbb{Z}$ because it contains 0, $\mathbb{N}$, and the negations of its elements, and it contains $\mathbb{Q}$ because it contains $\mathbb{Z}$ and the inverses and products of its elements. Thus every subfield of $\mathbb{R}$ is an extension of $\mathbb{Q}$.

The result above has some simple consequences that generate a huge collection of examples.

Corollary 2.3. *Let F be a field. Then:*

(a) If L is a subfield of F and K is a subfield of L, then K is a subfield of F.

(b) If $\mathcal{K}$ is a nonempty set of subfields of F, then $\bigcap_{K \in \mathcal{K}} K$ is a subfield of F.

(c) For any set $S \subset F$ there is a smallest subfield K of F that contains S. (In more detail, this means that K contains S and is contained in any other subfield of F that contains S.)

Proof. To prove (a) observe that because it is a subfield of L, K contains 0, 1, and all sums, products, and inverses of its elements, and by virtue of these properties K is a subfield of F. The proof of (b) is similar: if each $K \in \mathcal{K}$ contains all sums, products, and inverses of its elements, then so does $\bigcap_{K \in \mathcal{K}} K$, and if each $K \in \mathcal{K}$ contains 0 and 1, then so does $\bigcap_{K \in \mathcal{K}} K$ because $\mathcal{K}$ is nonempty. Finally (c) follows from (b) because we can construct K by taking the intersection of all subfields of F that contain S. (This collection of subfields is nonempty because it contains F itself.) □

If K is a subfield of F and $x_1, \ldots, x_r \in F$, then the smallest subfield of F containing $K \cup \{x_1, \ldots, x_r\}$ is denoted by $K(x_1, \ldots, x_r)$. To get a better sense of how these things work we'll look at $\mathbb{Q}(\sqrt{2})$ in some detail. Clearly any subfield of $\mathbb{R}$ that contains $\sqrt{2}$ must contain every number of the form

$$\frac{r + s\sqrt{2}}{t + u\sqrt{2}}$$

where $r, s, t, u \in \mathbb{Q}$ and either $t \neq 0$ or $u \neq 0$. (Since $\sqrt{2}$ is irrational, $t + u\sqrt{2}$ is different from 0 except when $t = 0 = u$.) The set of such numbers contains 0, 1, and all sums, products, and inverses of its elements, so it *is* $\mathbb{Q}(\sqrt{2})$. Now observe that

$$\frac{r + s\sqrt{2}}{t + u\sqrt{2}} = \frac{r + s\sqrt{2}}{t + u\sqrt{2}} \cdot \frac{t - u\sqrt{2}}{t - u\sqrt{2}} = \frac{(rt - 2su) + (st - ru)\sqrt{2}}{t^2 - 2u^2}.$$

Based on this calculation we can conclude that

$$\mathbb{Q}(\sqrt{2}) = \{\, r + s\sqrt{2} : r, s \in \mathbb{Q} \,\}.$$

In general an **algebraic number** is a possibly complex number (these are described below) α that satisfies some equation of the form

$$c_n\alpha^n + c_{n-1}\alpha^{n-1} + \cdots + c_1\alpha + c_0 = 0$$

where $n \geq 1$ and $c_0, c_1, \ldots, c_n$ are integers with $c_0 \neq 0 \neq c_n$. An **algebraic number field** is a field of the form $\mathbb{Q}(\alpha)$ where α is an algebraic number. (It turns out that if $\alpha_1, \ldots, \alpha_n$ are algebraic numbers, then there is an algebraic number β such that $\mathbb{Q}(\alpha_1, \ldots, \alpha_n) = \mathbb{Q}(\beta)$, so this definition is less restrictive than it seems.) A **transcendental number** is a complex number that is not algebraic. If α is a transcendental number, $\mathbb{Q}(\alpha)$ is said to be a **transcendental extension** of $\mathbb{Q}$.

So, when we have a big field and a small field, there are often lots of intermediate fields. An obvious choice for the small field is $\mathbb{Q}$, and $\mathbb{R}$ is perhaps the most obvious "big" field containing $\mathbb{Q}$. However, the field $\mathbb{C}$ of complex numbers is even bigger, and a much more natural choice because (as we'll prove in Chapter 3) it is **algebraically complete**: every polynomial

$$\alpha_n X^n + \cdots + \alpha_1 X + \alpha_0$$

whose coefficients $\alpha_0, \ldots, \alpha_n$ are in $\mathbb{C}$ has a root in $\mathbb{C}$, where a **root** is a number r such that $\alpha_n r^n + \cdots + \alpha_1 r + \alpha_0 = 0$.

For us a **complex number** is a symbol $x + iy$ where x and y are real numbers. (The standard notational convention is to write x rather than $x + i0$ and iy rather than $0 + iy$ when $y \neq 0$.) The sum and product of two such numbers $\beta = x + iy$ and $\gamma = w + iz$ are defined by the formulas

$$\beta + \gamma = (x + w) + i(y + z) \quad \text{and} \quad \beta\gamma = (xw - yz) + i(xz + yw)$$

which are obtained by treating i as a square root of -1.

Just by looking at these formulas carefully, it is easy to see that $(\mathbb{C}, +)$ is a commutative group, that multiplication is commutative, and that 1 is a multiplicative identity. Applying the formula for multiplication shows that the multiplicative inverse of a nonzero $x + iy$ is

$$\frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2}.$$

To check that multiplication is associative, and that the distributive law hold, we have the computations below, in which $\alpha = u + iv$ is a third complex number.

Before you look at them, though, I want to say something about how you should read such a calculation. It is a good idea to go slowly and convince yourself that every step is justified, but after that you shouldn't worry about "understanding" or "remembering" anything beyond the general method that led to the calculation. When a complicated calculation has some "meaning" it is usually a symptom of bad writing or of the possibility of introducing more concepts that would replace the calculation with a line of reasoning. For the calculations below such concepts exist, but we won't come to them until much later, and in the meantime we want to know that $\mathbb{C}$ is a field.

Getting down to work:

$$\begin{aligned}
\alpha(\beta\gamma) &= (u+iv)\big((xw-yz)+i(xz+yw)\big) \\
&= \big(u(xw-yz)-v(xz+yw)\big)+i\big(v(xw-yz)+u(xz+yw)\big) \\
&= \big((ux-vy)w-(uy+vx)z\big)+i\big((uy+vx)w+(ux-vy)z\big) \\
&= \big((ux-vy)+i(uy+vx)\big)(w+iz) \\
&= (\alpha\beta)\gamma;
\end{aligned}$$

$$\begin{aligned}
\alpha(\beta+\gamma) &= (u+iv)\big((x+w)+i(y+z)\big) \\
&= \big(u(x+w)-v(y+z)\big)+i\big(v(x+w)+u(y+z)\big) \\
&= \big((ux-vy)+(uw-vz)\big)+i\big((uy+vx)+(uz+vw)\big) \\
&= \big((ux-vy)+i(uy+vx)\big)+\big((uw-vz)+i(uz+vw)\big) \\
&= \alpha\beta+\alpha\gamma.
\end{aligned}$$

Our next example of a field is both quite surprising, if you haven't seen it before, and an application of an extremely important method for defining new mathematical objects. We'll begin by explaining this method in general. Let S be a set, and let $\cong$ be a binary relation on S. (As we explained in the last chapter, "formally" $\cong$ is a triple whose components are two copies of S and a subset of $S \times S$, but we'll follow the standard practice of writing '$s \cong t$' to indicate that s and t are related.) The relation $\cong$ is said to be:

(R) **reflexive** if $s \cong s$ for all $s \in S$;

(S) **symmetric** if, for all $s, t \in S$, $s \cong t$ implies $t \cong s$;

(T) **transitive** if, for all $s, t, u \in S$, $s \cong u$ whenever $s \cong t$ and $t \cong u$.

An **equivalence relation** is a relation that is reflexive, symmetric, and transitive. The granddaddy of all equivalence relations is equality itself, obviously.

If $\cong$ is an equivalence relation on S and $s \in S$, the **equivalence class** containing s is

$$[s] := \{\, t \in S : s \cong t \,\}.$$

A **partition** of S is a set $\mathcal{P}$ of nonempty subsets of S such that each element of S is an element of exactly one element of $\mathcal{P}$:

(a) $\emptyset \notin \mathcal{P}$;

(b) for all $C, D \in \mathcal{P}$, either $C = D$ or $C \cap D = \emptyset$;

(c) $\bigcup_{C \in \mathcal{P}} C = S$.

It may seem obvious that the set $\{\, [s] : s \in S \,\}$ of all equivalence classes is a partition of S, but let's walk through the details. For each $s \in S$ we have

$$s \in [s] \subset \bigcup_{s \in S} [s]$$

by reflexivity, so $[s] \neq \emptyset$ and $\bigcup_{s \in S} [s] = S$. Therefore (a) and (c) hold.

The proof of (b) illustrates one of the things that makes math "hard," namely that something can be fairly obvious, and at the same time a completely precise proof of it involves unexpected and, frankly, rather tedious details. (The argument below could actually be even more detailed, since we do not explicitly mention certain appeals to the symmetry of $\cong$, expecting that the reader will regard them as obvious.) Suppose that $[s] \cap [t] \neq \emptyset$. We need to show that $[s] = [t]$, which follows if we can show that $[s] \subset [t]$ and $[t] \subset [s]$. We will only prove the first inclusion since the proof of the second is obtained from the proof of the first by swapping s and t. Choose $r \in [s] \cap [t]$, and let u be any element of $[s]$. Since $r \cong s$ and $r \cong t$, transitivity implies that $s \cong t$. Since $u \cong s$ and $s \cong t$, transitivity implies that $u \cong t$, so that

$u \in [t]$. But u was an arbitrary element of $[s]$, so we have shown that every element of $[s]$ is an element of $[t]$, which is what we set out to do.

One of the most common ways to use set theory to define new mathematical objects is "passage to equivalence classes." Starting with a set we already know about, we define an equivalence relation on it and take the equivalence classes as the elements of the new set we are constructing. Often there are operations, or relations, or functions, we would like to define on the new set, and typically these are defined by saying that the sum (or product, or whatever) of two equivalence classes is the equivalence class of any sum of an element of one of the classes and an element of the other. There is then the (typically mundane and tedious) task of proving that what you get in this way is "independent of the choice of representatives," meaning that the equivalence class you end up with doesn't depend on which elements you chose from the two classes you are summing.

To see how this works in a concrete example, let n be a positive integer. We say that two integers a and b are **congruent** mod n, and we write

$$a \equiv b \bmod n,$$

if $a - b$ is a multiple of n, i.e., there is an integer k such that $a - b = kn$. Make sure that you can see for yourself that this relation is, indeed, reflexive, symmetric, and transitive. An equivalence class of the relation 'congruent mod n' is called an **integer mod** n, and the set of integers mod n is denoted by $\mathbb{Z}_n$. We define addition and multiplication of integers mod n by the formulas

$$[a] + [b] := [a + b] \quad \text{and} \quad [a][b] := [ab].$$

To check that these definitions are independent of the choice of representatives we suppose that $[a] = [a']$ (so that a and a' are both **representatives** of $[a] = [a']$) and that $[b] = [b']$. In order for our definitions of addition and multiplication to make sense it had better be the case that $[a+b] = [a' + b']$ and $[ab] = [a'b']$. We have $a' = a + kn$ and $b' = b + \ell n$ for some integers k and ℓ, and

$$a' + b' = a + b + (k + \ell)n \quad \text{and} \quad a'b' = ab + (a\ell + bk + k\ell n)n,$$

so this is indeed the case.

We have defined a set of equivalence classes $\mathbb{Z}_n$ and two binary operations on this set, that we have called "addition" and "multiplication." Now I want you to go back and look at (F1)-(F9) again, for each axiom asking whether it is satisfied by $(\mathbb{Z}_n, +, \cdot)$. Ideally, for each axiom you should either write out a proof that it holds or give an example that shows that it doesn't, and if you feel like doing all that work that's great. But for many of the axioms you'll probably be able to see whether it holds right away, in which case writing things down is not necessary so long as you think carefully about how to prove it.

Done? Actually, if you're like most of my students, I'd be pretty surprised. Go back and really do it this time!

Okay, here's the situation: $\mathbb{Z}_n$ satisfies (F1)-(F6), (F8), and (F9), but it can fail to satisfy (F7). For example, there is no inverse of $[2]$ in $\mathbb{Z}_4$. One way to think about this is that $\mathbb{Z}_4$ has zero divisors. In general $[a]$ is a **zero divisor** in $\mathbb{Z}_n$ if there is a $[b] \neq [0]$ such that $[a][b] = [ab] = [0]$. In $\mathbb{Z}_4$ we have $[2][2] = [0]$. If $[c]$ was an inverse of such an $[a]$ we would have the contradictory computation

$$[0] \neq [b] = [1][b] = ([c][a])[b] = [c]([a][b]) = [c][0] = [0],$$

so a zero divisor can't have an inverse.

Hopefully you already know a bit about prime numbers. Probably you think of a prime number as an integer whose only divisors are 1 and the number itself, but here a slightly different definition works better: a **prime number** is an integer $p > 1$ such that whenever p divides a product ab, either p divides a or p divides b. There cannot be any zero divisors in $\mathbb{Z}_p$ because if $[a][b] = [ab] = 0$, then p divides ab, so either p divides a, in which case $[a] = 0$, or p divides b, in which case $[b] = 0$.

Theorem 2.4. *If p is a prime, then $\mathbb{Z}_p$ is a field.*

Proof. You've already shown that $\mathbb{Z}_p$ satisfies (F1)-(F6), (F8), and (F9). Fix $[a] \in \mathbb{Z}_p^*$ (recall that whenever F is a field, $F^* = F \setminus \{0\}$) and consider the function $[b] \mapsto [a][b]$ from $\mathbb{Z}_p^*$ to itself. This function is injective because if $[a][b] = [a][b']$, then $[0] = [a][b] - [a][b'] = [a]([b] - [b'])$, whence $[b] - [b'] = 0$, i.e., $[b] = [b']$. Whenever a function from a finite set to itself is injective, it is also surjective, just because the domain and the image have the same number of elements, so there must be some $[b]$ such that $[a][b] = [1]$. Since $[a]$ was arbitrary, we have shown that $\mathbb{Z}_p$ satisfies (F7). □

Certainly $\mathbb{Z}_p$ seems quite different from the other fields ($\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$) we already know about, and it might strike you as silly and inconsequential. In order to provide some evidence that $\mathbb{Z}_p$ isn't silly we'll briefly describe one of the most celebrated theorems in all of mathematics.

From now on let p be an odd prime. That is, in addition to assuming that p is a prime, we also assume that $p \neq 2$. Just as is the case with $\mathbb{R}$, some numbers in $\mathbb{Z}_p^*$ are squares and others aren't. If a is not divisible by p and $[a] \in \mathbb{Z}_p^*$ has a square root—that is, there is some $[b] \in \mathbb{Z}_p$ such that $[b]^2 = [a]$—then we say that a is a **quadratic residue** mod p. For an odd prime p and any integer a the **Legendre symbol** $\left(\frac{a}{p}\right)$ is defined to be

$$\left(\frac{a}{p}\right) := \begin{cases} 1, & a \text{ is a quadratic residue mod } p, \\ -1, & a \text{ is neither divisible by } p \text{ nor a quadratic residue mod } p, \\ 0, & a \text{ is divisible by } p. \end{cases}$$

Theorem 2.5 (Law of Quadratic Reciprocity). *If p and q are odd primes, then*

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\cdot\frac{q-1}{2}}.$$

We should say a little about the right hand side of this equation. It will be 1 or -1 according to whether $(p-1)(q-1)/4$ is even or odd. Since p is odd, $(p-1)/2$ will be even or odd according to whether p is congruent to 1 or 3 mod 4, and similarly for q. After parsing all this, we see that the assertion is that $\left(\frac{p}{q}\right)\left(\frac{q}{p}\right)$ is 1 unless p and q are both congruent to 3 mod 4, in which case it's -1. To me the formula seems like an unnecessarily cryptic way of expressing this, even if it's quite compact, but in any event this formulation is now an unshakeable tradition.

The quadratic reciprocity theorem was stated without proof by Euler in 1783 and proved (after Adrien-Marie Legendre (1752-1833) published an incorrect proof) in 1796 by Carl Friedrich Gauss (1777-1855) who many regard as the greatest mathematician of all time. It was his favorite theorem among the many contributions he made to number theory, and he gave eight different proofs. Quadratic reciprocity has been an important theme in number theory ever since; among other things, one of the Hilbert Problems was to find generalizations of quadratic reciprocity that apply to algebraic number fields.

So $\mathbb{Z}_p$ is far from silly, at least in the eyes of Euler, Gauss, and Hilbert, but perhaps it's still inconsequential? Traditionally number theory has been thought of as devoid of practical applications, a subject studied entirely for its beauty and intellectual challenge. But recently all that has changed: during the last thirty years algorithms involving $\mathbb{Z}_p$ for large p have become the workhorses of cryptography and computer security.

2.2 Rings

It is very natural to wonder about the types of structures that satisfy some, but not all, of the axioms (F1)-(F9). There are many possibilities here, but it turns out that some are much more important than others.

Definition 2.6. *A* ***ring*** *is a triple* $(R, +, \cdot)$ *in which* R *is a set and*

$$+ : R \times R \to R \quad \text{and} \quad \cdot : R \times R \to R$$

are binary operations satisfying (R1)-(R6) below.

(R1) $x + (y + z) = (x + y) + z$ *for all* $x, y, z \in R$.

(R2) There is $0 \in R$ *such that* $x + 0 = x$ *for all* $x \in R$.

(R3) For each $x \in R$ *there is* $-x$ *such that* $x + (-x) = 0$.

(R4) $x + y = y + x$ *for all* $x, y \in R$.

(R5) $x(yz) = (xy)z$ *for all* $x, y, z \in R$.

(R6) $x(y + z) = xy + xz$ *and* $(x + y)z = xz + yz$ *for all* $x, y, z \in R$.

We say that R *is a* ***ring with unit*** *if*

(R7) There is $1 \in R^* := R \setminus \{0\}$ *such that* $x \cdot 1 = x = 1 \cdot x$ *for all* $x \in R$.

We say that R *is* ***commutative*** *if*

(R8) $xy = yx$ *for all* $x, y \in R$.

Here (R1)-(R5) are (F1)-(F5) and (R8) is (F8). In particular, $(R, +)$ is a commutative group. The difference between (R6) and (F9), and between (R7) and (F7), is that insofar as multiplication is not necessarily commutative, it is necessary to give two equations rather than one. The fancy way to state (R7) in words is to say that "there is a two sided identity element for multiplication."

In simpler words, a ring is a set whose elements can be added and multiplied. Addition has all the nice properties (associative, commutative, identity, and inverses) while multiplication is associative and satisfies the distributive law, but may lack other properties. After many years of elementary and secondary school mathematics you may have never seen the word "ring," but you have certainly seen a lot of objects that belong to rings. To emphasize this we begin with lots of examples. In each case make sure you understand why the relevant axioms are satisfied.

Example 2.7. *Any field is a commutative ring with unit, obviously.*

Example 2.8. *The set of integers $\mathbb{Z}$ is a commutative ring with unit.*

Example 2.9. $\mathbb{Z}[\sqrt{2}]$ *is the set of all numbers of the form $a + b\sqrt{2}$ where a and b are integers. It is a commutative ring with unit.*

Example 2.10. *If $n > 1$ is an integer, then $\mathbb{Z}_n$ is a commutative ring with unit, as we saw in the last section.*

Example 2.11. $3\mathbb{Z} = \{\dots, -6, -3, 0, 3, 6, \dots\}$ *is a commutative ring without a unit.*

This example somehow gives the feeling that when you have a commutative ring without a multiplicative identity, something has been left out. As a matter of general principle things are not that simple, but as a practical matter almost all the rings that come up are subsets of rings with units.

Example 2.12. *Let S be any set, let R be a ring, and let $\mathcal{F}_R(S)$ be the set of functions*

$$f : S \to R,$$

with addition and multiplication defined "pointwise:" for $f, g \in \mathcal{F}_R(S)$, $f+g$ and fg are the functions that takes each $s \in S$ to $f(s) + g(s)$ and $f(s)g(s)$ respectively. Then $\mathcal{F}_R(S)$ is a ring, it is commutative if R is commutative, and if R has a unit, then the function taking each $s \in S$ to 1 is a unit for $\mathcal{F}_R(S)$.

There are lots and lots of important rings of functions, as we'll see as we go along. Here's another ring that *looks* like a ring of functions, but is actually a bit different.

Example 2.13. *Let R be a commutative ring, and let X be a variable. A* ***polynomial*** *in X with coefficients in R is an expression of the form*

$$a_m X^m + a_{m-1} X^{m-1} + \cdots + a_1 X + a_0$$

where m is a nonnegative integer and $a_0, a_1, \dots, a_{m-1}, a_m$ are elements of R. Let $R[X]$ denote the set of such polynomials. Elements of $R[X]$ are added and multiplied "as if" they were functions: if $m \le n$, then

$$\begin{aligned}(a_m X^m + \cdots + a_0) + (b_n X^n + \cdots + b_0) &= \\ b_n X^n + \cdots + b_{m+1} X^{m+1} + (a_m + b_m) X^m &+ \cdots + (a_0 + b_0)\end{aligned}$$

and

$$(a_mX^m + \cdots + a_0)(b_nX^n + \cdots + b_0) =$$
$$a_mb_nX^{m+n} + (a_mb_{n-1} + a_{m-1}b_n)X^{m+n-1} + \cdots + (a_1b_0 + a_0b_1)X + a_0b_0.$$

We are free to *interpret* this example as a ring of functions that map R into itself. But sometimes other interpretations are interesting. For example, if $R = \mathbb{Z}$, then we can also interpret $\mathbb{Z}[X]$ as a ring of functions mapping $\mathbb{Q}$ to itself, or as a ring of functions mapping $\mathbb{R}$ to itself, or as a ring of functions mapping $\mathbb{C}$ to itself. For this reason the "proper" way to think about $R[X]$ is that its elements are agglomerations of symbols that recombine according to certain rules.

Matrices give rise to a wide range of examples. If m and n are positive integers and R is a ring, then an $m \times n$ **matrix** with entries in R is a rectangular array

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

whose entries are elements of R. We sometimes indicate the relationship between the notation for the matrix itself and the notation for its entries by saying, for example, that "$A = (a_{ij})$ is an $m \times n$ matrix."

Addition of matrices is defined "componentwise." That is, if $A = (a_{ij})$ and $A' = (a'_{ij})$ are $m \times n$ matrices, then $A + A'$ is, by definition, the $m \times n$ matrix whose ij-entry is $a_{ij} + a'_{ij}$. This operation is associative and commutative, and has an identity element (the matrix whose entries are all zero) and inverses, simply because R satisfies (R1)-(R4). (If you want to sling the lingo like a pro, you say that these properties are "inherited" from R.) So, the $m \times n$ matrices with entries in R are a commutative group with addition as the group operation.

Now suppose that $A = (a_{ij})$ is an $m \times n$ matrix with entries in R, and $B = (b_{jk})$ is $n \times p$ matrix with entries in R. We define the **product** AB of these two matrix to be the $m \times p$ matrix whose ik entry is

$$a_{i1}b_{1k} + \cdots + a_{in}b_{nk}.$$

(Think of picking up the i^{th} row of A, rotating it 90°, and dropping it on the k^{th} column of B.) In order for it to be possible to multiply A and B it must be the case that, as we assumed, the number of columns of A is the same as the number of rows of B. When this is so we say that A and B are **conformable**.

Matrix multiplication is associative: if $C = (c_{k\ell})$ is a $p \times q$ matrix with entries in R, then the $i\ell$-entry of $(AB)C$ is

$$\sum_{k=1}^{p} \big(\sum_{j=1}^{n} a_{ij} b_{jk} \big) c_{k\ell}$$

and the $i\ell$-entry of $A(BC)$ is

$$\sum_{j=1}^{n} a_{ij} \big(\sum_{k=1}^{p} b_{jk} c_{k\ell} \big).$$

One can use the distributive law (R6) to rewrite the first expression as a sum of np terms of the form $a_{ij}b_{jk}c_{k\ell}$, then use associativity and commutativity of addition to reorder them in a way that allows another application of the distributive law to give the second expression.

It is even easier to see that matrix addition and multiplication satisfy the distributive laws. Let $A = (a_{ij})$ and $A' = (a'_{ij})$ be $m \times n$ matrices, and let $B = (b_{jk})$ and $B' = (b'_{jk})$ be $n \times p$ matrices. Then the ik-entry of $(A + A')B$ is

$$\sum_{j=1}^{n} (a_{ij} + a'_{ij}) b_{jk}$$

while the ik-entry of $AB + A'B$ is

$$\sum_{j=1}^{n} a_{ij} b_{jk} + \sum_{j=1}^{n} a'_{ij} b_{jk},$$

and appropriate applications of the distributive laws and associativity and commutativity of addition show that these expressions are equal. The proof that $A(B + B') = AB + AB'$ is the mirror image of this.

For any positive integer n the set of $n \times n$ matrices with entries in R is denoted by $M_n(R)$. Any two elements of $M_n(R)$ are conformable, and we have shown that addition and multiplication of elements of $M_n(R)$ satisfies (R1)-(R6), so:

Example 2.14. *If R is a ring and n is a positive integer, then $M_n(R)$ is a ring. If R has a multiplicative identity, then the **identity matrix***

$$I := \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

is multiplicative identity for $M_n(R)$.

Multiplication in $M_n(R)$ is *not* commutative when $n > 1$, even if R is commutative. If this is not a familiar fact it would be a good idea to take the time to find $A, B \in M_2(\mathbb{R})$ such that $AB \neq BA$. If you don't know how to get started, just experiment: a "sufficiently random" choice will work.

2.3 Ring Homomorphisms

We now have a rich supply of objects, so a categorically minded person should be looking for some morphisms. As is very frequently the case, the most interesting and important maps are those that "commute" with the operations that define the structure.

Definition 2.15. *If R and S are rings, then a function $\varphi : R \to S$ is a **homomorphism** if*

$$\varphi(a+b) = \varphi(a) + \varphi(b) \quad \textit{and} \quad \varphi(ab) = \varphi(a)\varphi(b)$$

for all $a, b \in R$.

Usually anything called a "homomorphism" should preserve all elements of structure, so we should wonder about the additive identity and additive inverses. But the requirement that $\varphi(a+b) = \varphi(a) + \varphi(b)$ says precisely that φ is a homomorphism between the additive groups $(R, +)$ and $(S, +)$, and when we introduced homomorphisms of groups we showed that the requirement that the homomorphism "commutes" with the group operations implies that it also preserves identities and inverses.

Here are two basic examples of ring homomorphisms. First, for any integer n the map $a \mapsto [a]$ taking an integer a to its congruence class mod n is a homomorphism from $\mathbb{Z}$ to $\mathbb{Z}_n$. Make sure you see why this is a homomorphism; the details involved in verifying this are similar to what was involved in showing that addition and multiplication in $\mathbb{Z}_n$ are well defined. Second, if S is a set, R is a ring, and $s_0 \in S$, then the mapping $f \mapsto f(s_0)$ is a homomorphism from $\mathcal{F}_R(S)$ to R. Again, make sure you understand this.

As was the case with groups, the assertion that "rings and homomorphisms constitute a category" encompasses a lot of trivial observations. First, compositions of homomorphisms are homomorphisms: if $\varphi : R \to S$ and $\psi : S \to T$ are homomorphisms, and $a, b \in R$, then

$$\psi(\varphi(a+b)) = \psi(\varphi(a) + \varphi(b)) = \psi(\varphi(a)) + \psi(\varphi(b))$$

and

$$\psi(\varphi(ab)) = \psi(\varphi(a)\varphi(b)) = \psi(\varphi(a))\psi(\varphi(b)).$$

Even more trivially: i) composition of homomorphisms is associative, because it is just composition of functions; ii) for any ring R the identity function Id_R is a homomorphism; iii) $\mathrm{Id}_S \circ \varphi = \varphi = \varphi \circ \mathrm{Id}_R$ for any homomorphism $\varphi : R \to S$. All these things are not just obvious, but *painfully* obvious, so it feels a bit embarrassing to be dwelling on them, but of course any discussion involving homomorphisms applies them *all the time*.

If φ is a bijection, then it is said to be an **isomorphism**, and R and S are said to be **isomorphic**. To show that we are using the word 'isomorphism' in its proper categoric sense we need to verify that if φ is an isomorphism, then so is φ^{-1}. Of course it is a bijection, and to establish that it is a homomorphism we take two elements α and β of H, set $a := \varphi^{-1}(\alpha)$ and $b := \varphi^{-1}(\beta)$, and compute that

$$\varphi^{-1}(\alpha+\beta) = \varphi^{-1}(\varphi(a)+\varphi(b)) = \varphi^{-1}(\varphi(a+b)) = a+b = \varphi^{-1}(\alpha)+\varphi^{-1}(\beta)$$

and

$$\varphi^{-1}(\alpha\beta) = \varphi^{-1}(\varphi(a)\varphi(b)) = \varphi^{-1}(\varphi(ab)) = ab = \varphi^{-1}(\alpha)\varphi^{-1}(\beta).$$

Retracing the path we followed when we studied groups, a **subring** of a ring R is a nonempty set $R' \subset R$ that is closed under addition, negation, and multiplication, so that it is itself a ring[2] when endowed with the restrictions of addition and multiplication to $R' \times R'$. There are some obvious and basic examples: it is always the case that $\{0\}$ and R are subrings, and for any $r \in R$,

$$rR = \{\, rs : s \in R \,\}$$

is a subring. If A is any set and, for each $\alpha \in A$, R_α is a subring of R, then $\bigcap_{\alpha\in A} R_\alpha$ is a subring.

With groups, kernels of homomorphisms were important, and the same is true for rings. If $\varphi : R \to S$ is a homomorphism, the **kernel** of φ is

$$\ker(\varphi) := \varphi^{-1}(0) = \{\, r \in R : \varphi(r) = 0 \,\}.$$

For example, the kernel of the homomorphism $a \mapsto [a]$ from $\mathbb{Z}$ to $\mathbb{Z}_n$ is

$$n\mathbb{Z} = \{\, \dots, -2n, -n, 0, n, 2n, \dots \,\},$$

[2] Here are the details of the verification that R' is a ring: (R1), (R4), (R5), and (R6) hold for R' because they hold for R; (R3) holds because we have required that R' be closed under negation; (R2) holds because we have required that R' be nonempty and closed under negation and addition, so that for any $a \in R'$ we have $-a \in R'$ and $a + (-a) = 0 \in R'$.

and the kernel of the homomorphism $f \mapsto f(s_0)$ from $\mathcal{F}_R(S)$ to R is

$$\{\, f \in \mathcal{F}_R(S) : f(s_0) = 0 \,\}.$$

If $\varphi : R \to S$ is a homomorphism and S' is a subring of S, then $\varphi^{-1}(S')$ is a subring of R. Concretely, if $a, b \in \varphi^{-1}(S')$, then $a+b$ and ab are also in $\varphi^{-1}(S')$ because

$$\varphi(a+b) = \varphi(a) + \varphi(b) \quad \text{and} \quad \varphi(ab) = \varphi(a)\varphi(b)$$

are in S'. Since $\{0\}$ is always a subring of S, the kernel of a homomorphism is a subring, but not every subring can be a kernel.

Definition 2.16. *A subset I of a ring R is a* ***two sided ideal*** *of R if it is a subgroup of the additive group $(R, +)$ of R and $ra \in I$ and $as \in I$ whenever $a \in I$ and $r, s \in R$.*

The latter condition is stronger than closure under multiplication, so a two sided ideal of a ring is a subring. Note that for any ring R, $\{0\}$ is a two sided ideal. Of course it's an additive subgroup, so to show this we can observe that for any $r \in R$ we have $r \cdot 0 = r(0+0) = r \cdot 0 + r \cdot 0$, so that $r \cdot 0 = 0$, and $0 \cdot s = 0$ for any s by a symmetric argument. On the other hand there are many subrings that are not two sided ideals; for example, the integers $\mathbb{Z}$ are a subring, but not a two sided ideal, of the rationals $\mathbb{Q}$.

In abstract algebra two sided ideals are much more important than subrings, for various reasons, most of which flow directly or indirectly out of the following fact:

Proposition 2.17. *The kernel I of a homomorphism $\varphi : R \to S$ is a two sided ideal.*

Since $\{0\}$ is always a two sided ideal, this is a special case of the following result, which is no more difficult to prove.

Proposition 2.18. *If $\varphi : R \to S$ is a homomorphism and $J \subset S$ is a two sided ideal of S, then $I := \varphi^{-1}(J)$ is a two sided ideal of R.*

Proof. Since φ is a homomorphism of the underlying additive groups and J is a subgroup of $(S, +)$, I is a subgroup of $(R, +)$. (This point was explained in Section 1.2.) To complete the proof we observe that for any $a \in I$ and $r, s \in R$ we have $ra \in I$ and $as \in I$ because

$$\varphi(ra) = \varphi(r)\varphi(a) \in \varphi(r)J \subset J \quad \text{and} \quad \varphi(as) = \varphi(a)\varphi(s) \in J\varphi(s) \subset J.$$

□

Is every two sided ideal the kernel of some homomorphism? To answer this question we introduce another construction based on equivalence classes. Let R be a ring, and let I be a two sided ideal. We say that ring elements a and b are **congruent** mod I if $a - b \in I$. Since $a - a = 0 \in I$, any a is congruent to itself, so 'congruence mod I' is a reflexive relation. Since I is closed under negation, $a - b \in I$ if and only if $b - a \in I$, so 'congruence mod I' is a symmetric relation. Since I is closed under addition, if $a - b \in I$ and $b - c \in I$, then $a - c \in I$, so a is congruent to c mod I whenever a is congruent to b and b is congruent to c mod I. That is, 'congruence mod I' is transitive. Thus we have an equivalence relation. In this context the equivalence classes are called **cosets**, and the equivalence class containing a is usually denoted by $a + I$ because, after all, it is $\{\, a + i : i \in I \,\}$.

We define addition and multiplication of cosets by the formulas

$$(a + I) + (b + I) := (a + b) + I \quad \text{and} \quad (a + I)(b + I) := ab + I.$$

We need to show that these definitions are independent of the choices of representatives, so suppose that $a' \in a + I$ and $b' \in b + I$. Then $a' - a \in I$ and $b' - b \in I$, so

$$(a' + b') + I = (a' - a) + (b' - b) + (a + b) + I = (a + b) + I,$$

and $a'(b' - b) \in I$ and $(a' - a)b \in I$, so

$$a'b' + I = a'(b' - b) + (a' - a)b + ab + I = ab + I.$$

We now check that these operations define a ring whose elements are the cosets. Clearly (R1), (R4), (R5), and (R6) are satisfied by addition and multiplication of cosets because they are satisfied by addition and multiplication in R. In addition, $0 + I$ is an additive identity, and for any a, $(-a) + I$ is an additive inverse of $a + I$. We have defined a new ring, called the **quotient ring**, that is usually denoted by R/I. Actually, we have already seen an example of this construction: for any integer m,

$$m\mathbb{Z} = \{\dots, -2m, -m, 0, m, 2m, \dots\}$$

is an ideal of $\mathbb{Z}$, and $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$.

Let $\varphi : R \to R/I$ be the function $a \mapsto a + I$. The fact that this map is a homomorphism is an automatic consequence of our definitions. In addition, $\varphi(a) = 0$ if and only if $a \in I$. *Beginning with an arbitrary two sided ideal I, we have constructed a homomorphism whose kernel is I.* Thus a subset of R is the kernel of some homomorphism if and only if it is a two sided ideal.

The general idea of this construction is extremely important, and will appear again later, so we're going to repeat the whole thing in the context of group theory. Let G be a group with identity element e, and let H be a subgroup. A **right coset** with respect to H is a set of the form

$$Hg = \{\, hg : h \in H \,\}$$

where $g \in G$. We claim that 'is in the right coset of' is an equivalence relation on G. For all g we have $g = eg \in Hg$, so the relation is reflexive. If $g' \in Hg$, then $g' = hg$ for some $h \in H$ and consequently $g = h^{-1}g' \in Hg'$, so the relation is symmetric. If $g' \in Hh$ and $g'' \in Hg'$, then there are $h, h' \in H$ with $g' = hg$ and $g'' = h'g'$, so that $g'' = h'hg \in Hg$. Therefore the relation is transitive. Since it is an equivalence relation, and the cosets are the equivalence classes, the collection of right cosets is a partition of G.

We would now like to define multiplication of right cosets by the formula '$HgH\tilde{g} := Hg\tilde{g}$,' but in order for this to make sense the right hand side cannot depend on the particular elements g and $\tilde{g}$ that were chosen to represent the cosets of the left hand side. Put more concretely, for any $h, \tilde{h} \in H$ we have $H(hg) = Hg$ and $H(\tilde{h}\tilde{g}) = H\tilde{g}$, so it must be the case that $Hg\tilde{g} = Hhg\tilde{h}\tilde{g}$. Now observe that $Hhg\tilde{h}\tilde{g} = Hg\tilde{h}\tilde{g}$, and that multiplying both sides of the equation $Hg\tilde{g} = Hg\tilde{h}\tilde{g}$ on the right by $\tilde{g}^{-1}g^{-1}$ gives $H = Hg\tilde{h}g^{-1}$, which implies that $g\tilde{h}g^{-1} \in H$. We've shown that if the definition $HgH\tilde{g} := Hg\tilde{g}$ is independent of the choice of representatives, then

$$C_g(\tilde{h}) = g\tilde{h}g^{-1} \in H$$

for all $g \in G$ and $\tilde{h} \in H$. That is, H *is a normal subgroup of* G. Conversely, if H is a normal subgroup and $g \in G$, then $C_g|_H$ is a bijection between H and itself (its inverse is $C_{g^{-1}}|_H$) so

$$HgH\tilde{g} = HgHg^{-1}g\tilde{g} = HC_g(H)g\tilde{g} = HHg\tilde{g} = Hg\tilde{g}.$$

So, we'll assume that our given subgroup is normal, and we'll denote it by N in conformity with the notational conventions of group theory. Multiplication of right cosets is not only well defined, but in fact this operation on right cosets turns the set of right cosets into a group:

- This operation is associative because the group operation of G is associative.

- The coset Ne acts as a two sided identity element, obviously.

- For any g we have $NgNg^{-1} = Ne = Ng^{-1}Ng$.

The group consisting of the set of all right cosets, endowed with this operation, is called the **quotient group** and is denoted by G/N.

Let $\varphi : G \to G/N$ be the function $g \mapsto Ng$. This map is a homomorphism, automatically by virtue of the formula defining the group operation in G/N. Moreover, $\varphi(g) = Ne$ if and only if $g \in N$, so $\ker(\varphi) = N$. *Beginning with an arbitrary normal subgroup N, we have constructed a homomorphism whose kernel is N.* In the last chapter we showed that the kernel of a homomorphism (of groups) is always a normal subgroup, and now we have shown that any normal subgroup is the kernel of a homomorphism.

Perhaps we should give examples, and of course many could be mentioned. Cartesian products are a rather rich source: for any two groups G and H we can endow $G \times H$ with a natural group structure by defining the group operation to be $(g,h)(g',h') := (gg',hh')$. (If you check to make sure that this satisfies the definition of a group, as you should, you'll easily see that this construction actually works for any number of groups.) If G and H are abelian, then so is $G \times H$, in which case all subgroups are normal. You might enjoy thinking about the various subgroups of $\mathbb{Z} \times \mathbb{Z}$ and their associated quotient groups. (Figure 2.1 in Section 2.5 is a relatively complicated instance.) If you pursue this topic, which is surprisingly rich, you'll find that it leads naturally into the study of the subgroups and quotient groups of $\mathbb{Z}_m \times \mathbb{Z}_n$. (Here $\mathbb{Z}_m$ and $\mathbb{Z}_n$ are regarded as additive groups.)

Pursuing the analogy with group theory a little further, it would seem to make sense to define a **simple ring** to be a ring R whose only two sided ideals are $\{0\}$ and R itself, and indeed this is a standard definition. What kind of rings are simple? The first thing to say is that this is an excellent question! Simple groups have a rich theory, so this might also be true of simple rings. Even if the answer is not very complex, it might prove to be a stepping stone to more interesting questions. And it might be a fun challenge to work it out.

It turns out that simple rings are not so, umm, "simple," and in fact there are some pretty famous theorems that address this issue. (Even worse, there are actually things called "semisimple" rings.) We'll restrict our focus to the commutative case, so for the rest of this section all the rings are commutative. For a commutative ring there is no distinction between left ideals, right ideals (I think you can guess what the definitions of these concepts are) and two sided ideals, so from now on we'll use the word "ideal" to describe these things.

Let R be any commutative ring. There are two sorts of ideal we'll need to consider. For each $a \in R$, $aR = \{\, ar : r \in R \,\}$ is an ideal, because

$$ar + as = a(r+s) \in aR \quad \text{and} \quad s(ar) = a(rs) \in aR$$

for all $r, s \in R$. In the standard notation of ring theory this ideal is denoted by (a), it is called the **principal ideal** generated by a, and an ideal is said to be **principal** if it has this form. Of course $(0) = \{0\}$ is such an ideal. Next, consider

$$I_0 = \{\, a \in R : (a) = (0) \,\}.$$

This is an ideal: if $a, b \in I_0$, then $a + b \in I_0$ because

$$(a+b) = \{\, (a+b)r : r \in R \,\} \subset \{\, ar + bs : r, s \in R \,\} = (a) + (b) = (0),$$

and if r is any element of R, then $ar \in I_0$ because

$$(ar) = \{\, ars : s \in R \,\} \subset (a) = (0).$$

Now suppose that R is simple. Then either $I_0 = R$ or $I_0 = (0)$, and for each $a \in R$ we must have either $(a) = R$ or $(a) = (0)$, so if $I_0 = (0)$, then $(a) = R$ for all nonzero $a \in R$. We consider these two possibilities in turn.

If $I_0 = R$, then multiplication is identically zero. Indeed, whenever G is a commutative group with the group operation thought of as addition (by the way, this is standard practice in discussions of commutative groups) we can turn it into a commutative ring by defining the product of any two elements to be 0. (This multiplication is associative, distributive, and commutative, obviously.) We'll say that such "multiplication" is **trivial**. For a ring with trivial multiplication, the ideals will be the subgroups of the underlying commutative group: by definition any subgroup H is closed under addition and negation, and contains 0, so it also contains the product of any element of itself and any element of G. So, the ring derived from G will be simple if and only if the only subgroups of G are $\{0\}$ and G itself. Since any subgroup of a commutative group is normal, this is case if and only if G is a simple group. For any $g \in G$ there is a subgroup

$$\{\ldots, -g-g, -g, 0, g, g+g, \ldots\}$$

called the **cyclic subgroup generated by** g. We say that G is **cyclic** if it is equal to one of its cyclic subgroups, but the requirement that G have no subgroups other than $\{0\}$ and G itself implies something stronger, namely that G is equal to *each* of its nonzero cyclic subgroups. Since it's a bit to the side of the main thrust of our discussion, and because I think you might find it to be an interesting and enjoyable challenge, I'll leave it to you to work out what the simple cyclic groups are. Here's a hint: start out by writing down the simplest (in the usual sense of this word) examples of commutative groups you can think of.

Now suppose that $(a) = R$ for all $a \in R \setminus \{0\}$. That is, $aR = R$ for all $a \neq 0$. Recall that a ring element a is a **zero divisor** if $a \neq 0$ and there is a nonzero b such that $ab = 0$. It is natural to ask whether there can be any zero divisors, and easy to see that the answer is "no," since we would have

$$(0) = 0R = (ab)R = a(bR) = aR = R.$$

This is impossible unless $R = \{0\}$, but then $a \neq 0 \neq b$ is impossible.

Next, observe that for each nonzero a we have $a \in (a)$, so there is some $1_a \in R$ satisfying $a1_a = a$. Naturally we would like to prove that $1_a = 1_b$ for all nonzero a, b, so that there is a multiplicative identity. This actually takes a bit of cleverness. First, observe that $a(1_a1_a) = (a1_a)1_a = a1_a$, so $a(1_a^2 - 1_a) = 0$. As there are no zero divisors, it follows that $1_a^2 = 1_a$. There is now the computation

$$1_a(1_a - 1_b)1_b = 1_a^21_b - 1_a1_b^2 = 1_a1_b - 1_a1_b = 0.$$

Again, there are no zero divisors, so it must be the case that $1_a = 1_b$. Denoting the common identity element by 1 (of course!) observe that for each nonzero a we have $1 \in (a)$, so a has a multiplicative inverse. Now look back at the axioms for a field. Any ring satisfies (F1)-(F5) and (F9), we assumed also that R satisfies (F8), and we have just proved that it also satisfies (F6) and (F7). To celebrate, let's sum it all up:

Theorem 2.19. *A simple commutative ring is either a field or a simple cyclic group with trivial multiplication.*

2.4 Prime Factorization

The last result was interesting in and of itself, but it also suggests a general principal: the difference between fields and more general commutative rings is that commutative rings that aren't fields (or abelian groups with trivial multiplication) have nontrivial ideals, so understanding a ring is, to a large extent at least, a matter of understanding the ring's ideals. Now we'll apply this principal to the concept of unique factorization into primes. The main ideas are ones that many people learn in grade school, but by taking an abstract approach we'll be able to show that the principle of unique factorization into primes holds in $\mathbb{Z}[i]$ (elements of $\mathbb{Z}[i]$ are called **Gaussian integers** because Gauss made a deep study of the properties of this ring) and $k[X]$ for any field k.

All the rings considered in this section will be commutative. As we noticed earlier, a ring without a multiplicative identity is a bit of a strange beast, so we won't consider that possibility either. In a discussion of prime factorization, zero divisors would tend to just get in the way, so we'll only consider rings that don't have any. As it happens, a commutative ring with unit that has no zero divisors is called an **integral domain**. Let R be such a ring.

For $a, b \in R$ we say that a **divides** b, and write $a|b$, if $b = ar$ for some $r \in R$. A ring element that divides 1 is called a **unit**. The units in $\mathbb{Z}$ are 1 and -1, the units in $\mathbb{Z}[i]$ are 1, i, -1, and $-i$, and if k is a field, then the units in $k[X]$ are the nonzero constant polynomials because every element of k^* is a unit of k. Some authors extend our notational convention by letting R^* denote the set of units of R. This is a commutative group with multiplication as the group operation. To show this we observe that if $u_1v_1 = 1 = u_2v_2$, then $(u_1u_2)(v_1v_2) = 1$, so a product of two units is a unit. Of course 1 is a unit, and any unit has a multiplicative inverse that is also a unit, because this is just what the definition of a unit says.

A ring element a is **irreducible** if, whenever $a = bc$, either b is a unit or c is a unit. For most people this definition is what they think of as the definition of primality, but recall that in the last section we said that a is **prime** if, whenever $a|bc$, either $a|b$ or $a|c$.

Lemma 2.20. *If $a \in R$ is prime, then it is irreducible.*

Proof. Suppose $a = bc$. Then $a|bc$, so either $a|b$ or $a|c$, and (because we could interchange b and c) we may assume that $a|b$, i.e., $b = ad$ for some d. Then $a = (ad)c = a(dc)$, so $dc = 1$ and c is a unit. □

Eventually we'll see some examples of irreducible ring elements that are not prime.

An **irreducible factorization** of $a \in R$ is a representation of a of the form

$$a = p_1 \cdot p_2 \cdot \ldots \cdot p_k$$

where $p_1, \ldots, p_k$ are irreducible. We say that this representation is **unique** if any other irreducible factorization differs only in the ordering of the irreducibles and multiplication of the irreducibles by units. To explain exactly what we mean by this recall that the symmetric group S_k is the set of bijections

$$\sigma : \{1, \dots, k\} \to \{1, \dots, k\}.$$

Such a bijection is called a **permutation**. We will regard the factorization as unique if any other irreducible factorization

$$a = q_1 \cdot q_2 \cdot \ldots \cdot q_\ell$$

has the same number of factors, so $\ell = k$, and there are units $u_1, \dots, u_k$ and a permutation $\sigma \in S_k$ such that $q_i = u_i p_{\sigma(i)}$ for all $i = 1, \dots, k$.

An integral domain R is a **unique factorization domain**, or UFD, if every nonzero ring element has a unique irreducible factorization.

Proposition 2.21. *If R is a UFD and $a \in R$ is irreducible, then a is prime.*

Proof. Suppose $a|bc$, so that $bc = ad$ for some d. Let $b = p_1 \cdot \ldots \cdot p_k$, $c = q_1 \cdot \ldots \cdot q_\ell$, and $d = r_1 \cdot \ldots \cdot r_m$ be irreducible factorizations. Then

$$p_1 \cdot \ldots \cdot p_k \cdot q_1 \cdot \ldots \cdot q_\ell \quad \text{and} \quad a \cdot r_1 \cdot \ldots \cdot r_m$$

are irreducible factorizations of bc. By uniqueness, a must be the product of a unit and some element of the list $p_1, \dots, p_k, q_1, \dots, q_\ell$, and consequently $a|b$ or $a|c$. □

Our goal is to show that $\mathbb{Z}$, $\mathbb{Z}[i]$, and $k[X]$ are UFD's. To this end we will study the relationship between unique factorization in an integral domain R and the ideals of R.

A **principal ideal domain**, or PID, is an integral domain whose ideals are all principal, so that for any ideal I there is some ring element a with $(a) = I$. The examples of rings we have seen so far don't include any that aren't PID's (for $\mathbb{Z}[\sqrt{2}]$ this is far from obvious) but in the larger world of commutative rings, or even integral domains, the PID's are quite special.

In view of Proposition 2.21, the lemma below will be superfluous once we show that every PID is a UFD, but we need it in the proof of that result.

Lemma 2.22. *Irreducible elements of a principal ideal domain R are prime.*

Proof. Supposing that a is irreducible, and that a divides bc, we will show that a divides either b or c. Extending our notation for principal ideals, let

$$(a, b) := \{\, ra + sb : r, s \in R \,\}.$$

It is straightforward to check that (a, b) is an ideal of R, so, since R is a PID, $(a, b) = (d)$ for some d. There are now two possibilities: either d is a unit or it isn't.

In the first case there is a u such that $du = 1$, and for any $r \in R$ we have $r = d(ur) \in (d)$, so $(d) = R$. In particular, there are $x, y \in R$ such that $xa + yb = 1$. But then $c = xac + ybc$, after which $a|c$ follows from the assumption that $a|bc$.

Now suppose that d is not a unit. We have $a = ud$ for some u because $a \in (d)$, and if d is not a unit then u must be a unit because a is irreducible, so $a|d$. But $d|b$ because $b \in (d)$, so it follows that $a|b$. □

The next proof is probably the hardest in the book up to this point. In part this is simply because it's a bit long, but I'm sure that if you take things one step at a time you'll be able to convince yourself that the logic is ironclad. The other difficulty is that it's not so easy to see how someone would have thought of it, and in fact it may not be the case that someone did. Especially in abstract algebra, arguments evolve over time, and are sometimes transported from one context where things seem intuitive to another where the ideas are less obvious. But there is also a way of thinking about proofs which allows one to discover arguments like this one: instead of trying to prove everything all at once, think about how to prove something much weaker, or some small piece of the desired conclusion.

Proposition 2.23. *If R is a PID, then it is a UFD.*

Proof. Fix an $a \in R$ that is not a unit. The proof has three steps.

Step 1: a has an irreducible factor. If a is irreducible we are done, so suppose that $a = a_1 b_1$ where a_1 and b_1 are not units. Then $a_1 \notin (a)$ because otherwise $a_1 = ac$ for some c, and $a_1 = ac = (a_1 b_1)c = a_1(b_1 c)$ would imply that $b_1 c = 1$, so that b_1 was a unit after all. If a_1 is not irreducible, then the same argument with a_1 in place of a gives $a_1 = a_2 b_2$ where a_2 and b_2 are not units and $a_2 \notin (a_1)$. If continuing this process never arrives at an irreducible factor of a, then there is a strictly increasing sequence of ideals

$$(a) \subset (a_1) \subset (a_2) \subset \dots.$$

Let $I := \bigcup_{k=1}^{\infty} (a_k)$. If $i, j \in I$ and $r \in R$, then $i, j \in (a_k)$ for some k, so that I contains $i + j$ and ri because they are elements of (a_k). Thus I is an ideal, so $I = (c)$ for some c. But then $c \in (a_k)$ for some k, and $a_{k+1} \in (c) \subset (a_k)$, which is a contradiction.

Step 2: a has an irreducible factorization. Step 1 gives a factorization $a = b_1 c_1$ where b_1 is irreducible. If c_1 is a unit or irreducible we are done, and otherwise Step 1 implies that $c_1 = b_2c_2$ with b_2 irreducible. If $c_2 = dc_1$ for some d, then $c_2 = dc_1 = (db_2)c_2$, which is impossible because b_2 is not a unit, so $c_2 \notin (c_1)$. Again, if c_2 is a unit or irreducible we are done, and otherwise Step 1 implies that $c_2 = b_3c_3$ with b_3 irreducible, so that $c_3 \notin (c_2)$ because b_3 is not a unit. If this process continues forever then, as in Step 1, $I = \bigcup_{k=1}^{\infty}(c_k)$ is an ideal, $I = (d)$ for some d, there is k such that $d \in (c_k)$, and $c_{k+1} \in (d) \subset (c_k)$, which is a contradiction. Therefore the process must eventually halt at an irreducible factorization.

Step 3: The irreducible factorization of a is unique. Suppose that $p_1 \cdot \ldots \cdot p_k$ and $q_1 \cdot \ldots \cdot q_\ell$ are two distinct irreducible factorizations of a. We may assume that among all the elements of R that have multiple irreducible factorizations, and among all the pairs of distinct irreducible factorizations, this is one for which $\max\{k, \ell\}$ is minimal. Since $p_k|a$, and irreducible elements of a PID are prime, $p_k|q_i$ for some i. After reordering, we may assume that $i = \ell$, so we obtain $q_\ell = up_k$, where u must be a unit because q_ℓ is irreducible. Then

$$p_1 \cdot \ldots \cdot p_{k-1} = (uq_1) \cdot q_2 \cdot \ldots \cdot q_{\ell-1},$$

contradicting our assumption that $\max\{k, \ell\}$ is minimal. □

If we can show that $\mathbb{Z}$, $\mathbb{Z}[i]$, and $k[X]$ are PID's, it will follow that these rings are UFD's. To show that they're PID's, a viable method is to hook up with the Euclidean algorithm for computing greatest common divisors by repeated division with remainder. I was taught this in grade school, but maybe your grade school was different, so let's quickly review it. Beginning with two integers, say $3,640,227$ and $364,531$, whose greatest common divisor we'd like to compute, we divide the bigger one by the smaller one:

$$3,640,227 = 9 \times 364,531 + 359,448.$$

The key point is that any common factor of the two numbers is also a factor of the remainder, and in fact a common factor of the remainder and the smaller of the two numbers we started with, so we can divide again:

$$364,531 = 1 \times 359,448 + 5083.$$

Continuing in this manner:

$$359,448 = 70 \times 5083 + 3638;$$

$$5083 = 1 \times 3638 + 1445;$$

$$3638 = 2 \times 1445 + 748;$$

$$1445 = 1 \times 748 + 697;$$

$$748 = 1 \times 697 + 51;$$

$$697 = 13 \times 51 + 34;$$

$$51 = 1 \times 34 + 17;$$

$$34 = 2 \times 17.$$

Any common divisor of $3,640,227$ and $364,531$ is a common divisor of every number in the list of remainders, and the final nonzero remainder, namely 17, is a common divisor of $3,640,227$ and $364,531$, so it must be the greatest common divisor.

The following definition and result extract and exploit the key property of the integers that makes this work. It will all happen quickly, without much apparent effort, and in fact one reason for reviewing the Euclidean algorithm, as we did above, is that otherwise the algorithmic aspect might not be apparent.

Definition 2.24. *An integral domain R is **Euclidean** if there is a function $\nu : R \to \mathbb{Z}_{\geq}$ (where $\mathbb{Z}_{\geq} := \{0, 1, 2, , \ldots\}$ is the set of nonnegative integers) such that:*

(a) $\nu(a) = 0$ if and only if $a = 0$;

(b) for any $a, b \in R$ with $b \neq 0$ there exist $q, r \in R$ such that $a = qb + r$ and $\nu(r) < \nu(b)$.

Theorem 2.25. *If the integral domain R is Euclidean, then it is a principal ideal domain.*

Proof. Let I be an ideal of R. Since (0) is principal, we may assume that I has a nonzero element. Choose a nonzero $b \in I$ for which $\nu(b)$ is minimal. For any $a \in I$ there exist $q, r \in R$ with $r = a - qb$ and $\nu(r) < \nu(b)$, and since $r \in I$, this implies that $r = 0$, so that $b|a$. Since a was an arbitrary element of I, we have shown that $I \subset (b)$, but the definition of an ideal implies that $(b) \subset I$, so $I = (b)$. □

The remaining task is to show that $\mathbb{Z}$, $\mathbb{Z}[i]$, and $k[X]$ are Euclidean, after which it will follow that they are PID's and UFD's. We need to find suitable functions

$$\nu_{\mathbb{Z}} : \mathbb{Z} \to \mathbb{Z}_{\geq}, \;\; \nu_{k[X]} : k[X] \to \mathbb{Z}_{\geq}, \;\text{ and }\; \nu_{\mathbb{Z}[i]} : \mathbb{Z}[i] \to \mathbb{Z}_{\geq}.$$

For $\mathbb{Z}$ this is easy: let

$$\nu_{\mathbb{Z}}(a) = |a|.$$

Because it will occur in an argument below we mention that for any $a, b \in \mathbb{Z}$ with $b \neq 0$ it is actually possible to find an integer q such that $|a-qb| \leq |b|/2$.

There is also a simple and obvious idea that works for $k[X]$. The **degree** $\deg(P)$ of a nonzero $P \in k[X]$ is the largest power of X that appears in P with a nonzero coefficient. For example, $\deg(3) = 0$ and $\deg(4X^2-5X+1) = 2$. We set

$$\nu_{k[X]}(P) := \begin{cases} 0, & P = 0, \\ \deg(P) + 1, & \text{otherwise.} \end{cases}$$

If you had a decent algebra course in high school or before, then it's probably obvious to you that for any nonzero $P_0, P_1 \in k[X]$ with $P_1 \neq 0$ we can use polynomial division with remainder to produce $Q, R \in k[X]$ with $P_0 = QP_1 + R$ and $\nu_{k[X]}(R) < \nu_{k[X]}(P_1)$. If it's not obvious, think about it carefully until you understand it.

For $\mathbb{Z}[i]$ the situation is a bit more complicated. The **modulus** (or **complex norm**, or **absolute value**) of a complex number $a = x + iy$ is

$$|a| := \sqrt{x^2 + y^2}.$$

In view of the Pythagorean theorem, this is just the distance from the origin 0 to a when we identify it with the point (x, y) in the plane. We'll need to know that $|ab| = |a|\,|b|$, which is proven by the following calculation:

$$\begin{aligned} |ab|^2 &= |(x+iy)(z+iw)|^2 \\ &= |(xz-yw) + i(xw+yz)|^2 \\ &= (xz-yw)^2 + (xw+yz)^2 \\ &= x^2z^2 - 2xzwy + y^2w^2 + x^2w^2 + 2xwyz + y^2z^2 \\ &= x^2z^2 + y^2w^2 + x^2w^2 + y^2z^2 \\ &= (x^2+y^2)(z^2+w^2) = |a|^2|b|^2. \end{aligned}$$

The modulus will play an important role later, and eventually we will have a formula that allows a proof of this that is not just a "miraculous" computation.

The modulus is essentially the right concept, but, technically speaking, the definition of a Euclidean ring requires an integer valued function, so we'll use the square of this quantity. That is, for $a = x + iy \in \mathbb{Z}[i]$ we set

$$\nu_{\mathbb{Z}[i]}(a) := |a|^2 = x^2 + y^2.$$

Our goal now is to show that for any $a, b \in \mathbb{Z}[i]$ with $b \neq 0$ there is $q = u + iv \in \mathbb{Z}[i]$ such that $|a - qb|^2 < |b|^2$.

First suppose that $b \in \mathbb{Z}$, so that b is real, and let $a = x + iy$. We can find $u, v \in \mathbb{Z}$ such that $|x - ub| \leq \frac{1}{2}|b|$ and $|y - vb| \leq \frac{1}{2}|b|$, after which

$$|a - qb|^2 = |(x - ub) + i(y - vb)|^2 = (x - ub)^2 + (y - vb)^2 \leq (\tfrac{1}{4} + \tfrac{1}{4})|b|^2.$$

Turning to the general case, let $b = z + iw$. The **complex conjugate** of b is $\overline{b} := z - iw$. We use the fact that $b\overline{b} = z^2 + w^2$ is real, so that the special case with $a\overline{b}$ and $b\overline{b}$ in place of a and b gives a q such that $|a\overline{b} - qb\overline{b}| < |b\overline{b}|$. But now we have

$$|a - qb|^2 = \frac{|a - qb|^2|\overline{b}|^2}{|\overline{b}|^2} = \frac{|a\overline{b} - qb\overline{b}|^2}{|\overline{b}|^2} \leq \frac{|b\overline{b}|^2}{2|\overline{b}|^2} = \frac{|b|^2|\overline{b}|^2}{2|\overline{b}|^2} = \tfrac{1}{2}|b|^2.$$

2.5 Algebraic Integers and Modules

Now consider the following fact:

$$(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6 = 3 \cdot 2.$$

This looks suspiciously like a failure of unique factorization, but what is the ring? We're going to develop some rather advanced ideas to explain this, and you may find this section a bit more difficult that what has come before and what will come later, but the ideas will have important echos, and there are some interesting stories to tell.

Definition 2.26. *Let R be an integral domain. A polynomial $P \in R[X]$ is **monic** if the leading coefficient is 1, so it is of the form*

$$P(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0.$$

*A complex number is an **algebraic integer** if it is a root of monic polynomial in $\mathbb{Z}[X]$.*

Before you've seen it, this is not an obvious way to define the notion of an algebraic integer, and once you've seen it, it's not at all clear that it's "the right" definition. It works well enough as a way of distinguishing ordinary integers from other rational numbers. Any $a \in \mathbb{Z}$ is an algebraic integer because it is a root of the monic polynomial $X - a$. On the other hand, if b/c is a fraction in lowest terms with $c > 1$, then b/c is not an algebraic integer. Specifically, for the monic polynomial P we have $P(b/c) \neq 0$ because if p

is a prime factor of c, then it is not a factor of b, so the power of p in the denominator of $(b/c)^n$ is greater than the power of p in c^{n-1}, which is in turn at least as large as the power of p in the denominator of

$$a_{n-1}(b/c)^{n-1} + \cdots + a_1(b/c) + a_0 = \frac{a_{n-1}b^{n-1} + \cdots + a_1bc^{n-2} + a_0c^{n-1}}{c^{n-1}}$$

when this fraction is reduced to lowest terms.

But at first sight it seems strange to regard

$$\frac{3+\sqrt{5}}{2} \quad \text{and} \quad \frac{3-\sqrt{5}}{2}$$

as integers, even though they are the roots of the equation $X^2 - 3X + 1 = 0$. Historically, the definition of an algebraic integer was eventually accepted, primarily as a result of experience, because it led to a coherent theory and, in many other respects, proved extremely fruitful. We're going to explain one crucial piece of this:

Theorem 2.27. *If α and β are algebraic integers, then so are $\alpha + \beta$ and $\alpha\beta$.*

Corollary 2.28. *The set of algebraic integers is an integral domain.*

Proof. For any ring R, a nonempty subset S is a subring if it contains all additive inverses, sums, and products of its elements. (Noting that $0 \in S$ because $0 = -a + a$ for any $a \in S$, you should quickly check that Axioms (R1)-(R6) are satisfied by S because they hold in R.) If, in addition, R is an integral domain, then multiplication in S is commutative and S has no zero divisors, so if $1 \in S$, then S must also be an integral domain.

The set of algebraic integers is a subset of $\mathbb{C}$ that contains 1, obviously, and sums and products of its elements. To see that it also contains negations of its elements, observe that if α is an algebraic integer by virtue of being a root of our monic polynomial P, then $-\alpha$ is a root of

$$X^n - a_{n-1}X^{n-1} + \cdots + (-1)^{n-1}a_1X + (-1)^n a_0.$$

□

The proof of Theorem 2.27 provides an opportunity to introduce a very general and important concept.

Definition 2.29. *If R is a ring with unit, a* ***left R-module*** *is a commutative group M (with the group operation written additively) for which there is a binary operation $R \times M \to M$ (written multiplicatively) such that:*

(a) $1m = m$ *for all* $m \in M$*;*

(b) $r(m_1 + m_2) = rm_1 + rm_2$ *for all* $m_1, m_2 \in M$ *and all* $r \in R$*;*

(c) $(r_1 + r_2)m = r_1m + r_2m$ *for all* $m \in M$ *and* $r_1, r_2 \in R$*;*

(d) $r(sm) = (rs)m$ *for all* $m \in M$ *and* $r, s \in R$*.*

Right R-modules are defined similarly. When R is commutative, the distinction between left and right R-modules is purely typographical, and they are usually described simply as modules. In some books some modules are written as right modules, but we will always have scalars act from the left.

There are many different types of modules, and they are the central objects of interest in numerous areas of mathematics. If R is a field, then an R-module is called a **vector space**. Vector spaces are enormously important in themselves, and Chapter 4 is devoted to the most basic facts about them. Even now, before having done much of anything, it is easy to list several other examples:

(a) There is a trivial R-module whose only element is 0.

(b) Any ideal of R (including R itself) is an R-module.

(c) For any set S there is an R-module structure on $\mathcal{F}_R(S)$ given by defining $rf : S \to R$, for $f \in \mathcal{F}_R(S)$, to be the function $s \mapsto rf(s)$. For any integer $k \geq 1$ we can regard R^k as an R-module by identifying it with $\mathcal{F}_R(\{1, \dots, k\})$.

(d) If R is a subring of S, then S is an R-module. If we wish to call attention to the R-module structure of S, we say that S is an R-**algebra**. In particular, $R[X]$ is an R-algebra if we identify R with the subring of constant polynomials in $R[X]$.

Possibly you already sense what is coming next. There will be homomorphisms, a category, submodules, kernels, and quotient modules. The basic facts about these things will be straightforward, and quite similar to what we have seen in connection with groups and rings. The only problematic aspect of this is that, insofar as none of it is the least bit problematic, there is really no better approach than just plowing through it, which will make for a patch of rather dull reading.

Definition 2.30. *If* M *and* N *are* R*-modules, a function* $\varphi : M \to N$ *is an* R***-module homomorphism*** *if:*

(a) $\varphi(m_1 + m_2) = \varphi(m_1) + \varphi(m_2)$ *for all* $m_1, m_2 \in M$ *(i.e.,* φ *is a homomorphism of the underlying commutative groups);*

(b) $\varphi(rm) = r\varphi(m)$ *for all* $m \in M$ *and* $r \in R$.

If φ *is bijective, then it is an* R***-module isomorphism****, and* M *and* N *are said to be* ***isomorphic****.*

If φ is an R-module isomorphism, then so is φ^{-1}. The proof is an extension of analogous arguments we saw earlier. In our discussion of groups we showed that the inverse of an isomorphism of groups is a homomorphism, hence an isomorphism, so φ^{-1} is a homomorphism of the underlying abelian groups. To see that φ^{-1} also satisfies (b) consider $n \in N$ and $r \in R$, set $m := \varphi^{-1}(n)$, and compute that

$$\varphi^{-1}(rn) = \varphi^{-1}(r\varphi(m)) = \varphi^{-1}(\varphi(rm)) = rm = r\varphi^{-1}(n).$$

In what is, by now, the "usual" sort of way with homomorphisms, if M, N, and P are R-modules and $\varphi : M \to N$ and $\psi : N \to P$ are R-module homomorphisms, then so is $\psi \circ \varphi$ because it is a homomorphism of the underlying abelian groups and

$$\psi(\varphi(rm)) = \psi(r\varphi(m)) = r\psi(\varphi(m))$$

for all $m \in M$ and $r \in R$. Composition of R-module homomorphisms is associative because composition of functions is associative. In addition, Id_M is always an R-module isomorphism, and

$$\varphi \circ \mathrm{Id}_M = \varphi = \mathrm{Id}_N \circ \varphi.$$

Thus, most unsurprisingly, there is a category of R-modules and R-module homomorphisms.

A **submodule** of an R-module M is a subset $M' \subset M$ that is itself an R-module. Concretely, $M' \subset M$ is a submodule if: (a) it is nonempty; (b) $-m \in M'$ for all $m \in M'$; (c) $m + m' \in M'$ for all $m, m' \in M'$ (so $0 = -m + m \in M'$); (d) $rm \in M'$ for all $m \in M'$ and $r \in R$. In this circumstance M' is a subgroup of the underlying commutative group of M, and the quotient group M/M' can be regarded as an R-module if we define the product of a ring element r with a coset $m + M' = \{\, m + m' : m' \in M' \,\}$ to be

$$r(m + M') := rm + M'.$$

(Here $r(m+M')$ may be a proper superset of $\{\, r(m+m') : m' \in M' \,\}$ if rM' is a proper subset of M'.) Please check for yourself that (a)-(d) of Definition 2.29 hold.

It is always the case that $\{0\}$ and M are submodules, and for any $r \in R$, and $m \in M$,

$$rM = \{\, rm : m \in M \,\} \quad \text{and} \quad Rm = \{\, rm : r \in R \,\}$$

are submodules. If A is any set and, for each $\alpha \in A$, M_α is a submodule of M, then $\bigcap_{\alpha \in A} M_\alpha$ is a submodule. If M' and M'' are submodules of V, then so is

$$M' + M'' := \{\, m' + m'' : m' \in M' \text{ and } m'' \in M'' \,\}.$$

The proof is trivial: if $m' + m'', \tilde{m}' + \tilde{m}'' \in M' + M''$, then $M' + M''$ contains

$$-(m' + m'') = (-m') + (-m'')$$

and

$$(m' + m'') + (\tilde{m}' + \tilde{m}'') = (m' + \tilde{m}') + (m'' + \tilde{m}''),$$

and if, in addition, $r \in R$, then $M' + M''$ contains and $r(m' + m'') = (rm') + (rm'')$.

Suppose $\varphi : M \to N$ is an R-module homomorphism. The image $\varphi(M)$ of φ is a submodule of N because it is nonempty, it contains $-n = \varphi(-m)$ and $n + n' = \varphi(m + m')$ whenever $n = \varphi(m)$ and $n' = \varphi(m')$ are elements, and it contains $rn = \varphi(rm)$ whenever $n = \varphi(m)$ is an element and $r \in R$. If N' is a submodule of N, then $\varphi^{-1}(N')$ is a submodule of M: $\varphi^{-1}(N')$ is a subgroup of M because φ is a homomorphism of groups, and $\varphi(rm) = r\varphi(m) \in N'$ whenever $m \in \varphi^{-1}(N')$ and $r \in R$. The **kernel** of φ is

$$\ker(\varphi) := \varphi^{-1}(0).$$

Since $\{0\}$ is a submodule of N, $\ker(\varphi)$ is a submodule of M. For any submodule M' the map $\varphi : m \mapsto m + M'$ is an R-module homomorphism from M to M/M' because

$$\varphi(m_1 + m_2) = (m_1 + m_2) + M' = (m_1 + M') + (m_2 + M') = \varphi(m_1) + \varphi(m_2)$$

for all $m_1, m_2 \in M$ and

$$\varphi(rm) = rm + M' = r(m + M') = r\varphi(m)$$

for all $m \in M$ and $r \in R$. Clearly $\ker(\varphi) = M'$, so this construction shows that every submodule M' of M is the kernel of some R-module homomorphism.

The concept of an R-module is so general that one might expect that there wouldn't be much general theory, for more of less the same reason that there is little to say about categories or functions in general. Actually, things aren't this simple, but in our work R will almost always either be $\mathbb{Z}$ or a field. When a definition or argument can be phrased in general terms we will do so, but this will be incidental to the main thrust of our discussion.

We now focus on $\mathbb{Z}$-modules. The first point of interest is that scalar multiplication in a $\mathbb{Z}$-module M is determined by the group operation. That is, for any $m \in M$ we have $(-1)m = -m$ because

$$(-1)m = (-1)m + m - m = (-1+1)m - m = -m,$$

and for any positive integer r we have

$$rm = (1 + \cdots + 1)m = m + \cdots + m$$

where the two sums involve r copies of 1 and m respectively. Moreover, any commutative group can be made into a $\mathbb{Z}$-module by defining multiplication by elements of $\mathbb{Z}$ in this manner. In short, a $\mathbb{Z}$-module is just a commutative group, and vice versa.

At this point you might be wondering why we introduced the module concept, and laid out all the standard associated formalities, when all we really wanted to do was talk about commutative groups. There are several reasons. Our discussion will emphasize the $\mathbb{Z}$-module structure of the groups considered below. Modules occur frequently, so the definition itself conveys some useful sense of perspective. Most practically, if we didn't do the work here, we would have had to do it at the beginning of Chapter 4, so the costs are actually rather low.

We now begin to explain what this all has to do with algebraic numbers and integers.

Definition 2.31. *If M is an R-module, a set $G \subset M$ is a* ***set of generators*** *for M if every element of M is of the form $r_1g_1 + \cdots + r_pg_p$ where $r_1, \ldots, r_p \in R$ and $g_1, \ldots, g_p \in G$. We say that M is* ***finitely generated*** *if it has a finite set of generators.*

Theorem 2.32. *A number $\alpha \in \mathbb{C}$ is an algebraic integer if and only if the module $\mathbb{Z}[\alpha]$ is finitely generated.*

Proof. First suppose that α is an algebraic integer, so

$$\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_1\alpha + a_0 = 0$$

for some n and integers $a_0, \ldots, a_{n-1}$. We claim that $1, \alpha, \ldots, \alpha^{n-1}$ is a set of generators for $\mathbb{Z}[\alpha]$. Every element of $\mathbb{Z}[\alpha]$ is $P(\alpha)$ for some polynomial $P \in \mathbb{Z}[X]$, so it suffices to show that the submodule generated by $1, \alpha, \ldots, \alpha^{n-1}$ includes α^m for any $m \geq 0$. This is true automatically for $m \leq n-1$, and for $m \geq n$ we have

$$\alpha^m = -a_{n-1}\alpha^{m-1} - \cdots - a_0\alpha^{m-n}.$$

Now suppose that $\mathbb{Z}[\alpha]$ is a finitely generated $\mathbb{Z}$-module, so there are polynomials $P_1, \ldots, P_k \in \mathbb{Z}[X]$ such that $P_1(\alpha), \ldots, P_k(\alpha)$ is a set of generators for $\mathbb{Z}[\alpha]$. In particular, for any integer n there are $b_1, \ldots, b_k \in \mathbb{Z}$ such that

$$\alpha^n = b_1P_1(\alpha) + \cdots + b_kP_k(\alpha).$$

If n is greater than any power of X appearing in $P_1, \ldots, P_k$, then α is a root of the monic polynomial

$$X^n - b_1P_1 - \cdots - b_kP_k,$$

so α is an algebraic integer. □

If α and β are algebraic integers, then we can show that $\alpha + \beta$ and $\alpha\beta$ are also algebraic integers by showing that $\mathbb{Z}[\alpha + \beta]$ and $\mathbb{Z}[\alpha\beta]$ are finitely generated. Let $\mathbb{Z}[\alpha, \beta]$ be the smallest subring[3] of $\mathbb{C}$ containing $\mathbb{Z}$, α, and β. Suppose that $1, \alpha, \ldots, \alpha^{k-1}$ is a set of generators for $\mathbb{Z}[\alpha]$ while $1, \beta, \ldots, \beta^{\ell-1}$ is a set of generators for $\mathbb{Z}[\beta]$. Then

$$\{\, \alpha^i\beta^j : 0 \leq i \leq k-1 \text{ and } 0 \leq j \leq \ell - 1 \,\}$$

[3] In general, if S is a commutative ring with unit, R is a subring of S with $1 \in R$, and $A \subset S$, then $R[A]$ is the smallest subring of S that contains R and A. This definition makes sense because the intersection of all subrings that contain R and A is a subring of S that is contained in any subring of S that contains R and A. If $R[A] = T$, then we say that A is a **set of generators** for T over R. When $A = \{\alpha_1, \ldots, \alpha_k\}$ is finite we write $R[\alpha_1, \ldots, \alpha_k]$ rather than $R[\{\alpha_1, \ldots, \alpha_k\}]$, and we say that T is **finitely generated over** R. This is a much weaker condition than T being finitely generated as an R-module, it is important to be careful about distinguishing between the two concepts, and the terminology is unfortunately not very helpful about this.

is a set of generators for $\mathbb{Z}[\alpha, \beta]$ because any polynomial in α and β can be reduced, by the sorts of repeated substitutions we saw in the proof above, to an expression of the form

$$\sum_{i=0}^{k-1} \sum_{j=0}^{\ell-1} c_{ij} \alpha^i \beta^j.$$

Of course $\mathbb{Z}[\alpha + \beta]$ and $\mathbb{Z}[\alpha\beta]$ are submodules of $\mathbb{Z}[\alpha, \beta]$, so Theorem 2.27 will follow from Theorem 2.32 if we can prove that:

Proposition 2.33. *A submodule of a finitely generated $\mathbb{Z}$-module is finitely generated.*

Is this true? The answer is "yes," but perhaps not obviously so. Among other things, there are commutative rings R for which finitely generated R-modules can have submodules that are not finitely generated.

How should we think about trying to prove it? When you are getting started on thinking about a proof, one important idea is to reduce to a special case. We claim that *in order to show that any submodule of a finitely generated $\mathbb{Z}$-module is finitely generated, it suffices to show that for any k, any submodule of $\mathbb{Z}^k$ is finitely generated.*

Let M be a finitely generated $\mathbb{Z}$-module, say with generators $g_1, \ldots, g_k$, and let M' be a submodule. We will "pull the problem upstairs" to $\mathbb{Z}^k$. Consider the map $\varphi : \mathbb{Z}^k \mapsto M$ defined by the formula

$$\varphi(r_1, \ldots, r_k) := r_1 g_1 + \cdots + r_k g_k.$$

The verification that φ is a $\mathbb{Z}$-module homomorphism is straightforward. For any $(r_1, \ldots, r_k)$ and $(r_1', \ldots, r_k')$ we have

$$\begin{aligned} \varphi((r_1, \ldots, r_k) + (r_1', \ldots, r_k')) &= \varphi(r_1 + r_1', \ldots, r_k + r_k') \\ &= (r_1 + r_1')g_1 + \cdots + (r_k + r_k')g_k \\ &= (r_1 g_1 + \cdots + r_k g_k) + (r_1' g_1 + \cdots + r_k' g_k) \\ &= \varphi(r_1, \ldots, r_k) + \varphi(r_1', \ldots, r_k'). \end{aligned}$$

For any $(r_1, \ldots, r_k)$ and $q \in \mathbb{Z}$ we have

$$\varphi(q(r_1, \ldots, r_k)) = \varphi(qr_1, \ldots, qr_k) = qr_1 g_1 + \cdots + qr_k g_k$$

$$= q(r_1 g_1 + \cdots + r_k g_k) = q\varphi(r_1, \ldots, r_k).$$

Since φ is a homomorphism, $\varphi^{-1}(M')$ is a submodule of $\mathbb{Z}^k$. Suppose that it has a finite system of generators, say $h_1, \ldots, h_\ell$. Since M is generated by $g_1, \ldots, g_k$, φ is surjective, so $\varphi(\varphi^{-1}(M')) = M'$. Therefore

$$M' = \varphi(\varphi^{-1}(M')) = \varphi\bigl(\{\, r_1h_1 + \cdots + r_\ell h_\ell : r_1, \ldots, r_\ell \in R \,\}\bigr)$$
$$= \{\, r_1\varphi(h_1) + \cdots + r_\ell\varphi(h_\ell) : r_1, \ldots, r_\ell \in R \,\},$$

so $\varphi(h_1), \ldots, \varphi(h_\ell)$ generates M'. Thus, if we can show that any submodule of $\mathbb{Z}^k$ is finitely generated, it will follow that $\varphi^{-1}(M')$ is finitely generated, and in turn this implies that M' is finitely generated.

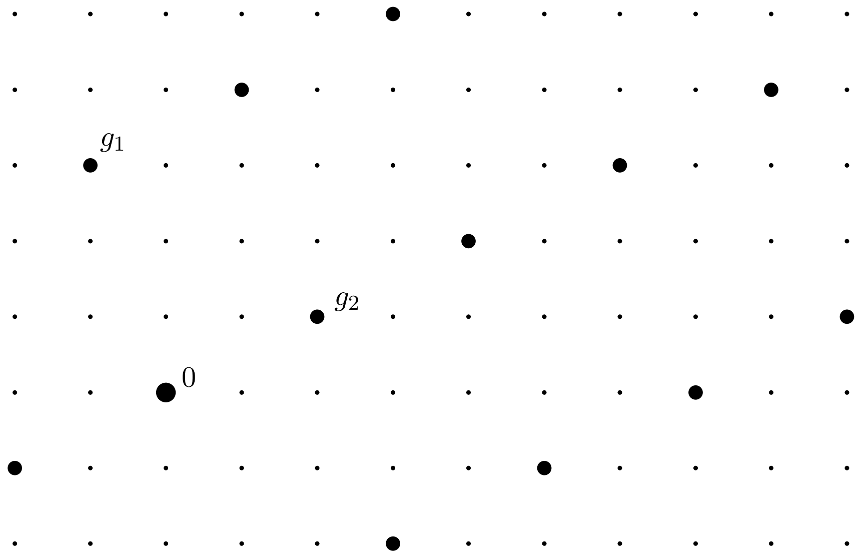


Figure 2.1

Another important way to think about proofs is to look at pictures. Figure 2.1 shows a typical submodule of $\mathbb{Z}^2$ that is clearly generated by g_1 and g_2. It seems that if $\mathbb{Z}^2$ had a submodule that was not finitely generated, you might have come across it before, so probably every submodule of $\mathbb{Z}^2$ is finitely generated, and in fact it looks like any submodule of $\mathbb{Z}^2$ is generated by two or fewer elements. Since Figure 2.1 looks simple, one is inclined to guess that any submodule of $\mathbb{Z}^k$ has a system of generators with k or fewer elements, and that this really shouldn't be too hard to prove.

A third method of thinking about proofs is to look at the simplest special cases. So consider the case $k = 1$. A submodule of $\mathbb{Z}$ is an ideal, and we have proved that $\mathbb{Z}$ is a PID, so any submodule is generated by a single element! So far, so good, but how to go forward from here? In this sort of situation there are two fundamentally different approaches. We could try to generalize the argument used to prove that $\mathbb{Z}$ is a PID, but this doesn't seem promising. Among other things, it is not clear how to generalize the Euclidean hypothesis. The second method is to use induction on k, taking advantage of the work we have already done to get the induction started. This is what we do.

Once you've completed a proof, it's a good idea to think about whether you've really proved more than you set out to prove. Can the assumptions be weakened or the conclusion strengthened. It turns out that the only fact about $\mathbb{Z}$ used in the argument below is that it is a PID, so we strengthen the statement accordingly.

Lemma 2.34. *Let R be a PID. Suppose the R-module M has a system of generators with k elements. Then any submodule M' has a system of generators with at most k elements.*

Proof. The argument we saw above (which works equally well with $\mathbb{Z}$ replaced by R) shows that it suffices to prove this in the special case $M = R^k$. When $k = 1$ a submodule of M is, in effect, an ideal of R, so it is either (0) or it is generated by a single element because R is a PID. By the principle of induction, we may assume that it has already been shown that any submodule of R^{k-1} has a system of generators with at most $k-1$ elements.

Let $\pi : R^k \to R$ be the function

$$\pi(r_1, \dots, r_k) := r_k.$$

Perhaps you're getting used to the standard maneuvers to the point that it seems obvious that π is a R-module homomorphism, but we write out the verification anyway:

$$\pi((r_1, \dots, r_k) + (r'_1, \dots, r'_k)) = \pi(r_1 + r'_1, \dots, r_k + r'_k)$$

$$= r_k + r'_k = \pi(r_1, \dots, r_k) + \pi(r'_1, \dots, r'_k)$$

and

$$\varphi(q(r_1, \dots, r_k)) = \varphi(qr_1, \dots, qr_k) = qr_k = q\varphi(r_1, \dots, r_k).$$

Let M' be a submodule of R^k. If $M' \subset \pi^{-1}(0) = R^{k-1}$, then the claim follows from the induction hypothesis, so we may assume that $\pi(M')$ is a nontrivial submodule (i.e., an ideal different from (0)) of R. Then $\pi(M') = (h_1)$ for some nonzero $h_1 \in R$, and we can choose $g_1 \in M'$ such that $\pi(g_1) = h_1$. Since $\ker(\pi) \cap M'$ is a submodule of $\ker(\pi) = R^{k-1}$ it has a system of generators $g_2, \dots, g_\ell$ where $\ell \leq k$.

To prove that $g_1, \dots, g_\ell$ is a system of generators for M' consider any $m \in M'$. For some r_1 we have $\pi(m) = r_1 h_1$. Since $\pi(m - r_1 g_1) = \pi(m) - r_1 \pi(g_1) = 0$, there are $r_2, \dots, r_\ell \in R$ such that $m - r_1 g_1 = r_2 g_2 + \cdots + r_\ell g_\ell$. Thus, as desired, we have

$$m = r_1 g_1 + r_2 g_2 + \cdots + r_\ell g_\ell.$$

□

2.6 Fermat's Last Theorem

At this point we've proved Theorem 2.27. Where does that leave us? The first point is that if α is an algebraic integer, then Theorem 2.27 implies that every element of $\mathbb{Z}[\alpha]$ is an algebraic integer. More generally, if $\alpha_1, \dots, \alpha_k$ are algebraic integers, then $\mathbb{Z}[\alpha_1, \dots, \alpha_k]$ is finitely generated, so every element of this ring is an algebraic integer. As it happens, a result with the exotic and suggestive name "The Theorem of the Primitive Element" implies that for any algebraic integers $\alpha_1, \dots, \alpha_k$ there is an algebraic integer β such that

$$\mathbb{Z}[\alpha_1, \dots, \alpha_k] = \mathbb{Z}[\beta].$$

It can easily happen that $\mathbb{Z}[\alpha]$ is not a UFD. To illustrate this we now complete the explanation of the apparent failure of prime factorization given by the calculation $(1 + \sqrt{-5})(1 - \sqrt{-5}) = 2 \cdot 3$. To begin with observe that $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$ are algebraic integers because they are elements of $\mathbb{Z}[\sqrt{-5}]$, and $\sqrt{-5}$ is an algebraic integer because it is a root of the monic polynomial $X^2 + 5$. For $\alpha = a + b\sqrt{-5} \in \mathbb{Z}[\sqrt{-5}]$ let[4]

$$N_{-5}(\alpha) := a^2 + 5b^2.$$

This function is multiplicative: if $\beta = c + d\sqrt{-5}$ is another element of

[4] Those who already know about determinants can understand $N_{-5}(\alpha)$ as the determinant of the matrix $\begin{pmatrix} a & -5b \\ b & a \end{pmatrix}$ representing multiplication by α in the sense that if $\gamma = x + y\sqrt{-5}$, then $\alpha\gamma = (ax - 5by) + (ay + bx)\sqrt{-5}$ and $\begin{pmatrix} a & -5b \\ b & a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax - 5by \\ ay + bx \end{pmatrix}$.

$\mathbb{Z}[\sqrt{-5}]$, then

$$\begin{aligned} N_{-5}(\alpha\beta) &= N_{-5}\big((ac-5bd)+(ad+bc)\sqrt{-5}\big) \\ &= (ac-5bd)^2+5(ad+bc)^2 \\ &= (a^2c^2-10abcd+25b^2d^2)+(5a^2d^2+10abcd+5b^2c^2) \\ &= a^2c^2+5a^2d^2+5b^2c^2+25b^2d^2 \\ &= (a^2+5b^2)(c^2+5d^2) \\ &= N_{-5}(\alpha)N_{-5}(\beta). \end{aligned}$$

Direct computation gives $N_{-5}(1+\sqrt{-5}) = N_{-5}(1-\sqrt{-5}) = 6$, $N_{-5}(2) = 4$, and $N_{-5}(3) = 9$. The only $\alpha \in \mathbb{Z}[\sqrt{-5}]$ with $N_{-5}(\alpha) = 1$ are the units 1 and -1, and it is easy to see that 2 and 3 are not possible values of N_{-5}. Since N_{-5} is multiplicative, it follows that $1+\sqrt{-5}$, $1-\sqrt{-5}$, 2, and 3 are all irreducible, and our example is indeed a failure of unique factorization. From this we conclude that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD, and consequently it cannot be a PID.

We now arrive at one of the most famous stories in mathematics. Pierre de Fermat (1601-1665) was a French lawyer who spent the bulk of his career as a councillor at the local parliament at Toulouse, and devoted his leisure time to mathematics. Although he made many important discoveries, much of his work was unpublished in his lifetime, and was found after his death in marginal annotations in his books. The most famous of these asserts that for any integer $n \geq 3$ there are no nonzero integers a, b, c such that

$$a^n + b^n = c^n,$$

going on to say that "I have a truly marvelous proof of this proposition which this margin is too narrow to contain." This assertion became known as **Fermat's Last Theorem** after all of his other proposed results were either proved or refuted by counterexamples.

Possibly Fermat first discovered that the result was true for $n = 3$, and his proof for the case $n = 4$ is extant. In 1753 Euler gave a proof of the case $n = 3$ that was incorrect, but in a way that could be fixed, and subsequently many distinguished mathematicians worked out special cases. Sophie Germain (1776-1831) proved that for $n < 100$ the equation cannot be satisfied when none of the numbers a, b, c is divisible by n. The remaining case (exactly one of a, b, and c is divisible by n) for $n = 5$ splits into two subcases for which proofs were given in 1825 by Lejeune Dirichlet (1805-1859) and Legendre respectively. Dirichlet gave a proof for the case $n = 14$

in 1832, and in 1839 Gabriel Lamé (1795-1870) gave a proof for the case $n = 7$. (Since $a^{rs} = (a^r)^s$ for any a, r, and s, the case $n = 14$ is the case $n = 7$ with the additional restriction that a, b, and c are squares.)

Famously, in 1847 Lamé announced to the Paris Académie des Sciences that he had proved Fermat's Last Theorem in its entirety. Here we will explain the starting point of his argument. If p is odd prime that is a factor of n, then $a^n = (a^{n/p})^p$, and similarly for b and c, so a counterexample to Fermat's conjecture with exponent n yields a counterexample with exponent p. Similarly, if the exponent is $n = 2^k$ for some $k \geq 2$, then a counterexample with exponent n yields a couterexample with exponent 4. Every integer greater than 2 is either a power of 2 or is divisible by an odd prime, so to establish Fermat's last theorem it suffices to establish it when the exponent is 4 (which Fermat did) or any odd prime p.

Let ζ be a p^{th} **root of unity**: that is, $\zeta \neq 1$ is a complex number such that $\zeta^p = 1$. Possibly you already know that $\exp(2\pi i/p)$ is a p^{th} root of unity; if you don't now, you will after you read the next chapter. For the discussion here it suffices to take the existence of such a ζ on faith. Since ζ is a root of the monic polynomial $X^p - 1$, it is an algebraic integer.

Now consider the polynomial

$$P(X,Y) := (X+Y)(X+\zeta Y)\cdots(X+\zeta^{p-1}Y).$$

Expanding this using the distributive law gives

$$P(X,Y) = X^p + c_{p-1}X^{p-1}Y + \cdots + c_1XY^{p-1} + c_0Y^p$$

for some complex numbers $c_0, \dots, c_{p-1}$. We have $c_0 = \zeta^e$ where[5] $e = \sum_{i=0}^{p-1} i = p(p-1)/2$, so $c_0 = \zeta^{p(p-1)/2} = 1^{(p-1)/2} = 1$. Now observe that if we substitute ζY for Y in the definition of P, then the factor $X - Y$ is replaced by the factor $X - \zeta^p Y$, but of course these are the same, so $P(X, \zeta Y) = P(X,Y)$. Expanding $X^p + Y^p$ from both sides of this equation gives

$$c_{p-1}X^{p-1}Y + \cdots + c_1XY^{p-1} = c_{p-1}X^{p-1}(\zeta Y) + \cdots + c_1X(\zeta Y)^{p-1},$$

which implies that $c_{p-j} = c_{p-j}\zeta^j$ for each $j = 1, \dots, p-1$. Since $\zeta^j \neq 1$ for each such j (why, precisely?) we have $c_1 = \cdots = c_{p-1} = 0$, and we conclude that $P(X,Y) = X^p + Y^p$. In particular, for a counterexample to Fermat's conjecture with exponent p we have

$$c^p = a^p + b^p = (a+b)(a+\zeta b)\cdots(a+\zeta^{p-1}b).$$

[5] To see that $1 + 2 + \cdots + (k-1) = k(k-1)/2$ we observe that this is true when $k = 1$, and $k(k-1)/2 + k = (k+1)k/2$, so it follows from induction.

Certainly this is a tantalizing beginning. I don't know anything about how Lamé's argument proceeded from there except that *it depended on an assumption that* $\mathbb{Z}[\zeta]$ *is a UFD*, and eventually Ernst Kummer (1810-1893) showed that this is not always true. These events gave enormous stimulation to the development of algebraic number theory, which studies the ideals of rings of algebraic integers and related matters, but although this field and nearby areas of mathematics have been active topics of research ever since, until recently there were no new methods that could be used to mount fresh attacks.

In 1986 Kenneth Ribet proved that Fermat's Last Theorem would follow from certain cases of another open problem known as the Shimura-Taniyama-Weil Conjecture. Upon learning about this, Andrew Wiles decided to attempt the proof of these cases, and worked in secret on this project for the next seven years. He disclosed his proof in a famous series of three lectures at Cambridge University that culminated with the announcement that Fermat's Last Theorem had been proven. However, examination of his argument uncovered one serious flaw, and in December 1993 Wiles withdrew his claim to have a proof. Although the mathematical community continued to believe that the parts of his argument that survived were major contributions, Wiles' disappointment must have been extraordinarily intense.

Everybody makes mistakes. If this is true even for Lamé and Wiles, what are ordinary mortals to do? In my experience the most effective guard against error is to develop the habit of looking at the question or phenomenon at issue from as many points of view as possible. The human mind is well adapted to processing somewhat discordant bundles of information, sensing apparent contradictions and seeking out resolutions. In mathematics there are usually several ways to solve a problem, or perform a calculation. Doing everything twice may sound like a waste of time, but sometimes one turns up new insights in this way. People who approach mathematics by focusing on learning a single method actually tend to calculate more slowly, in part because they get into the habit of proceeding cautiously, as they must, since one wrong step is fatal.

A second important idea is to work on expressing your reasoning clearly. Proofs start from rough sketches, after which one works to fill in the missing steps. A mistake that remains at the end is typically concealed by a piece of vague, misleading, or lazy language. Working to make the expression of the argument crystal clear will usually root out these sorts of errors.

In early 1994 Wiles invited his former student Richard Taylor (b. 1962) to come from Harvard to Princeton to help him work on the problem. They tried various approaches, but in August 1994 Wiles announced that they were no nearer to a solution than he had been nine months earlier. Shortly thereafter Taylor suggested an attempt that Wiles was sure would fail, but Wiles agreed to work on it anyway, mainly in order to convince Taylor that it was hopeless. Two weeks later there was a flash of inspiration, and in October they circulated a draft of a new argument that was eventually accepted.

So, we can add two more items to our list of techniques for offsetting human fallibility: a) don't give up; b) have really smart friends.

2.7 Ordered Fields

After the adventures of the preceeding sections, we return to the main focus of this chapter, namely laying out an axiomatic description of the real numbers. The next step is to introduce the notion of order.

Definition 2.35. *An* **ordered field** *is a field* $(R, +, \cdot)$ *with a binary relation* $<$ *(with the standard associated relations, so the symbols '*$>$*,', '*$\leq$*,' and '*$\geq$*' have their usual meanings) such that:*

(O1) For all $x, y \in R$*, exactly one of* $x < y$*,* $x = y$*, and* $y < x$ *holds.*

(O2) For all $x, y, z \in R$*, if* $x < y$ *and* $y < z$*, then* $x < z$*.*

(O3) For all $x, y, z \in R$*, if* $x < y$*, then* $x + z < y + z$*.*

(O4) For all $x, y, z \in R$*, if* $x < y$ *and* $0 < z$*, then* $xz < yz$*.*

There is surprisingly little to say about these additional axioms, at least in relation to the topics considered so far in this chapter. For any irrational number $\alpha \in \mathbb{R}$, $\mathbb{Q}(\alpha)$ is example of an ordered field, but seemingly not a particularly interesting example. More precisely, on the surface there doesn't seem to be much interest in the interaction of the order axioms with the algebraic properties of $\mathbb{Q}(\alpha)$. As usual, at more advanced levels things are much more complicated, but none of this needs to be discussed at the beginning.

In real analysis textbooks it is common to have one or more pages devoted to proofs of little facts that everyone knows about inequalities. The point of this is mainly to give some experience in the construction of detailed proofs. (If you feel like a little challenge, try giving a proof, with each step justified by an axiom, that if $y > x > 0$, then $x^{-1} > y^{-1} > 0$.) Exercises of this sort are somewhat outside the spirit of this book, but we give one example.

Theorem 2.36. *$1 > 0$.*

Proof. Axiom (F6) requires that $1 \neq 0$, so, by (O1), either the claim is correct or $1 < 0$. If $1 < 0$, then (O3) implies that $0 = -1+1 < -1+0 = -1$, after which (O4) gives $0 = 0 \cdot 0 < (-1) \cdot (-1) = 1$ after all. □

Of course I would be quite surprised if you didn't already know that the **absolute value** of $x \in R$ is denoted by $|x|$, and is x if $x \geq 0$, and otherwise it is $-x$. The following facts will be applied again and again.

Lemma 2.37. *If x and y are elements of an ordered field R, then*

$$|xy| = |x|\,|y| \quad \textit{and} \quad |x+y| \leq |x| + |y|.$$

Proof. To prove that $|xy| = |x|\,|y|$ there is really nothing easier than going through each of the four cases. There would be no real point in writing it out; as usual, you should make sure you understand it.

For any $z \in R$ we have $z \leq |z|$ and $-z \leq |z|$. The variant of (O3) in which strict inequality is replaced by weak inequality holds because $x + z = y + z$ when $x = y$. The asserted inequality is obtained by applying this four times:

$$x + y \leq |x| + y \leq |x| + |y| \quad \text{and} \quad -x - y \leq |x| - y \leq |x| + |y|.$$

□

The **characteristic** of a field k is either the smallest integer p such that $0 = 1 + \cdots + 1$ (p summands) or (by convention) 0 if no such integer exists. If the characteristic is $p \neq 0$, then p must be prime since if $p = qr$, where $q, r > 1$, then the product of $1 + \cdots + 1$ (q summands) and $1 + \cdots + 1$ (r summands) would be zero, even though neither factor is zero. Fields of nonzero characteristic play an important role in number theory, and have a rich and interesting theory, but we won't need to consider them here.

Lemma 2.38. *If R is an ordered field, then the characteristic of R is zero.*

Proof. If $1 + \cdots + 1 = 0$, then the fact that $0 < 1$ and repeated applications of (O4) would give

$$0 < 1 < 1 + 1 < \ldots < 1 + \cdots + 1 = 0,$$

and (O2) (transitivity) would imply that $0 < 0$, contrary to (O1). □

We will need to know that $\mathbb{Q}$ has only one ordering satisfying (O1)-(O4). Since $0 < 1$, for every integer n (O3) implies that $n < n+1$, and since $<$ is transitive (that is, (O2)) it is clear that any two orderings satisfying (O1)-(O4) order any two integers in the same way. Consider $p/q, r/s \in \mathbb{Q}$, where $p, q, r, s \in \mathbb{Z}$ with q and s positive, and suppose that $ps < rq$. If we knew that $0 < 1/qs$ we could use (O4) to conclude that

$$p/q = ps(1/qs) < rq(1/qs) = r/s,$$

so that the ordering of p/q and r/s is determined by the ordering of the integers ps and rq. But (O4) implies that $0 < qs$, and in general, if $0 < z$, then $0 < 1/z$ because $1/z = 0$ is impossible, and if $1/z < 0$, we could apply (O4) with $x = 1/z$ and $y = 0$ to obtain $1 < 0$, which we know to be false. Thus the ordering of any two elements of $\mathbb{Q}$ is determined by the ordering of the integers, and there is only one way to order the integers.

Let R be a field of characteristic zero. There is a unique homomorphism $\varphi : \mathbb{Q} \to R$ satisfying $\varphi(1) = 1$ that is defined by setting $\varphi(0) := 0$, $\varphi(n) := 1 + \cdots + 1$ (n summands) and $\varphi(-n) := (-1) + \cdots + (-1)$ (n summands) for each positive integer n, and setting $\varphi(m/n) := \varphi(m)/\varphi(n)$ for all nonzero integers m and n. If R is an ordered field, we can define an ordering $\prec$ on $\mathbb{Q}$ by specifying that $r \prec s$ if and only if $\varphi(r) < \varphi(s)$, and it is easy to see that since (O1)-(O4) are satisfied in R, $\prec$ is a relation on $\mathbb{Q}$ satisfying (O1)-(O4). But there is only one such relation, so $\prec$ agrees with $<$, and consequently φ must be order preserving in the sense that $\varphi(r) < \varphi(s)$ if and only if $r < s$. That is, there is only one way to embed $\mathbb{Q}$ in R as a subfield, and this embedding respects the order. Instead of explicitly carrying the homomorphism φ around, it is simpler to treat $\mathbb{Q}$ as a subfield of R, and we will do so from now on.

2.8 The Least Upper Bound Axiom

We now introduce the property that distinguishes the set of real numbers from all other ordered fields. Fix an ordered field R. If $S \subset R$ and $b \in R$, we say that b is an **upper bound** for S if $b \geq s$ for all $s \in S$.

Definition 2.39. *A **real number field** is an ordered field $(R, +, \cdot, <)$ that also satisfies the **least upper bound axiom:***

*(LUB) Every nonempty $S \subset R$ that is bounded above has a **least upper bound:** there is an upper bound $\underline{b}$ such that $\underline{b} \leq b$ whenever b is an upper bound for S.*

Roughly, (LUB) insures that a real number field has no "holes," so that when we go looking for a number to be in a particular place, we find one. Below we'll explain two of the most basic and important general manifestations of this idea. Each of these is the seed of a circle of ideas that will be developed much more fully in the next chapter, with important echoes throughout mathematics.

In preparation for that we introduce some important notation. If a and b are elements of R, then

$$[a, b] := \{\, t \in R : a \leq t \leq b \,\} \quad \text{and} \quad (a, b) := \{\, t \in R : a < t < b \,\}.$$

Sets like these are called **closed** and **open intervals** respectively. **Half open intervals** are those of the forms

$$(a, b] := \{\, t \in R : a < t \leq b \,\} \quad \text{and} \quad [a, b) := \{\, t \in R : a \leq t < b \,\}.$$

There are also intervals that are **unbounded** in one direction:

$$[a, \infty) := \{\, t \in R : a \leq t \,\}; \qquad (a, \infty) := \{\, t \in R : a < t \,\};$$

$$(-\infty, b] := \{\, t \in R : t \leq b \,\}; \qquad (-\infty, b) := \{\, t \in R : t < b \,\}.$$

Fix $a, b \in R$ with $a < b$, and let $f : [a, b] \to R$ be a function. Possibly you already know what it means for f to be continuous, but in case you don't, here is the definition. We say that f is **continuous** if, for each $s \in [a, b]$ and each $\varepsilon > 0$, there is $\delta > 0$ such that $|f(s') - f(s)| < \varepsilon$ for all $s' \in (s - \delta, s + \delta) \cap [a, b]$. That is, for any s we can force $f(s')$ to be as close to $f(s)$ as we like by requiring that s' be chosen from a sufficiently small interval around s. A common visual intuition is that you can draw the graph of f without lifting the pencil off the piece of paper.

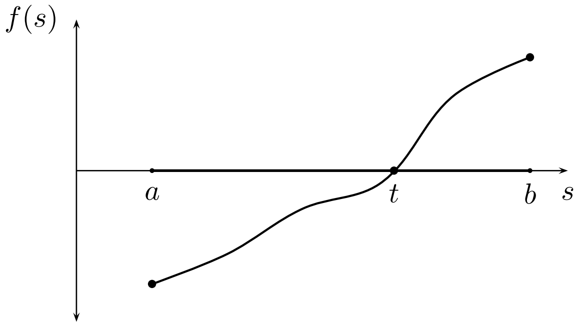


Figure 2.2

Theorem 2.40 (Intermediate Value Theorem)**.** *If R is a real number field, $f : [a,b] \to R$ is continuous, $f(a) < 0$, and $f(b) > 0$, then there is some t such that $f(t) = 0$.*

Proof. Let

$$S := \{\, s \in [a,b] : f(s) < 0 \,\}.$$

Then S contains a and is bounded above to b, so it has a least upper bound t. If $f(t) < 0$, then there is $\delta > 0$ such that $f(t') < 0$ for all $t' \in (t-\delta, t+\delta)$, contradicting the assumption that t is an upper bound on S. If $f(t) > 0$, then there is $\delta > 0$ such that $f(t') > 0$ for all $t' \in (t-\delta, t+\delta)$, contradicting the assumption that t is the least upper bound on S. In view of (O1), the only remaining possibility is that $f(t) = 0$. □

In the rest of this book this theorem will be invoked many times, with the following application being fairly typical. In our work up to this point we've proved that $\sqrt{2}$ is irrational, and we had a few things to say about the field $\mathbb{Q}(\sqrt{2})$. But how do we know that 2 actually has a square root?

Theorem 2.41. *If R is a real number field, then R contains a square root of 2.*

Proof. The claim follows if we can show that the hypotheses of the intermediate value theorem are satisfied by the function $f : [0,2] \to \mathbb{R}$ given by $f(s) := s^2 - 2$. Of course $f(0) = -2 < 0$ and $f(2) = 2 > 0$. To show that f is continuous we consider some $s \in [0,2]$ and $\varepsilon > 0$. Let $\delta := \varepsilon/4$. Since ${s'}^2 - s^2 = (s'-s)(s'+s)$, if $|s'-s| < \delta$, then

$$|f(s') - f(s)| \leq |s'-s|\,(|s'| + |s|) < (\varepsilon/4)(2+2) = \varepsilon.$$

□

The LUB axiom is often used to show that certain sequences have limits. Technically, for any set X a **sequence** in X is a function from the natural numbers (or, sometimes, the nonnegative integers) to X, but instead of thinking of it as a function, we think of it as an infinite list $x_1, x_2, x_3, \ldots$. To save space, we'll often denote such a sequence by $\{x_n\}_{n=1}^{\infty}$ or just $\{x_n\}$.

Fix an ordered field R, and let $s_1, s_2, s_3, \ldots$ be a sequence in R. We say that the sequence **converges** to a number s, and we write $s_n \to s$, if, for any $\delta > 0$, s_n is in the interval $(s-\delta, s+\delta)$ for all sufficiently large n. In symbols,

$$\big(\forall \delta > 0\big)\big(\exists N \in \mathbb{N}\big)\big(\forall n > N\big)\ |s - s_n| < \delta.$$

The sequence is **convergent** if it converges to some number; otherwise it is **divergent**. (Sometimes you'll see the notation "$s_n \to \infty$," meaning that for every $\Delta > 0$ there is a natural number N such that $s_n > \Delta$ whenever $n > N$, and you might hear someone say that $\{s_n\}$ "converges to infinity." There is no real harm in this, but it is more proper to say that $\{s_n\}$ "diverges to infinity.")

We say that $\{s_n\}$ is a **Cauchy sequence** if the distance between s_m and s_n becomes arbitrarily small as $\min\{m, n\}$ becomes large. Put precisely,

$$\big(\forall \delta > 0\big)\big(\exists N \in \mathbb{N}\big)\big(\forall m, n > N\big)\ |s_m - s_n| < \delta.$$

A Cauchy sequence "ought" to have a limit in a very practical sense: if you were working with a system of numbers within which Cauchy sequences might not have limits, there would be an irresistible temptation to create "virtual" numbers that were their limits. These might be defined as equivalence classes of Cauchy sequences, where two Cauchy sequences $\{s_m\}$ and $\{t_n\}$ are equivalent if they are eventually arbitrarily close to each other:

$$\big(\forall \delta > 0\big)\big(\exists N \in \mathbb{N}\big)\big(\forall m, n > N\big)\ |s_m - t_n| < \delta.$$

(This method of construction will be described in detail in Section 2.9, where it is used to construct the real numbers, and again in a more general context in Section 6.2.) Other methods might be used to define these virtual numbers. But you *would* define them, and eventually you would treat them as full fledged numbers. After all, when the discovery that $\sqrt{2}$ is irrational threw the Pythagorean school of philosophy into crisis, it was logically possible for them to say that there actually is no number whose square is 2, even if certain fancy constructions, e.g., continuing decimals, can mimic the behavior you would expect from $\sqrt{2}$. But they must have known that nobody would buy into such a convoluted and cumbersome way of doing mathematics.

Since we would feel that an ordered field was incomplete if there was a Cauchy sequence that didn't have a limit, the following terminology is natural.

Definition 2.42. *The ordered field R is* ***complete*** *if each of its Cauchy sequences is convergent.*

Proposition 2.43. *A real number field is complete.*

Proof. Suppose that R is a real number field and $\{s_n\}$ is a Cauchy sequence in R. Let S be the set of $s \in R$ such that $s_n > s$ for infinitely many n. If

$|s_m - s_n| < 1$ for all $m, n > N$, then $s_{N+1} - 1 \in S$ and $s_{N+1} + 1$ is an upper bound on S. Since S is nonempty and bounded above, it has a least upper bound b.

Consider a number $\delta > 0$. There cannot be infinitely many n such that $s_n \geq b + \delta$ because that would imply that $b + \delta/2 \in S$. There cannot be infinitely many n such that $s_n \leq b - \delta$ because the existence of an N such that $|s_m - s_n| < \delta/2$ for all $m, n > N$ would then imply that $b - \delta/2$ was an upper bound on S. Therefore $b - \delta < s_n < b + \delta$ for all sufficiently large n. Moreover, this is true for every $\delta > 0$, which means that $s_n \to b$. □

In preparation for the next section we'll now develop one more consequence of the least upper bound axiom which might seem rather surprising, since at first sight it is hard to imagine that it might not be true automatically.

Definition 2.44. *We say that R is* ***Archimedean*** *if, for every $s \in R$ with $s > 0$, there is some $q \in \mathbb{Q}$ with $0 < q < s$.*

There are ordered fields that are not Archimedean, but they're quite exotic creatures.

Proposition 2.45. *If R is a real number field, then R is Archimedean.*

Proof. Let Δ be the set of $s \in R$ such that $s > 0$ and $s \leq q$ for all positive $q \in \mathbb{Q}$. Then Δ is bounded above by (for example) 1, and if Δ was nonempty we could let δ be its least upper bound. If q was a rational number satisfying $\delta \leq q < 2\delta$, then $0 < q/2 < \delta$, so there could not be such a q, but this would imply that δ was *not* an upper bound of Δ since, for example, $3\delta/2 \in \Delta$. In view of this contradiction it must be the case that $\Delta = \emptyset$. □

The next result gives a useful alternative characterization of real number fields.

Theorem 2.46. *An ordered field is a real number field if and only if it is Archimedean and complete.*

In the proof it will be necessary to distinguish between different kinds of Cauchy sequences and different kinds of completeness. We say that $\{s_n\}$ is a **$\mathbb{Q}$-Cauchy sequence** if the distance between s_m and s_n becomes smaller than any rational number as $\min\{m, n\}$ becomes large. That is,

$$\big(\forall k \in \mathbb{N}\big)\big(\exists N \in \mathbb{N}\big)\big(\forall m, n > N\big)\ |s_m - s_n| < 1/k.$$

an ordered field R is **$\mathbb{Q}$-complete** if each of its $\mathbb{Q}$-Cauchy sequences is convergent. Since a Cauchy sequence is a $\mathbb{Q}$-Cauchy sequence, R is complete whenever it is $\mathbb{Q}$-complete. On the other hand, if R is Archimedean, then every $\mathbb{Q}$-Cauchy sequence is Cauchy, so if R is complete, then it is $\mathbb{Q}$-complete.

Proof of Theorem 2.46. We've already seen that a real number field is Archimedian and complete, so we need to show that if an ordered field R is Archimedean and complete, then it is a real number field. Let S be a nonempty subset of R that is bounded above. Choose $x_0 \in S$, and let b_0 be an upper bound for S. We are going to hunt down the least upper bound of S using repeated bisection, as implemented by a sequence $s_1, s_2, \dots$ that we define "inductively." To begin the process set

$$s_1 := \begin{cases} x_0 + (b_0 - x_0)/2, & x \geq x_0 + (b_0 - x_0)/2 \text{ for some } x \in S, \\ x_0, & \text{otherwise.} \end{cases}$$

The process continues according to the following rule: if we have already constructed s_{n-1}, then

$$s_n := \begin{cases} s_{n-1} + (b_0 - x_0)/2^n, & x \geq s_{n-1} + (b_0 - x_0)/2^n \text{ for some } x \in S, \\ s_{n-1}, & \text{otherwise.} \end{cases}$$

We now use induction to show that for each n, $s_n + (b_0 - x_0)/2^n$ is an upper bound for S and $x \geq s_n$ for some $x \in S$. It is easy to see that this is the case when $n = 1$, for both possibilities in the definition of s_1. Suppose, for some $n \geq 2$, that we have shown that it is true with n replaced by $n-1$. Again, it is easy to see that this implies that it is also true for n, regardless of which case occurs in the construction of s_n.

The sequences $\{s_n\}$ and $\{s_n + (b_0 - x_0)/2^n\}$ are $\mathbb{Q}$-Cauchy sequences, so they are Cauchy sequences because R is Archimedean, and they have limits because R is complete. The limit of the first sequence cannot be greater than the limit of the second sequence, and the difference between the two limits is less than any rational number, so (again because R is Archimedean) these limits coincide. Let b be the common limit. Since each $s_n + (b_0 - x_0)/2^n$ is an upper bound for S, b is an upper bound for S. For each n there an $x \in S$ with $x \geq s_n$, so S cannot have an upper bound that is less than b. We have shown that S has a least upper bound. □

By definition an ordered field R is a real number field if it satisfies (LUB), and we have shown that this is the case if and only if it is complete and Archimedean. Since a $\mathbb{Q}$-complete ordered field is complete, the next result gives a third chararacterization of real number fields.

Proposition 2.47. *If R is $\mathbb{Q}$-complete, then R is Archimedean.*

Proof. The sequence $1, 1/2, 1/3, \ldots$ is $\mathbb{Q}$-Cauchy, of course. Let δ be its limit. If there was an $\varepsilon > 0$ with $\varepsilon < q$ for all positive rational numbers q, then $\delta \geq \varepsilon$. But it is impossible for the sequence to converge to a positive δ because if $1/m$ is in the interval $(\delta/2, 3\delta/2)$, then $1/n$ is outside this interval for all $n > 3m$. □

2.9 Constructing the Real Numbers

Does a real number field exist? Is there more than one real number field? (What we are *really* asking here is whether any two real number fields are necessarily isomorphic, where an isomorphism $\iota : R \to R'$ of ordered fields R and R' is an isomorphism of fields that is order preserving in the sense that for all $r, s \in R$, $\iota(r) < \iota(s)$ if and only if $r < s$.) These are crucial questions because they test the adequacy of our axiom system. If there were no real number fields, our axiom system would be an axiomatization of nothing, and obviously not a sound basis for mathematical reasoning. If there were multiple (isomorphism types of) real number fields, we would need to add additional axioms to finish the job of giving a complete description of the real numbers, or at least there would be an unexpected opportunity to investigate how the various real number fields differ from each other.

The two main approaches to constructing the real numbers, pioneered by Cantor and Richard Dedekind (1831-1916), were both published in 1872. Here we'll focus on Cantor's construction because it is both easier, from a technical point of view, and applicable more generally, but first we'll give a brief description of Dedekind's method.

A **Dedekind cut** is a pair (A, B) where $A, B \subset \mathbb{Q}$ are both nonempty, $A \cap B = \emptyset$, $A \cup B = \mathbb{Q}$, $a < b$ for all $a \in A$ and $b \in B$, and A has no greatest element. The guiding intuition is that the function taking each Dedekind cut (A, B) to the least upper bound of A is a bijection between the set of Dedekind cuts and the set of real numbers. After defining the order relation and arithmetic operations on the set of Dedekind cuts correctly, one can show that the set of Dedekind cuts is a real number field, so a real number field actually exists. One can then show that any real number field is isomorphic to the one constructed in this manner. The main difficulty with this approach is that the definition of multiplication has four cases, according to the signs of the numbers being multiplied, and this makes the proofs lengthy and unpleasant. However, this approach does have one significant advantage, namely that there is no need to appeal to the axiom of choice.

Cantor's construction uses equivalence classes of Cauchy sequences of rational numbers. We will say that two Cauchy sequences $\{s_n\}$ and $\{t_n\}$ in $\mathbb{Q}$ are **equivalent** if, for every rational $\delta > 0$, there is an integer N such that $|s_n - t_n| < \delta$ for all $n > N$. Obviously this relation is reflexive and symmetric, and it is easy to show that it is transitive, hence an equivalence relation: if $\{s_n\}$ and $\{t_n\}$ are equivalent, and $\{t_n\}$ and $\{u_n\}$ are equivalent, then for any $\delta > 0$ there is N such that $|s_n - t_n| < \delta/2$ and $|t_n - u_n| < \delta/2$ for all $n > N$, so that

$$|s_n - u_n| \leq |s_n - t_n| + |t_n - u_n| < \delta/2 + \delta/2 = \delta$$

for all such n. The equivalence class of $\{s_n\}$ is denoted by $[\{s_n\}]$.

Let R_0 be the set of equivalence classes of Cauchy sequences in $\mathbb{Q}$. By identifying each $r \in \mathbb{Q}$ with the equivalence class of the sequence $r, r, \dots$ we may regard $\mathbb{Q}$ as a subset of R_0. We define the relation $[\{s_n\}] < [\{t_n\}]$ to mean that there is a rational $\delta > 0$ and a natural number N such that $s_n + \delta < t_n$ for all $n > N$. *This definition immediately implies that R_0 is Archimedean because whenever $[\{s_n\}] > 0$ there is a rational δ with* $0 < \delta/2 < [\{s_n\}]$. Addition and multiplication of equivalence classes of Cauchy sequences is defined in the obvious way:

$$[\{s_n\}] + [\{t_n\}] := [\{s_n + t_n\}] \quad \text{and} \quad [\{s_n\}] \cdot [\{t_n\}] := [\{s_n t_n\}].$$

Of course there is some work to do to demonstrate that these definitions don't depend on the choice of representatives, and that $\{s_n + t_n\}$ and $\{s_n t_n\}$ are Cauchy sequences whenever $\{s_n\}$ and $\{t_n\}$ are Cauchy sequences. It would be quite repetitive to write out every argument, so we will just do the verifications in connection with multiplication, expecting that you would not have any difficulty extending the underlying methods to generate the necessary arguments for order and addition, which are somewhat simpler. Let $\{s_n\}$ and $\{t_n\}$ be Cauchy sequences. There are positive rational numbers S and T such that $|s_n| < S$ and $|t_n| < T$ for all sufficiently large n. (For example we can let $S := s_{N+1} + 1$ where N is large enough that $|s_m - s_n| < 1$ for all $m, n > N$.) For any rational $\delta > 0$ we can choose N large enough that $|s_m - s_n| < \frac{1}{2}\delta/T$ and $|t_m - t_n| < \frac{1}{2}\delta/S$, so that

$$\begin{aligned} |s_m t_m - s_n t_n| &= |s_m(t_m - t_n) + (s_m - s_n)t_n| \\ &\leq |s_m(t_m - t_n)| + |(s_m - s_n)t_n| \\ &= |s_m|\,|t_m - t_n| + |s_m - s_n|\,|t_n| \\ &< S(\tfrac{1}{2}\delta/S) + (\tfrac{1}{2}\delta/T)T = \delta. \end{aligned}$$

Therefore $\{s_n t_n\}$ is a Cauchy sequence. Suppose that $\{s'_n\}$ is equivalent to $\{s_n\}$ and that $\{t'_n\}$ is equivalent to $\{t_n\}$. Replacing T with a slightly larger number, if need be, insures that $|t'_n| < T$ for all sufficiently large n. For sufficiently large n we have $|s'_n - s_n| < \frac{1}{2}\delta/T$ and $|t'_n - t_n| < \frac{1}{2}\delta/S$, so that

$$\begin{aligned}|s'_n t'_n - s_n t_n| &= |(s'_n - s_n)t'_n + s_n(t'_n - t_n)| \\ &\le |s'_n - s_n|\,|t'_n| + |s_n|\,|t'_n - t_n| \\ &< (\tfrac{1}{2}\delta/T)T + S(\tfrac{1}{2}\delta/S) = \delta.\end{aligned}$$

Thus $\{s'_n t'_n\}$ is equivalent to $\{s_n t_n\}$, which shows that the definition of multiplication doesn't depend on the choice of representatives.

We now wish to show that R_0 is a real ordered field, which means that we have to show that it satisfies (F1)-(F9), (O1)-(O4), and (LUB). If you want to practice writing some simple proofs, you can do some of this explicitly, but mostly the ideas will be very clear without belaboring the details, so we will describe how this works in general terms. It is easy to see that R_0 satisfies (F1)-(F9) because these axioms are satisfied by $\mathbb{Q}$. For example, for any $[\{s_n\}]$ and $[\{t_n\}]$ we have

$$[\{s_n\}] + [\{t_n\}] = [\{s_n + t_n\}] = [\{t_n + s_n\}] = [\{t_n\}] + [\{s_n\}],$$

which verifies (F4).

Only (F7) (existence of inverses) presents any difficulties at all. Suppose that $[\{s_n\}] \neq 0$. Then $s_1, s_2, \dots$ is not equivalent to $0, 0, \dots$, so there is some $\delta > 0$ such that for any N there is an $n > N$ such that $|s_n| \ge \delta$. If N is sufficiently large we have $|s_m - s_n| < \delta/2$ for all $m, n > N$, and in particular $|s_n| > \delta/2$ for all $n > N$. It works to set $[\{s_n\}]^{-1} = [\{t_n\}]$ where $\{t_n\}$ is a sequence with $t_n = 1/s_n$ for all $n > N$, provided that we can show that such a $\{t_n\}$ is a Cauchy sequence. But for any rational $\varepsilon > 0$ there is $M \ge N$ such that $|s_m - s_n| < \varepsilon/4\delta^2$ for all $m, n > M$, in which case

$$\left|\frac{1}{s_m} - \frac{1}{s_n}\right| = \frac{|s_m - s_n|}{|s_m|\,|s_n|} < \frac{\varepsilon/4\delta^2}{(\delta/2)(\delta/2)} = \varepsilon.$$

We'll prove (O1), but not (O2)-(O4) because you should have no difficulty seeing that these are straightforward consequences of the definition of the order relation on R_0. Consider two elements $[\{s_n\}]$ and $[\{t_n\}]$ of R_0. If you review the definition of inequality you will easily see that at most one of the three relations $[\{s_n\}] < [\{t_n\}]$, $[\{s_n\}] = [\{t_n\}]$, and $[\{t_n\}] < [\{s_n\}]$ holds. We want to show that at least one holds, so we may begin the argument by supposing that it is not the case that $[\{s_n\}] < [\{t_n\}]$, nor is it the case that

$[\{t_n\}] < [\{s_n\}]$. Then for any rational $\delta > 0$ and any N there are $m, n > N$ such that $t_m \le s_m + \delta$ and $s_n \le t_n + \delta$. By choosing N large enough we can also insure that $|s_p - s_q| < \delta$ and $|t_p - t_q| < \delta$ for all $p, q > N$. Then for any $p > N$ we have $|s_p - t_p| < 3\delta$ because

$$s_p < s_n + \delta \le t_n + 2\delta < t_p + 3\delta \text{ and } t_p < t_m + \delta \le s_m + 2\delta < s_p + 3\delta.$$

This is true for every positive rational δ, so $\{s_n\}$ and $\{t_n\}$ are equivalent, which means that $[\{s_n\}] = [\{t_n\}]$.

Proving that R_0 satisfies (LUB) is a bit more complicated.

Theorem 2.48. *R_0 is a real number field.*

Proof. If we regard (F1)-(F9) and (O1)-(O4) as established, it only remains to show that (LUB) is satisfied. Moreover, R_0 is Archimedean, so in view of Theorem 2.46 it suffices to show that R_0 is complete.

Let $u^1, u^2, \ldots$ be a Cauchy sequence in R_0, where $u^i = [\{s^i_n\}]$. We will show that this sequence has a limit in R_0 by picking out a suitable Cauchy sequence from the various numbers s^i_n. Specifically, for each $k = 1, 2, \ldots$ let $t_k = s^{i_k}_{n_k}$ where i_k is large enough that $|u^m - u^n| < 1/3k$ for all $m, n \ge i_k$ and n_k is large enough that $|s^{i_k}_m - s^{i_k}_{m'}| < 1/3k$ for all $m, m' \ge n_k$. Noting that $|t_k - u^{i_k}| \le 1/3k$, if $m, n \ge k$ we have

$$|t_m - t_n| \le |t_m - u^{i_m}| + |u^{i_m} - u^{i_n}| + |u^{i_n} - t_n| \le 1/k.$$

Therefore $\{t_k\}$ is a Cauchy sequence in $\mathbf{Q}$. Let $u := [\{t_k\}]$. Clearly $\left|t_k - u\right| \le 1/k$, so

$$\left|u^{i_k} - u\right| \le \left|u^{i_k} - t_k\right| + \left|t_k - u\right| \le 4/3k$$

for all k. Since R_0 is Archimedean, this implies that $u^i \to u$. □

We've shown that a real number field exists, because R_0 is such a field. The second step of our program is to show that there is (up to isomorphism) only one real number field. Specifically, we will show that any real number field R is isomorphic to R_0. Since R is Archimedean and complete, it is $\mathbf{Q}$-complete, so any Cauchy sequence $\{s_n\}$ in $\mathbf{Q}$ has a limit in R. If $\{s_n\}$ and $\{t_n\}$ are equivalent Cauchy sequences, then the absolute value of the difference between their limits must be smaller than any positive rational number, and since R is Archimedean this means that the two limits must be the same. Therefore we can define a function $\iota : R_0 \to R$ by setting

$$\iota\big([\{s_n\}]\big) := \lim_{n\to\infty} s_n.$$

We will show that ι is an order preserving isomorphism.

We now need the following elementary facts about convergent sequences.

Lemma 2.49. *If $\{s_n\}$ and $\{t_n\}$ are convergent sequences in R with $s_n \to s$ and $t_n \to t$, then*

$$s_n + t_n \to s + t \quad \textit{and} \quad s_n t_n \to st.$$

Proof. The definition of convergence implies that for any $\delta > 0$ there is N such that $|s_n - s| < \delta/2$ and $|t_n - t| < \delta/2$ whenever $n > N$. The first asserted convergence follows easily from this.

For the second equation there are four cases. If $s = 0 = t$, then for any $\delta > 0$ there is N such that $|s_n| < \delta$ and $|t_n| < 1$ for all $n > N$, so that

$$|s_n t_n - st| = |s_n t_n| = |s_n|\,|t_n| < \delta. \tag{$*$}$$

If $s = 0$ and $t \neq 0$, then for any $\delta > 0$ there is N such that $|s_n| < \delta/2|t|$ and $|t_n| < 2|t|$ whenever $n > N$, so that again $(*)$ holds. A similar argument works when $s \neq 0$ and $t = 0$. If $s \neq 0 \neq t$, then for any $\delta > 0$ there is N such that $|s_n| < 2|s|$, $|t_n - t| < \delta/4|s|$, and $|s_n - s| < \delta/2|t|$ whenever $n > N$, in which case

$$|s_n t_n - st| = |s_n(t_n - t) + (s_n - s)t| \leq |s_n|\,|t_n - t| + |s_n - s|\,|t| < \delta.$$

□

It is now easy to see that ι is a homomorphism: for any $[\{s_n\}]$ and $[\{t_n\}]$ we have

$$\begin{aligned} \iota\big([\{s_n\}] + [\{t_n\}]\big) &= \iota\big([\{s_n + t_n\}]\big) = \lim_{n\to\infty} s_n + t_n \\ &= \lim_{n\to\infty} s_n + \lim_{n\to\infty} t_n = \iota\big([\{s_n\}]\big) + \iota\big([\{t_n\}]\big) \end{aligned}$$

and

$$\begin{aligned} \iota\big([\{s_n\}] \cdot [\{t_n\}]\big) &= \iota\big([\{s_n \cdot t_n\}]\big) = \lim_{n\to\infty} s_n \cdot t_n \\ &= \big(\lim_{n\to\infty} s_n\big) \cdot \big(\lim_{n\to\infty} t_n\big) = \iota\big([\{s_n\}]\big) \cdot \iota\big([\{t_n\}]\big). \end{aligned}$$

In each case the first equality is the definition of the operation in R_0, the second and final equality is the definition of ι, and the remaining equality is from the result above.

When $[\{s_n\}] < [\{t_n\}]$ there is a rational number $\delta > 0$ and a natural number N such that $s_n + \delta < t_n$ for all $n > N$, so

$$\iota\big([\{s_n\}]\big) + \delta = \lim_{n\to\infty} s_n + \delta \leq \lim_{n\to\infty} t_n = \iota\big([\{t_n\}]\big)$$

and consequently $\iota\big([\{s_n\}]\big) < \iota\big([\{t_n\}]\big)$. That is, ι is order preserving, and in addition it is injective.

The only thing left to do is to show that ι is surjective. Fix $u \in R$. For each n there is a rational number $t_n \in (u, u + 1/n)$. (For example, some integer multiple of $1/2n$ lies in this interval.) Of course $\{t_n\}$ is a Cauchy sequence, and its limit cannot be different from u because R is Archimedean. Therefore $\iota\big([\{t_n\}]\big) = u$.

Finally, after arguments that were much harder (or at least more detailed) than anything that came before, and almost anything we will see later, we are entitled to let $\mathbb{R}$ denote the unique (up to order preserving isomorphism) real number field, and to call it *the* set of real numbers. This is fundamental to all the work we do from here on because *any property of the real numbers is a logical consequence of (F1)-(F9), (O1)-(O4), and (LUB).* When we work with the real numbers, we know *exactly* what we are talking about, and the axioms, and any of their logical consequences that might have been proven earlier, are always what we start with when we want to prove a theorem involving real numbers. In this sense we have a logically secure foundation for that part of mathematics (namely almost all of it) that is built on top of $\mathbb{R}$.

Chapter 3

Limits and Continuity

It's a bit hard to say just why continuity is such an important concept in mathematics, perhaps because there are so many reasons. Actually, the number of reasons has grown enormously in response to the abstract formulations of the concept that we'll describe in this chapter.

When one is first exposed to the concept of continuity, typically it is explained in terms of one or both of the following definitions.

Definition 3.1. *A function $f : \mathbb{R} \to \mathbb{R}$ is* ***continuous*** *at $t \in \mathbb{R}$ if $f(t_n) \to f(t)$ whenever $\{t_n\}$ is a sequence in $\mathbb{R}$ that converges to t.*

Definition 3.2. *A function $f : \mathbb{R} \to \mathbb{R}$ is* ***continuous*** *at $t \in \mathbb{R}$ if, for every $\varepsilon > 0$, there is $\delta > 0$ such that $|f(t') - f(t)| < \varepsilon$ for all $t' \in (t - \delta, t + \delta)$.*

In either case we say that f is **continuous** if it is continuous at each $t \in \mathbb{R}$.

First of all we need to show that these definitions are equivalent. It is a good idea to think about this visually, using our intuitive "definition" of a continuous function as one whose graph can be drawn without lifting the pencil off the paper. But however much intuition one develops, it remains the case that both of these definitions are, in a purely logical sense, rather complicated.

There is a trick for dealing with such concepts. The negation of a proposition of the form '$(\forall x)P(x)$' is '$(\exists x)\neg P(x)$,' and the negation of a proposition of the form '$(\exists y)Q(y)$' is '$(\forall y)\neg Q(y)$.' When one has a proposition in which a sequence of logical quantifiers (that is, clauses of the form '$\forall x$' and '$\exists y$') precede some unquantified proposition, a "zero thought" procedure for formulating its negation is to apply these transformations repeatedly. This amounts to replacing each '$\forall$' with '$\exists$' and each '$\exists$' with '$\forall$,' then negating the unquantified proposition. Applying this procedure, we find that the meaning of $f(t_n) \not\to f(t)$ is that

$$(\exists \varepsilon > 0)(\forall N \in \mathbb{N})(\exists n > N)\ |f(t_n) - f(t)| \geq \varepsilon. \tag{$*$}$$

Let's apply this procedure again, this time to the second definition of continuity. It turns out that failure of this definition means that

$$(\exists \varepsilon > 0)(\forall \delta > 0)(\exists t' \in (t - \delta, t + \delta))\ |f(t') - f(t)| \geq \varepsilon. \tag{$**$}$$

If you want more practice, try applying this procedure to $(*)$ and $(**)$, thereby recovering the meaning of $f(t_n) \to f(t)$ and Definition 3.2.

We now turn to the proof of equivalence. Since this is a matter of showing that each definition implies the other, the proof has two parts, and for each part it is simplest to use reductio ad absurdum, arguing that a failure of one definition implies that the other also fails to hold.

- First suppose that $f(t_n) \not\to f(t)$ for some sequence $\{t_n\}$ converging to t. Then $(*)$ holds, and we need to show that $(**)$ holds. But $(*)$ gives us ε, and for any $\delta > 0$ we can choose N such that $|t_n - t| < \delta$ for all $n > N$, then choose $n > N$ such that $|f(t_n) - f(t)| \geq \varepsilon$. That is, for any $\delta > 0$ we can find n such that $t' = t_n$ satisfies $t' \in (t - \delta, t + \delta)$ and $|f(t') - f(t)| \geq \varepsilon$.

- Now suppose that Definition 3.2 fails. Then $(**)$ gives an ε such that for each $n \in \mathbb{N}$ we choose $t_n \in (t - \frac{1}{n}, t + \frac{1}{n})$ with $|f(t_n) - f(t)| \geq \varepsilon$. Then $t_n \to t$ and $f(t_n) \not\to f(t)$.

This illustrates some important aspects of "doing" mathematics. Roughly, we can think of the process of proving something as having two parts: a) figuring out what needs to be done, in the sense of finding an overall plan for the argument; b) filling in the details. Although the culture of mathematics celebrates the exceptions, in the everyday work of a mathematician, and for students doing problem sets, b) is typically easy. In addition, although a) is often "truly hard" in some irreducible sense, *there are systematic mechanical methods for analyzing this task.* Paradoxically, the way to become "quick and clever" is to be a bit slower and more methodical than seems necessary when doing part a). Extra care tends to reveal the hidden nuances, and the quirky, counterintuitive, or paradoxical qualities of a concept. If these are fully absorbed, later your mind will be able to instantly grasp things that others have to work through one step at a time.

Returning from this little digression, how good is our understanding of continuity? From the point of view of functions from $\mathbb{R}$ to $\mathbb{R}$, all is well. The two definitions express a clear understanding of the meaning of continuity, and we are free to apply either, according to convenience. The problem is that there are *lots* of other settings in which a function might or might not be continuous. Most obviously, the domain or range may be $\mathbb{R}^m$, or the domain might be some subset of $\mathbb{R}^m$. Let $C([0,1])$ be the set of continuous functions $f : [0,1] \to \mathbb{R}$. For any $t \in [0,1]$ there is a function $f \mapsto f(t)$ from $C([0,1])$ to $\mathbb{R}$ that sure looks like it would be continuous if we could endow the notion of 'continuity' with some relevant meaning. Etc., etc.

This chapter describes abstractions of the two definitions of continuity above. These two concepts, and related notions, constitute the central subject matter of the field of topology, which is both an important field of research in its own right, and a defining feature of the spirit and character of 20th century mathematics. This is another illustration of the general principle that abstractions introduced to formalize important preexisting concepts often become objects of study themselves.

3.1 Metric Spaces

This section describes a general concept of distance. In order to talk about the distance between two things, say x and y, one should have a space, say X, in which they both live. The distance between x and y should be nonnegative, and it is natural to require that the distance from x to y should be the same as the distance from y to x, and that this quantity should be zero when $x = y$ and not otherwise.

If we only impose these requirements some unpleasant things can happen. For example, we might have two sequences $\{x_n\}$ and $\{y_n\}$ and a point z such that the distance from x_n to y_n and the distance from y_n to z both go to zero as $n \to \infty$, but the distance from x_n to z stays away from zero. The final condition in the following definition, which is called the **triangle inequality**, prevents this sort of thing. It can be motivated by saying that the length of a trip from x to z should never be decreased by constraining the traveller to choose a route that passes by y, but to be honest, explanations of why the big definitions in mathematics "ought" to impose some requirement have a fictional *ex post facto* quality. These definitions emerged from some process of trial and error, and became popular because in actual experience they gave rise to useful and interesting mathematics.

Definition 3.3. *A **metric space** is a pair (X,d) in which X is a set and*

$$d : X \times X \to [0,\infty)$$

is a function, called a ***metric****, such that for all* $x, y, z \in X$*:*

(a) $d(x, y) = 0$ *if and only if* $x = y$*;*

(b) $d(x, y) = d(y, x)$*;*

(c) $d(x, z) \leq d(x, y) + d(y, z)$*.*

Here are three simple but very important examples.

Example 3.4. *For any set* X *the* ***discrete metric*** *is defined by setting*

$$d(x, y) := \begin{cases} 0, & x = y, \\ 1, & \textit{otherwise.} \end{cases}$$

Example 3.5. *Let* $d : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ *be the function* $d(x, y) := |x - y|$*. Then* $d(x, y) = 0$ *if and only if* $x = y$ *because* 0 *is the only number whose absolute value is zero. Clearly*

$$d(x, y) = |x - y| = |y - x| = d(y, x),$$

and to prove the triangle inequality we apply Lemma 2.37:

$$d(x, z) = |(x - y) + (y - z)| \leq |x - y| + |y - z| = d(x, y) + d(y, z).$$

Example 3.6. *If* (X, d) *is a metric space and* $A \subset X$*, then* $(A, d|_{A \times A})$ *is a metric space. Among many other possibilities, it is worth noting that any subfield of* $\mathbb{R}$*, such as* $\mathbf{Q}$*, inherits a metric from* $\mathbb{R}$*.*

The definitions of convergence and continuity generalize to metric spaces in a straightforward manner. Let (X, d_X) be a metric space. If $x \in X$ and $r > 0$, the **open ball** of radius r around x is

$$\mathbf{U}_r(x) := \{\, x' \in X : d_X(x, x') < r \,\}.$$

A sequence $\{x_n\}$ in X **converges** to a point x if, for any $\delta > 0$, the sequence is eventually inside $\mathbf{U}_\delta(x)$. That is:

$$(\forall \delta > 0)(\exists N \in \mathbb{N})(\forall n > N)\; x_n \in \mathbf{U}_\delta(x).$$

We indicate this by writing $x_n \to x$.

Let (Y, d_Y) be another metric space, let $f : X \to Y$ be a function, and consider a point $x \in X$. As before, we can define what it means for f to be continuous at x in two different ways. First, we can say that f is

continuous at x if $f(x_n) \to f(x)$ whenever $\{x_n\}$ is a sequence converging to x. Alternatively, we can say that f is continuous at x if, for every $\varepsilon > 0$, there is $\delta > 0$ such that $f(\mathbf{U}_\delta(x)) \subset \mathbf{U}_\varepsilon(f(x))$. The proof that the two definitions of continuity at a point are equivalent is not in any significant sense different from the proof of this given earlier for functions from $\mathbb{R}$ to $\mathbb{R}$, so we won't repeat it. (It wouldn't be a bad idea to review it in order to see for yourself.) As before, we say that f is **continuous** if it is continuous at every point of X.

Of course metric spaces wouldn't be as important as they are if the Pythagorean distance $\sqrt{\sum_{i=1}^m (y_i - x_i)^2}$ between two points x and y in $\mathbb{R}^m$ wasn't a metric. That it is is intuitively obvious, since in this context the triangle inequality amounts to an assertion that a straight line is the shortest path between x and y, but we still need to prove it. This turns out to be less simple than one might expect, but in a rewarding way, in part because it provides an opportunity to introduce several important definitions, including the one below, and in part because the heart of the proof, the Cauchy-Schwartz inequality, is extremely important.

In this context the most salient feature of $\mathbb{R}^m$ is that it is an $\mathbb{R}$-module. In general, as we saw in the last chapter, if R is a commutative ring with unit and k is a natural number, then R^k is an R-module. In our system of definitions we defined an R-module structure on the set $\mathcal{F}_R(S)$ of R-valued functions on a general set S, then identified R^k with $\mathcal{F}_R(\{1, \ldots, k\})$. But instead of recalling how all this worked it is simpler to just define the module operations directly:

$$w + z = (w_1 + z_1, \ldots, w_k + z_k) \quad \text{and} \quad \beta w = (\beta w_1, \ldots, \beta w_k)$$

for $w, z \in R^k$ and $\beta \in R$.

The Pythagorean distance is "invariant under translation." That is, for any $a, x, y \in \mathbb{R}^m$ the distance between x and y is the same as the distance between $a+x$ and $a+y$. A metric with this property is completely determined by the function taking each x to its distance from the origin. As we'll explain in the proof of Lemma 3.8, properties (a) and (c) in the definition of a metric correspond exactly to (i) and (iii) in the next definition, and (b) corresponds to (ii) with $\alpha = -1$.

Definition 3.7. *A **norm** on an $\mathbb{R}$-module V is a function*

$$\|\cdot\| : V \to [0, \infty)$$

such that:

(i) for all $x \in V$, $\|x\| = 0$ if and only if $x = 0$;

(ii) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $x \in V$ and all $\alpha \in \mathbb{R}$;

(iii) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$.

The prototypical example is the **Euclidean norm** on $\mathbb{R}^m$, which is defined by setting

$$\|x\|_2 := \sqrt{x_1^2 + \cdots + x_m^2}.$$

Since squares and square roots of positive numbers are positive, that (i) holds is obvious. It is almost as obvious that (ii) holds, but we'll write out the calculation anyway:

$$\begin{aligned}\|\alpha x\|_2 &= \big(|\alpha x_1|^2 + \cdots + |\alpha x_m|^2\big)^{1/2} \\ &= \big(|\alpha|^2(|x_1|^2 + \cdots + |x_m|^2)\big)^{1/2} = |\alpha| \cdot \|x\|_2.\end{aligned}$$

That it also satisfies (iii) is the main point of our work over the next few pages.

After $\|\cdot\|_2$, the most important norms on $\mathbb{R}^m$ are

$$\|x\|_1 := |x_1| + \cdots + |x_m|$$

and

$$\|x\|_\infty := \max\{|x_1|, \ldots, |x_m|\}.$$

It is easy to see that these also satisfy (i) and (ii). For (iii) we have the calculations

$$\begin{aligned}\|x + y\|_1 &= |x_1 + y_1| + \cdots + |x_m + y_m| \\ &\leq |x_1| + |y_1| + \cdots + |x_m| + |y_m| = \|x\|_1 + \|y\|_1\end{aligned}$$

and

$$\begin{aligned}\|x + y\|_\infty &= \max\{|x_1 + y_1|, \ldots, |x_m + y_m|\} \\ &\leq \max\{|x_1| + |y_1|, \ldots, |x_m| + |y_m|\} \\ &\leq \max\{|x_1|, \ldots, |x_m|\} + \max\{|y_1|, \ldots, |y_m|\} \\ &= \|x\|_\infty + \|y\|_\infty.\end{aligned}$$

Especially for infinite dimensional spaces, norms are interesting for several reasons. But for us the only important point is:

Lemma 3.8. *If $\|\cdot\|$ is a norm on V, then there is a metric $d_{\|\cdot\|}$ on V defined by setting*

$$d_{\|\cdot\|}(x,y) := \|x-y\|.$$

Proof. We must show that $d_{\|\cdot\|}$ satisfies (a)-(c) of the definition of a metric space. Property (a) corresponds directly to (i), and for (b) we have

$$d_{\|\cdot\|}(y,x) = \|y-x\| = \|-(x-y)\| = |-1| \cdot \|x-y\| = d_{\|\cdot\|}(x,y).$$

The triangle inequality is a simple consequence of (iii):

$$d_{\|\cdot\|}(x,z) = \|(x-y)+(y-z)\| \le \|x-y\| + \|y-z\| = d_{\|\cdot\|}(x,y) + d_{\|\cdot\|}(y,z).$$

□

Our proof that $\|\cdot\|_2$ satisfies (iii) involves another extremely important geometric concept. The **inner product** of two points x and y in $\mathbb{R}^m$ is

$$x_1y_1 + \cdots + x_my_m.$$

It is denoted by either $x \cdot y$ (in which case it is often called the "dot product") or $\langle x, y\rangle$. For all $x, y, z \in \mathbb{R}^m$ and all $\alpha \in \mathbb{R}$ we have:

$$\langle x, y\rangle = \langle y, x\rangle;$$

$$\langle x+y, z\rangle = \langle x, z\rangle + \langle y, z\rangle;$$

$$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle;$$

$$\langle x, x\rangle^{1/2} = \|x\|_2.$$

Each of these is simple enough that I don't think there is any need for an explicit proof, but make sure that you understand why they are true. One reason these properties are so important is that the first three constitute the definition of an abstract inner product, which is a concept with many important applications. The method of passing from an inner product to a norm expressed by the fourth equation works for any inner product, for reasons we'll explain at the end of the section.

The following result is one of the most important basic facts of mathematics.

Theorem 3.9 (Cauchy-Schwartz Inequality). *For all $x, y \in \mathbb{R}^m$,*

$$|\langle x, y\rangle| \le \|x\|_2 \cdot \|y\|_2.$$

If $y \neq 0$, then this holds with equality if and only if $x = \alpha y$ for some $\alpha \in \mathbb{R}$.

The following calculation uses the properties of the inner product, and the Cauchy-Schwartz inequality, to show that $\|\cdot\|_2$ satisfies condition (iii) of Definition 3.7:

$$\begin{aligned}\|x+y\|_2^2 &= \langle x+y, x+y\rangle \\ &= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle \\ &= \|x\|_2^2 + 2\langle x,y\rangle + \|y\|_2^2 \\ &\le \|x\|_2^2 + 2\|x\|_2 \cdot \|y\|_2 + \|y\|_2^2 = \big(\|x\|_2 + \|y\|_2\big)^2.\end{aligned}$$

That is, the Cauchy-Schwartz inequality implies that $\|\cdot\|_2$ is a norm.

On the way to proving the Cauchy-Schwartz inequality, we introduce one more geometric concept. Two points w and z in $\mathbb{R}^m$ are said to be **perpendicular** or **orthogonal** if $\langle w,z\rangle = 0$. Often we express this by writing $w \perp z$. At first it might seem that "$\langle w,z\rangle = 0$ implies $w \perp z$" is something we should prove, but our general approach is to define geometric concepts algebraically, then validate these definitions by showing that expected geometric relations hold. In this case we should expect that if $w \perp z$, then the origin, w, and z should be the vertices of a right triangle whose hypotenuse is the line segment between w and z, and indeed the Pythagorean relationship between the side lengths does hold:

$$\begin{aligned}\|w-z\|_2^2 &= \langle w-z, w-z\rangle \\ &= \langle w,w\rangle - \langle w,z\rangle - \langle z,w\rangle + \langle z,z\rangle \\ &= \langle w,w\rangle + \langle z,z\rangle \\ &= \|w\|_2^2 + \|z\|_2^2.\end{aligned}$$

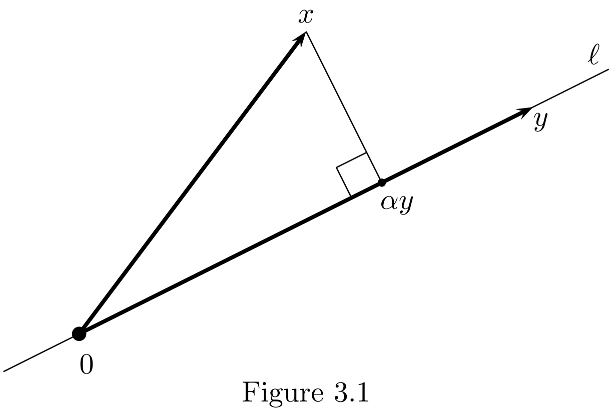


Figure 3.1

The Cauchy-Schwartz inequality holds automatically when $x \perp y$, simply because $\|x\|$ and $\|y\|$ are both nonnegative. Let's see what happens when we apply this special case to $x - \alpha y$ and y with α chosen so that $(x - \alpha y) \perp y$. (Such an α exists because when $y \neq 0$ we can set $\alpha := \langle x, y\rangle / \langle y, y\rangle$, and any α is satisfactory when $y = 0$.) We have

$$\langle x - \alpha y, y\rangle = 0 \leq \|x - \alpha y\| \cdot \|y\|.$$

Expanding the squares of the two sides of this inequality gives

$$\langle x - \alpha y, y\rangle^2 = \big(\langle x, y\rangle - \alpha\langle y, y\rangle\big)^2 = \langle x, y\rangle^2 - 2\alpha\langle x, y\rangle\langle y, y\rangle + \alpha^2\langle y, y\rangle^2$$

and

$$\begin{aligned}\langle x - \alpha y, x - \alpha y\rangle\langle y, y\rangle &= \big(\langle x, x\rangle - 2\alpha\langle x, y\rangle + \alpha^2\langle y, y\rangle\big)\langle y, y\rangle \\ &= \langle x, x\rangle\langle y, y\rangle - 2\alpha\langle x, y\rangle\langle y, y\rangle + \alpha^2\langle y, y\rangle^2.\end{aligned}$$

Therefore

$$0 \leq \big(\|x - \alpha y\| \cdot \|y\|\big)^2 - \langle x - \alpha y, y\rangle^2 = \langle x, x\rangle\langle y, y\rangle - \langle x, y\rangle^2.$$

We now have $\langle x, y\rangle^2 \leq \langle x, x\rangle\langle y, y\rangle$, and if we take the square root on both sides we have the Cauchy-Schwartz inequality! In addition, we see that the inequality holds with equality precisely when $\|x - \alpha y\| = 0$, so if it holds with equality, then x is a scalar multiple of y. But the converse is also true: if $y \neq 0$, x is a scalar multiple of y, and $x - \alpha y \perp y$, then $x - \alpha y = 0$.

Above we gave four "simple" properties of the inner product, the first three of which could be taken as the definition of an abstract inner product for $\mathbb{R}$-modules. Our proof of the Cauchy-Schwartz inequality relied on these four equations, without making any reference to the definition $\langle x, y\rangle := \sum_i x_i y_i$, as you can (and should) see for yourself by reviewing the discussion above. Consequently the Cauchy-Schwartz is a property of abstract inner products. In turn, our proof that $\|\cdot\|_2$ is a norm depended only on the fourth equation $\|x\|_2 = \sqrt{\langle x, x\rangle}$, the basic properties of the inner product, and the Cauchy-Schwartz inequality. The upshot of this is that the fourth equation defines a norm whenever we have an abstract inner product satisfying the first three equations.

3.2 Topological Spaces

Although occasionally one deals with situations where it is important that a function is continuous at a particular point and possibly not elsewhere, mostly one is concerned with functions that are continuous everywhere. If we restrict attention in this way, then the second definition of continuity in the last section can be rephrased in a simple and illuminating manner.

Definition 3.10. *If (X, d) is a metric space, a set $U \subset X$ is* ***open*** *if it contains an open ball around each of its points:*

$$(\forall x \in U)(\exists \delta > 0)\ \mathbf{U}_\delta(x) \subset U.$$

Proposition 3.11. *A function $f : X \to Y$ between metric spaces X and Y is continuous if and only if $f^{-1}(V)$ is open whenever $V \subset Y$ is open.*

The proof involves a point that deserves special emphasis, so we treat it separately. As much as anything else, the following fact is why the "right" definition of a metric space involves the triangle inequality.

Lemma 3.12. *If (X, d) is a metric space, $x \in X$, and $\delta > 0$, then $\mathbf{U}_\delta(x)$ is open.*

Proof. For any point $x' \in \mathbf{U}_\delta(x)$ we have $\mathbf{U}_{\delta - d(x,x')}(x') \subset \mathbf{U}_\delta(x)$ because if $d(x', x'') < \delta - d(x, x')$, then the triangle inequality gives

$$d(x, x'') \leq d(x, x') + d(x', x'') < \delta.$$

□

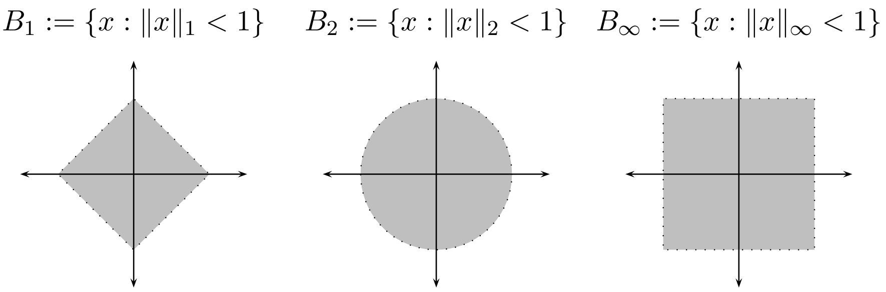
Proof of Proposition 3.11. First suppose that f is continuous. Let $V \subset Y$ be open, and consider an arbitrary $x \in f^{-1}(V)$. Since V is open, there is some $\varepsilon > 0$ such that $\mathbf{U}_\varepsilon(f(x)) \subset V$. Since f is continuous, there is some $\delta > 0$ such that $f(\mathbf{U}_\delta(x)) \subset \mathbf{U}_\varepsilon(f(x))$, so that

$$\mathbf{U}_\delta(x) \subset f^{-1}(\mathbf{U}_\varepsilon(f(x))) \subset f^{-1}(V).$$

Since x was arbitrary, this shows that $f^{-1}(V)$ is open.

Now suppose that $f^{-1}(V)$ is open whenever $V \subset Y$ is open. Consider a point $x \in X$ and a number $\varepsilon > 0$. Since $\mathbf{U}_\varepsilon(f(x))$ is open, $f^{-1}(\mathbf{U}_\varepsilon(f(x)))$ is open, and there is some $\delta > 0$ such that $\mathbf{U}_\delta(x) \subset f^{-1}(\mathbf{U}_\varepsilon(f(x)))$, i.e., $f(\mathbf{U}_\delta(x)) \subset \mathbf{U}_\varepsilon(f(x))$. Since x and ε were arbitrary, this shows that f is continuous. □

Whether or not $f : X \to Y$ is continuous is a matter of which subsets of X and Y are open, but *it can easily happen that two different metrics give rise to the same system of open sets.* Specifically, suppose that d and d' are two metrics on X. If, for every x and $\delta > 0$, there is some $\delta' > 0$ such that the open ball of radius δ' around x in (X, d') is contained in the open ball of radius δ around x in (X, d), then every set that is open in (X, d) is also open in (X, d'). It is not at all unusual that this is the case and that, at the same time, every set that is open in (X, d') is also open in (X, d).



$B_1 := \{x : \|x\|_1 < 1\}$ $B_2 := \{x : \|x\|_2 < 1\}$ $B_\infty := \{x : \|x\|_\infty < 1\}$

Figure 3.2

In Figure 3.2 we see the open balls of radius one centered at the origin in $\mathbb{R}^2$ for the metrics derived from $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$. It is visually obvious that $B_1 \subset B_2 \subset B_\infty$, and this is an algebraic consequence of the inequalities

$$|x_1| + |x_2| = \sqrt{|x_1|^2 + 2|x_1|\,|x_2| + |x_2|^2} \ge \sqrt{x_1^2 + x_2^2} \ge \max\{|x_1|, |x_2|\}.$$

In addition we have the inequality

$$\max\{|x_1|, |x_2|\} \ge \tfrac{1}{2}(|x_1| + |x_2|).$$

Therefore

$$B_1 \subset B_2 \subset B_\infty \subset 2B_1.$$

More generally, for any $x \in \mathbb{R}^2$ and any $\delta > 0$ we have

$$x + \delta B_1 \subset x + \delta B_2 \subset x + \delta B_\infty \subset x + 2\delta B_1.$$

Thus the metrics derived from $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ all determine the same system of open sets for $\mathbb{R}^2$.

If $f : X \to Y$ is a function between spaces X and Y that are endowed with "systems of open sets," then we can *define* continuity to mean that $f^{-1}(V)$ is an open subset of X whenever V is an open subset of Y. The question then becomes: what conditions should we impose on the systems of open sets? Ideally we would like to impose conditions that are restrictive enough to give us a coherent and meaningful theory, and are also loose enough to encompass all, or at least most, of the interesting applications. It turns out that the open sets of a metric space (X, d) have a great many properties, but three of them are pretty much indispensable to any useful theory of convergence and continuity.

Definition 3.13. *A **topology** for a set X is a collection τ of subsets of X with the following properties:*

(T1) $\emptyset, X \in \tau$*;*

(T2) for all $U_1, U_2 \in \tau$, $U_1 \cap U_2 \in \tau$*;*

(T3) $\bigcup_{\alpha \in A} U_\alpha \in \tau$ *whenever A is a set and, for each $\alpha \in A$, $U_\alpha \in \tau$.*

*Elements of τ are called **open sets**. A **topological space** is a pair (X, τ) in which X is a set and τ is a topology for X.*

It is easy to see that the open sets of a metric space (X, d) have these properties. The entire space X is open because $\mathbf{U}_r(x) \subset X$ for any x and $r > 0$, and $\emptyset$ is open because it has no points and consequently contains a ball around each of its points "trivially." If U_1 and U_2 are open subsets of X, then so is $U_1 \cap U_2$ because for any $x \in U_1 \cap U_2$ we have

$$\mathbf{U}_{\min\{\delta_1, \delta_2\}}(x) \subset U_1 \cap U_2$$

whenever $\mathbf{U}_{\delta_1}(x) \subset U_1$ and $\mathbf{U}_{\delta_2}(x) \subset U_2$. If $\{U_\alpha\}_{\alpha \in A}$ is any collection of open subsets of X, then $U := \bigcup_\alpha U_\alpha$ is open because for any $x \in U$ there is some α with $x \in U_\alpha$ and some $\delta > 0$ such that $\mathbf{U}_\delta(x) \subset U_\alpha \subset U$.

Something approximating Definition 3.13 first appeared in a 1914 paper by Felix Hausdorff (1868-1942), and a 1909 paper by Frigyes Riesz (1880-1956) is an important precursor. Given the dates, the people involved, and the nature of the definition itself, it is clear that it should be seen as one of the fruits of the set theory revolution. One might guess that this formulation emerged from experience with numerous applications, but the actual historical process was quite different. In one field after another since that time topologies have been noticed, studied, and incorporated into the field's basic organizing principles. In some cases, such as the one we describe in Section 3.4, the topology in question seemed bizarre at first, but proved quite useful in spite of this. Of all the concepts of mathematics, the notion of a topological space might be the one whose significance is most mystical.

It would now be possible to give a great many simple and useful definitions and results concerning topological spaces and functions between them. Instead of doing this, for the most part we will introduce basic topological concepts as we need them, and even in the end our coverage of the basics will be woefully incomplete. We restrict the discussion in the rest of this section to a quick exposition of some definitions and facts that are among the most crucial and fundamental.

Definition 3.14. *If (X, τ_X) and (Y, τ_Y) are topological spaces, a function $f : X \to Y$ is* ***continuous*** *if $f^{-1}(V)$ is open in X whenever V is an open subset of Y.*

The equivalence between the two definitions of continuity for functions between metric spaces does not carry through to this level of generality. To help explain this, and because it is extremely useful in many other circumstances, we introduce the following terminology. A set $A \subset X$ is said to be a **neighborhood** of a point x if it contains an open set that in turn contains x. Note that A need not be open itself, so we will need to use the longer phrase "open neighborhood" when that is what we want.

Definition 3.15. *If (X, τ) is a topological space, $\{x_n\}$ is a sequence in X, and $x \in X$, then we say that $\{x_n\}$* ***converges*** *to x, and write $x_n \to x$, if $\{x_n\}$ is eventually inside every open neighborhood of x:*

$$(\forall U \in \tau)\big[x \in U \Rightarrow (\exists N \in \mathbb{N})(\forall n > N)\; x_n \in U\big].$$

Proposition 3.16. *If (X, τ_X) and (Y, τ_Y) are topological spaces, $f : X \to Y$ is continuous, and $\{x_n\}$ is a sequence in X that converges to x, then $f(x_n) \to f(x)$.*

Proof. For any open $V \subset Y$ that contains $f(x)$, $f^{-1}(V)$ is open, so it contains x_n for all sufficiently large n, and consequently V contains $f(x_n)$ for all sufficiently large n. □

It can happen that f is not continuous even though $f(x_n) \to f(x)$ whenever $\{x_n\}$ is a sequence in X converging to x. Roughly speaking, it is possible that a point $x \in X$ has open neighborhoods that are so "diverse" that it is impossible for a single sequence $\{x_n\}$ to eventually get inside, and stay inside, all of them unless there is an N such that $x_n = x$ for all $n > N$. In this case the requirement that $f(x_n) \to f(x)$ whenever $x_n \to x$ has no bite, even though continuity can still impose nonvacuous restrictions on the relationship between $f(x)$ and the values of f at other points.

Proposition 3.17. *If (X, τ_X), (Y, τ_Y), and (Z, τ_Z) are topological spaces, and $f : X \to Y$ and $g : Y \to Z$ are continuous, then $g \circ f$ is continuous.*

Proof. If $W \subset Z$ is open, then $(g \circ f)^{-1}(W)$ is open because the continuity of g implies that $g^{-1}(W)$ is open, after which the continuity of f implies that $f^{-1}(g^{-1}(W))$ is open. □

Theorem 3.18. *There is a category whose objects are topological spaces and whose morphisms are the continuous functions between them.*

Proof. As usual, everything is trivial. We have just shown that compositions of continuous functions are continuous, and composition of continuous functions is an associative operation because this is a general property of composition of functions. It is obvious that the identity function from any topological space to itself is continuous, and $\mathrm{Id}_Y \circ f = f = f \circ \mathrm{Id}_X$ whenever (X, τ_X) and (Y, τ_Y) are topological spaces and $f : X \to Y$ is continuous, again simply because this a property of functions in general. □

After composition, perhaps the second most fundamental and important operation on functions is restriction to a subspace of the domain. It is also possible to restrict to a superset of the image in the range. In both cases we need to impose a topology on a subset of a topological space, and there is a natural (indeed, almost inevitable) way to do this.

Proposition 3.19. *If (X, τ) is a topological space and $A \subset X$, then*

$$\tau|_A := \{\, U \cap A : U \in \tau \,\}$$

is a topology on A.

Proof. We show that $\tau|_A$ satisfies (T1)-(T3). First of all, $\emptyset$ and A itself are open because $\emptyset = \emptyset \cap A$ and $A = X \cap A$. To show that the intersection of two open sets is open suppose that $V_1 = U_1 \cap A$ and $V_2 = U_2 \cap A$ for some $U_1, U_2 \in \tau$. Then

$$V_1 \cap V_2 = (U_1 \cap A) \cap (U_2 \cap A) = (U_1 \cap U_2) \cap A,$$

so $V_1 \cap V_2 \in \tau|_A$. Finally suppose that I is a set and, for each $i \in I$, $V_i \in \tau|_A$, so that there is some $U_i \in \tau$ such that $V_i = U_i \cap A$. Then $\bigcup_i V_i \in \tau|_A$ because

$$\bigcup_i V_i = \bigcup_i (U_i \cap A) = \Big(\bigcup_i U_i\Big) \cap A.$$

□

We call $\tau|_A$ the **subspace topology**, the **relative topology**, or (less frequently) the **induced topology** of A. Often A is open, in which case there are no new open sets: if $V \subset A$ is open in X, then it is relatively open because $V = V \cap A$, while if V is relatively open, for instance because $V = U \cap A$ where U is open in X, then V is open in X because U and A are both open.

The following basic fact about the relative topology is applied so frequently that it is usually taken for granted.

Proposition 3.20. *If (X, τ_X) and (Y, τ_Y) are topological spaces, $f : X \to Y$ is continuous, $A \subset X$, and $f(A) \subset B \subset Y$, then*

$$f|_A : A \to B$$

is continuous when A and B have their subspace topologies.

Proof. Our task is to show that if W is open in B, then $(f|_A)^{-1}(W)$ is open in A. The definition of the topology of B implies that $W = V \cap B$ for some open $V \subset Y$, and since $f(A) \subset B$, $(f|_A)^{-1}(W) = (f|_A)^{-1}(V)$. Of course $(f|_A)^{-1}(V) = f^{-1}(V) \cap A$, and $f^{-1}(V)$ is open in X because f is continuous, so $f^{-1}(V) \cap A$ is open in A. □

Mathematicians often say that a property of a space, or a function, or perhaps some other type of object, is **local**. What they mean by this is that the space (or function, or whatever) has this property whenever each point in the space (or the domain of the function, or ...) has a neighborhood with this property. The most useful way of expressing this involves the following important concept: an **open cover** of a topological space X is a collection $\{U_\alpha\}_{\alpha \in A}$ where A is some index set, each U_α is an open subset of X, and $\bigcup_{\alpha \in A} U_\alpha = X$. A property is local if a space (or function, or ...) has the property whenever every element of some open cover (or the restriction of the function to every element of some open cover, or ...) has the property.

Proposition 3.21. *Continuity is a local property.*

Proof. Suppose X and Y are topological spaces, $f : X \to Y$ is a function, $\{U_\alpha\}_{\alpha \in A}$ is an open cover of X, and each $f|_{U_\alpha}$ is continuous. If $V \subset Y$ is open, then

$$f^{-1}(V) = \bigcup_{\alpha \in A} (f|_{U_\alpha})^{-1}(V).$$

Each $(f|_{U_\alpha})^{-1}(V)$ is open in the relative topology of U_α, so it is open in X because U_α is open, and consequently $f^{-1}(V)$ is open because it is a union of open sets. □

3.3 Closed Sets

If (X, τ) is a topological space, a set $C \subset X$ is **closed** if $X \setminus C$ is open. The systems of closed and open sets are the yin and yang of a topological space. We have described a topology in terms of the properties of its open sets, but it should also be possible to give a description in terms of the properties of closed sets, and indeed this is easy. Since union (intersection) of open sets corresponds to intersection (union) of the complementary closed sets, a collection χ of subsets of X is the system of closed sets of a topology if:

(a) $\emptyset, X \in \chi$;

(b) $C_1 \cup C_2 \in \chi$ whenever $C_1, C_2 \in \chi$;

(c) $\bigcap_\alpha C_\alpha \in \chi$ whenever A is a set and, for each $\alpha \in A$, $C_\alpha \in \chi$.

To get a more direct and intuitive way of thinking about closed sets we introduce some more terminology, which (in the collective experience of the mathematical community) has proven very useful. If $A \subset X$, a point $x \in X$ is an **accumulation point** of A if every neighborhood of x has a nonempty intersection with A. Since any neighborhood of x contains x itself, x is an accumulation point of A if it is an element of A.

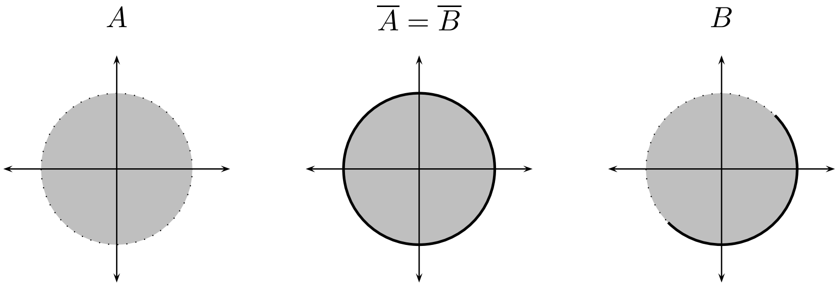


Figure 3.4

The **closure** of A, denoted by $\overline{A}$, is the set of all accumulation points of A. As we just mentioned, $A \subset \overline{A}$. If a sequence in A converges to a point x, then (by the definition of convergence) $x \in \overline{A}$. If X is a metric space, then the converse is true: if x is an accumulation point of A, then $\mathbf{U}_{1/r}(x) \cap A \neq \emptyset$ for every integer r, so we can construct a sequence $x_1, x_2, \ldots$ in A that converges to x by choosing $x_r \in \mathbf{U}_{1/r}(x)$. (This construction applies the axiom of choice!) In a general topological space $\overline{A}$ contains all the limits of sequences in A, but A can also have an accumulation point x that is not the limit of any sequence in A, roughly because x has so many neighborhoods that no sequence $\{x_n\}$ can eventually be inside all of them unless $x_n = x$ for all large n. Nonetheless, for "practical purposes" the most direct intuition for, and visualization of, this concept is that the closure of a set in a metric space is the set of all limits of sequences in the set.

In this system of terminology it is not really proper to treat 'close' as a verb (you *are* allowed to "take the closure" of a set) but if it were we could say that "a closed set is one that has already been closed:"

Lemma 3.22. *A set $C \subset X$ is closed if and only if $C = \overline{C}$.*

Proof. Suppose C is closed. As noted above, $C \subset \overline{C}$. On the other hand, all the points of $X \setminus C$ are contained in an open set (namely $X \setminus C$ itself) whose intersection with C is empty, so none of them are accumulation points of C. Therefore $X \setminus C \subset X \setminus \overline{C}$, so $\overline{C} \subset C$.

Now suppose that $\overline{C} = C$, and consider $x \in X \setminus \overline{C}$. Since x isn't an accumulation point of C, it has an open neighborhood V_x with $V_x \cap C = \emptyset$. We have

$$X \setminus C = \bigcup_{x \in X \setminus C} V_x,$$

so $X \setminus C$ is open, and consequently C is closed. □

Many of the basic facts of topology can be stated either in terms of open sets or in terms of closed sets, and a collection of reformulations is useful because it often happens in proofs that an application of a result described in terms of open (closed) sets would be indirect and confusing, while the same idea expressed in terms of closed (open) sets seems direct and straightforward. Without trying to be exhaustive, we now give some of the rephrasings that come up most frequently.

Lemma 3.23. *If (X, τ) is a topological space and $A \subset X$, then $C \subset A$ is closed in the relative topology of A if and only if there is a closed $D \subset X$ such that $C = D \cap A$.*

Proof. We have (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ (c) $\Leftrightarrow$ (d) where:

(a) C is closed in the relative topology of A;

(b) $A \setminus C$ is open in the relative topology of A;

(c) $A \setminus C = U \cap A$ for some open $U \subset X$;

(d) $C = D \cap A$ for some closed $D \subset X$.

□

Since a set is open if it is a neighborhood of each of its points, openness is a local property. This feels too obvious to state as an explicit formal result, but the flip side of this is somehow a bit less intuitive.

Proposition 3.24. *Closedness is a local property: if $C \subset X$, $\{U_\alpha\}_{\alpha \in A}$ is an open cover of X, and each $C \cap U_\alpha$ is closed in the relative topology of U_α, then C is closed.*

Proof. Each $U_\alpha \setminus C$ is relatively open, hence open in X because U_α is open, so $X \setminus C = \bigcup_{\alpha \in A}(U_\alpha \setminus C)$ is open. □

Continuity has a symmetric characterization in terms of closed sets.

Proposition 3.25. *A function $f : X \to Y$ is continuous if and only if $f^{-1}(D)$ is closed whenever $D \subset Y$ is closed.*

Proof. We have (a) $\Leftrightarrow$ (b) $\Leftrightarrow$ (c) $\Leftrightarrow$ (d) where:

(a) $f^{-1}(V)$ is open whenever $V \subset Y$ is open;

(b) $X \setminus f^{-1}(V)$ is closed whenever $V \subset Y$ is open;

(c) $f^{-1}(Y \setminus V)$ is closed whenever $V \subset Y$ is open;

(d) $f^{-1}(D)$ is closed whenever $D \subset Y$ is closed.

□

Proving that a function is continuous is one of the commonest tasks in mathematics. It is usually the case that the simplest and easiest arguments use set theoretic manipulations to show that the preimage of any open set is open, or that the preimage of any closed set is closed. Eventually such arguments seem clear and natural, but they take some getting used to because such a proof usually does not give a visual description of why the function is continuous. The following result, which is a sort of complement to Proposition 3.21 (but in this case the cover must be finite) illustrates these ideas.

Proposition 3.26. *Suppose $f : X \to Y$ is a function, where X and Y are topological spaces, and $X = \bigcup_{j=1}^m C_j$, where each C_j is closed. If each restriction $f|_{C_j}$ is continuous, then f is continuous.*

In thinking about why this should be true one might imagine a sequence $x_1, x_2, \ldots$ in X converging to a point x. A **subsequence** is a sequence of the form $x_{i_1}, x_{i_2}, \ldots$ where $i_1 < i_2 < \cdots$. If $f(x_1), f(x_2), \ldots$ did not converge to $f(x)$ there would be an open $V \subset Y$ containing $f(x)$ but not containing $f(x_i)$ for every sufficiently large i, so there would be a subsequence with $f(x_{i_h}) \notin V$ for all h. Since there are only finitely many C_j, there would have to be some j with $x_{i_h} \in C_j$ for infinitely many h, and by replacing our subsequence with a further subsequence of itself we could make it the case that $x_{i_h} \in C_j$ for all h. If a sequence converges to a point, then so does any of its subsequences, obviously, so $x_{i_h} \to x$. Therefore x is an accumulation point of the set $\{x_{i_1}, x_{i_2}, \ldots\}$, so $x \in C_j$ because C_j is closed. But $f|_{C_j}$ is continuous, so $f(x_{i_h}) \to f(x)$ and consequently $f(x_{i_h}) \in V$ for all sufficiently large h after all. This contradiction shows that $f(x_i) \to f(x)$.

Since we are working with general topological spaces, to prove that f is continuous it is not enough to show that $f(x_i) \to f(x)$ whenever $x_i \to x$, so this "explanation" of Proposition 3.26 is not a valid proof. But even if it were, for those with a certain amount of experience the following is preferable because it is brief and direct.

Proof. We apply Proposition 3.25 and Lemma 3.23. It suffices to show that $f^{-1}(D)$ is closed for any given closed set $D \subset Y$. For each $j = 1, \ldots, m$, $(f|_{C_j})^{-1}(D)$ is closed in the relative topology of C_j, so it is the intersection of C_j (which is closed) with a closed subset of X, and consequently it is closed in X. Therefore

$$f^{-1}(D) = \bigcup_{j=1}^{m} (f|_{C_j})^{-1}(D)$$

is closed because it is a finite union of closed sets. □

3.4 The Zariski Topology

It's extremely easy to give an example of a topology that is not derived from a metric. Let $X_n = \{1, \ldots, n\}$, and let $\tau_n = \{U_0, \ldots, U_n\}$ where

$$U_0 = \emptyset,\ U_1 = \{1\},\ U_2 = \{1,2\},\ \ldots,\ U_{n-1} = \{1, \ldots, n-1\},\ U_n = X.$$

For any collection $\{U_{i_\alpha}\}_{\alpha \in A}$ of open sets and any pair of open sets U_{i_1} and U_{i_2} we have

$$\bigcup_{\alpha \in A} U_{i_\alpha} = U_{\max_\alpha i_\alpha} \quad \text{and} \quad U_{i_1} \cap U_{i_2} = U_{\min\{i_1, i_2\}},$$

so τ_n satisfies (T1)-(T3).

To show that this topology is not derived from a metric we introduce the following important concept: a **Hausdorff space** is a topological space (X, τ) such that for any two distinct points $x, y \in X$ there are open sets U and V such that $x \in U$, $y \in V$, and $U \cap V = \emptyset$. Any metric space (X, d) is a Hausdorff space because

$$\mathbf{U}_{d(x,y)/2}(x) \quad \text{and} \quad \mathbf{U}_{d(x,y)/2}(y)$$

are open and disjoint (due to the triangle inequality) whenever x and y are distinct points in X. But any two nonempty elements of τ_n have a nonempty intersection, so (X_n, τ_n) isn't a Hausdorff space if $n \geq 2$.

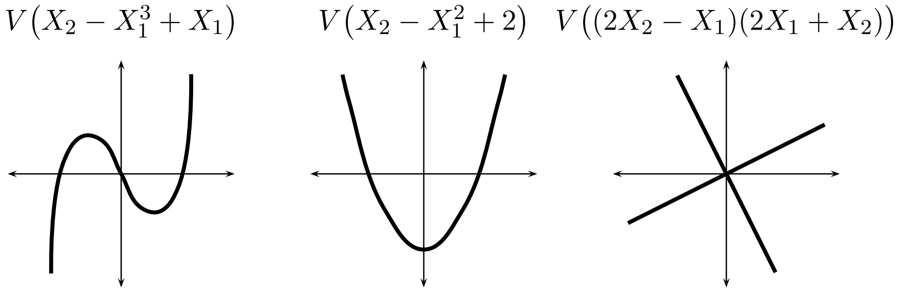
As an example of a topology that is not Hausdorff, and therefore not derived from a metric, (X_n, τ_n) suffers from the fact that it's trivial, giving no sense of why such a topology might be interesting or useful. So, for the sake of giving a better example, and also just because the topic is interesting, in this section we are going to describe a rather remarkable topology. The main result that we need, the Hilbert basis theorem, is relatively advanced, with a proof that's a bit more complicated than most of what we've seen so far. If it seems a bit difficult, well, that's because it is, but you shouldn't worry too much: the hard parts of this section won't be needed later in the book.

Fix a field k and an integer n. In the last chapter we saw the ring $R[X]$ of polynomials in a single variable with coefficients in a commutative ring R, and we paid particular attention to the ring $k[X]$. Now we'll be working with $k[X_1, \ldots, X_n]$, which is the ring of polynomials, like $1 + 2X_1X_2^2 - X_3^5$, in n variables with coefficients in k. To drive home the point that this is not really a new concept (and to prepare for an inductive argument below) observe that we can build this ring up one step at a time, treating $k[X_1, \ldots, X_n]$ as $k[X_1, \ldots, X_{n-1}][X_n]$.

Affine n-space over k, denoted by $\mathbf{A}^n(k)$, is the n-fold cartesian product of k with itself, i.e., the set of ordered n-tuples of elements of k. (That is, $\mathbf{A}^n(k)$ is just k^n, but regarded as a quite particular geometric object.) For any polynomial $f \in k[X_1, \ldots, X_n]$ there is an associated function, which we also denote by f, from $\mathbf{A}^n(k)$ to k, with the obvious definition: if, for example, $f = 1 + 2X_1X_2^2 - X_3^5$, then $f(a_1, a_2, a_3) = 1 + 2a_1a_2^2 - a_3^5$. An **affine algebraic set** is a set of the form

$$V(S) = \{\, a \in \mathbf{A}^n(k) : f(a) = 0 \text{ for all } f \in S \,\}$$

where S may be any subset of $k[X_1, \ldots, X_n]$. (If $S = \{f_1, \ldots, f_k\}$ is finite we usually write $V(f_1, \dots, f_k)$ in place of the more cumbersome $V(\{f_1, \dots, f_k\})$.) When $k = \mathbb{R}$ these are familiar objects of the sort shown in Figure 3.3.



$V\bigl(X_2 - X_1^3 + X_1\bigr)$ $V\bigl(X_2 - X_1^2 + 2\bigr)$ $V\bigl((2X_2 - X_1)(2X_1 + X_2)\bigr)$

Figure 3.3

A first definition of the field of mathematics called **algebraic geometry** is that it studies algebraic sets. (Actually this field has redefined its fundamental objects of study several times during its history, becoming highly abstract in the process, so this is very far from being a final or satisfactory description.) In the most important topology in algebraic geometry, called the **Zariski topology**, the closed sets of $\mathbf{A}^n(k)$ are the affine algebraic sets. (Oscar Zariski (1899-1986) did important work on the algebraic foundations of algebraic geometry.)

Let's check that these sets have the properties of the system of closed sets of a topology. If 0 and 1 denote the polynomials whose only terms are the corresponding constants, then $\emptyset = V(1)$ and $\mathbf{A}^n(k) = V(\emptyset) = V(0)$. If $S_1, S_2 \subset k[X_1, \dots, X_n]$, then we claim that $V(S_1) \cup V(S_2)$ is closed because

$$V(S_1) \cup V(S_2) = V(\{\, fg : f \in S_1, g \in S_2\}).$$

In concrete detail:

(i) if $a \in V(S_1) \cup V(S_2)$, then $fg(a) = 0$ whenever $f \in S_1$ and $g \in S_2$;

(ii) if $a \in \mathbf{A}^n(k) \setminus (V(S_1) \cup V(S_2))$, then there is an $f \in S_1$ such that $f(a) \neq 0$ and a $g \in S_2$ such that $g(a) \neq 0$, so $fg(a) \neq 0$.

If, for each $\alpha \in A$, $S_\alpha \subset k[X_1, \dots, X_n]$, then it is easy to see that

$$\bigcap_\alpha V(S_\alpha) = V\bigl(\bigcup_a S_\alpha\bigr).$$

So, we do have a topology.

We will show that the Zariski topology is not a Hausdorff topology when k is infinite, with the next result giving the key idea of the argument. First we need a bit of terminology. For any commutative ring R, if

$$f = c_d X^d + \cdots + c_1 X + c_0 \in R[X]$$

with $c_d \neq 0$, then $c_d X^d$ is the **leading term** of f, c_d is the **leading coefficient**, and d is the **degree**.

Proposition 3.27. *If k is infinite, $f \in k[X_1, \dots, X_n]$, and the function from $\mathbf{A}^n(k)$ to k defined by f is identically zero, then $f = 0$.*

We first deal with the case $n = 1$.

Lemma 3.28. *A nonzero polynomial $f \in k[X]$ of degree d has at most d distinct roots.*

Proof. Certainly this is true when $d = 0$, since f is then a nonzero constant. Arguing by induction, suppose that the result has been established for polynomials of degree $d - 1$, and that f has degree d. For any $r \in k$ division with remainder gives $f = (X - r) \cdot g + q$ where g is a polynomial of degree $d - 1$ and $q \in k$. If r is a root of f, then $q = 0$, so $f = (X - r) \cdot g$, and the roots of f are d and the roots of g. □

Proof of Proposition 3.27. The lemma establishes the claim in the case $n = 1$. Now suppose that $n > 1$, and that the result has already been established with $n - 1$ in place of n. Let $f = c_d X_n^d + \cdots + c_1 X_n + c_0$ where the coefficients are now elements of $k[X_1, \dots, X_{n-1}]$. If the function defined by f vanishes everywhere in $\mathbf{A}^n(k)$, then the univariate polynomial derived from any particular values of $X_1, \dots, X_{n-1}$ must be zero for all values of X_n, so the lemma imples that it is the zero polynomial. Therefore each of the functions defined by the coefficients vanishes everywhere in k^{n-1}, so each of the coefficients is the zero polynomial. Thus $f = 0$. □

Our application of Proposition 3.27 involves the following fact.

Lemma 3.29. *$k[X_1, \dots, X_n]$ is an integral domain.*

Proof. Obviously $k[X_1, \dots, X_n]$ is a commutative ring with unit, so the claim boils down to it not having any zero divisors. Let f and g be nonzero elements of $k[X_1, \dots, X_n]$. If we think of f and g as univariate polynomials in the variable X_n with coefficients in $k[X_1, \dots, X_{n-1}]$, then the leading term of fg is the product of the leading term of f and the leading term of g, and it suffices to show that it is nonzero. This follows by induction: when $n = 1$ the leading term of fg is a product of two nonzero elements of k; when $n > 1$ the leading term of fg is a product of two nonzero elements of $k[X_1, \ldots, X_{n-1}]$, and we may assume that the result has already been established with $k[X_1, \ldots, X_{n-1}]$ in place of $k[X_1, \ldots, X_n]$. □

To show that the Zariski topology is not Hausdorff we need to find distinct points $x, y \in \mathbf{A}^n(k)$ that do not have disjoint neighborhoods, but we will actually prove something stronger, namely that either $U = \emptyset$ or $V = \emptyset$ whenever U and V are open sets with $U \cap V = \emptyset$. This is the same as:

Proposition 3.30. *If k is infinite, C and D are closed subset of $\mathbf{A}^n(k)$, and $C \cup D = \mathbf{A}^n(k)$, then either $C = \mathbf{A}^n(k)$ or $D = \mathbf{A}^n(k)$.*

Proof. Suppose that $C = V(S_C)$ and $D = V(S_D)$. Then, as we saw above,

$$C \cup D = V\big(\{\, fg : f \in S_C \text{ and } g \in S_D \,\}\big),$$

so fg vanishes on all of $\mathbf{A}^n(k)$ whenever $f \in S_C$ and $g \in S_D$. Since k is infinite, it follows that $fg = 0$, and since $k[X_1, \ldots, X_n]$ doesn't have zero divisors, either $f = 0$ or $g = 0$. That is, there do not exist nonzero $f \in S_C$ and $g \in S_D$, so either $S_C \subset \{0\}$ or $S_D \subset \{0\}$. □

The power of the Zariski topology is harnessed by creating a suitable category. If $A \subset \mathbf{A}^m$ and $B \subset \mathbf{A}^n(k)$ are affine algebraic sets, a **regular function** from A to B is a function $\varphi : A \to B$ for which there exist

$$f_1, \ldots, f_n \in k[X_1, \ldots, X_m]$$

such that

$$\varphi(a) = (f_1(a), \ldots, f_n(a))$$

for all $a \in A$. Clearly Id_A is a regular function, since we can set $f_1 = X_1, \ldots, f_m = X_m$. Compositions of regular functions are regular because a composition of polynomial functions is the function given by the polynomials we obtain by substituting and expanding. (For example, if $y_1 = x_1^2 + x_2$, $y_2 = x_2^2$, and $z = 2y_1 + y_2^2$, then $z = 2(x_1^2 + x_2) + (x_2^2)^2 = 2x_1^2 + 2x_2 + x_2^4$.) As usual, composition of regular functions is associative, and $\mathrm{Id}_B \circ \varphi = \varphi = \varphi \circ \mathrm{Id}_A$, because these are properties of functions in general. Thus there is a category of affine algebraic sets and regular functions.

We endow each affine algebraic set $A \subset \mathbf{A}^m$ with the subspace topology it inherits as a subset of $\mathbf{A}^m$. In effect this means that the closed sets in A are just the Zariski-closed subsets of $\mathbf{A}^m$ that happen to be contained in A.

Theorem 3.31. *If $A \subset \mathbf{A}^m(k)$ and $B \subset \mathbf{A}^n(k)$ are affine algebraic sets and $\varphi : A \to B$ is a regular function, then φ is continuous.*

Proof. Choose $f_1, \ldots, f_n \in k[X_1, \ldots, X_m]$ such that

$$\varphi(a) = (f_1(a), \ldots, f_n(a))$$

for all $a \in A$, and let $A = V(S_A)$. If $C = V(S_C)$ is a closed subset of B, then

$$\varphi^{-1}(C) = V\big(S_A \cup \{\, h \circ (f_1, \ldots, f_n) : h \in S_C \,\}\big)$$

is Zariski-closed. □

Up to this point we haven't done anything profound. The only property of polynomial functions that we used in the verification that the Zariski topology is a topology is that the product of any two polynomial functions is a polynomial function, and the proof above depends only on the fact that a composition of polynomial functions is a polynomial function. The following result is much deeper.

Theorem 3.32. *For any $S \subset k[X_1, \ldots, X_n]$ there is a finite system of polynomials $f_1, \ldots, f_s \in k[X_1, \ldots, X_n]$ such that*

$$V(S) = V(f_1, \ldots, f_s).$$

There is a slightly different way of thinking about things, that would be more natural for an algebraic geometer, in which this is a result about the Zariski topology. As we defined the notion above, an affine algebraic set is the set of common zeros of any set of polynomials. However, if we had defined an affine algebraic set to be the set of common zeros of a *finite* system of polynomials, this result would then be interpreted as saying that the Zariski topology is, in fact, a topology, because arbitrary intersections of closed sets are closed.

One might visualize a category as a collection of electrodes (the objects) connected by wires (the morphisms). The Zariski topology gives a sense in which the morphisms of the category of affine algebraic sets and regular functions are continuous, so one gets some sense of how the Zariski topology provides a powerful tool for turning algebraic facts into useful geometric information flowing effortlessly through this vast network. But actually this image expresses only a small part of the influence the Zariski topology has had on algebraic geometry.

To a rather surprising extent mathematical research is guided by reasoning by analogy. Suppose we have two categories, one of which is well understood. If the second category exhibits phenomena that seem analogous to properties of the first, it is natural to attempt to develop theories in the second category that parallel successful existing theories in the first. Algebraic geometry studies objects that are very special, and have a great deal of structure. For this reason the categories studied in many other areas of mathematics have subcategories of algebraic objects, or have features that are, perhaps in a rough sense, mirrored in the categories studied in algebraic geometry. Historically, the development of the formal methods of algebraic geometry was largely a matter of importing ideas and methods from other subfields, and the Zariski topology was the heart and soul of this process.

The remainder of this section is devoted to the proof of Theorem 3.32. At first it seems remarkable that one could prove something like this. Given an arbitrary S, how could one sensibly search for suitable $f_1, \dots, f_r$? Instead of lunging head-on at the problem, let's first cultivate an appreciation of the relevant abstractions.

For any $S \subset k[X_1, \dots, X_n]$ we let $I(S)$ denote the ideal of $k[X_1, \dots, X_n]$ generated by the elements of S. That is, $I(S)$ is the smallest ideal that contains all the elements of S. Alternatively, $I(S)$ is the set of elements of $k[X_1, \dots, X_n]$ of the form

$$g_1 f_1 + \cdots + g_s f_s$$

where $f_1, \dots, f_s \in S$ and $g_1, \dots, g_s \in k[X_1, \dots, X_n]$. Concretely, any ideal containing S has to contain every $g_1 f_1 + \cdots + g_s f_s$, and the set of all such $g_1 f_1 + \cdots + g_s f_s$ is easily seen to be closed under addition and multiplication by elements of $k[X_1, \dots, X_n]$, hence an ideal.

Lemma 3.33. *For any $S \subset k[X_1, \dots, X_n]$, $V(S) = V(I(S))$.*

Proof. Of course $V(I(S)) \subset V(S)$ because S is a subset of $I(S)$. For the reverse inclusion observe that if $a \in V(S)$, then for any $f_1, \dots, f_s \in S$ and $g_1, \dots, g_s \in k[X_1, \dots, X_n]$ we have

$$g_1(a) f_1(a) + \cdots + g_s(a) f_s(a) = 0$$

because $f_1(a) = \cdots = f_s(a) = 0$, so $a \in V(I(S))$. □

Let R be any commutative ring with unit. Recall that in Chapter 2 we defined the principal ideal generated by $f \in R$ to be $(f) := \{\, rf : r \in R \,\}$.

We now extend this notation: if $f_1, \dots, f_s \in R$, let

$$(f_1, \dots, f_s) := \{ r_1 f_1 + \cdots + r_s f_s : r_1, \dots, r_s \in R \}^1.$$

It is easy to see that $(f_1, \dots, f_s)$ contains all sums and additive inverses of its elements, and it contains rg whenever if contains g and $r \in R$, so it is an ideal. An ideal $I \subset R$ is **finitely generated** if $I = (f_1, \dots, f_s)$ for some $f_1, \dots, f_s$, in which case we say that $f_1, \dots, f_s$ is a **system of generators** for I. That is, I is finitely generated if it is finitely generated as an R-module.

We started with the goal of showing that for any $S \subset k[X_1, \dots, X_n]$ there are $f_1, \dots, f_s$ such that $V(S) = V(f_1, \dots, f_s)$, and the last result tells us that this is the same as $V(I(S)) = V((f_1, \dots, f_s))$, so we are done if we can show that $I(S) = (f_1, \dots, f_s)$ for some $f_1, \dots, f_s$. Therefore it is enough to show *every* ideal of $k[X_1, \dots, X_n]$ is finitely generated, which at first sounds a bit too good to be true. But S could be any ideal of $k[X_1, \dots, X_n]$, so we actually can't avoid proving something at least this strong.

In general it is *always* the case that the only way to prove something is to prove something else that is at least as powerful, in the sense of having the desired proposition as an implication. We have proceeded through a sequence of propositions, each of which implies its predecessor, which is what mathematicians do when trying to prove something. Mathematical talent is largely a matter of good sense, incisive insight, and inspiration, concerning which of the propositions, among those implying some desired conclusion, is likely to be both true and provable. As laid out above, the process might have seemed rather straightforward, but we have now arrived at a pivotal concept in abstract algebra whose importance was initially far from obvious.

A commutative ring R is said to be **Noetherian** if each of its ideals is finitely generated. This terminology honors Emmy Noether (1882-1935) who (in addition to other very important contributions, including a profound theorem of physics) transformed the field of mathematics known as commutative algebra by using the Noetherian ring concept to recast the foundations of certain key results, simultaneously extending their scope and simplifying their proofs. Reexpressed in this terminology, our target is:

Theorem 3.34. *$k[X_1, \dots, X_n]$ is Noetherian.*

[1] This notation burdens the reader with the task of determining, from context, whether $(f_1, \dots, f_s)$ is an ideal or an element of R^s. In practice there is no real difficulty, so this is a small price to pay if we can avoid notational monstrosities like $I(\{f_1, \dots, f_s\})$.

There are several ways of thinking about Noetherian rings:

Lemma 3.35. *For a commutative ring with unit R the following are equivalent:*

(a) R is Noetherian;

(b) every increasing sequence of ideals $I_1 \subset I_2 \subset \dots$ "stabilizes": there is some K such that $I_k = I_K$ for all $k \geq K$;

(c) any collection of ideals of R has an element that is maximal in the sense of not being a subset of some other element of the collection.

Proof. Suppose that (a) holds, and let $I_1 \subset I_2 \subset \dots$ be an increasing sequence of ideals. Then (as in the proof of Proposition 2.23) $\bigcup_{i\geq 1} I_i$ is an ideal, so it must have a finite set of generators, and there must be an integer K such that all the generators are contained in I_K. Thus (a) implies (b).

Suppose (b) holds. If there was collection of ideals without a maximal element we could create an increasing sequence that did not stabilize by letting I_1 be an arbitrary element of the collection, letting I_2 be an element of the collection that was a proper superset of I_1, letting I_3 be an element of the collection that was a proper superset of I_2, and so forth. Thus (b) implies (c).

Suppose (c) holds, and let I be an ideal. The collection of finitely generated ideals contained in I has a maximal element, and the maximal element cannot be a proper subset of I because we could obtain a contradiction of maximality by appending another generator, so the maximal element must be I. Thus (c) implies (a). □

We now come to another major contribution of Hilbert.

Theorem 3.36 (Hilbert Basis Theorem)**.** *If R is a Noetherian ring, then so is $R[X]$.*

The only ideals of the field k are $\{0\}$ and k itself, so k is Noetherian. Applying the Hilbert Basis Theorem inductively shows that

$$k[X_1, \dots, X_n] = k[X_1, \dots, X_{n-1}][X_n]$$

is Noetherian for all n. Therefore the Hilbert Basis Theorem implies Theorem 3.34.

Let's think about what a proof of the Hilbert Basis Theorem might look like. An ideal I will be given to us, and we will need to find $f_1, \dots, f_s \in I$

such that the ideal $I' = (f_1, \dots, f_s)$ is equal to I. This means that for any $f \in I$ we will need to be able to find $g_1, \dots, g_s \in R[X]$ such that

$$f = g_1 f_1 + \cdots + g_s f_s.$$

Breaking things down a bit more, it suffices to be able to find $g_1, \dots, g_s$ such that the degree of $f - g_1 f_1 - \cdots - g_s f_s$ is less than the degree of f, since we can repeat this procedure until we get $f - g_1 f_1 - \cdots - g_s f_s = 0$. We can now see what kind of polynomials we need in I'. First of all, it should be the case that for every polynomial $f \in I$ there should be a polynomial $f' \in I'$ that has the same leading coefficient as f. But we need a bit more than this, insofar as our reduction technique requires that the degree of f' be no greater than the degree of f.

Proof of the Hilbert Basis Theorem. Let I be an ideal of $R[X]$. For each $d = 0, 1, 2, \dots$ let J_d be the union of $\{0\}$ and the set of leading coefficients of elements of I of degree d. To see that J_d is an ideal of R observe that:

(i) if f_1 and f_2 are polynomials in I of degree d, then the sum of the leading coefficients of f_1 and f_2 is either 0 or the leading coefficient of $f_1 + f_2$;

(ii) if the degree of $f \in I$ is d and $r \in R$, then r times the leading coefficient of f is either 0 (this might happen even when $r \neq 0$, because R might have zero divisors) or the leading coefficient of rf.

If $d < d'$, then $J_d \subset J_{d'}$ because if $f \in I$ has degree d, then its leading coefficient is the leading coefficient of $X^{d'-d} f$. Therefore $J_0 \subset J_1 \subset J_2 \subset \dots$, and since R is Noetherian, there is a $\overline{d}$ such that $J_d = J_{\overline{d}}$ for all $d \geq \overline{d}$. In addition, for each $0 \leq d \leq \overline{d}$ there are finitely many polynomials $f_{1,d}, \dots, f_{s_d,d} \in I$ of degree d whose leading coefficients generate J_d. For each $d \leq \overline{d}$ let $I_d := (f_{1,d}, \dots, f_{s_d,d})$, and let

$$I' = (f_{1,0}, \dots, f_{s_0,0}, \dots, f_{1,\overline{d}}, \dots, f_{s_{\overline{d}},\overline{d}}).$$

Of course $I' \subset I$, and we claim that $I' = I$. Aiming at a contradiction, let f be an element of $I \setminus I'$ of lowest degree, say d, and let $c \in J_d$ be the leading coefficient of f. If $d \leq \overline{d}$ and $c_1, \dots, c_{s_d}$ are the leading coefficients of $f_{1,d}, \dots, f_{s_d,d}$, then $c = a_1 c_1 + \cdots + a_{s_d} c_{s_d}$ for some $a_1, \dots, a_{s_d} \in R$, so $f - (a_1 f_1 + \cdots + a_{s_d} f_{s_d,d})$ is an element of $I \setminus I'$ of lower degree than f. If $d > \overline{d}$ and $c_1, \dots, c_{s_{\overline{d}}}$ are the leading coefficients of $f_{1,\overline{d}}, \dots, f_{s_{\overline{d}},\overline{d}}$, then $c = a_1 c_1 + \cdots + a_{s_{\overline{d}}} c_{s_{\overline{d}}}$ for some $a_1, \dots, a_{s_{\overline{d}}} \in R$ because $J_d = J_{\overline{d}}$, and again $f - (a_1 f_1 + \cdots + a_{s_{\overline{d}}} f_{s_{\overline{d}},\overline{d}}) X^{d-\overline{d}}$ is an element of $I \setminus I'$ of lower degree than f. In either case we have contradicted the choice of f. □

3.5 Compact Sets

The next definition is just weird. A topological space X is **compact** if, whenever $\{U_\alpha\}_{\alpha \in A}$ is an open cover of X, there is a finite subcollection $U_{\alpha_1}, \ldots, U_{\alpha_k}$ such that $U_{\alpha_1} \cup \ldots \cup U_{\alpha_k} = X$. The mantra of compactness is "every open cover has a finite subcover." Say this over and over until it sinks in. Compactness was unknown in the 19th century, but is now a fundamental idea that plays a role in a large percentage of proofs in most subfields of mathematics[2]. Throughout this book we discuss many topics that are then dropped, never to appear again, but compactness won't be like that. The results in this section and the next will be applied many times, and you'll probably need to review the material here more than once.

Our explanation begins with a simple observation that allows us to expand the usage of the term 'compact.' We will say that a subset $K \subset X$ is **compact** if the relative topology it inherits from X makes it a compact space. Extending another piece of terminology a little bit, we will say that a collection $\{U_\alpha\}_{\alpha \in A}$ of open subsets of X is an **open cover of** K if $K \subset \bigcup_{\alpha \in A} U_\alpha$. Then K is compact if and only if every open cover has a finite subcover. The proof that this is so has two parts.

First suppose that K is compact, and let $\{U_\alpha\}_{\alpha \in A}$ be an open cover of K. Then each $U_\alpha \cap K$ is open in the relative topology of K, so $\{U_\alpha \cap K\}_{\alpha \in A}$ is a cover of K by relatively open subsets, and consequently there are $\alpha_1, \ldots, \alpha_k$ such that

$$K \subset (U_{\alpha_1} \cap K) \cup \ldots \cup (U_{\alpha_k} \cap K) \subset U_{\alpha_1} \cup \ldots \cup U_{\alpha_k}.$$

Now suppose that every open cover of K has a finite subcover. If $\{V_\alpha\}_{\alpha \in A}$ is a collection of relatively open subsets of K that cover K, then for each α we can choose an open (in X) U_α such that $V_\alpha = U_\alpha \cap K$. There are $\alpha_1, \ldots, \alpha_k$ such that $K \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_k}$, and clearly this implies that $K \subset V_{\alpha_1} \cup \cdots \cup V_{\alpha_k}$. Thus K with its relative topology is a compact space.

What kinds of sets are compact? Of course any finite set is compact, but the concept wouldn't be worth anything if that was all there was to it. We will focus on compact spaces that are subsets (with their relative topologies) of $\mathbb{R}$ and $\mathbb{R}^n$. Here is the key example.

[2] Ironically, compactness is not a useful concept in algebraic geometry because *every* subset of $\mathbf{A}^n(k)$ is Zariski-compact! Specifically, suppose that $B \subset \bigcup_{\alpha \in A} U_\alpha \subset \mathbf{A}^n(k)$ where for each α there is $S_\alpha \subset k[X_1, \ldots, X_n]$ such that $U_\alpha = \mathbf{A}^n(k) \setminus V(S_\alpha)$. The Hilbert basis theorem implies that the ideal generated by $\bigcup S_\alpha$ has a finite system of generators, and each generator is of the form $g_1 f_1 + \cdots + g_m f_m$ where each f_i is an element of some S_α. Consequently there are $\alpha_1, \ldots, \alpha_k$ such that $S_{\alpha_1} \cup \cdots \cup S_{\alpha_k}$ generates the same ideal as $\bigcup_\alpha S_\alpha$, so $\bigcap V(S_\alpha) = V(S_{\alpha_1}) \cap \cdots \cap V(S_{\alpha_k})$ and $B \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_k}$.

Lemma 3.37. *For any numbers $a, b \in \mathbb{R}$ with $a \leq b$ the interval $[a, b]$ is compact.*

Proof. Let $\{U_\alpha\}_{\alpha \in A}$ be an open cover of $[a, b]$, and let S be the set of numbers $s \in [a, b]$ such that $[a, s]$ has a finite subcover. Then S is nonempty because it contains a, and it is bounded above by b, so it has a least upper bound $\overline{s}$, and there is an index β such that $\overline{s} \in U_\beta$. Since U_β contains an open interval around $\overline{s}$, $\overline{s} = a$ is impossible, and there is an $s < \overline{s}$ with $[s, \overline{s}] \subset U_\beta$. Combining U_β with a finite subcover of $[a, s]$ gives a finite subcover of $[a, \overline{s}]$, so $\overline{s} \in S$, and if $\overline{s}$ is less than b, then this is a finite subcover of $[a, s']$ for some $s' > \overline{s}$, contradicting the definition of $\overline{s}$, so $\overline{s} = b$. Thus $b \in S$. □

The next result now gives a rich supply of compact subsets of $\mathbb{R}$.

Theorem 3.38. *If K is a compact subset of a topological space X and $C \subset K$ is closed in the relative topology of K (this is necessarily the case, by Lemma 3.23, when C is closed in X) then C is compact.*

Proof. Let $\{U_\alpha\}_{\alpha \in A}$ be a collection of open subsets of X that covers C. Since $K \setminus C$ is relatively open, there is an open $U \subset X$ such that $K \setminus C = K \cap U$. Then $\{U\} \cup \{U_\alpha\}_{\alpha \in A}$ is a collection of open subsets of X that covers K, and K must have a finite subcover. If U is in this subcover we can throw it away, thereby obtaining a finite subset of $\{U_\alpha\}_{\alpha \in A}$ that covers C. □

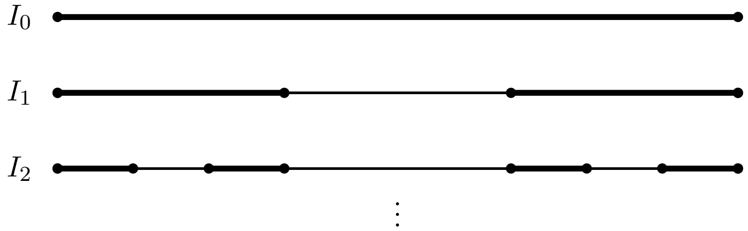


Figure 3.5

There is an example that illustrates the diversity of compact subsets of $\mathbb{R}$ rather vividly, and is also quite famous for many reasons. Start by setting $I_0 := [0, 1]$. Form

$$I_1 := [0, \tfrac{1}{3}] \cup [\tfrac{2}{3}, 1] = I_0 \setminus (\tfrac{1}{3}, \tfrac{2}{3})$$

by removing the open "middle third" of the interval. Then I_1 consists of two closed intervals, and we form

$$I_2 := [0, \tfrac{1}{9}] \cup [\tfrac{2}{9}, \tfrac{1}{3}] \cup [\tfrac{2}{3}, \tfrac{7}{9}] \cup [\tfrac{8}{9}, 1] = I_1 \setminus \Big((\tfrac{1}{9}, \tfrac{2}{9}) \cup (\tfrac{7}{9}, \tfrac{8}{9}) \Big)$$

by removing each of their middle thirds. Continue in this manner: if $I_n = [a_1, b_1] \cup \ldots \cup [a_{2^n}, b_{2^n}]$ where

$$a_1 < b_1 < a_2 < b_2 < \cdots < a_{2^n} < b_{2^n},$$

form I_{n+1} by removing the middle third of each of the intervals $[a_i, b_i]$:

$$I_{n+1} := \bigcup_{i=1}^{2^n} [a_i, \tfrac{2}{3}a_i + \tfrac{1}{3}b_i] \cup [\tfrac{1}{3}a_i + \tfrac{2}{3}b_i, b_i] = I_n \setminus \bigcup_{i=1}^{2^n} (\tfrac{2}{3}a_i + \tfrac{1}{3}b_i, \tfrac{1}{3}a_i + \tfrac{2}{3}b_i).$$

Finally, the **Cantor Set** is the infinite intersection

$$C := I_0 \cap I_1 \cap I_2 \cap \ldots.$$

Each I_n is a closed set, so C is closed because it is the intersection of a collection of closed sets. Since it is a subset of $[0, 1]$, it is compact. Each endpoint of an interval in I_n is also an endpoint of an interval in I_{n+1}, so each such endpoint is an element of C, and in particular C is nonempty. Does C contain any other points? Before answering this question we will show that C is, in a certain sense, quite small.

There is a standard method of measuring the one dimensional "volume" of sets like C that is beyond the scope of this book, but without going into any great detail we can see that any reasonable method of measuring volume cannot ascribe a positive volume to C. By "reasonable" we mean that: a) the volume of a disjoint union of finitely many intervals is the sum of their lengths, and b) if $A \subset B$, and the theory attributes a volume to both of these sets, then the volume of A cannot be greater than the volume of B. By induction, the volume of I_n is $(2/3)^n$, so if a reasonable theory assigns a volume to C, that volume must be zero.

In spite of C being small in this sense, it turns out that the cardinality of C is the same as the cardinality of $[0, 1]$! To show this we use the base 2 and base 3 decimal expansions of numbers in $[0, 1]$. The key point is that every sum

$$s = \tfrac{1}{3}b_1 + \tfrac{1}{9}b_2 + \cdots + \tfrac{1}{3^n}b_n$$

with $b_1, \ldots, b_n \in \{0, 2\}$ is a lower endpoint of one of the intervals constituting I_n. It is easy to see this by arguing inductively: 0 is the lower endpoint of I_0, and if s is the lower endpoint of the interval $[s, s + 1/3^n]$, then s and $s + 2/3^{n+1}$ are the lower endpoints of the two intervals obtained by removing the middle third.

Now consider the map

$$f : \tfrac{1}{2}a_1 + \tfrac{1}{4}a_2 + \tfrac{1}{8}a_3 + \cdots \ \mapsto \ \tfrac{2}{3}a_1 + \tfrac{2}{9}a_2 + \tfrac{2}{27}a_3 + \cdots$$

taking a point in $[0,1]$, written in base 2 as $0.a_1a_2a_3\ldots$, to a point written in base 3 using only 0's and 2's. There is a bit of ambiguity here insofar as, for example, in base 2 we have $0.01011111\ldots = 0.01100000\ldots$, but this isn't a problem: when two base 2 expansions are possible we always take the one with a tail consisting of 1's. If two base 3 decimal expansions represent the same number, then one ends with a tail of 2's and the other ends with a tail of 0's, and the latter is not an image of this map, so this map is injective. Of course the sequence of partial sums $2a_1/3 + \cdots + 2a_n/3^n$ converges because it's Cauchy, and its limit is in C because C is closed.

At this point we have defined an injective map $f : [0,1] \to C$. Let $g : C \to [0,1]$ be the inclusion. (Whenever $A \subset X$, the function $i : A \to X$ taking each $a \in A$ to itself is called the **inclusion**.) Of course g is injective, so it sure seems like C and $[0,1]$ ought to have the same cardinality. Nevertheless, to prove this we still need to actually produce a suitable bijection. Depending on your mood, this is either an annoying picayune detail or an opportunity to learn an important result from set theory with a cute proof.

Theorem 3.39 (Schroeder-Bernstein Theorem)**.** *If $f : X \to Y$ and $g : Y \to X$ are injections, then there is a bijection $h : X \to Y$.*

Proof. Let $A_0 := X \setminus g(Y)$, define $A_1, A_2, \ldots$ inductively by setting $A_{j+1} := g(f(A_j))$, and let $B := \bigcup_{j=0}^{\infty} A_j$. Then

$$g(Y \setminus f(B)) = g(Y) \setminus g(f(B))$$

because g is injective, so

$$g(Y \setminus f(B)) = (X \setminus A_0) \setminus g(f(A_0 \cup A_1 \cup \ldots)) = (X \setminus A_0) \setminus \bigcup_{j=0}^{\infty} g(f(A_j))$$

$$= X \setminus (A_0 \cup A_1 \cup A_2 \cup \ldots) = X \setminus B.$$

Therefore g restricts to a bijection between $Y \setminus f(B)$ and $X \setminus B$, and of course f restricts to a bijection between B and $f(B)$, so we can define h by setting $h(a) := f(a)$ if $a \in B$ and $h(a) := g^{-1}(a)$ if $a \in X \setminus B$. □

Cantor proved a version of this result that is weaker, insofar as his "proof" used the axiom of choice. In recognition of this it has recently become fashionable to call this result the Cantor-Schroeder-Bernstein theorem. I don't know whether the analogous result with 'injection' replaced with 'surjection' can be proved without the axiom of choice, but if we have the axiom of choice, then it follows easily. Suppose $F : X \to Y$ and $G : Y \to X$ are surjections. Then each $F^{-1}(y)$ is nonempty, so the axiom of choice gives us a function $g : Y \to X$ with $g(y) \in F^{-1}(y)$ for all y. Of course g is injective: if $y \neq y'$, then $g(y) \neq g(y')$ because $F(g(y)) \neq F(g(y'))$. Similarly, there is an injection $f : X \to Y$ with $f(x) \in G^{-1}(x)$ for all x, and we can apply the result above to f and g.

We can now give a precise argument showing that C contains points other than the endpoints in the intervals in the sets I_n. Let S be the set of such endpoints. The set of rational numbers is countable. (The set of nonzero rationals is a union of countably many disjoint countable sets because it is the union over all $j = 1, 2, \ldots$ of the set of nonzero rationals that have denominator j when reduced to lowest terms.) Every element of S is rational, so there is an injection from S to the set of natural numbers, and we can obviously construct an injection from the set of natural numbers to S, so the Cantor-Schroeder-Bernstein theorem implies that S is countable. Since C has the same cardinality as $[0, 1]$, and is consequently uncountable, the inclusion mapping S into C cannot be a bijection.

Theorems are basically assertions that certain things can't happen. The other side of the coin are examples showing the certain things are possible, e.g., a subset of $\mathbb{R}$ with the cardinality of the continuum can have zero volume. In fact a large number of examples in analysis begin with the Cantor set or some variant, to the point where a good rule is that if you are looking for an example of some seemingly bizarre phenomenon, first think about whether you can use the Cantor set to construct one.

3.6 More on Compactness

In the last section we saw that, in order to be a compact set, it suffices to be a closed set that is "small," in the somewhat circular sense of being a relatively closed subset of some other compact set. Are all compact subsets of X closed? In general no, but the majority of spaces that we care about are Hausdorff, so the answer is yes "for practical purposes:"

Theorem 3.40. *If X is a Hausdorff space and K is a compact subset of X, then K is closed.*

Proof. We'll show that $X \setminus K$ is open because an arbitrary $x \in X \setminus K$ has a neighborhood U that doesn't intersect K. Since X is Hausdorff, for each $y \in K$ there are disjoint open neighborhoods U_y and V_y of x and y. Since K is compact, there are $y_1, \ldots, y_k$ such that $K \subset V_{y_1} \cup \cdots \cup V_{y_k}$. Set

$U := U_{y_1} \cap \cdots \cap U_{y_k}$, and observe that $x \in U$, U is open, and

$$U \cap K \subset \Big(\bigcap_{i=1}^{k} U_{y_i}\Big) \cap \Big(\bigcup_{i=1}^{k} V_{y_i}\Big) = \bigcup_{i=1}^{k} \Big((\bigcap_{i=1}^{k} U_{y_i}) \cap V_{y_i}\Big) \subset \bigcup_{i=1}^{k} (U_{y_i} \cap V_{y_i}) = \emptyset.$$

□

In particular, any compact subset of $\mathbb{R}^n$ must be closed. The sets $\mathbf{U}_1(0), \mathbf{U}_2(0), \ldots$ cover $\mathbb{R}^n$, so any compact subset of $\mathbb{R}^n$ must have a finite subcover, which means that it is contained in $\mathbf{U}_r(0)$ for sufficiently large $r > 0$. A set with this property is said to **bounded**. Our next goal is to show that "closed and bounded" is an exact characterization of the compact subsets of $\mathbb{R}^n$: not only is every compact set necessarily closed and bounded, as we have shown, but also every closed and bounded set is compact.

Any closed bounded $K \subset \mathbb{R}^n$ is contained in a rectangle

$$[a_1, b_1] \times \cdots \times [a_n, b_n],$$

and if we can show that this rectangle is compact, then (by Theorem 3.38) any closed subset such as K is compact. The proof that the rectangle is compact is best undertaken in full generality, by showing that any cartesian product of compact sets is compact. Before we can say what we mean by this, we need a topology on the cartesian product $X \times Y$ of two topological spaces X and Y, and the most natural such topology on is the **product topology**, which is defined by specifying that $W \subset X \times Y$ is open if, for each $(x, y) \in W$, there are open sets $U \subset X$ and $V \subset Y$ such that

$$(x, y) \subset U \times V \subset W.$$

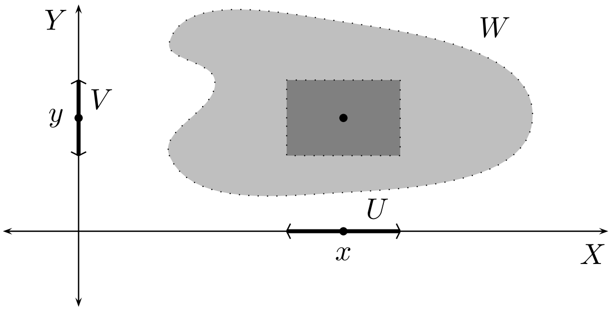


Figure 3.6

We need to check that this system of sets actually is a topology. It is easy to see that it contains $\emptyset$, $X \times Y$, and any union of its elements. To see that $W_1 \cap W_2$ is open whenever W_1 and W_2 are open subsets of $X \times Y$, consider a particular point $(x, y) \in W_1 \cap W_2$. There are open sets $U_1, U_2 \subset X$ and $V_1, V_2 \subset Y$ such that $(x, y) \in U_1 \times V_1 \subset W_1$ and $(x, y) \in U_2 \times V_2 \subset W_2$. Then

$$(x, y) \in (U_1 \cap U_2) \times (V_1 \cap V_2) = (U_1 \times V_1) \cap (U_2 \times V_2) \subset W_1 \cap W_2.$$

It is straightforward to extend our definition to a finite cartesian product

$$X_1 \times \cdots \times X_n.$$

One may do this directly by specifying that $W \subset X_1 \times \cdots \times X_n$ is open if, for each $(x_1, \ldots, x_n) \in W$, there are open sets $U_1 \subset X_1, \ldots, U_n \subset X_n$ such that

$$(x_1, \ldots, x_n) \in U_1 \times \cdots \times U_n \subset W.$$

Alternatively, one may proceed inductively, endowing $X_1 \times \cdots \times X_n$ with the product topology of the cartesian product of $X_1 \times \cdots \times X_{n-1}$ and X_n.

Here is a fact that comes up frequently, and is often treated as too obvious to mention explicitly.

Lemma 3.41. *If $f_1 : X_1 \to Y_1, \ldots, f_n : X_n \to Y_n$ are continuous functions, then the function $f : X_1 \times \cdots X_n \to Y_1 \times \cdots \times Y_n$ given by*

$$f(x_1, \ldots, x_n) := (f_1(x_1), \ldots, f_n(x_n))$$

is continuous.

Proof. Every open set in $Y_1 \times \cdots \times Y_n$ is a union of products $V_1 \times \cdots \times V_n$ where $V_1 \subset Y_1, \ldots, V_n \subset Y_n$ are open, and

$$f^{-1}(V_1 \times \cdots \times V_n) = f_1^{-1}(V_1) \times \cdots \times f_n^{-1}(V_n)$$

is open in $X_1 \times \cdots \times X_n$. □

It is visually obvious that a subset of $\mathbb{R}^2$ is open in the product topology if and only if it contains a ball around each of its points, but we will belabor the point a bit, in part to prepare for Chapter 6 where the general idea will be important. A set $W \subset \mathbb{R}^2$ is open in the product topology if, for each $(x, y) \in W$, we can find numbers a, b, c, d such that

$$(x, y) \in (a, b) \times (c, d) \subset W.$$

(Unfortunately the notation here is potentially confusing: (x, y) is an element of $\mathbb{R}^2$, but (a, b) and (c, d) are open intervals.) If this is the case, then W is open in the topology induced by the norm $\|\cdot\|_\infty$ because in this circumstance we have $(x, y) + \delta B_\infty \subset W$ where

$$\delta := \min\{x - a, b - x, y - c, d - y\}$$

and $B_\infty = \{\, (x, y) \in \mathbb{R}^2 : \|(x, y)\|_\infty < 1\,\}$. Conversely, if $(x, y) + \delta B_\infty \subset W$, then

$$(x, y) \in (x - \delta, x + \delta) \times (y - \delta, y + \delta) \subset W,$$

so W is open in the product topology whenever it is open in the topology induced by $\|\cdot\|_\infty$. Thus the product topology on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the topology induced by $\|\cdot\|_\infty$. We pointed out earlier that $\|\cdot\|_\infty$, $\|\cdot\|_1$, and $\|\cdot\|_2$ all induce the same topology, so the product topology on $\mathbb{R}^2$ is the "usual" topology on this space. Nothing here depends on there being only two dimensions: for any finite n the product topology on $\mathbb{R}^n$ is induced by $\|\cdot\|_\infty$, so it coincides with the "usual" topology induced by $\|\cdot\|_2$.

In general it is quite uncommon to endow a finite cartesian product with any topology other than the product topology. The Zariski topology is, perhaps, the most prominent example. Any Zariski-closed subset of $\mathbf{A}^1$ is the set of roots of a finite collection of polynomials. Except for $\mathbf{A}^1$ itself (which is the set of roots of the zero polynomial) any such set is finite. Each $c \in \mathbf{A}^1$ is a root of $X - c$, and finite unions of closed sets are closed, so any finite set is closed. Thus the closed subsets of $\mathbf{A}^1$ are $\emptyset$, $\mathbf{A}^1$ itself, and the finite subsets. The complement of a finite set is said to be **cofinite**, and the open subsets of $\mathbf{A}^1$ are $\emptyset$, $\mathbf{A}^1$ itself, and the cofinite sets.

We claim that the general form of an open set in the product topology of $\mathbf{A}^1 \times \mathbf{A}^1$ is $(C_1 \times C_2) \setminus F$ where $C_1, C_2 \subset \mathbf{A}^1$ are cofinite and F is finite. The verification of this is a bit tedious, both to write out and to read, and thinking through the details would be a good way to review and solidify your understanding, so we leave it as an exercise. The main point is that an affine algebraic set such as $V(X_2 - X_1^2)$ is closed in the Zariski topology of $\mathbf{A}^2$, by virtue of the definition of that topology, but not in the product topology of $\mathbf{A}^1 \times \mathbf{A}^1$ unless the given field k is finite.

The product topology is important for many reasons, but our motivation for introducing it here is:

Theorem 3.42. *If X and Y are topological spaces and $K \subset X$ and $L \subset Y$ are compact, then $K \times L$ is a compact subset of $X \times Y$ when this space is endowed with the product topology.*

Proof. Suppose $\{W_\alpha\}_{\alpha \in A}$ is an open cover of $K \times L$. For each $(x, y) \in K \times L$ choose $\alpha(x, y) \in A$ such that $(x, y) \in W_{\alpha(x,y)}$, and choose open neighborhoods $U_{(x,y)}$ and $V_{(x,y)}$ of x and y such that

$$U_{(x,y)} \times V_{(x,y)} \subset W_{\alpha(x,y)}.$$

For each $y \in L$, $\{\, U_{(x,y)} : x \in K \,\}$ is an open cover of K, so it has a finite subcover, say $\{\, U_{(x,y)} : x \in F_y \,\}$ where $F_y \subset K$ is finite. If we set $V_y := \bigcap_{x \in F_y} V_{(x,y)}$, then

$$K \times \{y\} \subset K \times V_y \subset \bigcup_{x \in F_y} U_{(x,y)} \times V_{(x,y)},$$

and the open cover $\{\, V_y : y \in L \,\}$ of L has a finite subcover, say $V_{y_1}, \ldots, V_{y_\ell}$. Obviously $\{\, W_{\alpha(x,y_i)} : i = 1, \ldots, \ell,\ x \in F_{y_i} \,\}$ is a finite cover of $K \times L$. □

We have assembled all of the required tools. The proof below simply recaps the relevant parts of the discussion to this point.

Theorem 3.43 (Heine-Borel). *A set $K \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof. We have already shown that K is closed and bounded if it is compact. Suppose that K is closed and bounded. Then K is contained in some rectangle $[a_1, b_1] \times \cdots \times [a_n, b_n]$. Lemma 3.37 implies that each $[a_i, b_i]$ is compact, so $[a_1, b_1] \times \cdots \times [a_n, b_n]$ is compact by Theorem 3.42, and Theorem 3.38 then implies that K is compact. □

Up to this point our work has focused on passing from the abstract definition of compactness to a concrete description of which subsets of $\mathbb{R}^n$ are compact. But we still don't have much insight into why compact sets are useful. There are diverse reasons, some of which are matters of certain types of objects existing.

Theorem 3.44. *If $x_1, x_2, \ldots$ is a sequence in a compact metric space (X, d), then there is a convergent subsequence $x_{i_1}, x_{i_2}, \ldots$.*

Proof. It cannot be the case that every $y \in X$ is contained in an open set U_y that contains x_i for only finitely many i because there would be a finite subcover $U_{y_1}, \ldots, U_{y_k}$, from which we would arrive at the absurd conclusion that $x_i \in X$ for only finitely many i. Therefore there is an x such that for every $\varepsilon > 0$ the ball $\mathbf{U}_\varepsilon(x)$ contains x_i for infinitely many i. Choose i_1 such that $x_{i_1} \in \mathbf{U}_1(x)$, choose $i_2 > i_1$ such that $x_{i_2} \in \mathbf{U}_{1/2}(x)$, choose $i_3 > i_2$ such that $x_{i_3} \in \mathbf{U}_{1/3}(x)$, and so forth. □

The following special case of Theorem 3.44 is famous (and you should know the name) presumably because it was discovered in the 19th century before metric spaces had been invented.

Theorem 3.45 (Bolzano-Weierstrass)**.** *If $B \subset \mathbb{R}^n$ is bounded (so its closure is compact) then every sequence in B has a convergent subsequence.*

We have already introduced Cauchy sequences of real numbers, but this is really a metric space concept: a sequence $x_1, x_2, \ldots$ in a metric space (X, d) is a **Cauchy sequence** if, for every $\delta > 0$, there is a natural number N such that $d(x_m, x_n) < \delta$ for all $m, n > N$. The space (X, d) is **complete** if each of its Cauchy sequences is convergent. Any limit of a subsequence of a Cauchy sequence is a limit of the sequence itself, so Theorem 3.44 implies that

Proposition 3.46. *A compact metric space (X, d) is complete.*

For an example of how this result is applied in practice, consider a continuous function $f : K \to \mathbb{R}$ where $K \subset \mathbb{R}^n$ is compact. We will prove that the image $f(K)$ of f is bounded above. Aiming at a contradiction, consider the alternative, which is that for each natural number n there is some $x_n \in K$ with $f(x_n) > n$. Then the sequence $x_1, x_2, \ldots$ has a subsequence $x_{i_1}, x_{i_2}, \ldots$ that converges to a point $x \in K$. For any $\varepsilon > 0$ there is $\delta > 0$ such that

$$f(x) - \varepsilon < f(x') < f(x) + \varepsilon$$

for all $x' \in \mathbf{U}_\delta(x)$, and this gives the desired contradiction because $x_{i_j} \in \mathbf{U}_\delta(x)$ and $i_j > f(x) + \varepsilon$ for sufficiently large j.

As it happens, another very useful property of compact sets gives an approach to the boundedness of f that is superior, with greater generality and simpler proofs.

Theorem 3.47. *If X and Y are topological spaces, $f : X \to Y$ is continuous, and $K \subset X$ compact, then $f(K)$ is compact.*

Proof. Let $\{V_\alpha\}_{\alpha \in A}$ be an open cover of $f(K)$. Then $\{f^{-1}(V_\alpha)\}_{\alpha \in A}$ is an open (by continuity) cover of K, so there are $\alpha_1, \ldots, \alpha_k$ such that

$$K \subset f^{-1}(V_{\alpha_1}) \cup \ldots \cup f^{-1}(V_{\alpha_k}) = f^{-1}(V_{\alpha_1} \cup \ldots \cup V_{\alpha_k}),$$

whence $f(K) \subset V_{\alpha_1} \cup \ldots \cup V_{\alpha_k}$. □

Since the preimages $f^{-1}(V)$ and $f^{-1}(D)$ of open sets $V \subset Y$ and closed sets $D \subset Y$ are (by virtue of the definition of continuity) open and closed respectively, this result is the reverse of what we might naively expect. Here is some related terminology. An **open mapping** is a continuous function $f : X \to Y$ such that $f(U)$ is open whenever $U \subset X$ is open, and a **closed mapping** is a continuous function $f : X \to Y$ such that $f(C)$ is closed whenever $C \subset X$ is closed. For reasons that are rooted neither in consistency nor any particular logic I am aware of, if $f : X \to Y$ is continuous and $f^{-1}(L)$ is compact whenever $L \subset Y$ is compact, then f is said to be **proper**.

Because the case of $Y = \mathbb{R}$ in the last result is so important, we collect what we have learned about it in the following result.

Theorem 3.48. *If X is a compact topological space and $f : X \to \mathbb{R}$ is continuous, then $f(X)$ is compact. In particular, $f(X)$ is bounded, and (provided $X \neq \emptyset$) there are points $\underline{x}, \overline{x} \in X$ such that for all $x \in X$,*

$$f(\underline{x}) \le f(x) \le f(\overline{x}).$$

Proof. Since $f(X)$ is compact, it is bounded above, and if it is nonempty it has a least upper bound $\overline{t}$. In addition $f(X)$ is closed, so it contains $\overline{t}$, which means that $f(\overline{x}) = \overline{t}$ for some $\overline{x} \in X$. A similar argument gives $\underline{x}$. □

There is some important related terminology and notation. If $f : X \to \mathbb{R}$ is a function, where X may be any sort of nonempty set, the **supremum** of f, which is denoted by

$$\sup_{x \in X} f(x),$$

is the least upper bound of $\{ f(x) : x \in X \}$ if this set is bounded, or ∞ if it is unbounded. The **maximum** of f, denoted by

$$\max_{x \in X} f(x),$$

is another term for the supremum of f, but it is used only in contexts in which there is a guarantee that there is some $x^* \in X$ such that $f(x^*) = \sup_{x \in X} f(x)$. That is, writing "$M = \max_{x \in X} f(x)$" asserts both that M is the supremum of f and that there is some $x^* \in X$ such that $f(x^*) = M$. Less frequently you will see the symbol

$$\operatorname*{argmax}_{x \in X} f(x) := \{ x^* \in X : f(x^*) = \max_{x \in X} f(x) \}$$

used to denote the set of **maximizers** of f; except when there is an explicit mention of the possibility that this set might be empty, generally there is an implicit assertion that it is nonempty.

Symmetrically, the **infimum** of f, denoted by $\inf_{x \in X} f(x)$, is the greatest lower bound of the image of f if the image is bounded, and $-\infty$ otherwise. The **minimum** of f is the infimum in a context in which there is known to be at least one $x_* \in X$ such that $f(x_*) = \inf_{x \in X} f(x)$, and $\operatorname{argmin}_{x \in X} f(x)$ is the set of **minimizers**.

3.7 Sequences and Series of Functions

In one sense the rest of the chapter is devoted to the proof of the fundamental theorem of algebra, with the central portion of the argument explained in the next section. But this work will involve several topics of considerable interest—sequences and series of functions, the exponential and trigonometric functions, an important topological concept called connectedness—and these are hardly less important. We now start down this road, studying convergence of infinite series $a_0 + a_1 + a_2 + \cdots$ of complex numbers, and the senses in which a series of complex valued functions $f_0, f_1, f_2, \ldots$ might converge, aiming ultimately at an understanding of infinite series of functions $g_0 + g_1 + g_2 + \cdots$.

Before anything else we need to impose a metric on $\mathbb{C}$. The **modulus** or **absolute value** of a complex number $z = x + iy$ is

$$|z| := \sqrt{x^2 + y^2}.$$

That is, $|z|$ is just the Euclidean norm $\|(x, y)\|_2$ of z under the identification of $\mathbb{C}$ with $\mathbb{R}^2$, so there is an associated metric $d(z, w) := |z - w|$. Recalling our discussion of the Cauchy-Schwartz inequality, we see immediately that the triangle inequality

$$|w + z| \le |w| + |z|$$

holds for all $w, z \in \mathbb{C}$, with strict inequality unless $w = 0$ or $z = tw$ for some nonnegative real number t. We endow $\mathbb{C}$ with the topology derived from the metric d, and we endow $\mathbb{C}^n$ with the product topology, which is, of course, the topology derived from the identification of $\mathbb{C}^n$ with $\mathbb{R}^{2n}$.

The **complex conjugate** of z is

$$\overline{z} := x - iy.$$

Geometrically, complex conjugation amounts to reflection across the x axis. Complex conjugation commutes with addition (that is, if $w = u + iv$, then $\overline{w + z} = \overline{w} + \overline{z}$) obviously, and also with multiplication:

$$\overline{wz} = \overline{(u + iv)(x + iy)} = \overline{(ux - vy) + i(uy + vx)}$$

$$= (ux - vy) - i(uy + vx) = (u - iv)(x - iy) = \overline{w}\,\overline{z}.$$

Together with the fact that

$$|z| = \sqrt{x^2 + y^2} = \sqrt{(x + iy)(x - iy)} = \sqrt{z\overline{z}},$$

this gives a simple proof that taking the modulus commutes with multiplication:

$$|wz| = \sqrt{wz\overline{wz}} = \sqrt{w\overline{w}} \cdot \sqrt{z\overline{z}} = |w| \cdot |z|.$$

For any natural number n and $a_0, \ldots, a_n, z \in \mathbb{C}$ the inequality above and this equation give the inequality

$$|a_n z^n + \cdots + a_1 z + a_0| \leq |a_n z^n| + \cdots + |a_1 z| + |a_0|$$

$$= |a_n|\,|z|^n + \cdots + |a_1|\,|z| + |a_0|,$$

which will be applied many times later on.

A series $a_0 + a_1 + a_2 + \cdots$ of complex numbers is said to **converge absolutely** if

$$|a_0| + |a_1| + |a_2| + \cdots < \infty.$$

When this is the case the sequence of partial sums $a_1 + \cdots + a_m$ is Cauchy: if $m < n$ then

$$\big|(a_1 + \cdots a_m) - (a_1 + \cdots + a_n)\big| \leq |a_{m+1}| + |a_{m+2}| + \cdots,$$

and the right hand side goes to 0 as $m \to \infty$. The word 'absolute' is meant to convey the idea that the limit does not depend on the order of summation. To see what is meant by this suppose that $\phi : \{0, 1, 2, \ldots\} \to \{0, 1, 2, \ldots\}$ is a bijection; we claim that

$$\lim_{m\to\infty} a_0 + \cdots + a_m = \lim_{m\to\infty} a_{\phi(0)} + \cdots + a_{\phi(m)}.$$

Since ϕ is bijective, for any integer N, we have $\{0, \ldots, N\} \subset \{\phi(0), \ldots, \phi(m)\}$ when m is sufficiently large, in which case

$$\{1, \ldots, m\} \setminus \{\phi(0), \ldots, \phi(m)\} \quad \text{and} \quad \{\phi(0), \ldots, \phi(m)\} \setminus \{1, \ldots, m\}$$

are disjoint subsets of $\{N + 1, N + 2, \ldots\}$, and

$$\big|(a_0 + \cdots + a_m) - (a_{\phi(0)} + \cdots + a_{\phi(m)})\big| \leq |a_{N+1}| + |a_{N+2}| + \cdots.$$

Since the sequence of partial sums $a_0 + \cdots + a_m$ is Cauchy, so is the sequence $a_{\phi(0)} + \cdots + a_{\phi(m)}$, and the two sequences have the same limit.

If $a_0 + a_1 + a_2 + \cdots$ and $b_0 + b_1 + b_2 + \cdots$ are two absolutely convergent series, then

$$\sum_m \sum_n |a_m b_n| = \sum_m \sum_n |a_m|\,|b_n| = \big(\sum_m |a_m|\big)\big(\sum_n |b_n|\big) < \infty.$$

Therefore the double summation $\sum_m \sum_n a_m b_n$ is defined and (by virtue of an argument like the one above) independent of the order of summation. The following **double summation formula** obtained by reordering is often useful:

$$\big(\sum_{m=0}^{\infty} a_m\big)\big(\sum_{n=0}^{\infty} b_n\big) = \sum_{k=0}^{\infty}\sum_{i=0}^{k} a_i b_{k-i}.$$

We now discuss convergence of sequences of functions. Much of what we have to say makes sense in a fairly general setting, so to start off with let X and Y be topological spaces. The most obvious notion of convergence is:

Definition 3.49. *A sequence of functions $f_1, f_2, \dots$ from X to Y* ***converges pointwise*** *to $f : X \to Y$ if, for each $x \in X$,*

$$\lim_{\ell\to\infty} f_\ell(x) = f(x).$$

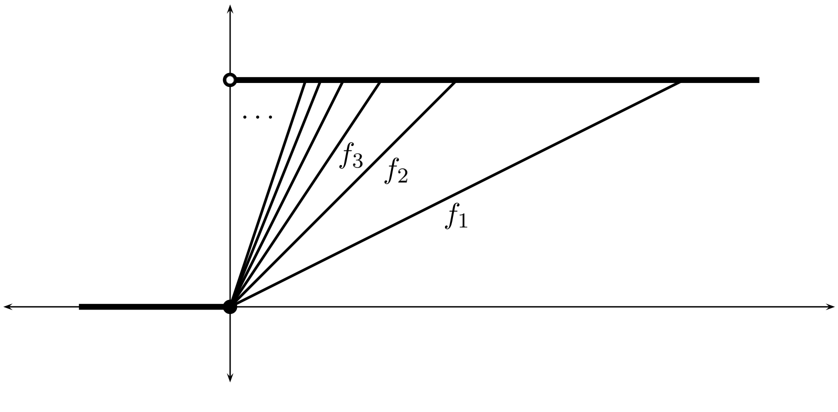


Figure 3.7

A simple example shows how a sequence of continuous functions can have a discontinuous pointwise limit. For $\ell = 1, 2, \dots$ let $f_\ell : \mathbb{R} \to \mathbb{R}$ be the function

$$f_\ell(t) := \begin{cases} 0, & t \le 0, \\ \ell t, & 0 \le t \le 1/\ell, \\ 1, & 1/\ell \le t. \end{cases}$$

Then the sequence $\{f_\ell\}$ converges pointwise to the function

$$f(t) := \begin{cases} 0, & t \le 0, \\ 1, & 0 < t. \end{cases}$$

Examples like this suggest that pointwise convergence is too permissive, and in fact it is not a very useful concept.

One possible reaction to this example is that although there is convergence at each point, the "true" distance between f_ℓ and f is always one. In response we would like to develop a notion in which we can say that f_ℓ is converging to f "simultaneously" at every point in X. There are ways to do this in a purely topological setting, but they're rather fancy. The basic idea can be expressed very clearly if we assume that (Y, d) is a metric space, which it will be in all our applications, so we will assume that this is the case from now on.

Definition 3.50. *A sequence of functions $f_1, f_2, \dots$ from X to Y* ***converges uniformly*** *to $f : X \to Y$ if, for each $\varepsilon > 0$, there is some L such that for all $\ell > L$,*

$$\sup_{x \in X} d(f_\ell(x), f(x)) < \varepsilon.$$

This convergence notion has an extremely pleasant and useful property.

Proposition 3.51. *If $\{f_\ell\}$ converges uniformly to f and each f_ℓ is continuous, then f is continuous.*

Proof. Fix $x_0 \in X$ and $\varepsilon > 0$. Choose ℓ such that $\sup_{x \in X} d(f_\ell(x), f(x)) < \varepsilon/3$. Since f_ℓ is continuous we can choose a neighborhood U of x_0 such that $d(f_\ell(x), f_\ell(x_0)) < \varepsilon/3$ whenever $x \in U$. Then for $x \in U$ we have

$$d(f(x), f(x_0)) \le d(f(x), f_\ell(x)) + d(f_\ell(x), f_\ell(x_0)) + d(f_\ell(x_0), f(x_0)) < \varepsilon.$$

□

The problem with uniform convergence is that it is too restrictive: if $f_\ell : \mathbb{C} \to \mathbb{C}$ is the function $f_\ell(z) := 1 + z + \frac{1}{2}z^2 + \cdots + \frac{1}{\ell!}z^\ell$, then $\{f_\ell\}$ does *not* converge uniformly to the exponential function (which is defined, in terms of this sequence, and analyzed in Section 3.9) because, for any ℓ, $f_\ell(z)$ and $\exp(z)$ are far apart when $|z|$ is large. As we mentioned in Section 3.2, continuity is a local concept (Proposition 3.21) in the sense that a function is continuous if it is continuous on a neighborhood of each point in the domain. A concept of convergence that only requires uniform convergence on some neighborhood of each point will still have the property that the limit of a sequence of continuous functions is continuous. The definition that is popular with analysts is superficially a bit different, but works out to the same thing in practice, as we explain below.

Definition 3.52. *A sequence of functions $f_1, f_2, \ldots$ from X to Y* ***converges to*** *$f : X \to Y$* ***uniformly on compacta*** *if, for each compact $D \subset X$, $\{f_\ell|_D\}$ converges uniformly to $f|_D$.*

Uniform convergence on compacta is, for our purposes, a "Goldilocks" concept, neither too weak nor too demanding.

Proposition 3.53. *If f and $f_1, f_2, \ldots$ are Y-valued functions on X and each $x \in X$ has a neighborhood U such that $\{f_\ell|_U\}$ converges uniformly to $f|_U$, then $\{f_\ell\}$ converges uniformly on compacta to f.*

Proof. Suppose that D is compact. The open sets on which $\{f_\ell\}$ converges uniformly to f are an open cover of D, so there is a finite subcover $U_1, \ldots, U_s$. For any $\varepsilon > 0$ we have

$$\sup_{x \in D} d(f_\ell(x), f(x)) \leq \sup_{x \in U_1 \cup \ldots \cup U_s} d(f_\ell(x), f(x)) < \varepsilon$$

when ℓ is large enough that $\sup_{x \in U_j} d(f_\ell(x), f(x)) < \varepsilon$ for each j. □

Thus uniform convergence on a neighborhood of each point implies uniform convergence on compacta. What about the converse? A topological space is **locally compact** if each neighborhood of each point in the space contains a compact neighborhood, and of course if X is locally compact, and $\{f_\ell\}$ converges to f uniformly on compacta, then $\{f_\ell\}$ converges to f uniformly on some neighborhood of each point in X. The spaces that are the center of attention in this book—most notably $\mathbb{R}^n$ and $\mathbb{C}^n$— are all locally compact.

Having talked about convergence of series of numbers and sequences of functions, we can now discuss the series of functions of greatest interest. A **power series** centered at $a \in \mathbb{C}$ is an infinite sum of the form

$$\sum_{n=0}^{\infty} c_n (z-a)^n$$

where a and the **coefficients** $c_0, c_1, c_2, \ldots$ are given complex numbers. In a rough sense it is clear that the asymptotic behavior of the sequence

$|c_0|, |c_1|, |c_2|, \dots$ determines whether or not this series converges absolutely at any particular z.

The precise quantitative expression of this intuition uses a pair of technical tools that are designed to handle situations in which a sequence $s_1, s_2, s_3, \dots$ of real numbers may not converge, but certain information about its asymptotic behavior is still important. The **limit inferior** of $\{s_n\}$, denoted by

$$\liminf_{n\to\infty} s_n,$$

is the least upper bound of the set of numbers ℓ such that $\ell < s_n$ for all but finitely many n, if the set of such ℓ is nonempty and bounded above. Otherwise there are two possibilities: (a) no such ℓ exists, in which case $\liminf_{n\to\infty} s_n := -\infty$; (b) the sequence diverges to ∞, in which case $\liminf_{n\to\infty} s_n := \infty$. Similarly, the **limit superior** of $\{s_n\}$, denoted by

$$\limsup_{n\to\infty} s_n,$$

is $-\infty$ if the sequence diverges to $-\infty$ and ∞ if there is a subsequence diverging to ∞, and otherwise it is the greatest lower bound of the set of numbers u such that $s_n < u$ for all but finitely many n. For example the limits inferior of the sequences

$$\tfrac{1}{2}, \tfrac{2}{3}, \tfrac{1}{4}, \tfrac{4}{5}, \tfrac{1}{6}, \tfrac{6}{7}, \tfrac{1}{8}, \tfrac{8}{9}, \dots \quad \text{and} \quad -\tfrac{1}{2}, \tfrac{4}{3}, -\tfrac{1}{4}, \tfrac{6}{5}, -\tfrac{1}{6}, \tfrac{8}{7}, -\tfrac{1}{8}, \tfrac{10}{9}, \dots$$

are both 0, and their limits superior are both 1.

The **radius of convergence** of the series $\sum_{n=0}^{\infty} c_n(z-a)^n$ is

$$R := \liminf_{n\to\infty} \frac{1}{\sqrt[n]{|c_n|}}.$$

This terminology is justified by the following result and Lemma 3.56 below.

Lemma 3.54. *If $0 < r < R$, then for all sufficiently large N it is the case that*

$$\sum_{n=N}^{\infty} |c_n(z-a)^n| \le \frac{(r/R)^{N/2}}{1-(r/R)^{1/2}}$$

for all z such that $|z-a| \le r$.

Proof. Of course $\sqrt{rR} < R$, so $\sqrt{rR} < |c_n|^{-1/n}$ and $|c_n| < (rR)^{-n/2}$ for all but finitely many n, and for all $n \ge N$ if N is large, in which case

$$|c_n(z-a)^n| = |c_n|\,|z-a|^n \le |c_n| r^n < (r/R)^{n/2}.$$

The claim follows from the formula $t^N + t^{N+1} + t^{N+2} + \cdots = t^N/(1-t)$ for the sum of a convergent geometric series. □

Thus the power series converges absolutely at each point in the open disk of radius R centered at a. A compact subset of this disk contains a point of maximum distance from a (Theorem 3.48) so it is contained in the closed ball of radius r centered at a for some $r < R$, and on this ball the convergence is uniform. We have shown that:

Proposition 3.55. *If $R := \liminf_{n\to\infty} 1/\sqrt[n]{|c_n|}$ is the radius of convergence of the power series $\sum_{n=0}^{\infty} c_n(z-a)^n$, then the series converges absolutely at each point of the disk*

$$D := \{\, z \in \mathbb{C} : |z-a| < R \,\}.$$

The convergence is uniform on compacta, so (Proposition 3.51) the function defined by the power series is continuous.

In Chapter 7 we will see that there may be sensible ways to extend the function defined by the power series to points outside of D, but this cannot be done by simply evaluating the infinite sums given by the series.

Lemma 3.56. *If $r := |z-a| > R$, then $\sum_{n=0}^{\infty} c_n(z-a)^n$ does not converge absolutely.*

Proof. There are infinitely many n such that $1/\sqrt[n]{|c_n|} < \sqrt{rR}$ and thus $|c_n(z-a)^n| > (r/R)^{n/2}$, so that $\sum_n |c_n(z-a)^n| = \infty$. □

A function $f : U \to \mathbb{C}$, where $U \subset \mathbb{C}$ is open, is said to be **holomorphic**, or **complex analytic**, if, for each $a \in U$, there is a power series centered at a that has a positive radius of convergence and that agrees with f on a neighborhood of a. An **entire function** is a holomorphic function $f : \mathbb{C} \to \mathbb{C}$. Holomorphic functions are extremely well behaved, with many interesting properties, and they are important to many subfields of mathematics. (The theory of holomorphic functions was developed in large part by Bernhard Riemann (1826-1866) whose work we'll feature in Chapters 8 and 9.) A basic issue in this theory is whether the function defined by the power series $\sum_{n=0}^{\infty} c_n(z-a)^n$ is holomorphic. That is, is it the case that for any $b \in D$ there is a power series $\sum_{n=0}^{\infty} c'_n(z-b)^n$ that has a positive radius of convergence and agrees with the function defined by $\sum_{n=0}^{\infty} c_n(z-a)^n$ on a neighborhood of b? The answer is affirmative (this is Theorem 7.11) but it makes sense to defer the proof until Section 7.4 where we treat it in the context of related issues. However, there is a special case that we will need in the next section:

Lemma 3.57. *A polynomial function $z \mapsto a_n z^n + \cdots + a_1 z + a_0$ is entire.*

Proof. Fix $a \in \mathbb{C}$. We claim that there are complex numbers $c_0, c_1, \dots, c_n$ such that

$$p(z) = c_0 + c_1(z-a) + c_2(z-a)^2 + \cdots + c_n(z-a)^n.$$

When $n = 0$ we can simply set $c_0 := a_0$, so, by induction, we may assume that the claim has already been established with $n-1$ in place of n. In particular, if $q(z) := p(z) - a_n(z-a)^n$, then

$$q(z) = c_0 + c_1(z-a) + \cdots + c_{n-1}(z-a)^{n-1}$$

for some $c_0, c_1, \dots, c_{n-1}$, and we can set $c_n := a_n$. □

There is a property of holomorphic functions called the **maximum modulus principle** that plays a role in many of the proofs in complex analysis, and which is at the heart of our proof of the fundamental theorem of algebra in the next section. Suppose $f : D \to \mathbb{C}$ is defined by the power series $\sum_{n=0}^{\infty} c_n(z-a)^n$. One possibility is that $c_n = 0$ for all $n \geq 1$, in which case f is a constant function. Otherwise we can write

$$f(z) = c_0 + c_k(z-a)^k + c_{k+1}(z-a)^{k+1} + \cdots$$

where $c_k \neq 0$. The key insight is that if $|z-a|$ is very small, but $z \neq a$, then $|c_k(z-a)^k|$ will be much larger than $\left|\sum_{j=k+1}^{\infty} c_j(z-a)^j\right|$, so $f(z)$ will be well approximated by $c_0 + c_k(z-a)^k$. By choosing z appropriately we can arrange for it to be the case that $|f(a)| < |f(z)|$ because $|c_0 + c_k(z-a)^k| > |c_0|$. In addition, if $c_0 \neq 0$ we can use the same method to find z near a with $|f(z)| < |f(a)|$.

The proof of this depends on the following fact:

Proposition 3.58. *For any $c \in \mathbb{C}$ and any integer $k \geq 1$ there is $w \in \mathbb{C}$ such that $w^k = c$.*

Although this is fairly simple, at least if you have some acquaintance with basic properties of the complex numbers, our proof of it, in the last two sections of this chapter, will be a drawn out affair. Instead of lunging at the quickest possible proof, we will use the task as a springboard for an exposition of several interesting concepts.

Theorem 3.59 (Maximum Modulus Principle). *Let $f : U \to \mathbb{C}$ be holomorphic, where $U \subset \mathbb{C}$ is open, and let a be an element of U. If f is not constant on any neighborhood of a, then $|f(a)| < \sup_{z \in U} |f(z)|$. If $f(a) \neq 0$, then $|f(a)| > \inf_{z \in U} |f(z)|$.*

Proof. Since f is holomorphic there is a power series $\sum_{k=0}^{\infty} c_k(z-a)^k$ that agrees with f in some neighborhood of a. There must be some $k \geq 1$ with $c_k \neq 0$ because f is not constant near a, so this series has the form

$$c_0 + c_k(z-a)^k + c_{k+1}(z-a)^{k+1} + \cdots$$

where $c_k \neq 0$. Let $r > 0$ be a number less than the radius of convergence such that f agrees with this series on the closed ball of radius r centered at a. The definition of the radius of convergence implies that $|c_n| < r^{-n}$ for all but finitely many n, so we can choose $A > 0$ large enough that $|c_n| < Ar^{-n}$ for all n.

If $c_0 = 0$ and $0 < t < r$, then

$$|f(a+t)| = \Big|\sum_{n=k}^{\infty} c_k t^k\Big| \geq |c_k|t^k - \sum_{n=k+1}^{\infty} (Ar^{-n})t^n = |c_k|t^k - \frac{A(t/r)^{k+1}}{1-t/r},$$

and the final quantity is positive when t is sufficiently small. For the remainder of the proof we assume that $f(a) = c_0 \neq 0$.

The last result gives a $w \in \mathbb{C}$ with $w^k = c_0/c_k$. If $t > 0$ is small enough that $|tw| < r$, then $c_0 + c_k(tw)^k = c_0(1+t^k)$ and

$$\begin{aligned}\big|f(a+tw)\big| &= \Big|c_0 + c_k(tw)^k + \sum_{n=k+1}^{\infty} c_n(tw)^n\Big| \geq |c_0|(1+t^k) - \sum_{n=k+1}^{\infty} |c_n|\,|tw|^n \\ &\geq |c_0|(1+t^k) - A\sum_{n=k+1}^{\infty} (|tw|/r)^n = |c_0|(1+t^k) - A\frac{(|tw|/r)^{k+1}}{1-|tw|/r},\end{aligned}$$

and the final expression is greater than $|f(a)| = |c_0|$ when $t > 0$ is sufficiently small.

The last result also gives $w \in \mathbb{C}$ with $w^k = -c_0/c_k$, and a similar calculation shows that if $|tw| < r$ and $0 < t < 1$, then

$$\big|f(a+tw)\big| \leq |c_0|(1-t^k) + A\frac{(|tw|/r)^{k+1}}{1-|tw|/r},$$

which is less than $|f(a)| = |c_0|$ when $c_0 \neq 0$ and $t > 0$ is sufficiently small. □

To explain how the maximum modulus principle is often understood we now introduce some more (quite important!) terminology from general topology. If X is a topological space and A is a subset of X, the **interior** of A is the union of all the open subsets of X that are contained in A. This union is itself open, so, in a well defined sense, the interior of A is the largest open set contained in A. A point is in the interior of A if and only if it is not an accumulation point of $X \setminus A$, so, if you feel like being a bit convoluted, you can say that the interior of A is the complement of the closure of the complement of A. The **boundary** of A is the set of points in the closure of A that are not in the interior of A. It is the intersection of the closure of A and the closure of the complement of A. Equivalently, it is the complement of the union of the interiors of A and A's complement[3].

Now suppose that $U \subset \mathbb{C}$ is open, $f : U \to \mathbb{C}$ is holomorphic, and $K \subset U$ is compact and nonempty. Let $\big|f|_K\big| : K \to [0, \infty)$ be the function $z \mapsto |f(z)|$. One common way of thinking about the maximum modulus principle is that $\big|f|_K\big|$ attains its maximum, and any maximizer is either an element of the boundary of K or has a neighborhood on which f is constant. Expressed in this fashion, the maximum modulus principle seems rather remarkable. In some sense I suppose this is true, but the sense of surprise is largely derived from compactness: the basic facts about maximization of a continuous real valued function on a compact set (Theorem 3.48) guarantee that the maximum is achieved somewhere, and the result above implies that it can't be attained at a point in the interior of K unless f is constant near that point.

3.8 The Fundamental Theorem of Algebra

A field k is **algebraically complete** if every polynomial

$$p = a_n X^n + \cdots + a_1 X + a_0 \in k[X]$$

that is not constant (because $n > 0$ and $a_n \neq 0$) has a root in k. The fundamental theorem of algebra asserts that the field $\mathbb{C}$ of complex numbers is algebraically complete. This result, more than any other, is what makes $\mathbb{C}$ the central object of mathematics, both in quantitative analysis and from the point of view of number theory.

To get a better understanding of algebraic completeness observe that for any $r \in k$ one can use division with remainder to obtain $q \in k[X]$ and $c \in k$ such that

$$p = (X - r)q + c.$$

[3] A standard exercise in real analysis, which you might enjoy, is to find all the (possibly) distinct sets that can be obtained from A using closure, complementation, union, intersection, and set difference.

If r is a root of p, then necessarily $c = 0$, and if k is algebraically complete then q must also have a root if it is not constant. Continuing in this fashion, one finds that algebraic completeness implies that p can be written as a product

$$p = a_n(X - r_1) \cdots (X - r_n)$$

of linear factors where $r_1, \ldots, r_n$ are all the roots of p. Thus k is algebraically complete if and only if each polynomial with coefficients in k is a product of linear elements of $k[X]$.

The fundamental theorem of algebra gave the mathematicians of the 18th century a great deal of trouble. Incomplete proofs were published by Jean le Rond d'Alembert (1717-1783), Daviet de Foncenex (1734-1798), Euler, Joseph Louis Lagrange (1736-1813), Pierre-Simon Laplace (1749-1827), and Gauss. The first fully complete proof was published by Jean-Robert Argand (1768-1822) in 1806. (A key breakthrough in Argand's work was his representation of $\mathbb{C}$ as a two dimensional plane[4], and the geometric properties of the modulus describe in the last section.) Given such a star-studded cast, one might expect the proof below to be lengthy, with subtle details, but in fact it is straightforward and easily understood.

To some extent the difficulties experienced by our forebears might be due to the lack of clear definitions and solid foundations. Specifically, the construction of $\mathbb{C}$ has two steps, the first being to append a square root i of -1 to $\mathbb{Q}$, while the second is the passage from $\mathbb{Q}[i]$ to $\mathbb{C}$ which we now understand as the topological operation of completion, i.e., appending limits of all Cauchy sequences. (The application of this procedure to a general metric space will be explained in Section 6.1.) Completion was not well understood at the time, so one might imagine that this was the source of the difficulty. However, our proof really uses completeness only insofar as it ultimately depends on the following simple fact.

Lemma 3.60. *For any $r \geq 0$ and any integer $n > 0$ there is a unique number $s \geq 0$ (which will be denoted by $\sqrt[n]{r}$, of course) such that $s^n = r$.*

Proof. The function $t \mapsto t^n$ is continuous, with $0^n = 0$ and $t^n \to \infty$ as $t \to \infty$. The intermediate value function implies that a suitable s exists, and there can be only one such s because the function is strictly increasing. □

Even though the understanding of continuity in the 18th century was (at least from our perspective) imperfect, the intermediate value theorem was known, and this consequence of it was certainly uncontroversial.

[4] What is now known as the **Argand plane** was actually first described by Caspar Wessel (1745-1818) in a 1799 paper that was unnoticed at the time.

Proposition 3.58 (that is, the special case of the fundamental theorem of algebra for the equation $X^k = c$) is another key ingredient in our proof. We'll prove it in the next two sections using the special case above, properties of the exponential and trigonometric functions, and topological reasoning. Although the argument won't be terribly complicated, some of the ingredients are modern, which might suggest that incomplete understanding of the roots of the equation $X^n = c$ was the critical stumbling block for mathematicians prior to Argand, but Euler certainly knew that this equation had solutions in $\mathbb{C}$.

Not being an historian of mathematics, I am hesitant to venture a guess about the critical impediment. Euler, Laplace, Lagrange, and Gauss were, shall we say, not exactly stupid people, and they thought about complex numbers a lot, so each of them presumably had a bag of tricks for dealing with them. The facts about the Argand plane given in the last section are now pretty much the first things people learn about $\mathbb{C}$, so it's hard for us to imagine how complex numbers were understood before these facts were known.

What we can say with confidence is that the mathematicians of the 18th century did not know about the application of compactness via the minimization maneuver in the proof below, which is familiar to any modern mathematician.

Theorem 3.61 (Fundamental Theorem of Algebra)**.** *$\mathbb{C}$ is algebraically complete.*

We separate out one step of the argument because it is computationally intensive. The intuition is simple—as $|z|$ becomes large, the leading term of p dominates the sum of all the other terms—and if you think the main idea is clear you can skim, or just skip, the details.

Lemma 3.62. *If $p = a_nX^n + \cdots + a_1X + a_0 \in \mathbb{C}[X]$ is a nonconstant polynomial, then for any $M > 0$ there is $R > 0$ such that $|p(z)| \geq M$ whenever $|z| \geq R$.*

Proof. When $z \neq 0$ the basic facts about the modulus and the triangle inequality allow us to compute that

$$\begin{aligned}
|p(z)| &\geq |a_nz^n| - |a_{n-1}z^{n-1} + \cdots + a_1z + a_0| \\
&\geq |a_n|\,|z|^n - \big(|a_{n-1}|\,|z|^{n-1} + \cdots + |a_0|\big) \\
&= |z|^n\Big(|a_n| - \Big(\frac{|a_{n-1}|}{|z|} + \cdots + \frac{|a_0|}{|z|^n}\Big)\Big).
\end{aligned}$$

If R is large enough, then

$$\frac{|a_{n-1}|}{R} + \cdots + \frac{|a_0|}{R^n} \leq \tfrac{1}{2}|a_n|,$$

so that when $|z| \geq R \geq \sqrt[n]{2M/|a_n|}$ we have

$$|p(z)| \geq R^n\Big(|a_n| - \Big(\frac{|a_{n-1}|}{R} + \cdots + \frac{|a_0|}{R^n}\Big)\Big) \geq \tfrac{1}{2}R^n|a_n| \geq M.$$

□

Proof of the Fundmental Theorem of Algebra. Let

$$p = a_nX^n + \cdots + a_1X + a_0 \in \mathbb{C}[X]$$

be a nonconstant polynomial. Regarded as a function on $\mathbb{C}$, p is entire (Lemma 3.57) and consequently continuous (Proposition 3.55). The function $z \mapsto |z|$ is continuous (due to the triangle inequality for the modulus) so the function $z \mapsto |p(z)|$ is continuous. For any $R > 0$ the disk

$$\mathbf{B}_R(0) := \{\, z \in \mathbb{C} : |z| \leq R \,\}$$

is closed and bounded, hence compact, so Theorem 3.48 implies that there is $z_0 \in \mathbf{B}_R(0)$ such that $|p(z_0)| \leq |p(z)|$ for all $z \in \mathbf{B}_R(0)$. If p has no roots, then the maximum modulus principle (Theorem 3.59) implies that there are two possibilities. The first is that p is constant on some neighborhood of z_0, which is impossible because it would imply that p is a constant polynomial, contrary to hypothesis. (In somewhat pedantic detail, the equation $p(z) = p(z_0)$ has a most n solutions, and any neighborhood of z_0 contains infinitely many points.) The second possibility is that z_0 is in the boundary of $\mathbf{B}_R(0)$, but the last result implies that if $R > 0$ large enough, then $|p(z)| > |a_0|$ whenever $|z| \geq R$, so that we have the contradictory inequality $|p(z_0)| > |a_0| = |p(0)| \geq |p(z_0)|$. This contradiction completes the proof. □

3.9 The Exponential and Trigonometric Functions

Our proof of the fundamental theorem of algebra has one remaining loose end, namely Proposition 3.58. This will be proved in Section 3.10 by combining an important topological concept with the basic properties of the exponential and trigonometric functions

$$\exp(z) := e^z := \sum_{n=0}^{\infty} \frac{z^n}{n!}, \quad \cos z := \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}, \quad \sin z := \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!},$$

which are developed in this section. The analysis has a quite different character from most of the material in this book, with many "heavy" computations which for some people (e.g., yours truly) makes for slower and more tedious reading. On the other hand, the theory developed here is, of course, beautiful and extremely important mathematics, so I think it's worth an extra effort.

To begin with observe that $1/\sqrt[n]{1/n!} = \sqrt[n]{n!} \to \infty$[5], so the series defining the exponential and trigonometric functions have infinite radii of convergence and are consequently everywhere absolutely convergent, and uniformly convergent on compacta. Consequently we can perform algebraic manipulations without worrying about divergence or the possibility that the order of summation might matter. For example, the binomial theorem and the double summation formula give

$$\begin{aligned}\exp(w+z) &= \sum_{k=0}^{\infty} \frac{(w+z)^k}{k!} = \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i} w^i z^{k-i} \\ &= \sum_{k=0}^{\infty} \sum_{i=0}^{k} \frac{1}{i!} w^i \frac{1}{(k-i)!} z^{k-i} \\ &= \Big(\sum_{m=0}^{\infty} \frac{w^m}{m!}\Big)\Big(\sum_{n=0}^{\infty} \frac{z^n}{n!}\Big) = \exp(w)\exp(z).\end{aligned}$$

Euler's famous equation is a matter of splitting the relevant sum into two parts:

$$\begin{aligned}\exp(iy) = \sum_{n=0}^{\infty} \frac{i^n y^n}{n!} &= \sum_{m=0}^{\infty} \Big(\frac{i^{2m} y^{2m}}{(2m)!} + \frac{i^{2m+1} y^{2m+1}}{(2m+1)!}\Big) \\ &= \sum_{m=0}^{\infty} \frac{(-1)^m y^{2m}}{(2m)!} + i \sum_{m=0}^{\infty} \frac{(-1)^m y^{2m+1}}{(2m+1)!} = \cos y + i \sin y.\end{aligned}$$

Together the last two equations give

$$\begin{aligned}\cos(\phi+\theta) + i\sin(\phi+\theta) &= \exp(i(\phi+\theta)) = \exp(i\phi)\exp(i\theta) \\ &= (\cos\phi + i\sin\phi)(\cos\theta + i\sin\theta),\end{aligned}$$

from which we obtain the **angle addition formulas**:

$$\cos(\phi+\theta) = \cos\phi\cos\theta - \sin\phi\sin\theta, \quad \sin(\phi+\theta) = \cos\phi\sin\theta + \sin\phi\cos\theta.$$

The following famous formula has a bulkier verification.

[5] When n is even we have $\sqrt[n]{n!} > \sqrt[n]{(n/2+1)\cdots n} > \sqrt[n]{(n/2)^{n/2}} = \sqrt{n/2}$, and it is obvious that similar, messier, inequalities pertain when n is odd.

Proposition 3.63. *For all $z \in \mathbb{C}$, $\cos^2 z + \sin^2 z = 1$.*

Proof. Using the double summation formula, we compute that

$$\begin{aligned}\cos^2 z &= \left(\sum_{m=0}^{\infty} \frac{(-1)^m z^{2m}}{(2m)!}\right)\left(\sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}\right)\\ &= \sum_{k=0}^{\infty}\sum_{i=0}^{k} \frac{(-1)^i z^{2i}}{(2i)!} \cdot \frac{(-1)^{k-i} z^{2(k-i)}}{(2k-2i)!}\\ &= 1 + \sum_{k=1}^{\infty} \Big((-1)^k \sum_{i=0}^{k} \frac{1}{(2i)!} \cdot \frac{1}{(2k-2i)!}\Big) z^{2k}\\ &= 1 + \sum_{k=1}^{\infty} \Big(\frac{(-1)^k}{(2k)!} \sum_{i=0}^{k} \binom{2k}{2i}\Big) z^{2k}\end{aligned}$$

and

$$\begin{aligned}\sin^2 z &= \left(\sum_{m=0}^{\infty} \frac{(-1)^m z^{2m+1}}{(2m+1)!}\right)\left(\sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}\right)\\ &= \sum_{k=0}^{\infty}\sum_{i=0}^{k} \frac{(-1)^i z^{2i+1}}{(2i+1)!} \cdot \frac{(-1)^{k-i} z^{2(k-i)+1}}{(2k-2i+1)!}\\ &= \sum_{k=0}^{\infty} \Big((-1)^k \sum_{i=0}^{k} \frac{1}{(2i+1)!} \cdot \frac{1}{(2k-2i+1)!}\Big) z^{2(k+1)}\\ &= \sum_{k=0}^{\infty} \Big(\frac{(-1)^k}{(2k+2)!} \sum_{i=0}^{k} \binom{2k+2}{2i+1}\Big) z^{2(k+1)}\\ &= -\sum_{k=1}^{\infty} \Big(\frac{(-1)^k}{(2k)!} \sum_{i=0}^{k-1} \binom{2k}{2i+1}\Big) z^{2k}.\end{aligned}$$

We now apply the binomial theorem to compute that, for each $k \geq 1$,

$$\sum_{i=0}^{k} \binom{2k}{2i} - \sum_{i=0}^{k-1} \binom{2k}{2i+1} = \sum_{\ell=0}^{2k} (-1)^\ell \binom{2k}{\ell} = (1 + (-1))^{2k} = 0.$$

□

3.10 Connectedness

Let's take stock, bearing in mind the goal of proving Proposition 3.58. The main idea is to show that any $c \in \mathbb{C}$ is $r\exp(i\theta)$ for some $r \geq 0$ and $\theta \in \mathbb{R}$, after which Lemma 3.60 gives a k^{th} root of r, and $\sqrt[k]{r}\exp(i\theta/k)$ is a k^{th} root of c, which is what Proposition 3.58 asks us to produce. Let

$$C = \{\, z \in \mathbb{C} : |z| = 1 \,\}$$

be the unit circle in $\mathbb{C}$. The map $z \mapsto (|z|, z/|z|)$ from $\mathbb{C}^*$ to $(0,\infty) \times C$ has the inverse $(r,c) \mapsto rc$, so any point in $\mathbb{C}$ is rc for some $r \geq 0$ and $c \in C$. Therefore Proposition 3.58 will follow if we can show that every $c \in C$ is $\exp(i\theta)$ for some $\theta \in \mathbb{R}$.

The equation $|wz| = |w|\,|z|$ implies that products and inverses of elements of C are contained in C, and of course $1 \in C$, so C is a subgroup of $\mathbb{C}^*$. Because the coefficients in the power series defining the exponential and trigonometric functions are all real, these functions map $\mathbb{R}$ into $\mathbb{R}$. Therefore the equations $\exp(i\theta) = \cos\theta + i\sin\theta$ and $\cos^2\theta + \sin^2\theta = 1$ imply that the function $\theta \mapsto \exp(i\theta)$ maps $\mathbb{R}$ into C. The equation

$$\exp(i(\theta + \phi)) = \exp(i\theta)\exp(i\phi)$$

states that this function is actually a homomorphism from $\mathbb{R}$ (thought of as a group with addition as the group operation) to C, and the image of any homomorphism is a subgroup of the range. We would like to use these facts to show that the image is all of C, but C has many subgroups, e.g. $\{1,-1\}$ or $\{1,\omega,\omega^2\}$ where $\omega = \frac{1}{2}(-1+i\sqrt{3})$ is a cube root of 1, so in itself this isn't quite enough to give us what we want.

This sort of situation comes up fairly frequently in the analysis of foundational questions. We "know" something is true, but it still needs to be proved. It seems like it ought to be easy, but somehow it isn't, and it's a bit confusing to think about because it's hard to keep track of what we can and can't use in a proof. For the question at hand various types of "brute force" arguments might work, but that would feel like an admission of defeat.

It turns out that there is another important idea from topology that can be applied. A subset A of a topological space X is **disconnected** if there are open sets $U_1, U_2 \subset X$ with $U_1 \cap U_2 = \emptyset$, $A \subset U_1 \cup U_2$, and $U_1 \cap A \neq \emptyset \neq U_2 \cap A$. That is, A decomposes into two nonempty pieces that are, in a certain sense, topologically separated from each other. We say that A is **connected** if it is not disconnected. Connectedness means pretty much what you might expect, based on the ordinary usage of the work 'connected,' to the point where a figure would be superfluous: a circle is a connected (as we'll prove below) subset of $\mathbb{R}^2$ but a subset consisting of two nonintersecting circles is not.

It's probably not immediately evident how one might use this definition to prove that a space is connected, but after you've seen a couple arguments the typical pattern—assume that it's not connected and derive consequences until a contradiction emerges—becomes clear. In contrast, the following proof is a bit less indirect.

Proposition 3.64. *If $\{Y_\alpha\}_{\alpha \in A}$ is a collection of connected subsets of X with $\bigcap_{\alpha \in A} Y_\alpha \neq \emptyset$, then $Y := \bigcup_{\alpha \in A} Y_\alpha$ is connected.*

Proof. Suppose that U_1 and U_2 are disjoint open subsets of X with $Y \subset U_1 \cup U_2$. Fixing a point $x \in \bigcap_{\alpha \in A} Y_\alpha$, suppose that $x \in U_1$. For each α we have $Y_\alpha \subset Y \subset U_1 \cup U_2$ and $x \in Y_\alpha \cap U_1$, so $Y_\alpha \cap U_2 = \emptyset$ because Y_α is connected. Therefore $Y \cap U_2 = \emptyset$. □

For each $x \in X$ the union of all connected subsets of X containing x is the **connected component** of X containing x. The last result implies that it is connected, so it is the largest connected set containing x. The connected components cover X, and the intersection of any two of them is empty, so they constitute a partition of X.

Intervals in $\mathbb{R}$ are the prototypical connected sets. Although the concept of a (possibly open, closed, half open, bounded, or unbounded) interval is presumably well understood, for the following proof we need a formal definition. We'll say that $A \subset \mathbb{R}$ is an **interval** if $[r, t] \subset A$ whenever $r, t \in A$ and $r \leq t$.

Lemma 3.65. *Let A be a subset of $\mathbb{R}$. Then A is connected if and only if it is an interval.*

Proof. If A is not an interval there are numbers $r < s < t$ with $r, t \in A$ and $s \notin A$. Setting $U_1 := (-\infty, s)$ and $U_2 := (s, \infty)$ shows that A is disconnected.

Now let A be an interval. Aiming at a contradiction, suppose A is disconnected: there are open sets U_1, U_2 with $U_1 \cap U_2 = \emptyset$, $A \subset U_1 \cup U_2$, and $U_1 \cap A \neq \emptyset \neq U_2 \cap A$. Choose $r \in U_1 \cap A$ and $t \in U_2 \cap A$. We can interchange U_1 and U_2, so we may assume that $r < t$. Then $U_1 \cap [r, t]$ is nonempty and bounded above; let s be its least upper bound. Of course $s \in A$ because A is an interval, and $s \notin U_2$ because U_2 is open and there are elements of U_1 arbitrarily close to s, so $s \in U_1$ and $s < t$. But since U_1 is open, $[s, s + \delta] \subset U_1 \cap A$ for sufficiently small $\delta > 0$, contradicting the definition of s. □

The next result is one of the most common vehicles for deriving useful consequences of connectedness.

Proposition 3.66. *If X and Y are topological spaces, X is connected, and $f : X \to Y$ is continuous, then the image of f is connected.*

Proof. Suppose the assertion is false: then there exist open sets $V_1, V_2 \subset Y$ with $V_1 \cap V_2 = \emptyset$, $f(X) \subset V_1 \cup V_2$, and $V_1 \cap f(X) \neq \emptyset \neq V_2 \cap f(X)$. For $i = 1, 2$ let $U_i := f^{-1}(V_i)$. Then U_1 and U_2 are open because f is continuous, $U_1 \cap U_2 = \emptyset$ because V_1 and V_2 are disjoint, $X = U_1 \cup U_2$, and $U_1 \neq \emptyset \neq U_2$, contrary to the assumption that X is connected. □

We have shown that $\mathbf{R}$ is connected, and (by Proposition 3.55) the exponential function is continuous, so the last result implies that the image of $\theta \mapsto e^{i\theta}$ is connected. Using the power series for $\cos\theta$ and $\sin\theta$ (as in the proof of the maximum modulus principle) one can easily show that $e^{i\theta} \neq 1$ when $\theta \neq 0$ and $|\theta|$ is very small, so the image of $\theta \mapsto e^{i\theta}$ is a connected subgroup of C that is different from $\{1\}$. The following result implies that it is C (so we have shown that C is connected!) thereby completing the proofs of Proposition 3.58 and the fundamental theorem of algebra.

Proposition 3.67. *If G is a connected proper subgroup of C, then $G = \{1\}$.*

Proof. Let $z = x + iy$ be an element of C that is not in G. Then $z^{-1} \notin G$ because G contains the inverse of each of its elements, and $z^{-1} = \overline{z}$ because $C = \{\, w \in \mathbf{C} : w\overline{w} = 1 \,\}$. Let U_1 be the set of elements of C with real part less than x, and let U_2 be the set of elements of C with real part greater than x. Clearly U_1 and U_2 are open and disjoint, and $G \subset U_1 \cup U_2$ because z and $\overline{z}$ are the only elements of C with real part x. Since $1 \in U_2$ and G is connected, $U_1 \cap G$ must be empty, and in particular $-1 = e^{i\pi} \notin G$. Since G is a subgroup, -1 can't have an n^{th} root in G for any n. Repeating our argument with $e^{i\pi/n}$ in place of z shows that the real part of every element of G is greater than $\cos \pi/n$ for every n. Since (by Proposition 3.55) the cosine function is continuous, 1 is the only element of C satisfying this condition. □

Chapter 4

Linear Algebra

For the system of equations

$$\begin{aligned} 3x + 4y + 5z &= 7,\\ 4x + 2y + 3z &= 6,\\ 2x + 3y + 2z &= 4, \end{aligned}$$

determine whether there are no solutions, a unique solution, or infinitely many solutions. If there is a unique solution, find it.

If you've had a linear algebra course it's quite likely that you think of the subject as all about problems like this. Pretty boring, no?

Linear systems of equations arise in pretty much every part of mathematics, and systematic ways of dealing with them can be seen in Chinese texts from 2000 years ago. At the same time the approach taken here is distinctly modern, with concepts that didn't exist two hundred years ago. It emphasizes generality and axiomatics, placing the actual process of finding a solution in the background. But to a certain extent it isn't new mathematics, just a new way of talking about ideas that have been understood for a long time.

And this chapter actually is a bit boring. In studying mathematics, and especially the sorts of fundamental topics described in this book, it is easy to get the feeling that the effort is entirely a matter of building infrastructure in preparation for the "real" mathematics that will come later. (Later one learns that the mathematics that is the most interesting, conceptually, and in relation to other subfields, is to a large extent precisely the material that eventually becomes part of the subject's infrastructure.) For some topics the infrastructure involves results that are themselves significant theorems, but we won't see anything like that here. Mostly it will be nuts-and-bolts definitions, with a few simple results describing basic facts. In an attempt to endow it with it with some independent interest we will emphasize two themes: a) coordinate-free notation; b) classification. In the remainder of the introduction we'll say a little about the first of these.

At the elementary level people are usually taught to think of vectors as tuples of numbers like $(1.5, 3.2, 4.7)$, and we're taught certain operations on these tuples that amount to recombining the components to get new tuples of numbers, or just a number, or perhaps an answer to some geometric question. The idea of representing points in the plane by pairs of numbers, or points in space with triples of numbers, is due to René Descartes (1596-1650), and is arguably the most important advance in our understanding of geometry since Euclid.

But when we apply Descartes' idea, the coordinate system is something we impose on the world, and for most purposes there are many different coordinate systems that are equally valid. More to the point, it's not how we actually think about things. If, for example, you imagine throwing a ball into the air, your brain is perfectly capable of creating a vivid and accurate picture of what will happen without ever (consciously at least) manipulating triples of numbers. Somehow you brain thinks about these things in a manner that seems more direct, and considerably more effective.

The general idea of coordinate-free notation is to create languages that express in some direct way the key operations of mathematical structures, or physical theories, without referring to the underlying bundles of numbers. We then study the formal properties of the language, thereby testing its adequacy as a representation of the phenomenon of interest and, in the event that it works well, developing useful concepts and theorems within this framework.

This turns out to be an extraordinarily powerful and fruitful idea. Perhaps this seems counterintuitive, since after all any one coordinate system is adequate, and the ability to shift to another one, or to combine the numbers in bundles that can be dealt with in more abstract ways, is a convenience that doesn't change the underlying reality. A computational analogy might help here. In principle everything a computer does is a matter of manipulating bits, and any piece of software can be written in a language consisting of a few elementary procedures for computing a new bit from one or two given bits. But as a practical matter the power of computers is amplified across many orders of magnitude by high level programming languages that allow people to talk to computers in terms that people understand, in terms of abstractions.

The development of abstract linear algebra is one of the more curious chapters in the history of mathematics. Hermann Grassman (1809-1877) did not excel in his studies when he was young to the same extent as other mathematicians whose work we have described, and he taught in secondary schools for much of his career and never obtained a university professorship. His masterpiece, *Die Lineale Ausdehnungslehre ein neuer Zweig der Mathematik*, describes the basic concepts of linear algebra along very modern lines similar to the presentation below. First published in 1845, it received very little recognition, and in 1862 Grassmann published a thoroughly rewritten version, but again it was almost completely ignored during his life. In later years he turned to historical linguistics, and received considerable recognition for work in that field that is still remembered. Eventually other mathematicians discovered and were influenced by his work, both in linear algebra and in the foundations of arithmetic.

The human mind tends to resist abstractions in the absence of a compelling case for their utility; it is easy to define a new abstraction, but difficult to know in advance which of the many possible definitions will be effective tools. August Möbius (1790-1868) (about whom we'll hear more later) encouraged Grassmann, but is also on record criticizing him for introducing abstract notions without giving the reader any intuition as to why they were valuable. It did not help that Grassmann had a rather indirect and florid prose style. In modern research a common test of a new concept is to ask whether it helps to solve a single concrete problem, and perhaps this was Grassmann's greatest failing as a salesman of his work. Although his concepts eventually proved enormously fruitful, it is probably correct to say that this was more because they constituted a powerful language and pointed in the direction of new classes of problems, and entirely new perspectives, than because they resolved existing conundrums.

4.1 Vector Spaces

Addition of points $v, w \in \mathbb{R}^n$, and multiplication of a point $v \in \mathbb{R}^n$ by a number $\alpha \in \mathbb{R}$, are defined by the formulas

$$v + w = (v_1 + w_1, \dots, v_n + w_n) \quad \text{and} \quad \alpha v = (\alpha v_1, \dots, \alpha v_n).$$

The general plan of our work is to define an abstract concept based on the most obvious properties of these operations, then study the relation between the abstract notion and the concrete phenomenon it models. Whereas the notion of a topological space turned out to be much more general than the examples that motivated it, here the opposite will be the case: there will be an exact correspondence between vector spaces and examples like $\mathbb{R}^n$, except that a vector space may not be finite dimensional.

Fix a field k. We are primarily interested in the cases $k = \mathbb{R}$ and $k = \mathbb{C}$, but there are numerous applications in which k is a different field. In addition, it is important to understand that everything that happens below depends only on the properties shared by all fields. In particular, although many of the examples will have natural topologies, topological notions will not figure in the analytic work of this chapter.

Definition 4.1. *A* ***vector space*** *over k is a commutative group V whose group operation is written additively and called* ***vector addition****, or just* ***addition****, together with an operation $(\alpha, v) \mapsto \alpha v$ from $k \times V$ to V, called* ***scalar multiplication****, satisfying:*

(a) $\alpha(v + w) = (\alpha v) + (\alpha w)$ for all $\alpha \in k$ and all $v, w \in V$.

(b) $(\alpha + \beta)v = (\alpha v) + (\beta v)$ for all $\alpha, \beta \in k$ and all $v \in V$.

(c) $(\alpha\beta)v = \alpha(\beta v)$ for all $\alpha, \beta \in k$ and all $v \in V$.

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

This should look rather familiar: a vector space is just a k-module. Due to its importance and relatively simple character, in comparison with the general theory of modules, the theory of vector spaces is usually developed without any mention of modules over other rings, and it has its own system of terminology. In particular, in this context elements of k are usually called **scalars**.

The rest of the section just presents a few examples.

To begin with, it's possible that $V = \{0\}$ with $0 + 0 = 0$ and $\alpha 0 = 0$ for all $\alpha \in k$. Of course this possibility is trivial, but it comes up frequently in proofs. It is worth emphasizing that, by requiring V to be a commutative group, so that it is automatically nonempty, we arranged for $\emptyset$ to *not* be a vector space.

Polynomials can be added and multiplied by scalars in the usual way. Thus $k[X_1, \ldots, X_n]$ is a vector space. Functions taking values in k (including the functions defined by evaluating polynomials) can be added and multiplied by scalars. Specifically, if S is any set and $f, g \in \mathcal{F}_k(S)$, then $f + g$ is the function

$$s \mapsto f(s) + g(s),$$

and for any $\alpha \in k$, αf is the function

$$s \mapsto \alpha f(s).$$

(The sophisticated way to explain all this is to say that "vector addition and scalar multiplication are defined **pointwise**.")

If X is a metric space, $C(X) \subset \mathcal{F}_{\mathbb{R}}(X)$ is the space of continuous real valued functions with domain X. Since $f+g$ and αf are continuous whenever $f, g \in C(X)$ and $\alpha \in \mathbb{R}$, $C(X)$ is a vector space over $\mathbb{R}$. If all this is relatively new to you it would be a good idea to make sure that you actually know how to prove that, in fact, $f+g$ and αf are continuous. Here is a hint: the quantities $\varepsilon/2$ and $\min\{\delta_1, \delta_2\}$ figure in the proof for $f+g$, and ε/α (with due attention to the case $\alpha = 0$!) appears in the proof for αf.

There are many important vector spaces of infinite sequences. To begin with, given sequences of real numbers $s_1, s_2, \ldots$ and $t_1, t_2, \ldots$, we can form the sum $s_1 + t_1, s_2 + t_2, \ldots$, and for any real number α we can form the sequence $\alpha s_1, \alpha s_2, \ldots$. It is easy to verify directly that the space of all real valued sequences is a vector space, or we can observe that it is the space of functions $\mathcal{F}_{\mathbb{R}}(\{1, 2, 3, \ldots\})$.

A sequence $\{s_n\}$ of real numbers is said to be **bounded** if there is some $M > 0$ such that $|s_n| < M$ for all n. Let $\ell_\infty \subset \mathcal{F}_{\mathbb{R}}(\{1, 2, 3, \ldots\})$ be the space of all bounded sequences of real numbers. Since a sum of two bounded sequences is bounded, and any scalar product of a bounded sequence is bounded (hint: the quantities $M_1 + M_2$ and αM figure in the proofs) ℓ_∞ is a vector space over $\mathbb{R}$.

The sequence $\{s_n\}$ is said to be **summable** if

$$|s_1| + |s_2| + \cdots < \infty.$$

Let ℓ_1 be the space of summable sequences. It is obvious that any scalar multiple of a summable sequence is summable, and that the sum of two summable sequences is summable. (Again, if the way to prove these facts isn't obvious, please stop and think about it, and this time make up your own hint.) Therefore ℓ_1 is a vector space over $\mathbb{R}$.

The sequence $\{s_n\}$ is said to be **square summable** if

$$s_1^2 + s_2^2 + \cdots < \infty.$$

Let ℓ_2 be the space of square summable sequences. As above, it is obvious that any scalar multiple of a square summable sequence is square summable. To see that the sum of two square summable sequences $\{s_n\}$

and $\{t_n\}$ is square summable we observe that, for each $n = 1, 2, \ldots$, the triangle inequality for the norm $\|\cdot\|_2$ gives

$$\big((s_1+t_1)^2+\cdots+(s_n+t_n)^2\big)^{\frac{1}{2}} \le \big(s_1^2+\cdots+s_n^2\big)^{\frac{1}{2}} + \big(t_1^2+\cdots+t_n^2\big)^{\frac{1}{2}}.$$

Since the right hand side is bounded as $n \to \infty$, the same is true for the left hand side. If a sequence $\{s_n\}$ is summable, then it is square summable because $|s_n| < 1$, and thus $s_n^2 < |s_n|$, for all sufficiently large n. Therefore $\ell_1 \subset \ell_2$. We now have the following vector spaces over $\mathbb{R}$:

$$\ell_1 \subset \ell_2 \subset \ell_\infty \subset \mathcal{F}_{\mathbb{R}}(\{1,2,3,\ldots\}).$$

Evidently a vector space can have additional structure over and above the vector operations. For example, $k[X_1, \ldots, X_n]$ is a ring, and a vector space can be endowed with an inner product, or a metric, or a topology. The last idea is particularly important: a **topological vector space** is a vector space over $\mathbb{R}$ endowed with a Hausdorff topology that makes the vector operations continuous. This definition is the starting point of the subfield of mathematics called functional analysis, which studies topological vector spaces, of various sorts, and functions between them. Functional analysis has grown enormously over the last several decades, and has pretty much swallowed most of what used to be called analysis. Roughly, analysts consider issues having to do with functions, e.g., in what senses might we speak of a sequence of functions converging, and under what circumstances is convergence guaranteed. For the most part the functions in question can be thought of as elements of vector spaces, and many of the relevant concepts, such as convergence, can be expressed topologically. This point of view helps solve some problems, and often provides a unifying perspective. Since it usually doesn't do any damage, over time the tendency has been in the direction of adopting the "functional analytic viewpoint" as a matter of course.

4.2 Bases and Dimension

What should our agenda be? At a minimum we have to describe the parts of the theory that we will need later, and it would be desirable to convey other important information. "Important" here means "important, as revealed by the experience of working mathematicians," and the reader is currently in no position to appreciate this. Perhaps for this reason, the approach of most books is to simply lay out a bunch of definitions and theorems without much auxiliary explanation, hoping that the reader will eventually come to appreciate the appropriateness of the material through her own experience.

An alternative is to pursue the subject with certain objectives in mind. These may be fictional in a sense, representing neither the actually historical development of the ideas nor a comprehensive approach to what the reader "needs" to know. But these defects can be addressed later, and in the meantime this approach endows our work with some tangible sense of purpose.

What might it mean to "understand" vector spaces, or any mathematical object for that matter? There are many reasonable or valid answers to this question, but a rather minimal criterion of "understanding" is that we should be able to tell whether two vector spaces are "the same" or "different."

As we mentioned in Chapter 1, in any category there is the concept of isomorphism: a morphism $f : X \to Y$ is an **isomorphism** if there is a morphism $g : Y \to X$ such that $g \circ f = \mathrm{Id}_X$ and $f \circ g = \mathrm{Id}_Y$. In this circumstance we say that the objects X and Y are **isomorphic**. In their lives beyond the category, X and Y might be as different as apples and oranges, but as far as the category is concerned, X is "just like" Y and vice versa. An attribute of objects in the category is **invariant under isomorphism**, or an **isomorphism invariant**, or simply an **invariant**, if isomorphic objects always have the same value of the attribute. A **complete set of invariants** is a collection of invariants that **classify** the objects up to invariance, in the sense that for any two objects X and Y that are *not* isomorphic, there is at least one invariant whose value for X is different from its value for Y.

Perhaps this sounds like completely unmotivated gobbledygook. Well, all I can say is that, after a certain amount of experience, you may well come to appreciate how the last paragraph expresses, succinctly and directly, an important aspect of mathematical thought. In the meantime here is the main idea pursued below, described more concretely: *the concept of dimension classifies finite dimensional vector spaces over k.* That is, two finite dimensional vector spaces over k are isomorphic if and only if they have the same dimension.

So, we need to define the concept of dimension. The intuition here is completely familiar: the dimension of a space is the number of numbers required to describe a point. The line $\mathbb{R}$ is one dimensional, a point in the plane is described by two numbers, a point in space by three numbers, and so forth. But in order to attain a coordinate-free expression of this idea we need a framework in which the numbers that determine a point in k^n may be different from its components as an element of $k^n = k \times \cdots \times k$.

The **standard unit basis** of k^n is $\mathbf{e}_1, \ldots, \mathbf{e}_n$ where, for each $i = 1, \ldots, n$,

$$\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 1)$$

is the element of k^n whose i^{th} component is 1 and whose other components are all 0. Any $v = (v_1, \ldots, v_n) \in k^n$ can be written as

$$v = v_1\mathbf{e}_1 + \cdots + v_n\mathbf{e}_n.$$

Moreover, there is only one way to write v as such an expression in the sense that if, in addition to this equation, we also have $v = v_1'\mathbf{e}_1 + \cdots + v_n'\mathbf{e}_n$, then necessarily $v_1' = v_1, \ldots, v_n' = v_n$.

In order to give a more conceptual explanation of what is going on here we need to expand our terminology. Fix a vector space V over k. For any set $S \subset V$, a **linear combination** of elements of S is an expression of the form $\sum_{s \in S} \alpha_s s$ in which each α_s is an element of k and there are only finitely many s such that $\alpha_s \neq 0$. The gist of the last paragraph is that each element of k^n is expressed in one and only one way as a linear combination of $\mathbf{e}_1, \ldots, \mathbf{e}_n$. In this sense $\mathbf{e}_1, \ldots, \mathbf{e}_n$ impose a coordinate system on k^n. This coordinate systems happens to coincide with the coordinate system given by the definition of k^n, but the general method leads to other coordinate systems as well.

Definition 4.2. *A* ***basis*** *for V is a set $B \subset V$ such that each element of V can be expressed in a unique way as a linear combination of the elements of B.*

For a concrete example, consider $\mathbf{b}_1 = (2, 1)$ and $\mathbf{b}_2 = (1, 2)$. We would like to show that $\mathbf{b}_1, \mathbf{b}_2$ is a basis of k^2. (According to the definition, "$\{\mathbf{b}_1, \mathbf{b}_2\}$ is a basis of k^2" is the logically exact way to say this, but in real life everybody leaves out the '{' and '}'.) This boils down to the assertion that for any $v = (v_1, v_2) \in k^2$ the system of equations

$$2\alpha_1 + \alpha_2 = v_1, \quad \alpha_1 + 2\alpha_2 = v_2$$

has a unique solution. A direct calculation verifies that

$$\alpha_1 = \frac{2v_1 - v_2}{3}, \quad \alpha_2 = \frac{-v_1 + 2v_2}{3}$$

is a solution[1]. To see that it is unique, suppose that α_1' and α_2' is another solution. Subtracting the system of equations above from

$$2\alpha_1' + \alpha_2' = v_1, \quad \alpha_1' + 2\alpha_2' = v_2$$

[1] In order for this calculation to be valid, the characteristic of k must be different from 3. If the characteristic of k is 3, then $\mathbf{b}_2 = -\mathbf{b}_1$ and $\mathbf{b}_1, \mathbf{b}_2$ actually isn't a basis.

gives

$$2(\alpha_1' - \alpha_1) + (\alpha_2' - \alpha_2) = 0, \ \ (\alpha_1' - \alpha_1) + 2(\alpha_2' - \alpha_2) = 0.$$

Putting $\alpha_1' - \alpha_1$ and $\alpha_2' - \alpha_2$ on opposite sides in both equations, then combining, leads to

$$\alpha_1' - \alpha_1 = -2(\alpha_2' - \alpha_2) = 4(\alpha_1' - \alpha_1).$$

Unless the characteristic of k is 3 this implies that $\alpha_1' - \alpha_1 = 0$, and a symmetric argument shows that $\alpha_2' - \alpha_2 = 0$.

In this example, and in general, a basis B imposes a **coordinate system** on V: for each $v \in V$ there is a unique system of scalars $\alpha_{\mathbf{b}}$ for $\mathbf{b} \in B$, only finitely many of which are nonzero, such that $v = \sum_{\mathbf{b} \in B} \alpha_{\mathbf{b}} \mathbf{b}$. Many calculations and arguments can be simplified by choosing an appropriate basis, even when the underlying vector space is k^n.

You can probably sense where this is headed: if a vector space has a finite basis, then all bases have the same finite number of elements, and we can define the **dimension** of the space to be the number of elements in any basis. To facilitate our explanation we introduce terminology corresponding to the two properties of a basis.

Definition 4.3. *A set of vectors $S \subset V$ is **linearly independent** if the only linear combination $\sum_{s \in S} \alpha_s s$ equal to 0 has $\alpha_s = 0$ for all s.*

Definition 4.4. *The **span** of a set of vectors $S \subset V$ is the set of all linear combinations of the elements of S.*

The definition of a basis for V states that a basis B spans V, and that it is linearly independent because there can be only one way to represent 0 as a linear combination of the elements of B. The converse is also true: if B is linearly independent and spans V, then it is a basis, and the only thing we need to show to prove it is that *every* element of V (not just 0) has only one representation as a linear combination of elements of B. But if some v can be represented as a linear combination of the elements of B in two different ways, say

$$\sum_{\mathbf{b} \in B} \alpha_{\mathbf{b}} \mathbf{b} = v = \sum_{\mathbf{b} \in B} \alpha_{\mathbf{b}}' \mathbf{b},$$

then subtraction gives

$$\sum_{\mathbf{b} \in B} (\alpha_{\mathbf{b}}' - \alpha_{\mathbf{b}}) \mathbf{b} = 0,$$

which is a violation of linear independence.

We should make a remark here about the space $\{0\}$. Specifically, we follow a convention according to which there is something called the "null sum" that has no terms and whose value is 0. Thus 0 is in the span of every subset of $\{0\}$, *including* $\emptyset$, and in fact $\emptyset$ is the unique basis of $\{0\}$.

The proof of the next result is long and "algebra intensive," but you should study it carefully, both because the result is very important and because it expresses a fundamental idea known as **Gaussian elimination**. (The material in this chapter is mostly straightforward, but even so Gauss manages to get a mention!) This is the basic idea that is applied in most concrete calculations in linear algebra: use one of the equations to express one of the variables in terms of the others, then substitute this into the other equations, thereby achieving a system with one fewer equation and one fewer unknown.

Lemma 4.5. *If S and T are linearly independent subsets of V with T finite, say $T = \{t_1, \dots, t_n\}$, and the span of S is contained in the span of T, then S has at most n elements.*

Proof. The result in the case $n = 0$ is clear: the span of $T = \emptyset$ is $\{0\}$, and if the span of S is also $\{0\}$, and S is linearly independent, then $S = \emptyset$. Arguing inductively, we may assume that the result has already been established with $n-1$ in place of n. Aiming at a contradiction, suppose that S has more than n elements, and choose some particular list $s_1, \dots, s_{n+1}$ of $n+1$ elements.

Since the span of T contains the span of S, which in turn contains S itself, there are scalars $\alpha_{i,j}$ such that

$$\begin{aligned} s_1 &= \alpha_{1,1}t_1 + \cdots + \alpha_{1,n}t_n, \\ &\vdots \\ s_n &= \alpha_{n,1}t_1 + \cdots + \alpha_{n,n}t_n, \\ s_{n+1} &= \alpha_{n+1,1}t_1 + \cdots + \alpha_{n+1,n}t_n. \end{aligned}$$

If $\alpha_{1,n}, \dots, \alpha_{n+1,n}$ were all zero the equations would involve only $t_1, \dots, t_{n-1}$, and we could apply the induction hypothesis. Therefore we may assume that this at least one of these scalars is not zero, and after reindexing, if necessary, it will be the case that $\alpha_{n+1,n} \neq 0$. Then

$$t_n = \frac{1}{\alpha_{n+1,n}}\big(s_{n+1} - \sum_{j=1}^{n-1} \alpha_{n+1,j}t_j\big).$$

Substituting this into the first n equations gives

$$\begin{aligned} s_1 - \frac{\alpha_{1,n}}{\alpha_{n+1,n}} s_{n+1} &= \alpha'_{1,1} t_1 + \dots + \alpha'_{1,n-1} t_{n-1}, \\ &\vdots \\ s_n - \frac{\alpha_{n,n}}{\alpha_{n+1,n}} s_{n+1} &= \alpha'_{n,1} t_1 + \dots + \alpha'_{n,n-1} t_{n-1}, \end{aligned}$$

where

$$\alpha'_{i,j} = \alpha_{i,j} - \frac{\alpha_{n,j}}{\alpha_{n+1,n}} \alpha_{i,n}$$

for all $i = 1, \dots, n$ and $j = 1, \dots, n-1$. This is a system of the same sort with n replaced by $n-1$, so the vectors $s_i - (\alpha_{i,n}/\alpha_{n+1,n}) s_{n+1}$ cannot be linearly independent and there must be scalars $\beta_1, \dots, \beta_n$, not all of which are zero, such that

$$\beta_1 \bigl(s_1 - \frac{\alpha_{1,n}}{\alpha_{n+1,n}} s_{n+1}\bigr) + \dots + \beta_n \bigl(s_n - \frac{\alpha_{n,n}}{\alpha_{n+1,n}} s_{n+1}\bigr) = 0.$$

This is a linear dependence of $s_1, \dots, s_{n+1}$ if the coefficient

$$\beta_1 \frac{\alpha_{1,n}}{\alpha_{n+1,n}} + \dots + \beta_n \frac{\alpha_{n,n}}{\alpha_{n+1,n}}$$

of s_{n+1} is nonzero, and otherwise it reduces to the linear dependence $\beta_1 s_1 + \dots + \beta_n s_n = 0$. Either way there is a contradiction, so the proof is complete. □

We now have the tools we need to give a precise explanation of dimension.

Theorem 4.6. *If V has a finite basis, then any two bases have the same number of elements.*

Proof. Let $B := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ be a finite basis of V, and let C be another basis. Since C is linearly independent and its span is contained in the span of B, Lemma 4.5 implies that C has at most n elements. In particular, C is finite, so now we can apply the same argument with B and C reversed, finding that C has at least n elements. □

Definition 4.7. *A vector space V is **finite dimensional** if it has a finite basis; otherwise V is **infinite dimensional**. The **dimension** of a finite dimensional vector space is the number of elements of any one of its bases.*

Theorem 4.6 is the real gist of the classification of finite dimensional vector spaces, but before we can state and prove the classification result formally we need to define the appropriate concept of isomorphism. This will be done in the next section.

So far our definitions have avoided the issue of whether vector spaces actually have bases, in a rather sneaky way: a finite dimensional vector space has a basis, simply because that's part of what it means to be finite dimensional. But we haven't yet proved some rather important and obvious things that ought to be true. The next result handles most of these issues.

Lemma 4.8. *Suppose that S is a linearly independent subset of V. If v is a point in V that is not contained in the span of S, then $S \cup \{v\}$ is linearly independent. If V is finite dimensional, then S is a subset of a basis of V. If V is infinite dimensional, then S is a subset of an infinite linearly independent set.*

Proof. Aiming at a contradiction, suppose that there is a linear dependence:

$$0 = \alpha v + \sum_{s \in S} \beta_s s,$$

where at least one of the coefficients is nonzero. We cannot have $\alpha = 0$ because S is linearly independent, but if $\alpha \neq 0$, then we could solve for v, arriving at

$$v = -\sum_{s \in S} \frac{\beta_s}{\alpha} s.$$

Since v is not spanned by S, this is impossible. We have shown that $S \cup \{v\}$ is linearly independent.

Starting with S, consider repeatedly adding points that are not already spanned. If V is finite dimensional this process must arrive at a basis after finitely many steps because no linearly independent set can have more elements than one of V's bases. If V is infinite dimensional, and S is not already infinite, the process cannot halt after finitely many steps, because it can only halt at a basis. The result of continuing it indefinitely is an infinite linearly independent set. □

Actually, a linearly independent set S is always a subset of a basis, even if V is not finite dimensional. The proof involves a technique called "transfinite induction" that extends the idea in the proof above: we go through the elements of $V \setminus S$ "one by one," adding the ones that have not already been spanned to our prospective basis. At the end of this process the prospective basis is still linearly independent, and every point is either in it or spanned by it, so it is in fact a basis. This makes perfect sense when V is countable, since, by definition, countability means that we can write down a list $v_1, v_2, \ldots$ that has all the elements of V. For uncountable V there is a sophisticated application of the axiom of choice that allows one to "line up" the elements of V in a suitable ordering.

4.3 Linear Transformations

By this point it should be almost instinctive: given a collection of objects (in this case, vector spaces) we should expect a corresponding concept of morphism.

Definition 4.9. *If V and W are vector spaces over k, a* ***linear transformation*** *from V to W is a function $\ell : V \to W$ satisfying*

$$\ell(v + v') = \ell(v) + \ell(v') \quad \textit{and} \quad \ell(\alpha v) = \alpha\ell(v)$$

for all $v, v' \in V$ and all $\alpha \in k$.

That is, a linear transformation is just a k-module homomorphism. Since, for any commutative ring with unit R, there is a category whose objects are R-modules and whose morphisms are R-module homomorphisms, *we already know that k-vector spaces and linear transformations constitute a category.*

Pretty much everything we need to know about linear transformations can be boiled down to the relationship between linear transformations and bases. In this sense the next three results are the technical underpinning of the entire theory of finite dimensional linear algebra. Unfortunately, this will be a patch of rather dry reading, with simple lemmas proved by prosaic arguments. If it is any consolation, beyond this level you will never have to see these arguments again, since in all "more advanced" mathematical literature these facts are invariably taken for granted.

A linear transformation can be defined by specifying the image of a basis.

Lemma 4.10. *If V and W are vector spaces over k, $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is a basis of V, and $w_1, \ldots, w_n \in W$, then there is a unique linear transformation $\ell : V \to W$ such that*

$$\ell(\mathbf{b}_1) = w_1, \ldots, \ell(\mathbf{b}_n) = w_n.$$

Proof. In view of the requirements that define a linear transformation, if ℓ is a linear transformation satisfying the given condition, then for any $\alpha_1, \ldots, \alpha_n$ it must be the case that

$$\begin{aligned}\ell(\alpha_1\mathbf{b}_1 + \cdots + \alpha_n\mathbf{b}_n) &= \ell(\alpha_1\mathbf{b}_1) + \cdots + \ell(\alpha_n\mathbf{b}_n)\\ &= \alpha_1\ell(\mathbf{b}_1) + \cdots + \alpha_n\ell(\mathbf{b}_n)\\ &= \alpha_1 w_1 + \cdots + \alpha_n w_n.\end{aligned}$$

Since each element of V can be written as a linear combination of the elements of $\mathbf{b}_1, \ldots, \mathbf{b}_n$ in exactly one way, this formula defines the function ℓ unambiguously.

We still have to show that, in fact, this function is linear. Consider $v, v' \in V$ and $\beta \in k$. There are scalars $\alpha_1, \ldots, \alpha_n, \alpha'_1, \ldots, \alpha'_n$ such that

$$v = \alpha_1\mathbf{b}_1 + \cdots + \alpha_n\mathbf{b}_n \text{ and } v' = \alpha'_1\mathbf{b}_1 + \cdots + \alpha'_n\mathbf{b}_n.$$

That ℓ satisfies the conditions in the definition of a linear transformation is confirmed by the calculations

$$\begin{aligned}\ell(v + v') &= \ell\big((\alpha_1 + \alpha'_1)\mathbf{b}_1 + \cdots + (\alpha_n + \alpha'_n)\mathbf{b}_n\big)\\ &= (\alpha_1 + \alpha'_1)w_1 + \cdots + (\alpha_n + \alpha'_n)w_n\\ &= (\alpha_1 w_1 + \cdots + \alpha_n w_n) + \cdots + (\alpha'_1 w_1 + \cdots + \alpha'_n w_n)\\ &= \ell(v) + \ell(v')\end{aligned}$$

and

$$\begin{aligned}\ell(\beta v) &= \ell\big(\beta(\alpha_1\mathbf{b}_1 + \cdots + \alpha_n\mathbf{b}_n)\big)\\ &= \ell\big((\beta\alpha_1)\mathbf{b}_1 + \cdots + (\beta\alpha_n)\mathbf{b}_n\big)\\ &= (\beta\alpha_1)w_1 + \cdots + (\beta\alpha_n)w_n\\ &= \beta(\alpha_1 w_1 + \cdots + \alpha_n w_n)\\ &= \beta\ell(v).\end{aligned}$$

□

The following notation will often be useful. If $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a basis of V and $w_1, \ldots, w_n \in W$, let

$$[w_1, \ldots, w_n]_{\mathbf{b}} : V \to W$$

be the linear transformation satisfying $[w_1, \ldots, w_n]_{\mathbf{b}}(\mathbf{b}_i) = w_i$ for $i = 1, \ldots, n$. Every linear transformation $\ell : V \to W$ whose domain is finite dimensional can be written in this way because for any basis $\mathbf{b}$ we have

$$\ell = [\ell(\mathbf{b}_1), \ldots, \ell(\mathbf{b}_n)]_{\mathbf{b}}.$$

The injectivity (surjectivity) of a linear transformation can be diagnosed from the linear independence (span) of the image of a basis.

Proposition 4.11. *If V and W are vector spaces over k, $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a basis of V, $w_1, \ldots, w_n \in W$, and $\ell = [w_1, \ldots, w_n]_{\mathbf{b}}$, then:*

(a) ℓ is injective if and only if $w_1, \ldots, w_n$ are linearly independent;

(b) ℓ is surjective if and only if $w_1, \ldots, w_n$ span W.

Proof. (a) If ℓ is not injective then $\ell(v) = \ell(v')$ for some distinct $v, v' \in V$. There are scalars $\alpha_1, \ldots, \alpha_n$ such that

$$v' - v = \alpha_1 \mathbf{b}_1 + \cdots + \alpha_n \mathbf{b}_n,$$

and since $v' - v \neq 0$, $\alpha_i \neq 0$ for some i. But then

$$0 = \ell(v') - \ell(v) = \ell(v' - v) = \alpha_1 w_1 + \cdots + \alpha_n w_n,$$

which shows that $w_1, \ldots, w_n$ are not linearly independent. Conversely, if $w_1, \ldots, w_n$ are not linearly independent, then $0 = \alpha_1 w_1 + \cdots + \alpha_n w_n$ for some scalars $\alpha_1, \ldots, \alpha_n$, not all of which are zero, so $\alpha_1 \mathbf{b}_1 + \cdots + \alpha_n \mathbf{b}_n \neq 0$, but

$$\ell(0) = 0 = \ell(\alpha_1 \mathbf{b}_1 + \cdots + \alpha_n \mathbf{b}_n).$$

(b) The image of ℓ is the set of points of the form $\alpha_1 w_1 + \cdots + \alpha_n w_n$, which is, by definition, the span of $w_1, \ldots, w_n$. □

Combining the two parts of this result, if V is finite dimensional and $\ell : V \to W$ is a linear transformation, then ℓ is a bijection if and only if ℓ maps any basis of V to a basis of W. In this case it is an isomorphism because in Section 2.5 we showed that for any commutative ring with unit R, if φ is a bijective R-module homomorphism, then so is φ^{-1}.

Now, at long last, we can see that dimension classifies finite dimensional vector spaces.

Theorem 4.12. *Two finite dimensional vector spaces V and W are linearly isomorphic if and only if they have the same dimension.*

Proof. Let $\mathbf{b} = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ be a basis of V. If $\ell : V \to W$ is a linear isomorphism, then Proposition 4.11 implies that $\ell(\mathbf{b}_1), \ldots, \ell(\mathbf{b}_n)$ is a basis of W, so W is n-dimensional. For the converse suppose that W is n-dimensional, and let $\mathbf{c}_1, \ldots, \mathbf{c}_n$ be a basis. Proposition 4.11 implies that $[\mathbf{c}_1, \ldots, \mathbf{c}_n]_{\mathbf{b}}$ is a linear bijection, so it is a linear isomorphism by the result above. □

4.4 Linear Subspaces

Taking the classification of finite dimensional vector spaces as our goal, in the last two sections we were led along a rather natural path that involved a number of definitions and results that, it turns out, are of great general importance. What next? Well, a natural impulse is to try to classify linear transformations. We will say that two linear transformations

$$\ell : V \to W \quad \text{and} \quad \ell' : V' \to W'$$

are **equivalent** if there are isomorphisms

$$\iota_V : V \to V' \quad \text{and} \quad \iota_W : W \to W'$$

such that $\ell' = \iota_W \circ \ell \circ \iota_V^{-1}$, or, equivalently, $\ell' \circ \iota_V = \iota_W \circ \ell$, or, equivalently, $\ell = \iota_W^{-1} \circ \ell' \circ \iota_V$. In order to lay out the information in a way that allows the eye to take it all in in one fell swoop, mathematicians like to express an equation like $\ell' \circ \iota_V = \iota_W \circ \ell$ by saying that "the diagram

$$\begin{tikzcd} V \arrow[r, "\ell"] \arrow[d, "\iota_V"'] & W \arrow[d, "\iota_W"] \\ V' \arrow[r, "\ell'"] & W' \end{tikzcd}$$

commutes."

As always, we have to verify that 'equivalence' is actually an equivalence relation, and, as usual, this is quite simple. The isomorphisms Id_V and Id_W can obviously be used to show that ℓ is equivalent to itself, and in the setting of the last paragraph the isomorphisms ι_V^{-1} and ι_W^{-1} can be used to show that ℓ' is equivalent to ℓ. To demonstrate transitivity, suppose that, in addition, ℓ' is equivalent to ℓ'' by virtue of the commutative diagram

$$\begin{tikzcd} V' \arrow[r, "\ell'"] \arrow[d, "\iota_{V'}"'] & W' \arrow[d, "\iota_{W'}"] \\ V'' \arrow[r, "\ell''"] & W'' \end{tikzcd}$$

in which $\iota_{V'}$ and $\iota_{W'}$ are isomorphisms. Then $\iota_{V'} \circ \iota_V$ and $\iota_{W'} \circ \iota_W$ are isomorphisms, and ℓ is equivalent to ℓ'' because

$$\iota_{W'} \circ \iota_W \circ \ell = \iota_{W'} \circ \ell' \circ \iota_V = \ell'' \circ \iota_{V'} \circ \iota_V .$$

In our general description of classification, the things being classified were the objects of some category. It is not particularly difficult to create a category of linear transformations in which two linear transformations are isomorphic if and only if they are equivalent in this sense, and in which our characterization of equivalence fits the general paradigm of classification described earlier, but the construction is a bit artificial, and we won't bother. And in fact the idea of classification isn't really restricted to isomorphism in a category, but makes sense for any equivalence relation. That is, given a collection of objects, some attributes of these objects, and an equivalence relation, if equivalent objects have the same values of all attributes, and any two inequivalent objects have different values of at least one attribute, then the attributes are said to **classify the objects up to equivalence**.

Eventually we'll find that the linear transformations between finite dimensional spaces are classified by the dimensions of the domain, the range, and the image. In order to explain this properly we need to develop some basic facts about linear subspaces.

Definition 4.13. *If V is a vector space over k, a nonempty set $P \subset V$ is a **linear subspace** if it is "closed" under addition and scalar multiplication:*

(a) $v + v' \in P$ whenever $v, v' \in P$.

(b) $\alpha v \in P$ whenever $v \in P$ and $\alpha \in k$.

That is, a linear subspace is just a submodule of the k-module V, and general properties of submodules hold here as well. For example, the restrictions to P of the vector operations satisfy all the conditions defining a vector space, so P is itself a vector space over k.

It is important to have a clear visual sense of what it means to be a linear subspace. Note that $\{0\}$ is always a linear subspace of V, and V is always a linear subspace of itself. In addition to $\{0\}$ and $\mathbb{R}^2$ itself, the linear subspaces of $\mathbb{R}^2$ are the lines through the origin. In addition to $\{0\}$ and $\mathbb{R}^3$ itself, the linear subspaces of $\mathbb{R}^3$ are:

(a) the lines through the origin;

(b) the two dimensional planes containing the origin.

The span of any $S \subset V$ is a linear subspace of V because a) the sum of two linear combinations of the elements of S is a linear combination of the elements of S, and b) a scalar multiple of a linear combination of the elements of S is a linear combination of the elements of S. In Chapter 2 we saw that for any commutative ring R, intersections and sums of submodules of an R-module are submodules, so if P and P' are linear subspaces of V, then so are $P \cap P'$ and

$$P + P' := \{ v + v' : v \in P \text{ and } v' \in P' \}.$$

We say that P and P' are **complementary subspaces** of V if $P \cap P' = \{0\}$ and $P + P' = V$.

Lemma 4.14. *If P and P' are finite dimensional complementary subspaces of V, then*

$$\dim V = \dim P + \dim P'.$$

Proof. Suppose $\mathbf{b}_1, \ldots, \mathbf{b}_p$ is a basis of P and $\mathbf{b}_{p+1}, \ldots, \mathbf{b}_{p+p'}$ is a basis of P'. Then $\mathbf{b}_1, \ldots, \mathbf{b}_{p+p'}$ span V because they span P and P' and $P+P' = V$. If $\alpha_1 \mathbf{b}_1 + \cdots + \alpha_{p+p'} \mathbf{b}_{p+p'} = 0$, then

$$\alpha_1 \mathbf{b}_1 + \cdots + \alpha_p \mathbf{b}_p = -\alpha_{p+1} \mathbf{b}_{p+1} - \cdots - \alpha_{p+p'} \mathbf{b}_{p+p'} \in P \cap P' = \{0\},$$

so $\alpha_1, \ldots, \alpha_p$ are all zero because $\mathbf{b}_1, \ldots, \mathbf{b}_p$ are linear independent and $\alpha_{p+1}, \ldots, \alpha_{p+p'}$ are all zero because $\mathbf{b}_{p+1}, \ldots, \mathbf{b}_{p+p'}$ are linearly independent. This shows that $\mathbf{b}_1, \ldots, \mathbf{b}_{p+p'}$ are linearly independent, so they are a basis of V. Consequently $\dim V = p + p'$. □

Lemma 4.15. *Suppose that V is a finite dimensional vector space and P is a linear subspace. Then there is a linear subspace P' such that P and P' are complementary subspaces of V.*

Proof. Using Lemma 4.8, we see that P must be finite dimensional because an infinite linearly independent set in P would also be linearly independent in V, contrary to our assumption that V is finite dimensional. Lemma 4.8 also implies that P has a basis $\mathbf{b}_1, \ldots, \mathbf{b}_p$, and that we can extend it to a basis $\mathbf{b}_1, \ldots, \mathbf{b}_n$ of V by choosing suitable $\mathbf{b}_{p+1}, \ldots, \mathbf{b}_n$. Let P' be the span of $\mathbf{b}_{p+1}, \ldots, \mathbf{b}_n$. To see that $P \cap P' = \{0\}$, consider that if

$$\alpha_1 \mathbf{b}_1 + \cdots + \alpha_p \mathbf{b}_p = \alpha_{p+1} \mathbf{b}_{p+1} + \cdots + \alpha_n \mathbf{b}_n,$$

then

$$\alpha_1 \mathbf{b}_1 + \cdots + \alpha_p \mathbf{b}_p - \alpha_{p+1} \mathbf{b}_{p+1} - \cdots - \alpha_n \mathbf{b}_n = 0,$$

so that $\alpha_1, \ldots, \alpha_n$ are all zero because the basis is a linearly independent set. Of course $P + P' = V$ because $\mathbf{b}_1, \ldots, \mathbf{b}_n$ spans V. □

If $\ell : V \to W$ is a linear transformation, its **kernel** is

$$\ker(\ell) := \ell^{-1}(0) = \{\, v \in V : \ell(v) = 0 \,\} \subset V,$$

and its **image** is

$$\mathrm{image}(\ell) := \ell(V) = \{\, \ell(v) : v \in V \,\} \subset W.$$

These are linear subspaces of V and W because (as we explained in Chapter 2) whenever R is a commutative ring with unit, the kernel and image of any R-module homomorphism are submodules of the domain and range respectively. If V is finite dimensional, then the dimension of the image of ℓ is called the **rank** of ℓ, and is denoted by $\mathrm{rank}(\ell)$, while the dimension of the kernel of ℓ is called the **nullity**, and is denoted by $\mathrm{null}(\ell)$.

Theorem 4.16 (Rank-Nullity Theorem)**.** *If $\ell : V \to W$ is a linear transformation and V is finite dimensional, then* $\mathrm{rank}(\ell) + \mathrm{null}(\ell) = \dim V$.

Proof. As we saw in the proof of the last result, there is a basis $\mathbf{b}_1, \dots, \mathbf{b}_n$ of V with $\mathbf{b}_1, \dots, \mathbf{b}_p$ a basis of $\ker(\ell)$. We claim that $\ell(\mathbf{b}_{p+1}), \dots, \ell(\mathbf{b}_n)$ is a basis of $\mathrm{image}(\ell)$. Of course this collection spans $\mathrm{image}(\ell)$ because $\ell(\mathbf{b}_1), \dots, \ell(\mathbf{b}_n)$ spans $\mathrm{image}(\ell)$ and $\ell(\mathbf{b}_1) = \cdots = \ell(\mathbf{b}_p) = 0$. This collection is linearly independent because if $\alpha_{p+1}\ell(\mathbf{b}_{p+1}) + \cdots + \alpha_n\ell(\mathbf{b}_n) = 0$, then

$$0 = \ell(\alpha_{p+1}\mathbf{b}_{p+1}) + \cdots + \ell(\alpha_n\mathbf{b}_n) = \ell(\alpha_{p+1}\mathbf{b}_{p+1} + \cdots + \alpha_n\mathbf{b}_n),$$

so $\alpha_{p+1}\mathbf{b}_{p+1} + \cdots + \alpha_n\mathbf{b}_n \in \ker(\ell)$ and consequently $\alpha_{p+1} = \cdots = \alpha_n = 0$. □

The result classifying linear transformations up to equivalence is:

Theorem 4.17. *If V, W, V', and W' are finite dimensional vector spaces over k, then two linear transformations $\ell : V \to W$ and $\ell' : V' \to W'$ are equivalent if and only if:*

(a) $\dim V = \dim V'$,

(b) $\dim W = \dim W'$,

(c) $\mathrm{null}(\ell) = \mathrm{null}(\ell')$, *and*

(d) $\mathrm{rank}(\ell) = \mathrm{rank}(\ell')$.

Proof. First suppose that ℓ and ℓ' are equivalent, so that there are linear isomorphisms $\iota_V : V \to V'$ and $\iota_W : W \to W'$ such that $\iota_W \circ \ell = \ell' \circ \iota_V$. Then $\dim V = \dim V'$ and $\dim W = \dim W'$ because isomorphic finite dimensional vector spaces have the same dimension. In addition, ι_V restricts to a bijection between the kernel of ℓ' and the kernel of $\ell' \circ \iota_v$, and the kernel of $\iota_W \circ \ell$ is just the kernel of ℓ, so

$$\ker(\ell') = \iota_V(\ker(\ell' \circ \iota_V)) = \iota_V(\ker(\iota_W \circ \ell)) = \iota_V(\ker(\ell)).$$

Therefore $\ker(\ell)$ and $\ker(\ell')$ are isomorphic and consequently have the same dimension. We have shown that (a), (b), and (c) hold, and in this circumstance the rank-nullity theorem implies that (d) also holds.

Now suppose that $\dim V = \dim V'$, $\dim W = \dim W'$, and $\mathrm{null}(\ell) = \mathrm{null}(\ell')$. There are only the most obvious constraints on the construction of the desired isomorphisms. Lemma 4.15 implies that V and V' have linear subspaces that are complementary to $\ker(\ell)$ and $\ker(\ell')$ respectively; let $\mathbf{b}_1, \ldots, \mathbf{b}_{m-p}$ and $\mathbf{b}'_1, \ldots, \mathbf{b}'_{m-p}$ be bases of these subspaces. Let $\mathbf{b}_{m-p+1}, \ldots, \mathbf{b}_m$ and $\mathbf{b}'_{m-p+1}, \ldots, \mathbf{b}'_m$ be bases of $\ker(\ell)$ and $\ker(\ell')$. As we have seen before, $\mathbf{b}_1, \ldots, \mathbf{b}_m$ and $\mathbf{b}'_1, \ldots, \mathbf{b}'_m$ are bases of V and V'.

For $i = 1, \ldots, m-p$ let $\mathbf{c}_i := \ell(\mathbf{b}_i)$ and $\mathbf{c}'_i := \ell(\mathbf{b}'_i)$. The restriction of ℓ to the span of $\mathbf{b}_1, \ldots, \mathbf{b}_{m-p}$ is injective, so $\mathbf{c}_1, \ldots, \mathbf{c}_{m-p}$ are linearly independent. Similarly, $\mathbf{c}'_1, \ldots, \mathbf{c}'_{m-p}$ are linearly independent. Applying Lemma 4.8, there exist $\mathbf{c}_{m-p+1}, \ldots, \mathbf{c}_n$ and $\mathbf{c}'_{m-p+1}, \ldots, \mathbf{c}'_n$ such that $\mathbf{c}_1, \ldots, \mathbf{c}_n$ and $\mathbf{c}'_1, \ldots, \mathbf{c}'_n$ are bases of W and W'. Let

$$\iota_V := [\mathbf{b}'_1, \ldots, \mathbf{b}'_m]_{\mathbf{b}} : V \to V' \quad \text{and} \quad \iota_W := [\mathbf{c}'_1, \ldots, \mathbf{c}'_n]_{\mathbf{c}} : W \to W'.$$

Proposition 4.11 implies that ι_V and ι_W are isomorphisms. Finally observe that

$$\ell = [\mathbf{c}_1, \ldots, \mathbf{c}_{m-p}, 0, \ldots, 0]_{\mathbf{b}} \quad \text{and} \quad \ell' = [\mathbf{c}'_1, \ldots, \mathbf{c}'_{m-p}, 0, \ldots, 0]_{\mathbf{b}'},$$

so that

$$\iota_W \circ \ell = [\mathbf{c}'_1, \ldots, \mathbf{c}'_{m-p}, 0, \ldots, 0]_{\mathbf{b}} = \ell' \circ \iota_V.$$

□

Let's be completely concrete about what this result says. If m, n, and p are nonnegative integers with $n \geq m - p$, then there is the linear transformation

$$L := [\mathbf{f}_1, \ldots, \mathbf{f}_{m-p}, 0, \ldots, 0]_{\mathbf{e}} : k^m \to k^n$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_m$ and $\mathbf{f}_1, \ldots, \mathbf{f}_n$ are the standard bases of k^m and k^n. If $\ell : V \to W$ is a linear transformation from an m-dimensional vector space to an n-dimensional vector space and $\text{null}(\ell) = p$, then ℓ is equivalent to L, so there are bases for V and W with respect to which ℓ "looks just like" L.

4.5 Matrices

Reader with some prior exposure to linear algebra know that a linear transformation can be represented by a matrix, and may well think of linear algebra as almost entirely a matter of doing various computations with matrices. Up to this point we have downplayed matrices, mainly in order to keep the focus on the linear transformations rather than the computational devices used to represent them, but they also simply would not have been very useful. Nevertheless our discussion of the fundamentals of linear algebra would not be complete without a brief discussion of the relationship between matrices and linear transformations.

Let $\ell : V \to W$ be a linear transformation between vector spaces V and W over k. If $\mathbf{b}_1, \ldots, \mathbf{b}_m$ and $\mathbf{c}_1, \ldots, \mathbf{c}_n$ are bases of V and W respectively, and

$$\ell(\mathbf{b}_i) = a_{1i}\mathbf{c}_1 + \cdots + a_{ni}\mathbf{c}_n$$

for each $i = 1, \ldots, m$, then we say that

$$A := \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

is the matrix of ℓ with respect to the bases $\mathbf{b}_1, \ldots, \mathbf{b}_m$ and $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Note that the bases $\mathbf{b}_1, \ldots, \mathbf{b}_m$ and $\mathbf{c}_1, \ldots, \mathbf{c}_n$ induce a bijection between linear transformations from V to W and $n \times m$ matrices: every linear transformation has a matrix, and equally, for any $n \times m$ matrix A there is a corresponding linear transformation

$$\Big[\sum_{j=1}^{n} a_{j1}\mathbf{c}_j, \ldots, \sum_{j=1}^{n} a_{jm}\mathbf{c}_j\Big]_{\mathbf{b}}.$$

The computation of the effect of ℓ, using A, is a matter of matrix multiplication. Recall that, in general, if

$$B := \begin{pmatrix} b_{11} & \cdots & b_{1s} \\ \vdots & \ddots & \vdots \\ b_{r1} & \cdots & b_{rs} \end{pmatrix} \quad \text{and} \quad C := \begin{pmatrix} c_{11} & \cdots & c_{1t} \\ \vdots & \ddots & \vdots \\ c_{s1} & \cdots & c_{st} \end{pmatrix}$$

are $r \times s$ and $s \times t$ matrices, then the matrix product BC is defined to be the $r \times t$ matrix whose ik-entry is

$$b_{i1}c_{1k} + \cdots + b_{is}c_{sk}.$$

That is, we compute the ik-entry by taking the inner product of the i^{th} row of B and the k^{th} column of C.

To see how A "computes" the effect of ℓ, suppose that

$$\ell(\beta_1\mathbf{b}_1 + \cdots + \beta_m\mathbf{b}_m) = \gamma_1\mathbf{c}_1 + \cdots + \gamma_n\mathbf{c}_n.$$

Then

$$\begin{aligned}\ell(\beta_1\mathbf{b}_1 + \cdots + \beta_m\mathbf{b}_m) &= \beta_1\sum_{j=1}^{n} a_{j1}\mathbf{c}_j + \cdots + \beta_m\sum_{j=1}^{n} a_{jm}\mathbf{c}_j \\ &= \big(\sum_{i=1}^{m} a_{1i}\beta_i\big)\mathbf{c}_1 + \cdots + \big(\sum_{i=1}^{m} a_{ni}\beta_i\big)\mathbf{c}_n,\end{aligned}$$

so, for each $j = 1, \ldots, n$,

$$\gamma_j = a_{j1}\beta_1 + \cdots + a_{jm}\beta_m.$$

We can express this result as the matrix multiplication $\gamma = A\beta$:

$$\begin{pmatrix}\gamma_1\\ \vdots\\ \gamma_n\end{pmatrix} = \begin{pmatrix}a_{11} & \cdots & a_{1m}\\ \vdots & \ddots & \vdots\\ a_{n1} & \cdots & a_{nm}\end{pmatrix}\begin{pmatrix}\beta_1\\ \vdots\\ \beta_m\end{pmatrix}.$$

One can think about what we're doing here in the following way. Let $r : V \to k^m$ and $s : W \to k^n$ be the functions

$$r(\beta_1\mathbf{b}_1 + \cdots + \beta_m\mathbf{b}_m) := (\beta_1, \ldots, \beta_m)$$

and

$$s(\gamma_1\mathbf{c}_1 + \cdots + \gamma_n\mathbf{c}_n) := (\gamma_1, \ldots, \gamma_n)$$

that "compute" the coordinates of points in V and W in the coordinate systems induced by our bases. Then for any $v \in V$ we have $\ell(v) = s^{-1}(Ar(v))$. If your education in linear algebra consists largely of concrete matrix calculations, it's easy to lose sight of the distinction between the linear transformation $\ell : v \mapsto s^{-1}(Ar(v))$ and the computation $\beta \mapsto A\beta$, but in any kind of scientific application an awareness of this distinction is a prerequisite of any sort of clear understanding of the situation. Elements of V and W are things like the electromagnetic field at a point in space, or a bundle of commodities consisting of certain amounts of iron ore, rubber, and coal. Elements of k^m and k^n are just tuples of elements of k.

Now suppose that X is a p-dimensional vector space over k with basis $\mathbf{d}_1, \ldots, \mathbf{d}_p$, and that $m : W \to X$ is a second linear transformation whose matrix, with respect to the bases $\mathbf{c}_1, \ldots, \mathbf{c}_n$ and $\mathbf{d}_1, \ldots, \mathbf{d}_p$, is the matrix B introduced above, so that

$$m(\mathbf{c}_j) = b_{1j}\mathbf{d}_1 + \cdots + b_{nj}\mathbf{d}_p$$

for each $j = 1, \ldots, n$. We compute the effect of the composition $m \circ \ell$:

$$\begin{aligned} m(\ell(\mathbf{b}_i)) &= m(a_{1i}\mathbf{c}_1 + \cdots + a_{ni}\mathbf{c}_n) \\ &= a_{1i}m(\mathbf{c}_1) + \cdots + a_{ni}m(\mathbf{c}_n) \\ &= \sum_{j=1}^{n} a_{ji}\big(b_{1j}\mathbf{d}_1 + \cdots + b_{nj}\mathbf{d}_p\big) \\ &= \big(\sum_{j=1}^{n} b_{1j}a_{ji}\big)\mathbf{d}_1 + \cdots + \big(\sum_{j=1}^{n} b_{pj}a_{ji}\big)\mathbf{d}_p. \end{aligned}$$

In this way we see that BA is the matrix of $m \circ \ell$ with respect to the bases $\mathbf{b}_1, \ldots, \mathbf{b}_m$ and $\mathbf{d}_1, \ldots, \mathbf{d}_p$.

This gives a new way of thinking about the associativity of matrix multiplication when the entries of the matrix are field elements. (When we dealt with this issue before, in Chapter 2, the proof was a calculation without much conceptual content.) Consider a fourth finite dimensional vector space Y over k and a third linear transformation $n : X \to Y$ that we assume is represented (with respect to $\mathbf{d}_1, \ldots, \mathbf{d}_p$ and some basis for Y) by the matrix C. Then

$$C(BA) = (CB)A$$

because $C(BA)$ represents the linear transformation $n \circ (m \circ \ell)$, $(CB)A$ represents the linear transformation $(n \circ m) \circ \ell$, and

$$n \circ (m \circ \ell) = (n \circ m) \circ \ell$$

because composition of functions is associative.

Now suppose that V and W are both n-dimensional, and $\ell : V \to W$ is a linear transformation. We say that ℓ is **nonsingular** if it is a linear isomorphism, and otherwise it is **singular**. Suppose that ℓ is nonsingular, and that A and B are the matrices of ℓ and ℓ^{-1} with respect to some bases for V and W. Then $BA = I = AB$, where I is the $n \times n$ identity matrix, because BA and AB are the matrices of $\ell^{-1} \circ \ell = \mathrm{Id}_V$ and $\ell \circ \ell^{-1} = \mathrm{Id}_W$ respectively.

In general, if A and B are $n \times n$ matrices such that $BA = I$, then we say that B is a **left inverse** of A, and that A is a **right inverse** of B. Suppose that A and B are the matrices of $\ell : V \to W$ and $m : W \to V$, with respect to some bases. If $BA = I$, then BA is the matrix of Id_V, and ℓ and m must be inverse isomorphisms, so $AB = I$ because AB is the matrix of $\ell \circ m = \mathrm{Id}_W$. Thus a left inverse of a square matrix is also a right inverse, and a right inverse is necessarily a left inverse. Moreover, A has only one left inverse because a left inverse is necessarily the matrix of ℓ^{-1}. In sum:

Theorem 4.18. *For any $n \times n$ matrix A either:*

(a) there is a unique matrix B that is the only left inverse of A and also the only right inverse of A;

(b) A has neither a left inverse nor a right inverse.

If (a) holds, then we say that A is **invertible** or **nonsingular**, and we denote the inverse by A^{-1}. If (b) holds, then we say that A is **singular**.

Chapter 5

The Determinant

My own education concerning the determinant was a haphazard affair. I was taught how to compute the determinants of 2×2 and 3×3 matrices, but initially, at least, it wasn't clear why one would want to do so. Various facts about determinants were introduced in piecemeal fashion, with little in the way of proofs, so eventually I knew the basics, sort of, but there were many important things I didn't learn until years later.

In this chapter we're going to approach the subject from a purely theoretical perspective, pursuing an inquiry that begins with a desire to explore a particular phenomenon and proceeds to certain intuitive considerations that become axioms. It will turn out that the axioms can be satisfied in a unique way, and that all the properties of the determinant can be derived from them. Because the theory is abstract and highly structured, it will require careful reading. But I think it is a quite satisfying piece of mathematics, addressing an important issue and arriving at a theory that is far from trivial, but which derives a certain ex post simplicity from its coherence. If your education to date has been like what I went through, it might be quite a revelation.

5.1 Positive and Negative Volume

In the last chapter the theme of classification was fruitful in two senses. Judged on its own terms, it gave us a clear picture of how vector spaces and linear transformations are structured. In addition, it forced us to develop vocabulary and basic technical facts that are used all the time in any sort of discussion involving linearity. But in the end the problems we used to guide our work turned out to be easy, and frankly, in my opinion, the analysis lacked mathematical depth.

From this point on we will aim at a more difficult classification problem. Fix a field k and a vector space V over k. A linear transformation from V to itself is called a **linear endomorphism**. (Recall that in any category an **endomorphism** is a morphism from an object to itself.) Let $\mathrm{End}(V)$ be the space of linear endomorphisms from V to itself. In what follows we will say that two linear endomorphisms $\ell \in \mathrm{End}(V)$ and $\ell' \in \mathrm{End}(V')$ are **similar** if there is a linear isomorphism $\iota : V \to V'$ such that $\iota \circ \ell = \ell' \circ \iota$, i.e., the diagram

$$\begin{tikzcd}
V \arrow[r, "\ell"] \arrow[d, "\iota"'] & V \arrow[d, "\iota"] \\
V' \arrow[r, "\ell'"] & V'
\end{tikzcd}$$

commutes. The proof that this is an equivalent relation follows the argument given in Section 4.4, with obvious modifications. An attribute of endomorphisms is an **invariant** if its value for ℓ is the same as its value for ℓ' whenever ℓ and ℓ' are similar. For us a solution of the classification problem will be a collection of invariants such that for any two distinct equivalence classes, at least one invariant is different.

When $V' = V$ we can think of the relation between ℓ and ℓ' as a matter of replacing the matrix of ℓ with the matrix of $\iota \circ \ell \circ \iota^{-1}$. In effect this notion of similarity forces us to choose the same coordinate system for the domain and range, making it much more difficult than before to find a canonical form that can always be achieved by some choice of basis. The determinant will be an invariant, and it will point in the direction of additional invariants, but these will not quite constitute a solution of the classification problem. As we will see in this chapter's final section, the invariants that solve the problem are quite subtle.

Although the theory we develop will be valid for any field, our motivation will be derived from the case of $k = \mathbb{R}$. Let V be an n-dimensional vector space over $\mathbb{R}^n$. In some sense that is, at this point, quite vague, it is intuitive that there should be a factor by which an endomorphism $\ell : V \to V$ expands or contracts "oriented" (we'll say more about this the term shortly) volume. To be concrete, think of this factor as the oriented volume of the image $\ell(C)$ of a cube $C \subset V$ of unit volume. Based on experience, one has the feeling that this quantity shouldn't depend on the coordinate system used to measure it. What is almost the same thing, this quantity should be invariant, depending only on the similarity class of ℓ. From the point of view of contemporary research methods this seems like a very natural quantity to study, and in fact our investigation will develop this intuition along very natural and straightforward lines until we arrive at the theory of the determinant. Based on this picture of its logical structure, one might guess that the theory emerged at a single point in time, when some gifted researcher noticed the issue and pursued it systematically.

In fact the theory of the determinant emerged over a long period of time, in fits and starts. The two dimensional case was considered by Gerolamo Cardano (1501-1576) and higher dimensional cases were considered by Leibniz. During the 18th century contributions to the theory were made by Gabriel Cramer (1704-1752), Étienne Bezout (1730-1783), Alexandre-Théophile Vandermonde (1735-1796), Laplace, and Lagrange, and at the beginning of the 19th century Gauss applied the determinant to issues in number theory, but it was not until 1811 that Jacques Binet (1786-1856) and Augustin Louis Cauchy (1789-1857) independently stated and proved what we might think of as the most fundamental fact about the determinant, namely the multiplicative property (Theorem 5.13). Now many things that seem simple and logical in retrospect are more complicated and far less obvious than we imagine, but I think this also says something about how powerful the axiomatic method is in research, by virtue of the questions it suggests and the methods it presents for addressing them. In a world with modern tools and hordes of hungry graduate students, theories like the one we'll see below do not go undeveloped for long.

In pursuing the geometric intuition described above, the first point to clarify is the word "oriented." Orientation is generally regarded as an advanced mathematical concept, and in some sense this is, perhaps, correct. But orientation is also a matter of universal everyday experience.

Consider a mirror. To be very explicit, suppose that the mirror is the xz-coordinate plane for a coordinate system in which the x-axis extends from left to right along the floor, the y-axis passes under your feet and straight ahead along the floor toward the horizon, and the z-axis is vertical. Then the mirror image of a point (x, y, z) with $y \leq 0$ appears to be located at $(x, -y, z)$.

Things look different in the mirror than when you look at them directly, with left shoes becoming right shoes and so forth. The technical way of saying this is that the map $(x, y, z) \mapsto (x, -y, z)$ "reverses the orientation of space," but we are now a long way from being able to say precisely what might be meant by this, much less justifying such language by providing a general theory. The main idea to absorb now is that our theory of the determinant will involve negative volume. In the example in question the volume of the image of the unit cube under the map $(x, y, z) \mapsto (x, -y, z)$ is

-1. In general, the volume of the image of the unit cube will be negative for a linear transformation that "reverses orientation" and positive for a linear transformation that "preserves orientation."

At this point we could try to compute the volume of the image of the unit cube in some examples, hoping to get some clue about how to go forward. One imagines the inventor of the determinant, having realized that it was invariant and hence an important quantity, doing extensive computations in search of inspiration. Well, as we saw above the actual historical process was quite different, and such computational experiments would get pretty complicated pretty quickly. We're not going to beat around the bush. Instead, we'll point out certain properties of the volume of the image of the unit cube, then show that there is a unique function with these properties.

Let $\mathbf{b} = (\mathbf{b}_1, \dots, \mathbf{b}_n)$ be a basis of V that, for the most part, will be fixed throughout our discussion. For $w_1, \dots, w_n \in V$ consider

$$\ell = [w_1, \dots, w_n]_{\mathbf{b}} \in \mathrm{End}(V).$$

Without knowing precisely what we mean by volume, we will try to discover properties of the ratio of the volume of the parallelepiped

$$\ell(C) := \{\, \delta_1 w_1 + \cdots + \delta_n w_n : 0 \le \delta_1, \dots, \delta_n \le 1 \,\}$$

to the volume of the unit cube

$$C := \{\, \delta_1 \mathbf{b}_1 + \cdots + \delta_n \mathbf{b}_n : 0 \le \delta_1, \dots, \delta_n \le 1 \,\}.$$

That is, we would like to define a function

$$\Delta_{\mathbf{b}} : \mathrm{End}(V) \to \mathbb{R}$$

whose properties correspond to the interpretation that $\Delta_{\mathbf{b}}(\ell)$ is this ratio. What properties should we expect $\Delta_{\mathbf{b}}$ to have?

First of all, and most obviously, the identity function on V maps the unit cube to itself, so we should have

$$\Delta_{\mathbf{b}}(\mathrm{Id}_V) = 1.$$

This point really requires no further comment.

If we interchange w_i and w_j, the result is to reverse the orientation of the image. Consistent with our interpretation of orientation reversal as negating volume, we should have

$$\begin{aligned}\Delta_{\mathbf{b}}([w_1, \dots, w_j, \dots, &w_i, \dots, w_n]_{\mathbf{b}}) \\ &= -\Delta_{\mathbf{b}}([w_1, \dots, w_i, \dots, w_j, \dots, w_n]_{\mathbf{b}}). \qquad (*)\end{aligned}$$

In the case of $n = 2$, this is illustrated in Figure 5.1. We look at the image of the unit square in the two cases. By including a test shape that is different from its mirror image, we can visually detect orientation reversal. For general n this idea is harder to visualize, and indeed we will soon have to do some combinatoric work just to show that it makes sense.

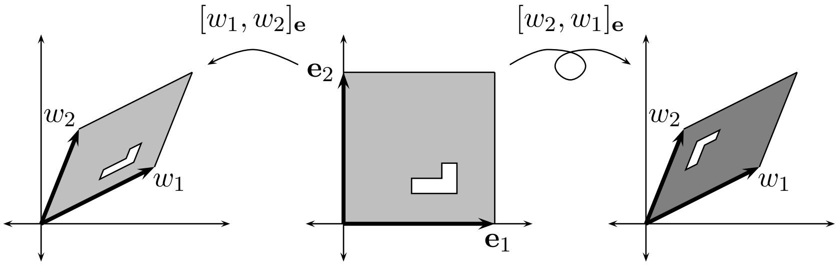


Figure 5.1

If we multiply some w_i by a scalar α, this should result in the volume being expanded by the same factor. That is,

$$\Delta_{\mathbf{b}}([w_1, \ldots, \alpha w_i, \ldots, w_n]_{\mathbf{b}}) = \alpha\Delta_{\mathbf{b}}([w_1, \ldots, w_i, \ldots, w_n]_{\mathbf{b}}). \qquad (**)$$

This idea is easy to accept when $\alpha > 0$, as shown in Figure 5.2.

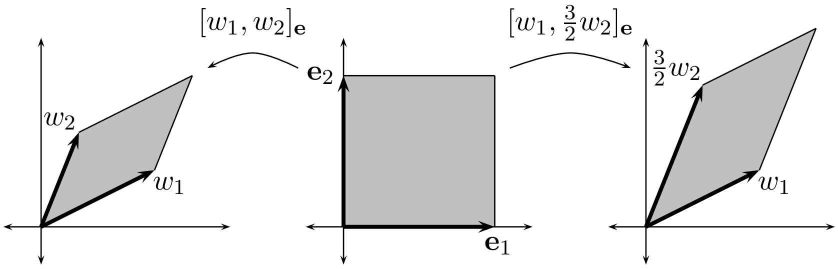

$[w_1, w_2]_{\mathbf{e}}$
$[w_1, \frac{3}{2}w_2]_{\mathbf{e}}$
w_2
w_1
$\mathbf{e}_2$
$\mathbf{e}_1$
$\frac{3}{2}w_2$
w_1

Figure 5.2

When $\alpha < 0$ we can think of this operation as having two parts: a) first multiply w_i by the absolute value of α; b) now multiply w_i by -1. We already understand a), and b) is an orientation reversal that turns positive volume into negative volume and vice versa. The combined effect is shown in Figure 5.3.

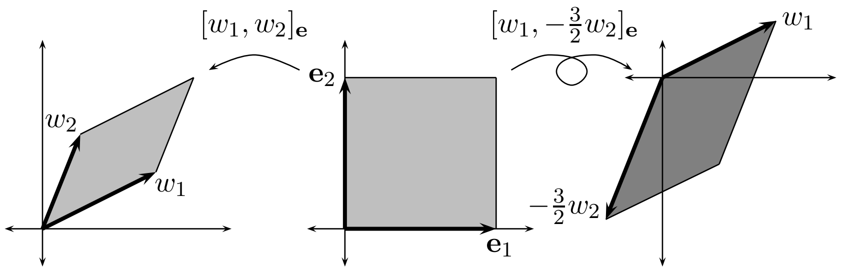

$[w_1, w_2]_{\mathbf{e}}$
$[w_1, -\frac{3}{2}w_2]_{\mathbf{e}}$
w_2
w_1
$\mathbf{e}_2$
$\mathbf{e}_1$
w_1
$-\frac{3}{2}w_2$

Figure 5.3

Suppose we add some multiple of w_j to w_i where $i \neq j$. As Figure 5.4 suggests, this doesn't change the total volume, so we should have

$$\Delta_{\mathbf{b}}([w_1, \ldots, w_i + \alpha w_j, \ldots, w_n]_{\mathbf{b}}) = \Delta_{\mathbf{b}}([w_1, \ldots, w_n]_{\mathbf{b}}) \qquad (***)$$

for all $i \neq j$ and all $\alpha \in \mathbb{R}$.

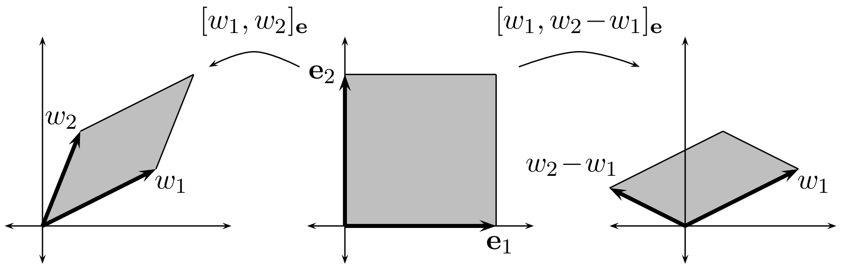


Figure 5.4

Here is a different way to visualize this idea. Put a deck of playing cards on a table in the usual way, with its sides at 90° angles. We think of its volume as $\Delta_{\mathbf{e}}([\ell\mathbf{e}_1, w\mathbf{e}_2, h\mathbf{e}_3]_{\mathbf{e}})$ where ℓ, w, and h are the length, width, and height, respectively, and $\mathbf{e} = \{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is the standard basis of $\mathbb{R}^3$. Now push against the deck with your hand in a way that leaves the bottom card fixed and slants the deck in some direction. The effect of this is to leave $\ell\mathbf{e}_1$ and $w\mathbf{e}_2$ fixed while replacing $h\mathbf{e}_3$ with $h\mathbf{e}_3 + \delta_\ell\mathbf{e}_1 + \delta_w\mathbf{e}_2$ for some numbers δ_ℓ and δ_w. Since the volume of the deck is unchanged, we should have

$$\Delta_{\mathbf{e}}([\ell\mathbf{e}_1, w\mathbf{e}_2, h\mathbf{e}_3 + \delta_\ell\mathbf{e}_1 + \delta_w\mathbf{e}_2]_{\mathbf{e}}) = \Delta_{\mathbf{e}}([\ell\mathbf{e}_1, w\mathbf{e}_2, h\mathbf{e}_3]_{\mathbf{e}}).$$

The final formula governing $\Delta_{\mathbf{b}}$ is

$$\Delta_{\mathbf{b}}([w_1, \ldots, w_i + w_i', \ldots, w_n]_{\mathbf{b}}) =$$

$$\Delta_{\mathbf{b}}([w_1,\dots,w_i,\dots,w_n]_{\mathbf{b}}) + \Delta_{\mathbf{b}}([w_1,\dots,w_i',\dots,w_n]_{\mathbf{b}}).$$

This is illustrated in the two dimensional case in Figure 5.5, but I think that in higher dimensions it is not so easy to visualize directly. Here is a different explanation. If the dimension of the span V' of $w_1,\dots,w_{i-1},w_{i+1},\dots,w_n$ is less than $n-1$, then the formula will hold because all its terms will be zero. If V' is $(n-1)$-dimensional, then take some $x \in V \setminus V'$. Since x and V' span V, we have $w_i = \alpha x + w$ and $w_i' = \alpha' x + w'$ where α and α' are scalars and $w, w' \in V'$. Then $(**)$ and $(***)$ imply that

$$\Delta_{\mathbf{b}}([w_1,\dots,w_i,\dots,w_n]_{\mathbf{b}}) = \alpha\Delta_{\mathbf{b}}([w_1,\dots,x,\dots,w_n]_{\mathbf{b}}),$$

$$\Delta_{\mathbf{b}}([w_1,\dots,w_i',\dots,w_n]_{\mathbf{b}}) = \alpha'\Delta_{\mathbf{b}}([w_1,\dots,x,\dots,w_n]_{\mathbf{b}}),$$

and

$$\Delta_{\mathbf{b}}([w_1,\dots,w_i+w_i',\dots,w_n]_{\mathbf{b}}) = (\alpha+\alpha')\Delta_{\mathbf{b}}([w_1,\dots,x,\dots,w_n]_{\mathbf{b}}),$$

and these combine to give the equation above.

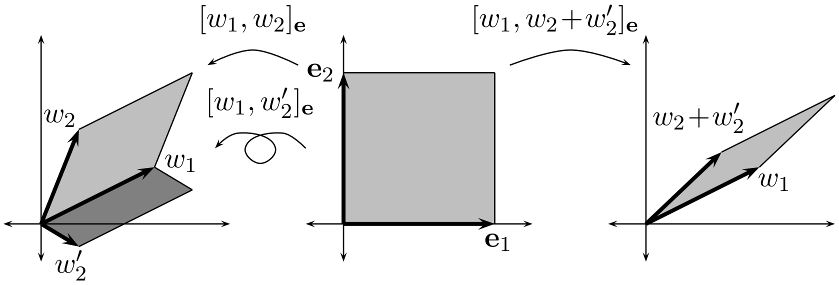

$[w_1, w_2]_{\mathbf{e}}$
$[w_1, w_2+w_2']_{\mathbf{e}}$
$\mathbf{e}_2$
$[w_1, w_2']_{\mathbf{e}}$
w_2
w_1
w_2'
$\mathbf{e}_1$
w_2+w_2'
w_1

Figure 5.5

We have just seen that the properties described above are not all independent, even though each has some distinct psychological importance as part of a functional understanding of the concept. The following two definitions boil them down to the logical minimum.

Definition 5.1. *A function* $\Delta_{\mathbf{b}} : \mathrm{End}(V) \to k$ *is* ***multilinear*** *if*

$$\Delta_{\mathbf{b}}([w_1,\dots,\alpha w_i+w_i',\dots,w_n]_{\mathbf{b}}) =$$

$$\alpha\Delta_{\mathbf{b}}([w_1,\dots,w_i,\dots,w_n]_{\mathbf{b}}) + \Delta_{\mathbf{b}}([w_1,\dots,w_i',\dots,w_n]_{\mathbf{b}})$$

for all $w_1,\dots,w_n \in V$, $i = 1,\dots,n$, $w_i' \in V$, *and* $\alpha \in k$.

That is, for any i and $w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n$,

$$w_i \mapsto \Delta_{\mathbf{b}}([w_1, \dots, w_{i-1}, w_i, w_{i+1}, \dots, w_n]_{\mathbf{b}})$$

is a linear function from V to k.

Definition 5.2. *A function $\Delta_{\mathbf{b}} : \mathrm{End}(V) \to k$ is* ***alternating*** *if*

$$\Delta_{\mathbf{b}}([w_1, \dots, w_n]_{\mathbf{b}}) = 0$$

whenever $w_i = w_j$ for some $i \neq j$.

If $\Delta_{\mathbf{b}}$ is multilinear and alternating, and $\Delta_{\mathbf{b}}(\mathrm{Id}_V) = 1$, then it satisfies all the equations above. This is clear except, perhaps, for the equation that says that interchanging w_i and w_j negates $\Delta_{\mathbf{b}}$. The way to obtain this equation is to note that, because $\Delta_{\mathbf{b}}$ is alternating,

$$0 = \Delta_{\mathbf{b}}([w_1, \dots, w_i + w_j, \dots, w_i + w_j, \dots, w_n]_{\mathbf{b}}).$$

Using multilinearity, we can expand the right hand side as a sum of four terms, two of which are zero because $\Delta_{\mathbf{b}}$ is alternating. What remains is a rearrangement of $(*)$.

Does a multilinear alternating $\Delta_{\mathbf{b}}$ exist at all? Is there a unique such $\Delta_{\mathbf{b}}$ satisfying $\Delta_{\mathbf{b}}(\mathrm{Id}_V) = 1$? The answers are affirmative, but will involve a brief adventure in group theory. In the remainder of this section we set up the key question, which is answered in the next section.

Let an $n \times n$ matrix $A = (a_{ij})$ be given. For $j = 1, \dots, n$ let

$$w_j = a_{1j}\mathbf{b}_1 + \cdots + a_{nj}\mathbf{b}_n,$$

and let

$$\ell_{\mathbf{b}}(A) := [w_1, \dots, w_n]_{\mathbf{b}} = \Big[\sum_{i=1}^{n} a_{i1}\mathbf{b}_i, \dots, \sum_{i=1}^{n} a_{in}\mathbf{b}_i\Big]_{\mathbf{b}}$$

be the element of $\mathrm{End}(V)$ whose matrix with respect to $\mathbf{b}$ is A. Applying multilinearity to w_1 gives

$$\Delta_{\mathbf{b}}(\ell_{\mathbf{b}}(A)) = \sum_{i_1=1}^{n} a_{i_1 1}\Delta_{\mathbf{b}}([\mathbf{b}_{i_1}, w_2, \dots, w_n]_{\mathbf{b}}).$$

This can be repeated for $w_2, \dots, w_n$, and eventually we arrive at

$$\Delta_{\mathbf{b}}(\ell_{\mathbf{b}}(A)) = \sum_{i_1,\dots,i_n=1}^{n} a_{i_1 1} \cdots a_{i_n n}\Delta_{\mathbf{b}}([\mathbf{b}_{i_1}, \dots, \mathbf{b}_{i_n}]_{\mathbf{b}}).$$

This simplifies a bit if we recognize that $\Delta_{\mathbf{b}}([\mathbf{b}_{i_1},\dots,\mathbf{b}_{i_n}]_{\mathbf{b}}) = 0$ whenever $i_h = i_j$ for some $h \neq j$. That is, the only terms in the sum that can be nonzero are those for which the function $j \mapsto i_j$ is injective, and consequently bijective because it maps a finite set to itself. Recall that the symmetric group S_n is the set of permutations of $\{1,\dots,n\}$. That is, an element of S_n is a bijection $\sigma : \{1,\dots,n\} \to \{1,\dots,n\}$. So,

$$\Delta_{\mathbf{b}}(\ell_{\mathbf{b}}(A)) = \sum_{\sigma \in S_n} a_{\sigma(1)1} \cdots a_{\sigma(n)n} \Delta_{\mathbf{b}}([\mathbf{b}_{\sigma(1)},\dots,\mathbf{b}_{\sigma(n)}]_{\mathbf{b}}).$$

We now see that $\Delta_{\mathbf{b}}$ is completely determined by its restriction

$$[\mathbf{b}_{\sigma(1)},\dots,\mathbf{b}_{\sigma(n)}]_{\mathbf{b}} \mapsto \Delta_{\mathbf{b}}([\mathbf{b}_{\sigma(1)},\dots,\mathbf{b}_{\sigma(n)}]_{\mathbf{b}})$$

to bases obtained by permuting $\mathbf{b}_1,\dots,\mathbf{b}_n$. This restriction needs to be alternating, in the obvious sense, and we need to have $\Delta_{\mathbf{b}}([\mathbf{b}_1,\dots,\mathbf{b}_n]_{\mathbf{b}}) = 1$. In the next section we will show that there is exactly one function satisfying these conditions.

5.2 Even and Odd Permutations

For $1 \le i, j \le n$ with $i \neq j$ the **swap** of i and j is the permutation $\tau_{ij} \in S_n$ given by

$$\tau_{ij}(h) = \begin{cases} j, & h = i, \\ i, & h = j, \\ h, & \text{otherwise.} \end{cases}$$

In this section we will mainly be concerned with compositions of several swaps, and compositions of swaps with other permutations, so to save space we will write compositions multiplicatively, e.g., $\tau_{k\ell}\tau_{ij}$ rather than $\tau_{k\ell} \circ \tau_{ij}$.

In thinking about permutations it can help to have a concrete image. Suppose we have n objects, labeled $1,\dots,n$, and n buckets, which are also labeled $1,\dots,n$. We identify a permutation σ with the arrangement in which, for each $k = 1,\dots,n$, the bucket labeled with k has the object labeled with $\sigma(k)$ in it. Suppose that we now interchange the contents of the buckets labeled i and j. Then bucket i has object $\sigma(j)$ in it, and bucket j has object $\sigma(i)$ in it, so the new arrangement is the one corresponding to the permutation $\sigma\tau_{ij} = \tau_{\sigma(i)\sigma(j)}\sigma$.

The following result is now a matter of concrete everyday experience: one can obtain any assignment of objects to buckets by swapping the contents of pairs of buckets until each object is where it should be. (In fact one can do this with $n-1$ or fewer swaps.)

Lemma 5.3. *Every $\sigma \in S_n$ can be written as a composition of swaps.*

Proof. Let ρ be a composition of swaps that is maximal, among all compositions of swaps, for the number of i such that $\rho(i) = \sigma(i)$. If $\rho \neq \sigma$ there is some j with $\rho(j) \neq \sigma(j)$, and there is some k such that $\rho(k) = \sigma(j)$. Then $\rho\tau_{jk}$ takes j to $\sigma(j)$, and $\rho(\tau_{jk}(i)) = \sigma(i)$ for all i such that $\rho(i) = \sigma(i)$. This contradicts the definition of ρ, so we must have $\rho = \sigma$. □

Actually, any permutation σ can be written as a composition of swaps of the form $\tau_{i,i+1}$. This is fairly obvious in the sense that everyone knows that any assignment of objects to buckets can be attained, eventually, by repeatedly swapping the contents of adjacent buckets. Alternatively, we can observe that if $i < j$, then

$$\tau_{ij} = \tau_{i,i+1}\tau_{i+1,i+2}\cdots\tau_{j-2,j-1}\tau_{j-1,j}\tau_{j-2,j-1}\cdots\tau_{i+1,i+2}\tau_{i,i+1}.$$

That is, starting with the assignment corresponding to e, we move the i^{th} object up one step at a time until it is in the j^{th} bucket, then move the j^{th} object down one step at a time until it is in the i^{th} bucket, after which every other item ends up where it began. Observe that $\tau_{j-1,j}$ appears once in this composition, and every other swap appears twice, so the total number $2(j-i)-1$ of swaps is odd.

If the function $\Delta_{\mathbf{b}}$ is alternating, then it must be the case that

$$\Delta_{\mathbf{b}}([\mathbf{b}_{\tau_{ij}\sigma(1)},\dots,\mathbf{b}_{\tau_{ij}\sigma(n)}]_{\mathbf{b}}) = -\Delta_{\mathbf{b}}([\mathbf{b}_{\sigma(1)},\dots,\mathbf{b}_{\sigma(n)}]_{\mathbf{b}})$$

for all $\sigma \in S_n$ and all swaps τ_{ij}. Provided that $\Delta_{\mathbf{b}}([\mathbf{b}_1,\dots,\mathbf{b}_n]_{\mathbf{b}}) = 1$, $\Delta_{\mathbf{b}}([\mathbf{b}_{\sigma(1)},\dots,\mathbf{b}_{\sigma(n)}]_{\mathbf{b}})$ must be 1 if σ can be written as the composition of an even number of swaps, and it must be -1 if σ can be written as the composition of an odd number of swaps. Our principle objective is to show that there is no σ that can be written in both these ways.

The method of proof is rather clever. Consider

$$\Delta_n(X_1,\dots,X_n) := \prod_{1\le i<j\le n}(X_j - X_i) \in \mathbb{Z}[X_1,\dots,X_n].$$

This polynomial is called the **Vandermonde determinant** for reasons that will be explained a bit later. For any $\sigma \in S_n$ we have either

$$\Delta_n(X_{\sigma(1)},\dots,X_{\sigma(n)}) = \Delta_n(X_1,\dots,X_n)$$

or

$$\Delta_n(X_{\sigma(1)},\dots,X_{\sigma(n)}) = -\Delta_n(X_1,\dots,X_n)$$

because $\Delta_n(X_{\sigma(1)},\dots,X_{\sigma(n)})$ can be obtained from $\Delta_n(X_1,\dots,X_n)$ by multiplying each of the factors $X_j - X_i$ by 1 or -1, so there is a function $\mathrm{sgn} : S_n \to \{-1,1\}$ such that

$$\Delta_n(X_{\sigma(1)},\dots,X_{\sigma(n)}) = \mathrm{sgn}(\sigma)\cdot\Delta_n(X_1,\dots,X_n).$$

We call $\mathrm{sgn}(\sigma)$ the **sign** of σ, and we say that σ is an **odd permutation** or an **even permutation** according to whether $\mathrm{sgn}(\sigma)$ is -1 or 1.

We now compare the sign of each factor in

$$\text{(a)} \prod_{1\le i<j\le n}(X_{\sigma\tau_{k,k+1}(j)} - X_{\sigma\tau_{k,k+1}(i)}) \quad\text{and}\quad \text{(b)} \prod_{1\le i<j\le n}(X_{\sigma(j)} - X_{\sigma(i)}).$$

If $\{i,j\}\cap\{k,k+1\}=\emptyset$, then $\sigma(\tau_{k,k+1}(i))=\sigma(i)$ and $\sigma(\tau_{k,k+1}(j))=\sigma(j)$, so $X_{\sigma(j)} - X_{\sigma(i)}$ has the same sign in (a) and (b). If $1\le i<k$, then

$$X_{\sigma\tau_{k,k+1}(k)} - X_{\sigma\tau_{k,k+1}(i)} = X_{\sigma(k+1)} - X_{\sigma(i)}$$

and

$$X_{\sigma\tau_{k,k+1}(k+1)} - X_{\sigma\tau_{k,k+1}(i)} = X_{\sigma(k)} - X_{\sigma(i)}$$

have the same signs in (a) and (b). Similarly, if $k+1<j\le n$, then

$$X_{\sigma\tau_{k,k+1}(j)} - X_{\sigma\tau_{k,k+1}(k)} = X_{\sigma(j)} - X_{\sigma(k+1)}$$

and

$$X_{\sigma\tau_{k,k+1}(j)} - X_{\sigma\tau_{k,k+1}(k+1)} = X_{\sigma(j)} - X_{\sigma(k)}$$

have the same signs in (a) and (b). Since

$$X_{\sigma\tau_{k,k+1}(k+1)} - X_{\sigma\tau_{k,k+1}(k)} = X_{\sigma(k)} - X_{\sigma(k+1)}$$

occurs with opposite signs in (a) and (b), there is exactly one sign reversal, and we can conclude that $\mathrm{sgn}(\sigma\tau_{k,k+1}) = -\mathrm{sgn}(\sigma)$.

Above we saw that any swap is a composition of an odd number of swaps of the form $\tau_{k,k+1}$, so $\mathrm{sgn}(\sigma\tau) = -\mathrm{sgn}(\sigma)$ for any permutation σ and any swap τ. Since every permutation can be written as a composition of swaps, we conclude that $\mathrm{sgn}(\sigma)=-1$ if σ can be written as a composition of an odd number of swaps and $\mathrm{sgn}(\sigma)=1$ if σ can be written as a composition of an even number of swaps, so, as desired, *it is never possible to write a permutation in both ways.*

In view of all this, $\mathrm{sgn}(\sigma\sigma') = \mathrm{sgn}(\sigma)\mathrm{sgn}(\sigma')$ for all $\sigma,\sigma'\in S_n$, so

$$\mathrm{sgn} : S_n \to \{-1,1\}$$

is a homomorphism if we regard $\{1, -1\}$ as a group with multiplication as the group operation. It won't figure in our subsequent work, but it is nonetheless worth mentioning that $A_n := \operatorname{sgn}^{-1}(1)$ is a normal (because it is the kernel of a homomorphism) subgroup of S_n called the **alternating group** on n letters. These groups figure prominently in the theory of simple groups because A_n is simple with one oddball exception: A_4 is not simple. (There is an elementary proof that A_n is simple when $n \geq 5$, but unfortunately it is a bit too long to include here.)

5.3 The Determinant of a Matrix

Let's review the situation. In the section before last we showed that if $\mathbf{b}$ is a basis of V, $\Delta_{\mathbf{b}} : \operatorname{End}(V) \to k$ is multilinear and alternating, and A is an $n \times n$ matrix, then

$$\Delta_{\mathbf{b}}(\ell_{\mathbf{b}}(A)) = \sum_{\sigma \in S_n} a_{\sigma(1)1} \cdots a_{\sigma(n)n} \Delta_{\mathbf{b}}([\mathbf{b}_{\sigma(1)}, \dots, \mathbf{b}_{\sigma(n)}]_{\mathbf{b}}).$$

Since $\Delta_{\mathbf{b}}$ is alternating, if $\Delta_{\mathbf{b}}(\mathrm{Id}_V) = 1$, then the results of the last section imply that $\Delta_{\mathbf{b}}([\mathbf{b}_{\sigma(1)}, \dots, \mathbf{b}_{\sigma(n)}]_{\mathbf{b}}) = \operatorname{sgn}(\sigma)$ for all σ. In sum, *if* $\Delta_{\mathbf{b}} : \operatorname{End}(V) \to k$ is multilinear and alternating with $\Delta_{\mathbf{b}}(\mathrm{Id}_V) = 1$, then

$$\Delta_{\mathbf{b}}(\ell_{\mathbf{b}}(A)) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n}.$$

But we still need to show that the function defined by this formula actually satisfies the stated conditions.

Both in order to be very clear about what our results depend on, and because there are important applications that depend on the additional generality, in this section we will work with matrices whose entries lie in a general commutative ring with unit R. Let

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$

be an $n \times n$ matrix with entries in R. The **determinant** of A is

$$|A| = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} := \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots a_{\sigma(n)n}.$$

In this section and the next the analysis will refer only to matrices, developing the key properties of the determinant using purely algebraic computations. Of course the relation between square matrices and linear endomorphisms is the underlying motivation, and the key properties are closely related to the formulas for $\Delta_{\mathbf{b}}([w_1, \ldots, w_n]_{\mathbf{b}})$ developed earlier, but from a logical point of view nothing depends on that.

We begin with perhaps the most basic property of the determinant. Any product $a_{\sigma(1)1} \cdots a_{\sigma(n)n}$ includes a factor from each row and a factor from each column, so:

Proposition 5.4. *If all the entries in one of the rows, or one of the columns, of A are zero, then $|A| = 0$.*

A slightly more sophisticated version of this idea occurs when exactly one of the entries of a column or row are nonzero. At this point we present only the simplest version of this: a_{11} is the nonzero entry in question. The other results developed in this and the next section will make it easy to generalize to an arbitrary a_{ij} later.

Proposition 5.5. *If $a_{12}, \ldots, a_{1n}$ are all zero, or $a_{21}, \ldots, a_{n1}$ are all zero, then*

$$|A| = a_{11} \cdot \begin{vmatrix} a_{22} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{n2} & \cdots & a_{nn} \end{vmatrix}.$$

Proof. In either case the only nonzero products $a_{\sigma(1)1} \cdots a_{\sigma(n)n}$ are those with $\sigma(1) = 1$. If, for such a σ, the permutation $\tau \in S_{n-1}$ is given by $\tau(j) := \sigma(i+1) - 1$, then $\operatorname{sgn}(\tau) = \operatorname{sgn}(\sigma)$ because any representation of σ as a composition of swaps can be "translated" into a representation of τ as a composition of the same number of swaps. Therefore

$$|A| = \sum_{\sigma \in S_n, \sigma(1)=1} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{\sigma(i)i} = a_{11} \sum_{\tau \in S_{n-1}} \operatorname{sgn}(\tau) \prod_{j=1}^{n-1} a_{\tau(j)+1,j+1}.$$

□

We now develop the properties of the determinant that correspond to the conditions we want $\Delta_{\mathbf{b}}$ to satisfy. The first result requires no comment and is, I think, obvious enough that there is no need to present a proof.

Proposition 5.6. *If I is the $n \times n$ identity matrix, then*

$$|I| = 1.$$

The next two results imply that the determinant of A is a multilinear function, and the third states that is alternating. The proofs reflect, in a straightforward manner, ideas we have seen earlier. But the formula defining the determinant is bulky, so computations involving it are necessarily rather messy.

Proposition 5.7. *Suppose $A = (a_{ij})$, $A' = (a'_{ij})$, and $A'' = (a''_{ij})$ are $n \times n$ matrices such that, for some h, $a''_{ih} = a_{ih} + a'_{ih}$ for all $i = 1, \dots, n$ and $a_{ij} = a'_{ij} = a''_{ij}$ for all $1 \leq i, j \leq n$ with $j \neq h$. Then $|A''| = |A| + |A'|$.*

Proof. This is a straightforward calculation:

$$\begin{aligned}
|A''| &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a''_{\sigma(1)1} \cdots a''_{\sigma(n)n} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots (a_{\sigma(h)h} + a'_{\sigma(h)h}) \cdots a_{\sigma(n)n} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots a_{\sigma(n)n} + \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a'_{\sigma(1)1} \cdots a'_{\sigma(n)n} \\
&= |A| + |A'|.
\end{aligned}$$

□

Proposition 5.8. *If $A = (a_{ij})$ and $A' = (a'_{ij})$ are $n \times n$ matrices such that, for some h and $\alpha \in R$,*

$$a'_{ij} = \begin{cases} \alpha a_{ij}, & j = h, \\ a_{ij}, & j \neq h, \end{cases}$$

for all $1 \leq i, j \leq n$, then $|A'| = \alpha |A|$.

Proof. This is another simple calculation:

$$\begin{aligned}
|A'| &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a'_{\sigma(1)1} \cdots a'_{\sigma(n)n} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots (\alpha a_{\sigma(h)h}) \cdots a_{\sigma(n)n} \\
&= \alpha \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots a_{\sigma(n)n} = \alpha |A|.
\end{aligned}$$

□

Proposition 5.9. *Suppose $A = (a_{ij})$ is an $n \times n$ matrix and $A' = (a'_{ij})$ is obtained from A by interchanging columns k and ℓ, for some $1 \leq k < \ell \leq n$, so that*

$$a'_{ij} = a_{i\tau_{k\ell}(j)}$$

for all $1 \leq i, j \leq n$. Then $|A'| = -|A|$.

Proof. For any $\sigma \in S_n$ we have

$$\prod_{h=1}^{n} a'_{\sigma(h)h} = \prod_{h=1}^{n} a_{\sigma(h)\tau_{k\ell}(h)} = \prod_{h'=1}^{n} a_{\sigma\tau_{k\ell}(h')h'}$$

where the second equality comes from the substitution $h' := \tau_{k\ell}(h)$. The function $\sigma \mapsto \sigma\tau_{k\ell}$ is a bijection from S_n to itself (in fact it is its own inverse) so

$$\begin{aligned}
|A'| &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a'_{\sigma(1)1} \cdots a'_{\sigma(n)n} \\
&= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma\tau_{k\ell}(1)1} \cdots a_{\sigma\tau_{k\ell}(n)n} \\
&= -\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma\tau_{k\ell}) \cdot a_{\sigma\tau_{k\ell}(1)1} \cdots a_{\sigma\tau_{k\ell}(n)n} \\
&= -\sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots a_{\sigma(n)n} = -|A|.
\end{aligned}$$

□

The most frequent operation in numerical computation of determinants—adding a scalar multiple of one column to another column—is a combination of the more elementary operations described above. Starting with A, distinct indices g, h, and $\alpha \in R$, form A' by replacing the g^{th} column of A with α times the h^{th} column. Then $|A'| = 0$ because $|A'|$ is α times the determinant of a matrix with two identical column, and the determinant of such a matrix must be zero because the last result implies that it is equal to its negation. If we then form A'' by replacing the g^{th} column of A with the sum of the g^{th} column and the g^{th} column of A' (which is α times the h^{th} column of A) then Proposition 5.7 gives $|A''| = |A| + |A'| = |A|$. We restate this conclusion with A' in place of A'':

Proposition 5.10. *Suppose $A = (a_{ij})$ and $A' = (a'_{ij})$ are $n \times n$ matrices such that, for some $1 \leq g, h \leq n$ with $g \neq h$ and some $\alpha \in R$, $a'_{ig} = a_{ig} + \alpha a_{ih}$ for all $i = 1, \ldots, n$, and $a_{ij} = a'_{ij}$ for all $1 \leq i, j \leq n$ with $j \neq g$. Then $|A'| = |A|$.*

When R is a field the last result and Propositions 5.5 and 5.9 combine to give a systematic procedure for computing determinants numerically. Let's suppose that $a_{11} \neq 0$. (If all the entries in the first row are zero, then the determinant is zero, and otherwise we can bring this about by interchanging columns to get a matrix whose determinant is the same or, in the case $n = 2$, its negation.) For each $i = 2, \ldots, n$ we can subtract a_{1i}/a_{11} times the first column from column i, thereby obtaining

$$|A| = \begin{vmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} - \frac{a_{12}}{a_{11}} a_{21} & \cdots & a_{2n} - \frac{a_{1n}}{a_{11}} a_{21} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} - \frac{a_{12}}{a_{11}} a_{n1} & \cdots & a_{nn} - \frac{a_{1n}}{a_{11}} a_{n1} \end{vmatrix}.$$

Applying Proposition 5.5,

$$|A| = a_{11} \cdot \begin{vmatrix} a_{22} - \frac{a_{12}}{a_{11}} a_{21} & \cdots & a_{2n} - \frac{a_{1n}}{a_{11}} a_{21} \\ \vdots & & \vdots \\ a_{2n} - \frac{a_{12}}{a_{11}} a_{n1} & \cdots & a_{nn} - \frac{a_{1n}}{a_{11}} a_{n1} \end{vmatrix}.$$

We've reduced the computation of the determinant of an $n \times n$ matrix to the computation of the determinant of an $(n-1) \times (n-1)$ matrix. The reduction took on the order of n^2 arithmetical operations, so if we repeatedly reduce in this fashion the total number of operations required to compute the determinant is on the order of

$$n^2 + (n-1)^2 + \cdots + 1 = \tfrac{1}{6} n(n+1)(2n+1).$$

(The expression on the right hand side is obviously correct when $n = 1$, and you are invited to check that $(n+1)(n+2)(2n+3) - n(n+1)(2n+1) = 6(n+1)^2$.) This number grows much less rapidly than $n!$, so this procedure is much more practical than computing the formula defining the determinant directly. A computer can use this method to compute determinants of matrices with hundreds or even thousands of rows and column. (Actually, there is something called the "fast Fourier transform" that is much faster still when the matrices are large.)

5.4 Transposes and Products

In addition to the properties directly motivated by the function $\Delta_{\mathbf{b}}$, there are two additional facts about the determinant that are quite important.

The **transpose** of an $m \times n$ matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

is the $n \times m$ matrix

$$A^T = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{mn} \end{pmatrix}$$

obtained by "turning A over," so that rows become columns and vice versa, and the ij-entry of A becomes the ji-entry of A^T. The results above describe how the determinant is affected by certain operations on the columns. The use of such operations in computations is made much more flexible by the following result, which shows that the same results pertain to operations on rows.

Proposition 5.11. *For any $n \times n$ matrix A,*

$$|A^T| = |A|.$$

Proof. For any $\sigma \in S_n$ the list of pairs $(1, \sigma^{-1}), \ldots, (n, \sigma^{-1}(n))$ is a reordering of the list $(\sigma(1), 1), \ldots, (\sigma(n), n)$, and $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)^{-1} = \operatorname{sgn}(\sigma)$ because sgn is a homomorphism. For any group G the function $g \mapsto g^{-1}$ is a bijection because it is its own inverse. Therefore

$$\begin{aligned} |A| &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{\sigma(1)1} \cdots a_{\sigma(n)n} \\ &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) \cdot a_{1\sigma^{-1}(1)} \cdots a_{n\sigma^{-1}(n)} \\ &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \cdot a_{1\sigma(1)} \cdots a_{n\sigma(n)} = |A^T|. \end{aligned}$$

□

Now that this result has been established, for each of the column operations studied in the last section, the corresponding result for row operations has now been "officially" established. In order to be able to refer to these results easily we summarize what we've learned.

Theorem 5.12. *Let A be an $n \times n$ matrix. If A' is obtained by interchanging two of the columns of A, then $|A'| = -|A|$, so $|A| = 0$ whenever A has two identical columns. If A' is obtained from A by multiplying one of A's columns by a scalar α, then $|A'| = \alpha|A|$. If A' is obtained from A by adding a linear combination of the other columns to one of the columns of A, then $|A'| = |A|$. If all the entries in column j other than a_{ij} are zero, then*

$$|A| = (-1)^{i+j} a_{ij} \cdot \begin{vmatrix} a_{11} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{nn} \end{vmatrix}.$$

All these results hold equally with 'column' replaced by 'row.'

Proof. The only claim that might not be clear, after all the other results above, is the one concerning all entries in column j, other than a_{ij}, being zero. The idea is to interchange rows i and $i-1$, then interchange rows $i-1$ and $i-2$, and so forth until row i becomes row 1, then interchange columns j and $j-1$, then interchange columns $j-1$ and $j-2$, and so forth until column j becomes column 1. The total number of swaps is $(i-1)+(j-1)$. □

As an illustration of how row and column operation can be combined, we will now verify the formula from which the Vandermonde determinant derives its name. The **Vandermonde matrix** is

$$V(X_1, \ldots, X_n) := \begin{pmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^{n-1} \\ 1 & X_2 & X_2^2 & \cdots & X_2^{n-1} \\ 1 & X_3 & X_3^2 & \cdots & X_3^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^{n-1} \end{pmatrix}.$$

We will show that

$$|V(X_1, \ldots, X_n)| = \Delta_n(X_1, \ldots, X_n) := \prod_{1 \le i < j \le n} (X_j - X_i).$$

The determinant of $V(X_1, \ldots, X_n)$ is unaffected if we subtract the first row from each of the other rows, so

$$|V(X_1,\dots,X_n)| = \begin{vmatrix} 1 & X_1 & X_1^2 & \cdots & X_1^{n-1} \\ 0 & X_2 - X_1 & X_2^2 - X_1^2 & \cdots & X_2^{n-1} - X_1^{n-1} \\ 0 & X_3 - X_1 & X_3^2 - X_1^2 & \cdots & X_3^{n-1} - X_1^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & X_n - X_1 & X_n^2 - X_1^2 & \cdots & X_n^{n-1} - X_1^{n-1} \end{vmatrix}.$$

Since the only entry in the first column is the 1 in the upper left hand corner, we have

$$|V(X_1,\dots,X_n)| = \begin{vmatrix} X_2 - X_1 & X_2^2 - X_1^2 & \cdots & X_2^{n-1} - X_1^{n-1} \\ X_3 - X_1 & X_3^2 - X_1^2 & \cdots & X_3^{n-1} - X_1^{n-1} \\ \vdots & \vdots & & \vdots \\ X_n - X_1 & X_n^2 - X_1^2 & \cdots & X_n^{n-1} - X_1^{n-1} \end{vmatrix}.$$

For each relevant i and j there is the factorization

$$X_i^j - X_1^j = (X_i - X_1)(X_i^{j-1} + X_i^{j-2}X_1 + \cdots + X_iX_1^{j-2} + X_1^{j-1}).$$

Therefore

$$|V(X_1,\dots,X_n)| = |W| \cdot \prod_{i=2}^{n}(X_i - X_1) \qquad (*)$$

where

$$W := \begin{pmatrix} 1 & X_2 + X_1 & X_2^2 + X_2X_1 + X_1^2 & \cdots & X_2^{n-2} + \cdots + X_1^{n-2} \\ 1 & X_3 + X_1 & X_3^2 + X_3X_1 + X_1^2 & \cdots & X_3^{n-2} + \cdots + X_1^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n + X_1 & X_n^2 + X_nX_1 + X_1^2 & \cdots & X_n^{n-2} + \cdots + X_1^{n-2} \end{pmatrix}.$$

In evaluating the determinant of W we can subtract X_1 times the first column from the second column, subtract X_1^2 times the first column from the third column, and so forth, finally subtracting X_1^{n-2} times the first column from the last column, arriving at

$$|W| := \begin{vmatrix} 1 & X_2 & X_2^2 + X_2X_1 & \cdots & X_2^{n-2} + \cdots + X_2X_1^{n-3} \\ 1 & X_3 & X_3^2 + X_3X_1 & \cdots & X_3^{n-2} + \cdots + X_3X_1^{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 + X_nX_1 & \cdots & X_n^{n-2} + \cdots + X_nX_1^{n-3} \end{vmatrix}.$$

We can now subtract X_1 times the second column from the third column, subtract X_1^2 times the second column from the fourth column, and so on until X_1^{n-3} times the second column is subtracted from the last column, so that

$$|W| := \begin{vmatrix} 1 & X_2 & X_2^2 & \cdots & X_2^{n-2} + \cdots + X_2^2 X_1^{n-4} \\ 1 & X_3 & X_3^2 & \cdots & X_3^{n-2} + \cdots + X_3^2 X_1^{n-4} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^{n-2} + \cdots + X_n^2 X_1^{n-4} \end{vmatrix}.$$

Continuing in this manner leads eventually to

$$|W| := \begin{vmatrix} 1 & X_2 & X_2^2 & \cdots & X_2^{n-2} \\ 1 & X_3 & X_3^2 & \cdots & X_3^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_n & X_n^2 & \cdots & X_n^{n-2} \end{vmatrix}.$$

Induction now implies the claim. Evidently $\Delta_1(X_1) = 1$ is the determinant of $V(X_1)$ because this is a 1×1 matrix whose only entry is 1. If the desired formula has already been established with $n-1$ in place of n, then

$$|W| = \Delta_{n-1}(X_2, \dots, X_n) = \prod_{2 \le j < k \le n} (X_k - X_j),$$

and the desired formula for $\Delta_n(X_1, \dots, X_n)$ is obtained by substituting this in equation $(*)$:

$$|V(X_1, \dots, X_n)| = \Big(\prod_{i=2}^{n} (X_i - X_1)\Big) \cdot \Big(\prod_{2 \le j < k \le n} (X_k - X_j)\Big)$$

$$= \prod_{1 \le i < j \le n} (X_j - X_i) = \Delta_n(X_1, \dots, X_n).$$

This formula has an interesting geometric interpretation. Obviously the rows of the Vandermonde matrix are linearly dependent whenever there are distinct i and j such that $X_i = X_j$. This formula implies that this is the *only* way there can be a linear dependence.

Perhaps the most commonly used result concerning the determinant is the fact that taking the determinant commutes with matrix multiplication. Thinking in terms of composition of linear functions leads us to expect this, since, for example, if $\ell : V \to V$ expands volume by a factor of 3, and $m : V \to V$ compresses volume according to a factor of $1/2$, then $m \circ \ell$ should expand volume by a factor of $3/2$. Even so, the complexity of the formula defining the determinant suggests that the proof might be a nightmare. But it turns out to be about as simple and straightforward as one might hope.

Theorem 5.13. *If A and B are $n \times n$ matrices, then*

$$|AB| = |A|\,|B|.$$

Proof. We start out with the definition of $|AB|$, then massage it in what seems like a promising direction:

$$\begin{aligned}|AB| &= \sum_{\sigma\in S_n} \operatorname{sgn}(\sigma)\big(\sum_{j_1=1}^{n} a_{\sigma(1)j_1}b_{j_1 1}\big)\cdots\big(\sum_{j_n=1}^{n} a_{\sigma(n)j_n}b_{j_n n}\big)\\ &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\sum_{j_1,\dots,j_n=1}^{n} a_{\sigma(1)j_1}b_{j_1 1}\cdots a_{\sigma(n)j_n}b_{j_n n}\\ &= \sum_{j_1,\dots,j_n=1}^{n}\Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{\sigma(1)j_1}\cdots a_{\sigma(n)j_n}\Big)b_{j_1 1}\cdots b_{j_n n}.\end{aligned}$$

Now observe that

$$\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{\sigma(1)j_1}\cdots a_{\sigma(n)j_n}$$

is the determinant of the matrix whose i^{th} column is the ${j_i}^{\text{th}}$ column of A. This is 0 if $j_i = j_{i'}$ for some $i \neq i'$, and if there is $\rho \in S_n$ such that $j_i = \rho(i)$ for all i, then it is $\operatorname{sgn}(\rho)|A|$. Therefore

$$|AB| = |A|\sum_{\rho\in S_n}\operatorname{sgn}(\rho)b_{\rho(1)1}\cdots b_{\rho(n)n} = |A|\,|B|.$$

□

5.5 Back to Linear Transformations

We now revert to working over a field k. Let V be an n-dimensional vector space over k with basis $\mathbf{b}$. Combining the results of the last three sections yields:

Theorem 5.14. *There is a unique multilinear alternating function $\Delta_{\mathbf{b}} : \operatorname{End}(V) \to k$ with $\Delta_{\mathbf{b}}(\mathrm{Id}_V) = 1$ given by the formula*

$$\Delta_{\mathbf{b}}(\ell_{\mathbf{b}}(A)) = |A|.$$

It's time to take what we've learned about determinants of matrices and use it to put the finishing touches on the theory of determinants of linear transformations.

Proposition 5.15. *An endomorphism $\ell \in \operatorname{End}(V)$ is singular if and only if $\Delta_{\mathbf{b}}(\ell) = 0$.*

Proof. Let $\ell = [w_1, \ldots, w_n]_{\mathbf{b}}$ where $w_j = \sum_{i=1}^n a_{ij}\mathbf{b}_i$, so that $A = (a_{ij})$ is the matrix of ℓ. If ℓ is singular, then $w_1, \ldots, w_n$ are linearly dependent (Proposition 4.11) which implies that one of the columns of $A = (a_{ij})$ can be expressed as a linear combination of the other columns, so that Theorem 5.12 implies that $|A| = 0$. If ℓ is nonsingular, then it is invertible, and we can let B be the matrix of ℓ^{-1}. Then BA is the matrix of Id_V, so $BA = I$ and $|A|\,|B| = |AB| = |I| = 1$, and consequently $|A| \neq 0$. □

The applicability of this insight is not restricted to endomorphisms. Suppose that W and X are n-dimensional vector spaces and A is the matrix of $m : W \to X$, say with respect to bases $\mathbf{c}$ and $\mathbf{d}$. If $|A| = 0$, then the endomorphism $\ell_{\mathbf{b}}(A) \in \operatorname{End}(V)$ is singular, and consequently A cannot be invertible, which in turn implies that m is singular. On the other hand, if $|A| \neq 0$, then $\ell_{\mathbf{b}}(A)$ is nonsingular, so A is invertible, and consequently m is nonsingular. Thus:

Theorem 5.16. *If W and X are n-dimensional vector spaces and A is the matrix of the linear transformation $m : W \to X$, then the following are equivalent:*

(a) A is invertible;

(b) m is an isomorphism;

(c) $|A| \neq 0$.

We will now show that similar endomorphisms have the same determinant. Suppose $\ell \in \operatorname{End}(V)$ and $\ell' \in \operatorname{End}(V')$ are similar, so there is a linear isomorphism $\iota : V \to V'$ such that the diagram

$$\begin{tikzcd} V \arrow[r, "\ell"] \arrow[d, "\iota"'] & V \arrow[d, "\iota"] \\ V' \arrow[r, "\ell'"] & V' \end{tikzcd}$$

commutes. Let $\mathbf{b}$ be a basis of V, let $\mathbf{c}$ be a basis of V', and let A and A' be the matrices of ℓ and ℓ' with respect to these bases. Let C be the matrix of ι. Since $\iota^{-1} \circ \iota = \mathrm{Id}_V$, C^{-1} is the matrix of ι^{-1}. We have $CAC^{-1} = A'$ because $\iota \circ \ell \circ \iota^{-1} = \ell'$, and $|C|\,|C^{-1}| = |CC^{-1}| = |I| = 1$, so

$$\Delta_{\mathbf{b}}(\ell) = |A| = |C|\,|A|\,|C^{-1}| = |A'| = \Delta_{\mathbf{c}}(\ell').$$

An important special case of the situation considered in the last paragraph is that $V' = V$, $\iota = \mathrm{Id}_V$, and $\ell' = \ell$, so that A is the matrix of ℓ with respect to $\mathbf{b}$ and A' is the matrix with respect to $\mathbf{c}$. Then

$$\Delta_{\mathbf{b}}(\ell) = |A| = |A'| = \Delta_{\mathbf{c}}(\ell).$$

That is, the determinant of an endomorphism doesn't depend on the basis used to compute it, as we should expect if the determinant of ℓ is the factor by which ℓ expands or contracts volume. Now, at long last, we can define the **determinant** of $\ell \in \mathrm{End}(V)$, denoted by $\det(\ell)$ or $|\ell|$, to be $\Delta_{\mathbf{b}}(\ell)$, where $\mathbf{b}$ may be *any* basis of V.

Here are the main things we know about the function $\det : \mathrm{End}(V) \to k$ at this point:

(a) $\det(\mathrm{Id}_V) = 1$;

(b) $\det(\cdot)$ is multilinear and alternating, in the sense that for any basis $\mathbf{b}$ the function

$$(w_1, \dots, w_n) \mapsto \det([w_1, \dots, w_n]_{\mathbf{b}})$$

from V^n to k has these properties;

(c) For all $\ell \in \mathrm{End}(V)$, if A is the matrix of ℓ with respect to some basis (for both the domain and range) then $\det(\ell) = |A|$. Consequently:

(i) $\det(m \circ \ell) = \det(m)\det(\ell)$ for all $\ell, m \in \mathrm{End}(V)$;

(ii) for all $\ell \in \mathrm{End}(V)$, $\det(\ell) = 0$ if and only if ℓ is singular.

In addition, det is the unique function satisfying (a) and (b).

5.6 The Characteristic Polynomial

Sometimes in mathematics one works very hard to attain conclusions that can be stated in a few lines. There are other times when a simple observation unearths an abundance of interesting consequences. Having labored to develop the theory of the determinant, we will now start with a given $\ell \in \mathrm{End}(V)$ and study

$$p_\ell(t) := \det(\ell - t \cdot \mathrm{Id}_V).$$

This is called the **characteristic polynomial** of ℓ, and we can think of it as a function from k to k.

Suppose that V' is another n-dimensional vector space over k and $\ell' \in \mathrm{End}(V')$ is similar to ℓ, so that $\ell' = \iota \circ \ell \circ \iota^{-1}$ for some linear isomorphism $\iota : V \to V'$. Then, for any t,

$$\ell' - t \cdot \mathrm{Id}_{V'} = \iota \circ \ell \circ \iota^{-1} - t \cdot (\iota \circ \mathrm{Id}_V \circ \iota^{-1}) = \iota \circ (\ell - t \cdot \mathrm{Id}_V) \circ \iota^{-1}.$$

Therefore $\ell - t \cdot \mathrm{Id}_V$ and $\ell' - t \cdot \mathrm{Id}_{V'}$ are similar and consequently have the same determinant, so $p_{\ell'}(t) = p_\ell(t)$ for all t. That is, *the characteristic polynomial is invariant.* Hopefully you don't need to be told that this should lead you to suspect that it's interesting and important.

If A is the matrix of ℓ with respect to any basis, then

$$p_\ell(t) = |A - t \cdot I|,$$

and the right hand side can be thought of as the determinant of a matrix with entries in $k[t]$, in which case $p_\ell(t)$ is itself an element of $k[t]$. This formula can be expressed in a way that is easier to manipulate algebraicly if we introduce a piece of notation that is often useful. For $1 \le i, j \le n$ let

$$\delta_{ij} := \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

This is called the **Kronecker delta** in honor of Leopold Kronecker. The $n \times n$ identity matrix is now $I = (\delta_{ij})$, so the entries of $A - t \cdot I$ are $a_{ij} - t\delta_{ij}$, and we have

$$p_\ell(t) = |A - t \cdot I| = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma)(a_{\sigma(1)1} - t\delta_{\sigma(1)1}) \cdots (a_{\sigma(n)n} - t\delta_{\sigma(n)n}). \quad (*)$$

Let

$$p_\ell(t) = c_n t^n + c_{n-1} t^{n-1} + \cdots + c_1 t + c_0 \in k[t].$$

Somehow we started with a single invariant, the determinant, and rather magically turned it into what seem to be $n+1$ invariants $c_0, \ldots, c_{n+1}$. Let's look at a few of these a bit more closely. Imagine expanding the right hand side of $(*)$ using the distributive law, obtaining $n!2^n$ terms, and then gathering together the terms involving like powers of t, thereby obtaining expressions for the coefficients $c_0, \ldots, c_n$ of $p_\ell(t)$. The terms that don't include a power of t (or include t^0, if you prefer to think of it that way) are precisely the terms in the expansion of the determinant of A, so $c_0 = |A| = \det(\ell)$, which we've seen before. In order to get a term in the expansion that involves t^n, and is nonzero, we need to have $\delta_{\sigma(i)i} = 1$ for all i, so that σ is the identity permutation e. Clearly there will be exactly one such term, namely $\operatorname{sgn}(e)(-t)^n = (-1)^n t^n$, so $c_n = (-1)^n$ is not an interesting invariant.

But the other coefficients $c_1, \ldots, c_{n-1}$ are all new and nontrivial invariants of ℓ. Of these, the last is by far the simplest and most important. To evaluate it we consider each term

$$\operatorname{sgn}(\sigma)(a_{\sigma(1)1} - t\delta_{\sigma(1)1}) \cdots (a_{\sigma(n)n} - t\delta_{\sigma(n)n})$$

in our formula for $p_\ell(t)$. If the degree of this polynomial is $n-1$ or n, then there must be at least $n-1$ indices i such that $\sigma(i) = i$. But the only permutation with this property is the identity permutation e. Therefore c_{n-1} is the coefficient of t^{n-1} in the polynomial

$$(a_{11} - t) \cdots (a_{nn} - t).$$

If we use the distributive law to expand this into 2^n terms, the ones that have t raised to the power $n-1$ are $a_{11}(-t)^{n-1}, \ldots, a_{nn}(-t)^{n-1}$. Thus c_{n-1} is $(-1)^{n-1}$ times the **trace** of A, which is, by definition,

$$a_{11} + \cdots + a_{nn}.$$

If $0 \neq v \in V$ and $\ell(v) = rv$ for some scalar r, then we say that v is an **eigenvector** of ℓ, and that r is the associated **eigenvalue**. ('Eigen' is the German word for "self.") In this circumstance $\ell - r \cdot \mathrm{Id}_V$ is singular, so

$$p_\ell(r) = \det(\ell - r \cdot \mathrm{Id}_V) = 0.$$

Conversely, if r is a root of p_ℓ, then $\ell - r \cdot \mathrm{Id}_V$ is singular, so there is a nonzero $v \in V$ such that $(\ell - r \cdot \mathrm{Id}_V)(v) = 0$, i.e., v is an eigenvector. Thus the eigenvalues of ℓ are precisely the roots of the characteristic polynomial, and for each eigenvalue there is at least one associated eigenvector. More precisely, for each eigenvalue r the associated **eigenspace** is the kernel of $\ell - r\mathrm{Id}_V$, which has positive dimension.

The next result gives a partial answer to the classification problem that originally motivated our discussion of the determinant, now rather long ago.

Theorem 5.17. *Suppose that $p_\ell(t)$ has n distinct roots in k, so that*

$$p_\ell(t) = (-1)^n (t - r_1) \cdots (t - r_n),$$

where $r_1, \ldots, r_n$ are all distinct. For each $i = 1, \ldots, n$ let v_i be an eigenvector of ℓ associated with r_i. Then $v_1, \ldots, v_n$ are linearly independent, hence a basis of V. If $p_{\ell'} = p_\ell$, where V' is another n-dimensional vector space over k and $\ell' \in \mathrm{End}(V')$, then ℓ and ℓ' are similar.

Proof. Aiming at a contradiction, suppose that $v_1, \dots, v_n$ are linearly dependent, and let $0 = \alpha_1 v_1 + \cdots + \alpha_n v_n$ be a linear dependence that is minimal in the sense of having as few nonzero coefficients as any other linear dependence. Each v_i is nonzero because it is an eigenvector, so at least two of the scalars $\alpha_1, \dots, \alpha_n$ are nonzero. In addition, the eigenvalues are distinct, so, after reindexing if necessary, we can have $\alpha_1 \neq 0 \neq \alpha_2$ and $r_1 \neq 0$.

Applying ℓ to our linear dependence gives

$$0 = \ell(\alpha_1 v_1 + \cdots + \alpha_n v_n) = r_1 \alpha_1 v_1 + \cdots + r_n \alpha_n v_n.$$

Dividing this linear dependence by r_1 and subtracting it from the one we started with yields

$$0 = (1 - r_2/r_1)\alpha_2 v_2 + \ldots + (1 - r_n/r_1)\alpha_n v_n.$$

Since $(1 - r_2/r_1)\alpha_2 \neq 0$, this is a linear dependence, and it has fewer nonzero terms than the one we started with, contrary to our assumption of minimality. In view of this contradiction we now know that $v_1, \dots, v_n$ are linearly independent.

For each $i = 1, \dots, n$ let $v'_i \in V'$ be an eigenvector of ℓ' with associated eigenvalue r_i, so that $v'_1, \dots, v'_n$ is a basis of V'. Let $\iota : V \to V'$ be the linear transformation that takes each v_i to v'_i. Then $\iota \circ \ell \circ \iota^{-1} = \ell'$, so that ℓ and ℓ' are similar, because for any scalars $\alpha_1, \dots, \alpha_n$ there is the computation

$$\begin{aligned}\iota(\ell(\iota^{-1}(\alpha_1 v'_1 + \cdots + \alpha_n v'_n))) &= \iota(\ell(\alpha_1 v_1 + \cdots + \alpha_n v_n)) \\ &= \iota(r_1 \alpha_1 v_1 + \cdots + r_n \alpha_n v_n) \\ &= r_1 \alpha_1 v'_1 + \cdots + r_n \alpha_n v'_n = \ell'(\alpha_1 v'_1 + \cdots + \alpha_n v'_n).\end{aligned}$$

□

To what extent does this result fail to fully solve the problem of classifying linear endomorphisms? First of all, p_ℓ can fail to be a product of n linear factors if k is not algebraically complete. In connection with $\mathbb{R}$ and $\mathbb{C}$ this possibility is one instance of an important general principle of mathematics that flows out of the fundamental theorem of algebra: *the world of complex objects is fairly simple and orderly, but reality is complicated and messy.*

The second problem is that even when k is algebraically complete, p_ℓ can have fewer than n roots. Consider the linear transformation $\ell : k^2 \to k^2$ with matrix

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Then $p_\ell(t) = (1-t)^2$, which is the same as p_{Id_V}, and the only root of p_ℓ is 1. An eigenvector (x, y) is a solution to the system

$$x + y = x; \quad y = y,$$

so the set of eigenvectors $\{\, (\alpha, 0) : \alpha \in k \,\}$ is one dimensional. There cannot possibly be a basis whose elements are all eigenvectors. (Such a basis is called an **eigenbasis**.) In contrast *any* basis of k^2 is an eigenbasis for Id_V. In general, if $\ell' \in \mathrm{End}(V')$ is similar to ℓ because $\iota : V \to V'$ is an isomorphism such that $\ell' \circ \iota = \iota \circ \ell$, then $v \in V$ is an eigenvector of ℓ if and only if $\iota(v)$ is an eigenvector of ℓ', as you can easily verify for yourself. Therefore ℓ and Id_V cannot be similar, even though they have the same characteristic polynomial.

5.7 The Cayley-Hamilton Theorem

We now study a beautiful theorem of Arthur Cayley (1821-1895) and William Rowan Hamilton (1805-1865) that will play an important role in the next section's analysis of the classification problem. Before diving in, I should say that the material in this section and the next is quite a bit more advanced than what we've done so far, and it won't play a role in the rest of the book, so you are certainly free to skip it if that is your inclination. But it is also quite beautiful, and one of the high points of 19^{th} mathematics, so if you choose to skip ahead now, I hope you'll come back sometime when you are in the mood for a challenge.

Suppose k is a field, V is a vector space over k, and $\ell \in \mathrm{End}(V)$. The Cayley-Hamilton theorem considers the result of substituting ℓ for t in the characteristic polynomial of ℓ. If $p(t) = a_m t^m + \cdots a_1 t + a_0 \in k[t]$ is any univariate polynomial, then p "evaluated" at ℓ is

$$p(\ell) = a_m \ell^m + \cdots + a_1 \ell + a_0 \mathrm{Id}_V \in \mathrm{End}(V)$$

where each ℓ^i as the i-fold composition $\ell \circ \cdots \circ \ell$ of ℓ with itself.

But what, precisely, are we doing here? We are quite accustomed to evaluating a polynomial $f(t) \in R[t]$, where R is a commutative ring, at an element of R. Can we think of $p(\ell)$ in this way?

Well, $\mathrm{End}(V)$ is a ring with unit if we define multiplication in the ring to be functional composition. To see this you should mentally check that (R1)-(R7) hold, or observe that once we fix a basis for V, $\mathrm{End}(V)$ can be identified with the ring $M_n(k)$ of $n \times n$ matrices with entries in k. But it's not commutative, except when $n = 1$, which suggests there might be some problems. In particular, we haven't defined a determinant for matrices with entries in a noncommutative ring. Fortunately, for any fixed ℓ the set E_ℓ of all expressions of the form $c_m\ell^m + \cdots + c_1\ell + c_0\mathrm{Id}_V$ is a subring of $\mathrm{End}(V)$ that *is* commutative. (Again, think about how to prove (R8); it's not hard, but there is a rather bulky calculation.) Even so, it may seem strange to evaluate an element of $k[t]$ by substituting an element of some different ring, but there is actually a simple way of thinking about this. If we identify each field element $\alpha \in k$ with the linear transformation $v \mapsto \alpha v$, then we can think of k as a subring of E_ℓ, so that we are really evaluating a polynomial in $E_\ell[t]$ at an element of E_ℓ, in the usual way, except that the polynomial happens to lie in $k[t] \subset E_\ell[t]$.

In particular, let

$$p_\ell(t) = (-1)^n t^n + c_{n-1}t^{n-1} + \cdots + c_1 t + c_0$$

be the characteristic polynomial of ℓ. Suppose that v is an eigenvector of ℓ with associated eigenvalue r. Then

$$\begin{aligned} p_\ell(\ell)v &= (-1)^n\ell^n(v) + c_{n-1}\ell^{n-1}(v) + \cdots + c_1\ell(v) + c_0\mathrm{Id}_V(v) \\ &= (-1)^n r^n v + c_{n-1}r^{n-1}v + \cdots + c_1 rv + c_0 v \\ &= ((-1)^n r^n + c_{n-1}r^{n-1} + \cdots + c_1 r + c_0)v \\ &= p_\ell(r)v = 0. \end{aligned}$$

If ℓ has an eigenbasis, then $p_\ell(\ell)$ is zero because it maps each element of the eigenbasis to zero.

When $k = \mathbb{C}$, $\mathrm{End}(V)$ is a finite dimensional vector space that has a natural topology derived from any norm. Since $\mathbb{C}$ is algebraically complete, the characteristic polynomial of any element of $\mathrm{End}(V)$ is a product of n linear factors, and it seems intuitively reasonable to guess that the characteristic polynomial of a "typical" or "generic" element of $\mathrm{End}(V)$ will have n distinct eigenvalues, or that a "random" element will have n distinct eigenvalues "with probability one." It isn't necessary to explain these concepts in detail—they are advanced and beyond the scope of this book—because here we are only concerned with developing intuition in support of a weaker concept that we will define precisely. A subset of a general topological space X is said to be **dense** if its closure is all of X, and hopefully it seems plausible that the set of elements of $\mathrm{End}(V)$ with n distinct eigenvalues is dense. If ℓ has n distinct eigenvalues, then $p_\ell(\ell) = 0$, so the set of such ℓ is a subset of

$$S := \{\, \ell \in \mathrm{End}(V) : p_\ell(\ell) = 0 \,\},$$

and we should expect S to be dense. It is not terribly difficult to show that $\ell \mapsto p_\ell(\ell)$ is a continuous function from $\mathrm{End}(V)$ to itself, and $\{0\}$ is closed in $\mathrm{End}(V)$, so S is a closed subset of $\mathrm{End}(V)$, and if it is also dense, then it must be all of $\mathrm{End}(V)$. This line of reasoning leads us to expect that $p_\ell(\ell) = 0$ even when ℓ does not have an eigenbasis, and in fact this is the case:

Theorem 5.18 (Cayley-Hamilton Theorem). *For all $\ell \in \mathrm{End}(V)$,*

$$p_\ell(\ell) = 0.$$

It is possible (using some high level results) to prove this when $k = \mathbb{C}$ by fleshing out the ideas described above, but for an algebraist such a proof would be highly unsatisfactory. A guiding principle of research in algebra is that if a theorem has a purely algebraic statement, then there should be an algebraic proof, and when a proof using topology or other techniques of analysis is known, the search for an algebraic proof becomes an important goal of algebraic research. This is much more than a matter of wanting to extend the result to fields other than $\mathbb{C}$ or $\mathbb{R}$: when there is an algebraic theorem with an analytic proof, but no algebraic proof (such situations have sometimes persisted for years or decades) there *must* be a gap in our understanding.

Below we give an algebraic proof, at the heart of which is a fact called **Cramer's rule**, after Gabriel Cramer (1704-1752). Consider an $n \times n$ matrix C with entries in R where R is commutative ring with unit. The **adjugate** or **classical adjoint** of C is the $n \times n$ matrix $\mathrm{adj}(C)$ whose ji-entry $\mathrm{adj}_{ji}(C)$ is the determinant

$$\sum_{\sigma \in S_n, \sigma(j)=i} \mathrm{sgn}(\sigma) c_{\sigma(1)1} \cdots c_{\sigma(j-1)j-1} \cdot 1 \cdot c_{\sigma(j+1)j+1} \cdots c_{\sigma(n)n}$$

of the matrix obtained from C by replacing c_{ij} with 1 and replacing all other entries of the i^{th} row and the j^{th} column by 0. (Note the reversal of row and column indices!)

Theorem 5.19 (Cramer's Rule). *If R is a commutative ring with unit, and C is an $n \times n$ matrix with entries in R, then*

$$\mathrm{adj}(C)C = \det(C)I.$$

Proof. We compute each of the entries of $\mathrm{adj}(C)C$. First observe that the j^{th} diagonal entry, namely the product $\sum_{i=1}^{n} \mathrm{adj}_{ji}(C) c_{ij}$ of the j^{th} row of

$\mathrm{adj}(C)$ and the j^{th} column of C, is

$$\sum_{i=1}^{n}\Big(\sum_{\sigma\in S_n,\sigma(j)=i} \mathrm{sgn}(\sigma)c_{\sigma(1)1}\cdots c_{\sigma(j-1)j-1}\cdot 1\cdot c_{\sigma(j+1)j+1}\cdots c_{\sigma(n)n}\Big)c_{ij}$$

$$= \sum_{\sigma\in S_n} \mathrm{sgn}(\sigma)c_{\sigma(1)1}\cdots c_{\sigma(j-1)j-1}\cdot c_{\sigma(j)j}\cdot c_{\sigma(j+1)j+1}\cdots c_{\sigma(n)n} = \det(C).$$

Now suppose that $k \neq j$. Then the product $\sum_{i=1}^n \mathrm{adj}_{ji}(C)c_{ik}$ of the j^{th} row of $\mathrm{adj}(C)$ and the k^{th} column of C is

$$\sum_{i=1}^{n}\Big(\sum_{\sigma\in S_n,\sigma(j)=i} \mathrm{sgn}(\sigma)c_{\sigma(1)1}\cdots c_{\sigma(j-1)j-1}\cdot 1\cdot c_{\sigma(j+1)j+1}\cdots c_{\sigma(n)n}\Big)c_{ik}$$

$$= \sum_{\sigma\in S_n} \mathrm{sgn}(\sigma)c_{\sigma(1)1}\cdots c_{\sigma(j-1)j-1}\cdot c_{\sigma(j)k}\cdot c_{\sigma(j+1)j+1}\cdots c_{\sigma(n)n},$$

and this quantity is zero because it is the determinant of a matrix with two identical columns, namely the matrix obtained by replacing the j^{th} column of C with the k^{th} column. □

In textbooks Cramer's rule is usually presented as a method for computing the determinant "by hand," and is written as the formula

$$C^{-1} = \mathrm{adj}(C)/|C|$$

for the inverse of C. Of course this version is less general, since it doesn't make sense unless R is a field, and even then its validity depends on $|C| \neq 0$. In addition, it is customary to express Cramer's rule in terms of the **cofactors** $M_{ij} := (-1)^{i+j}\mathrm{adj}_{ij}(C)$, reflecting a feeling that it is psychologically easier to compute M_{ij} because it is the determinant of the matrix obtained from C by deleting the j^{th} row and the i^{th} column.

Proof of Theorem 5.18. Let $\mathbf{b}_1,\ldots,\mathbf{b}_n$ be a basis of V, let A be the matrix of ℓ with respect to this basis, and set

$$B := \begin{pmatrix} a_{11}\mathrm{Id}_V - \ell & \cdots & a_{1n}\mathrm{Id}_V \\ \vdots & \ddots & \vdots \\ a_{n1}\mathrm{Id}_V & \cdots & a_{nn}\mathrm{Id}_V - \ell \end{pmatrix}.$$

Then $p_\ell(\ell)$ is the determinant of B, so our goal is to show that $\det(B) = 0 \in \mathrm{End}(V)$.

Since A is the matrix of ℓ,

$$a_{i1}\mathbf{b}_1 + \cdots + a_{in}\mathbf{b}_n - \ell(\mathbf{b}_i) = 0$$

for each $i = 1, \ldots, n$. This equation can be rewritten as

$$B \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix} = 0 \in V^n.$$

(This formula is a bit different than anything we've seen before, insofar as we are multiplying a matrix of endomorphisms of V with a column vector whose entries are elements of V, but it makes perfect sense if we interpret the product of an endomorphism and a vector as the result of applying the endomorphism to the vector.) Multiplying both sides of this equation by the adjugate of B gives

$$\operatorname{adj}(B)\Big(B \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}\Big) = 0.$$

Below we will show that we can pass from this to

$$\big(\operatorname{adj}(B)B\big) \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix} = 0,$$

after which Cramer's rule yields

$$0 = \Big(\det(B)\begin{pmatrix} \mathrm{Id}_V & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathrm{Id}_V \end{pmatrix}\Big) \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix} = \begin{pmatrix} \det(B) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \det(B) \end{pmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}.$$

Since $\mathbf{b}_1, \ldots, \mathbf{b}_n$ is a basis, it follows that $\det(B)$ is the origin in $\operatorname{End}(V)$.

The remaining step is to show that $M(Nw) = (MN)w$ whenever M and N are $n \times n$ matrices with entries in E_ℓ and w is a column vector whose entries are elements of V. (The only *real* difficulty here is recognizing that this needs to be proved!) That the i^{th} entries of $M(Nw)$ and $(MN)w$ are the same is shown by the calculation

$$\sum_j m_{ij}\Big(\sum_k n_{jk}(w_k)\Big) = \sum_k \Big(\sum_j m_{ij}(n_{jk}(w_k))\Big) = \sum_k \Big(\sum_j m_{ij} \circ n_{jk}\Big) w_k$$

in which the first equality combines linearity and commutativity of addition, and the second equality is simply the definition of addition of endomorphisms. □

5.8 Canonical Forms

We now explain the approach that gives the best understanding of the classification problem. The material will be a bit more technical than the rest of the chapter in two senses. First, our approach is rather abstract, taking advantage of the material on rings and modules introduced in Chapter 2. But this aspect is not (at least as I see things) very difficult. Quite the contrary: this section exemplifies how the abstract approach can simplify and clarify ideas that might seem much more complex if they were presented only in relation to the particular application under consideration.

But there are also two theorems whose proofs are rather technical. Since that sort of material is contrary to the spirit of this book, I have decided not to include these arguments in the main text. Rather paradoxically, it seems better to leave them for you to work out, and each is sketched in a sequence of problems, among those at the end of the book related to this chapter. You don't have to do them if you prefer not to, but (after being broken down into smaller steps) they are not that difficult; the arguments are "technical" primarily in the sense that there are many details that need to be attended to. Working these problems would certainly strengthen and deepen your understanding and appreciation of this topic.

Throughout this section we work with a fixed n-dimensional vector space V and a given endomorphism $\ell \in \mathrm{End}(V)$. The general idea will be to show that we can find a basis with respect to which the matrix of ℓ has a certain "canonical" form. If the canonical matrices of two linear transformations are the same, then the two linear transformations are similar, of course. The main point will be that if the canonical forms of two linear transformations are different, then the transformations are not similar, so the canonical form solves the classification problem.

What might a canonical form look like, and how might one find one? The usual approach to thinking about a question like this is to first think about special cases that suggest some sort of answer, hoping that one can generalize its main features. If ℓ has an eigenbasis $\mathbf{b}_1, \dots, \mathbf{b}_n$, then its matrix with respect to this basis has zeros off the main diagonal (such a matrix is called a **diagonal matrix**) and the diagonal entries are the eigenvalues. For each eigenvalue λ, the associated eigenspace includes all the $\mathbf{b}_i$ such that $\ell(\mathbf{b}_i) = \lambda \mathbf{b}_i$. It obviously includes the span of these $\mathbf{b}_i$, and by writing an arbitrary element of V as a linear combination of $\mathbf{b}_1, \dots, \mathbf{b}_n$, then applying ℓ, one can easily see that this span includes every element of the eigenspace, so it is the eigenspace. Note that ℓ maps each eigenspace to itself; a subspace with this property is called an **invariant subspace** of ℓ. Every element of V can be written as a sum of elements of the eigenspaces, so we have decomposed V into a sum of invariant subspaces.

In general ℓ won't have an eigenbasis, but we can always think about decomposing V into a sum of invariant subspaces. If such a decomposition is, in some sense, determined by the similarity class of ℓ, we can try to develop a canonical form for ℓ by looking for canonical forms for the restrictions of ℓ to each of the invariant subspaces in the decomposition. On the other hand, if there is no nontrivial decomposition of a particular sort, that information can be used to guide and refine our search for a canonical form for ℓ. For each $v \in V$ the span of $v, \ell(v), \ell^2(v), \dots$ is an invariant subspace, and it turns out that these subspaces are especially useful.

There are some general remarks that will play a role in what follows. Let A be a k-algebra. (Recall that this means that A is a ring that has k as a subring.) Then each $a \in A$ induces a homomorphism $\varphi_a : k[X] \to A$ given by the formula

$$\varphi_a(p) := p(a).$$

The formal verification that this is a homomorphism is, as usual, trivial and mechanical: if $p, q \in k[X]$, then

$$\varphi_a(p+q) = (p+q)(a) = p(a) + q(a) = \varphi_a(p) + \varphi_a(q)$$

and

$$\varphi_a(pq) = (pq)(a) = p(a)q(a) = \varphi_a(p)\varphi_a(q).$$

As we saw already in the last section, it can be important that the image of this homomorphism, namely the set of polynomial functions of a, is a commutative ring even when A is not commutative.

Now recall that in Chapter 2 we showed that $k[X]$ is Euclidean, hence a PID. The kernel of φ_a is an ideal, of course, hence a principal ideal, and any generator is called a **minimal polynomial** for a. If the ideal is (0), then 0 is the unique minimal polynomial. Otherwise we can divide a minimal polynomial by its leading coefficient to obtain a polynomial that is minimal (since it generates the same ideal) and monic. There is a terminological convention that facilitates our discussion, namely agreeing (unless stated otherwise, or when the kernel is (0)) that a minimal polynomial should be understood to be monic. There cannot be two distinct monic minimal polynomials because they would have to have the same degree (each is an element of the ideal generated by the other) and their difference would be a nonzero element of the kernel of φ_a that had lower degree than either (putatively) minimal polynomial, and was consequently outside the ideal they each generate. Therefore there is a unique monic minimal polynomial that will be called *the* minimal polynomial of a.

Let $\varphi : S \to R$ be a ring homomorphism between rings with unit that takes the unit $1_S \in S$ to the unit $1_R \in R$, and let M be a left R-module. Then φ induces an S-module structure on M given by the formula $sm := \varphi(s)m$. This is fairly obvious, but it's "virtuous" to go through the verification anyway. Of course

$$1_S \cdot m = \varphi(1_S)m = 1_R \cdot m = m$$

for any $m \in M$. If $s \in S$ and $m_1, m_2 \in M$, then

$$s(m_1 + m_2) = \varphi(s)(m_1 + m_2) = \varphi(s)m_1 + \varphi(s)m_2 = sm_1 + sm_2.$$

If $s_1, s_2 \in S$ and $m \in M$, then

$$\begin{aligned}(s_1 + s_2)m = \varphi(s_1 + s_2)m &= (\varphi(s_1) + \varphi(s_2))m \\ &= \varphi(s_1)m + \varphi(s_2)m = s_1 m + s_2 m\end{aligned}$$

and

$$(s_1 s_2)m = \varphi(s_1 s_2)m = (\varphi(s_1)\varphi(s_2))m = \varphi(s_1)(\varphi(s_2)m) = s_1(s_2 m).$$

Let's now apply these generalities to the problem at hand. As we explained in the last section, if we define multiplication to be functional composition, then $\operatorname{End}(V)$ is a ring with unit $1 = \mathrm{Id}_V$, and in fact $\operatorname{End}(V)$ is a k-algebra if we identify each $\alpha \in k$ with the the endomorphism $v \mapsto \alpha v$. It is easy to see (but you should check the details for yourself) that the vector space structure of V extends to an $\operatorname{End}(V)$-module structure if we define the scalar product of $l \in \operatorname{End}(V)$ and $v \in V$ to be $l(v)$. For this reason (but also simply because it makes things more attractive visually and easier to read) we will often write lv rather than $l(v)$. As an element of $\operatorname{End}(V)$, ℓ induces a homomorphism

$$\varphi_\ell : k[X] \to \operatorname{End}(V),$$

and φ_ℓ induces a $k[X]$-module structure on V given by the function $(q, v) \mapsto \varphi_\ell(q)v = q(\ell)v$. We will sometimes write R_ℓ in place of $k[X]$ when we wish to emphasize this structure.

We now use the fact that V is finite dimensional. Any basis of V induces an isomorphism between $\operatorname{End}(V)$ and the ring $M_n(k)$ of $n \times n$ matrices with entries in k, so $\operatorname{End}(V)$ is n^2-dimensional. Since $k[X]$ is infinite dimensional and φ_ℓ is linear (as a map between vector spaces over k) its kernel cannot be (0), and consequently the minimal polynomial of ℓ cannot be equal to 0. Concretely, the minimal polynomial is a polynomial

$$p = a_0 + a_1 X + \cdots + a_{m-1}X^{m-1} + X^m$$

whose coefficients are the coefficients of a linear dependence

$$a_0 \mathrm{Id}_V + a_1\ell + \cdots + a_{m-1}\ell^{m-1} + \ell^m = 0$$

with the additional property that for any $m' < m$, $\ell^{m'}$ cannot be expressed as a linear combination of $\mathrm{Id}_V, \ell, \ldots, \ell^{m'-1}$.

We claim that the minimal polynomial of ℓ is an invariant, i.e., it depends only on the similarity class of ℓ. Let $\iota : V \to V'$ be an isomorphism. From a technical point of view the key is the following lemma, which implies that if $\ell' = \iota \circ \ell \circ \iota^{-1}$, then $q(\ell') = \iota \circ q(\ell) \circ \iota^{-1}$ for every $q \in k[X]$, so $q(\ell) = 0$ if and only if $q(\ell') = 0$, which is to say that φ_ℓ and $\varphi_{\ell'}$ have the same kernel.

Lemma 5.20. *If W and W' are vector spaces over k, $\alpha : W \to W'$ is linear, and $l \in \mathrm{End}(W)$ and $l' \in \mathrm{End}(W')$ satisfy $\alpha \circ l = l' \circ \alpha$, then $\alpha \circ q(l) = q(l') \circ \alpha$ for all $q \in k[X]$.*

Proof. For any nonnegative integer h and any $c \in k$ we have

$$\alpha \circ cl^h = c(\alpha \circ l^h) = c(l' \circ \alpha \circ l^{h-1}) = \cdots = c({l'}^{h-1} \circ \alpha \circ l) = c{l'}^h \circ \alpha.$$

Thus the claim holds when q is a monomial, and it holds in general because if $l_1, \ldots, l_k \in \mathrm{End}(W)$ and $l'_1, \ldots, l'_k \in \mathrm{End}(W')$ with $\alpha \circ l_j = l'_j \circ \alpha$ for all j, then

$$\alpha \circ (l_1 + \cdots + l_k) = \alpha \circ l_1 + \cdots + \alpha \circ l_k = l'_1 \circ \alpha + \cdots + l'_k \circ \alpha = (l'_1 + \cdots + l'_k) \circ \alpha.$$

□

It's not really germaine to the discussion here, but it is interesting to note that the minimal polynomial p diagnoses whether ℓ is singular or nonsingular. If the constant term a_0 is different from 0, then ℓ is invertible because

$$\mathrm{Id}_V = \ell\big(- (a_1/a_0) - (a_2/a_0)\ell - \cdots - (a_m/a_0)\ell^{m-1}\big).$$

If, on the other hand, $a_0 = 0$, then ℓ cannot be invertible because if it was the computation

$$0 = \ell^{-1}p(\ell) = a_1 + a_2\ell + \cdots + a_m\ell^{m-1}$$

would contradict the minimality of p.

The map $q + \ker \varphi_\ell \mapsto \varphi_\ell(q)$ is a vector space isomorphism between $k[X]/\ker \varphi_\ell$ and the image of φ_ℓ. The dimension of $k[X]/\ker \varphi_\ell$ is m ($1 + \ker \varphi_\ell, \ldots, X^{m-1} + \ker \varphi_\ell$ is a basis) and the dimension of $\mathrm{End}(V)$ is n^2, so $m \le n^2$. But actually *the Cayley-Hamilton theorem is precisely the assertion that the characteristic polynomial of ℓ is in the kernel of φ_ℓ.* The characteristic polynomial has degree n, so $m \le n$. This is crucial! A quick browse of this section might suggest that what we are doing here is completely unrelated to the theory of the determinant, but without this "little fact" it would be impossible to develop the theory. In this sense everything we have learned about the determinant in earlier sections is not just "relevant." It is indispensable.

In general a left R-module M is **cyclic** if there is some $m \in M$ such that $M = Rm$. We now study the case in which V is a cyclic R_ℓ-module, so $V = R_\ell v$ for some $v \in V$. Note that $V = R_\ell v$ if and only if $v, \ell(v), \ell^2(v), \ldots$ span V. We saw above that $m \le n$, and when V is a cyclic R_ℓ-module, the reverse inequality also holds, so that $m = n$. To see this observe that for any $\mu \ge m$ we have

$$\ell^\mu(v) = -a_0 \ell^{\mu-m}(v) - \cdots - a_{m-1}\ell^{\mu-1}(v)$$

because this equation holds when $\mu = m$ since $p(\ell)v = 0$, and the general case can be obtained from this case by applying $\ell^{\mu-m}$ to both sides. Therefore (by induction) $\ell^m(v), \ell^{m+1}(v), \ldots$ are spanned by $v, \ell(v), \ldots, \ell^{m-1}(v)$. Since $n \ge m$, the vectors $v, \ell(v), \ldots, \ell^{m-1}(v)$ must constitute a basis.

Simply by considering the image under ℓ of each element of this basis, we find that the matrix of ℓ with respect to it is

$$C(p) := \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -a_{m-2} \\ 0 & 0 & 0 & \cdots & 1 & -a_{m-1} \end{pmatrix}.$$

In the cyclic case the minimal polynomial is a complete invariant: if two linear transformations have the same minimal polynomial p and are cyclic, in the sense that for each there is an element of the vector space such that a basis is generated by repeatedly applying the transformation to this element, then the matrix of each with respect to this basis is $C(p)$, so the two transformations are similar. It is not obvious at this point, but eventually it will become apparent that in general two elements of $\mathrm{End}(V)$ that have the same minimal polynomial need not be similar, so the minimal polynomial does not completely solve the general classification problem.

If M is a left R-module, then the **annihilator** of $m \in M$ is

$$\mathrm{Ann}(m) = \{\, a \in R : am = 0 \,\}.$$

Clearly $\mathrm{Ann}(m)$ is a left ideal of R. We say that M is the **internal direct sum** of submodules $M_1, \dots, M_k$, and we write

$$M = M_1 \oplus \cdots \oplus M_k,$$

if each $m \in M$ has a unique representation of the form $m = m_1 + \cdots + m_k$ with $m_i \in M_i$ for each i. The key to extending our analysis beyond the cyclic case is the first of the two "technical" results mentioned at the outset:

Theorem 5.21 (Structure Theorem for Principal Ideal Domains)**.** *If R is a PID and M is a finitely generated R-module, then M is a finite internal direct sum $Rg_1 \oplus \cdots \oplus Rg_r$ of cyclic submodules where*

$$\mathrm{Ann}(g_1) \supset \mathrm{Ann}(g_2) \supset \cdots \supset \mathrm{Ann}(g_r).$$

If $g_1', \dots, g_s'$ is another system of generators with these properties, then $s = r$ and $\mathrm{Ann}(g_i') = \mathrm{Ann}(g_i)$ for all i. In addition r is the minimal number of elements of any system of generators.

In view of this result we can choose $v_1, \dots, v_r$ such that

$$V = R_\ell v_1 \oplus \cdots \oplus R_\ell v_r$$

with $\mathrm{Ann}(v_1) \supset \cdots \supset \mathrm{Ann}(v_r)$. For each $j = 1, \dots, r$ let $V_j := R_\ell v_j$, let $\ell_j := \ell|_{V_j}$, and let $\pi_j : V \to V_j$ be the linear map $w_1 + \cdots + w_r \mapsto w_j$. Then $\ell_j \circ \pi_j = \pi_j \circ \ell$, and consequently (Lemma 5.20) $q(\ell_j) \circ \pi_j = \pi_j \circ q(\ell)$ for all $q \in k[X]$. In particular,

$$q(\ell)v_j = 0 \iff \pi_j \circ q(\ell) = 0 \iff q(\ell_j) \circ \pi_j = 0 \iff q(\ell_j) = 0,$$

so the monic generator of $\mathrm{Ann}(v_j) = \ker \varphi_{\ell_j}$ is the minimal polynomial of ℓ_j. Denote this polynomial by p_j. Then $(p_j) = \mathrm{Ann}(v_j) \supset \mathrm{Ann}(v_{j+1}) = (p_{j+1})$, so p_j divides p_{j+1}. Since $q(\ell) = 0$ if and only if $q(\ell)v_j = 0$ for all j, and this is the case if and only if $q(\ell_j) = 0$ for all j, the minimal polynomial of ℓ is the least common multiple of $p_1, \dots, p_r$, which is p_r.

Since each V_j is a cyclic R_{ℓ_j}-module, by choosing, for each j, a basis of V_j consisting of the appropriate initial segment of $v_j, \ell(v_j), \ell^2(v_j), \ldots$, we obtain a basis of V with respect to which the matrix of ℓ is

$$\begin{pmatrix} C(p_1) & & & \\ & C(p_2) & & \\ & & \ddots & \\ & & & C(p_r) \end{pmatrix}.$$

This is called the **coarse canonical form** of ℓ.

We now wish to show that the coarse canonical form is an invariant of ℓ. The objects that determine the coarse canonical form, namely the ideals $\mathrm{Ann}(v_1), \ldots, \mathrm{Ann}(v_r)$, are determined by the R_ℓ-module structure of V, so it suffices to show that if $\iota : V \to V'$ is an isomorphism and $\ell' = \iota \circ \ell \circ \iota^{-1}$, then V and V' are isomorphic as $k[X]$-modules. But Lemma 5.20 implies that $\iota \circ q(\ell) = q(\ell') \circ \iota$ for all $q \in R_\ell$. Formally, the verification that ι is a $k[X]$-module homomorphism consists of the linearity of ι, which implies that ι is a homomorphism of the underlying commutative groups, together with the fact that for all $q \in k[X]$ and $v \in V$ we have

$$\iota(qv) = \iota(q(\ell)v) = q(\ell')(\iota(v)) = q\iota(v).$$

Since two linear transformations with the same coarse canonical form are similar, the coarse canonical form is a complete invariant. In a certain sense we have solved the classification problem.

In another sense this resolution of the issue is not fully satisfactory, or at least it leaves a desire for results that give a fuller and more descriptive picture of the structure of ℓ. When ℓ had an eigenbasis we obtained a related decomposition of V as an internal direct sum of invariant subspaces, but even in that special case the relation between the coarse canonical form and the eigenvalues and eigenspaces of ℓ is at least a bit murky. It would be nice to have canonical forms that were more closely related to the eigenvalues.

One consequence of the Cayley-Hamilton theorem is that every root of the minimal polynomial is an eigenvalue, because it is a root of the characteristic polynomial. The converse is true as well—any eigenvalue is a root of the minimal polynomial—by virtue of a rather clever and elegant argument. Suppose that λ is an eigenvalue, so $\ell(v) = \lambda v$ for some nonzero v. Then $\ell^2(v) = \ell(\lambda v) = \lambda\ell(v) = \lambda^2 v$, $\ell^3(v) = \lambda^3 v$, and so forth. Multiplying these equations by field elements and adding them together, we find that $q(\ell)v = q(\lambda)v$ for any polynomial $q \in k[X]$. In particular, for the minimal polynomial we have $0 = p(\ell)v = p(\lambda)v$, so $p(\lambda) = 0$.

But when k is not algebraicly complete it can easily happen that ℓ has no eigenvalues. Every polynomial in $k[X]$ factors as a product of linear factors when k is algebraicly complete, and when k is not algebraicly complete it is still the case that $k[X]$ is a unique factorization domain, as we saw in Chapter 2. The next result, which is the second of the two mentioned at the beginning of this section, shows us how to take advantage of this.

Theorem 5.22. *Suppose the minimal polynomial of ℓ has the prime factorization*

$$p = p_1^{e_1} \cdots p_k^{e_i}.$$

For each $i = 1, \dots, k$ let $q_i := \prod_{j \neq i} p_j^{e_j}$, and let U_i be the image of $q_i(\ell)$. Then $V = U_1 \oplus \cdots \oplus U_k$. In addition, each U_i is the kernel of $p_i^{e_i}(\ell)$ and an invariant subspace, and the minimal polynomial of $\ell|_{U_i}$ is $p_i^{e_i}$.

Let's apply the earlier analysis to one of the U_i. We have

$$U_i = W_{i1} \oplus \cdots \oplus W_{ir_i}$$

where each W_{ij} is a cyclic R_ℓ-submodule of U_i, say with generator v_{ij}. We saw above that the annihilator of v_{ij} is the minimal polynomial of $\ell|_{W_{ij}}$, and that the least common multiple of these minimal polynomials is the minimal polynomial of $\ell|_{U_i}$, which is $p_i^{e_i}$. In particular, since p_i is irreducible, the annihilator of v_{ij} is $(p_i^{f_{ij}})$ for some integer $f_{ij} \leq e_i$, and since $\mathrm{Ann}(v_{i1}) \supset \cdots \supset \mathrm{Ann}(v_{ir_i})$ we have $f_{i1} \leq \ldots \leq f_{ir_i} = e_i$. We have shown that we can find a basis of U_i with respect to which ℓ_i has the block diagonal matrix

$$R_i := \begin{pmatrix} C(p_i^{f_{i1}}) & & & \\ & C(p_i^{f_{i2}}) & & \\ & & \ddots & \\ & & & C(p_i^{f_{ir_i}}) \end{pmatrix},$$

and doing this for every i gives a basis for V with respect to which the matrix of ℓ is

$$\begin{pmatrix} R_1 & & & \\ & R_2 & & \\ & & \ddots & \\ & & & R_k \end{pmatrix}.$$

This matrix is called the **rational canonical form** of ℓ. The objects that determine the rational canonical form, namely the minimal polynomial $p = p_1^{e_1} \cdots p_k^{e_i}$, the integers $r_1, \dots, r_k$, and the integers f_{ij} for $1 \leq i \leq k$ and

$1 \leq j \leq r_i$, are all derived from the R_ℓ-module structure of V, so they are invariants of ℓ, and consequently the rational canonical form is a complete invariant.

In comparison with the coarse canonical form, the rational canonical form has one significant advantage. Suppose that (with respect to some basis) the matrix of ℓ is $C(q^e)$ where q is an irreducible polynomial. Then the dimension of V is the degree of q^e, which is e times the degree of q. It turns out that it is impossible to write V as a nontrivial internal direct sum $W_1 \oplus W_2$ of invariant subspaces because, for each $i = 1, 2$, the minimal polynomial of $\ell|_{W_i}$ must divide q^e, so it is q^{e_i} for some integer $e_i > 0$. The minimal polynomial of ℓ would then be $q^{\max\{e_1,e_2\}}$, and consequently the dimension of $W_1 \oplus W_2$ would be $e_1 + e_2 = e + \min\{e_1, e_2\} > e$ times the degree of q, which is impossible. In this sense the rational canonical form gives a decomposition of V as an internal direct sum of invariant subspaces that is as fine as possible, because each of the summands cannot be further decomposed.

The rational canonical form depends on the field k because in an extension of k it may no longer be the case that $p_1, \ldots, p_k$ are all irreducible. Passing to the extension field may or may not be simplifying, but if we have access to an algebraically complete field that contains k, there is at least a sense in which working with that field is a bit more "canonical." For this reason (and also just because it is simple) we should be particularly interested in what happens when each $p_i = X - \lambda_i$ is a linear monic polynomial, as is the case necessarily when k is algebraically complete. Each such λ_i is an eigenvalue because any root of the minimal polynomial is a root of the characteristic polynomial. As we saw earlier, every eigenvalue is a root of the minimal polynomial, so $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of ℓ.

Suppose that $(X - \lambda_i)^{f_{ij}}$ is the minimal polynomial of

$$\ell_{ij} := \ell|_{W_{ij}} : W_{ij} \to W_{ij}.$$

There is a vector $v \in W_{ij}$ such that $(\ell_{ij} - \lambda_i)^{f_{ij}-1}v \neq 0$, and for such a v the vectors $v, (\ell_{ij} - \lambda_i)v, \ldots, (\ell_{ij} - \lambda_i)^{f_{ij}-1}v$ are linearly independent. (If there was a linear dependence we could apply an appropriate power of $\ell_{ij} - \lambda_i$ to obtain a linear dependence with only one term, which is impossible.) We know that the dimension of W_{ij} is f_{ij}, so this collection of vectors is actually a basis, and the matrix of $\ell_{ij} - \lambda_i$ with respect to this basis has a 1 in each slot right below the main diagonal and 0's elsewhere. Consequently the matrix of $\ell_{ij} = (\ell_{ij} - \lambda_i) + \lambda_i$ with respect to this basis is

$$J_{f_{ij}}(\lambda_i) := \begin{pmatrix} \lambda_i & 0 & 0 & \cdots & 0 & 0 \\ 1 & \lambda_i & 0 & \cdots & 0 & 0 \\ 0 & 1 & \lambda_i & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i & 0 \\ 0 & 0 & 0 & \cdots & 1 & \lambda_i \end{pmatrix}.$$

This matrix is a called the f_{ij}-dimensional **basic Jordan block** belonging to λ_i.

For each i let

$$S_i := \begin{pmatrix} J_{f_{i1}}(\lambda_i) & & & \\ & J_{f_{i2}}(\lambda_i) & & \\ & & \ddots & \\ & & & J_{f_{ir_i}}(\lambda_i) \end{pmatrix}.$$

We have shown that if p is a product of linear factors, then we can find a basis for V with respect to which the matrix of ℓ is

$$\begin{pmatrix} S_1 & & & \\ & S_2 & & \\ & & \ddots & \\ & & & S_k \end{pmatrix}$$

This matrix is called the **Jordan canonical form** of ℓ. Since the similarity class of ℓ determines the minimal polynomial p, its roots $\lambda_1, \ldots, \lambda_k$ (which are the eigenvalues of ℓ) and the numbers f_{ij}, the Jordan canonical form is also a complete invariant for ℓ when its minimal polynomial is a product of linear factors.

Chapter 6

The Derivative

Sir Isaac Newton (1643-1727) was the most important scientist ever. That's not to say that he was the cleverest, or in some other sense the most intelligent. One can mount a respectable argument that Gauss, Riemann, and a few others were much more talented mathematicians, and the 20th century saw a host of extremely brilliant physicists. Other mathematicians, say Hilbert or Euler, have been more prolific. Newton was, by all accounts, very far from being the nicest scientist ever. It's just that his accomplishments, both in mathematics and in physics, are more important than anyone else's.

In this chapter we'll study the calculus, which is Newton's most important contribution to mathematics, and also Leibniz's. Newton worked out his method years before Leibniz, but published almost nothing about it at that time. (Later he claimed that his reluctance to publish grew out of a fear of being mocked.) Leibniz published a full account in 1684, almost a decade before Newton's first publication on the subject in 1693. A few years later members of the Royal Society accused Leibniz of stealing Newton's ideas, and the subsequent priority dispute marred both their lives, with severe and lingering damage to the collegial spirit of the mathematics profession. Nowadays the consensus among historians is that the discoveries were independent, and both men are regarded as more or less equally the founders of the subject.

Put very generally, the concept we are going to study is as follows: *the derivative of a function at a point is a linear function that, together with the value of the function at the point, gives an affine approximation of the function that is asymptotically accurate near the point.* Over the next few pages we will slowly untangle this description, aiming at a precise definition. After that we will lay out the most important properties of this concept. Our treatment has a rather paradoxical quality: although the focus is entirely theoretical, with little attention paid to developing the particular knowledge and skills that constitute the ability to compute derivatives, the most basic and important results are precisely the ones establishing that such computations are possible in a wide variety of circumstances.

First of all, what do we mean by "affine" here? If V and W are vector spaces, an **affine function** from V to W is a function of the form

$$v \mapsto w_0 + \ell(v)$$

where $w_0 \in W$ and $\ell : V \to W$ is linear. An **affine subspace** of V is a subset of the form $v_0 + L$ where L is a linear subspace, so the affine functions from V to W map affine subspaces of V to affine subspaces of W. Imagine that you were trying to do physics in some nice vector space, but you didn't know where the origin was, and (what is more or less the same thing) the laws of physics in this space didn't depend on where the origin was. (The technical terminology for this is that they are **invariant under translations**. This means that if you pick up a collection of particles, or fields, or whatever, and move them to a different part of the space, they will behave in the same way.) The path of high principle would seemingly be to rewrite all of linear algebra in terms of "affine spaces" and "affine functions," but it wouldn't be very much fun, and nothing truly new would come out of the project. The practical approach is to say that affine functions are just linear functions with constant terms tacked on, and have all the "obvious" properties. (Even more bluntly, you could say that because nothing depends on where the origin is, you are free to put it anywhere you like.)

Next, we should say a bit about what we mean by an "asymptotically accurate approximation." The basic idea is concrete and familiar. The simplest expression of it is that a curve is well approximated near a point, and very well approximated very near the point, by a line that is tangent to the curve at that point. The "flat earth" theory of the structure of the Universe is pretty accurate within a few kilometers of your house or apartment. On a somewhat larger scale, Newtonian physics works very well for low velocities and weak gravitational fields.

We now have the gist of the intuition: we are going to be giving an affine approximation of a function between vector spaces that is asymptotically accurate near some point. But the concepts won't make sense for vector spaces over an arbitrary field of scalars because we need to say what 'aymptotic' means, so we will need to restrict attention to fields satisfying certain conditions, and the next section gives an abstract description of the required properties. The definition of the derivative also requires that the vector spaces have norms. (This concept was defined for vector spaces over $\mathbb{R}$ in Section 3.1.) For the sake of simplicity we will only consider finite dimensional vector spaces, and it turns out that one can impose a variety of norms on any finite dimensional space over a field of the allowed type, but (and this will be important!) in our work the choice of norm won't matter, as a consequence of certain topological facts. These issues are discussed in Section 6.3. After these preparations the precise definition of the derivative is given in Section 6.4.

6.1 Assumptions on Scalars

The word 'asymptotically' tips us off that limiting processes will be involved, so we are going to be working with vector spaces that are endowed with topologies. In the one dimensional case such a topology amounts to a topology on the field of scalars. Let's fix a field k with a topology once and for all. Practically speaking, k is either $\mathbb{R}$ or $\mathbb{C}$. There are other topological fields for which calculus makes sense, but you'll probably never encounter that sort of calculus unless you study high level analysis or even higher level number theory, and you should feel free to (in fact encouraged to) think exclusively in terms of $k = \mathbb{R}$ except when we are explicitly considering some other field.

Whenever one imposes a topology on an algebraic structure it is natural to (more precisely, very unnatural not to) require that the algebraic operations are continuous. Thus we should expect that the field operations (addition, multiplication, negation, inversion) will be continuous, but in fact we will actually insist that k have a certain structure of the sort enjoyed by $\mathbb{R}$ and $\mathbb{C}$. In the case of $\mathbb{C}$ there is the modulus or absolute value

$$|x + iy| = \sqrt{x^2 + y^2},$$

and for $\mathbb{R}$ there is the restriction of this function, which is, of course, the absolute value. The definition of the derivative, and everything that follows, depends on certain properties of these functions that were established for $\mathbb{C}$ in Section 3.8. The following definition abstracts them.

Definition 6.1. *A **valuation** on k is a function $|\cdot| : k \to [0, \infty)$ such that for all $s, t \in k$:*

(i) $|s| = 0$ if and only if $s = 0$;

(ii) $|st| = |s|\,|t|$;

(iii) $|s+t| \le |s| + |t|$.

Before anything else, here are a couple basic properties of a valuation. From (ii) we have $|1| = |1 \cdot 1| = |1|\,|1|$, so $|1| = 1$. Necessarily $|-1| = 1$, because $|-1|$ is a positive number whose square is 1: $|-1|^2 = |(-1)^2| = |1| = 1$. Therefore $|-s| = |-1|\,|s| = |s|$ for all s.

The valuation induces a metric on k given by the formula

$$d(s,t) := |s-t|.$$

Specifically, (i) implies that $d(s,t) = 0$ if and only if $s = t$, the fact just mentioned gives $d(t,s) = d(s,t)$, and (iii) implies the triangle inequality.

With respect to the topology of k induced by this metric, $|\cdot|$ is a continuous function from k to $[0,\infty)$. To see this suppose that $\{s_n\}$ is a sequence converging to s, so that $|s_n - s| \to 0$. Then (iii) gives

$$|s_n| \le |s_n - s| + |s| \quad \text{and} \quad |s| \le |s - s_n| + |s_n|,$$

so that

$$\big||s_n| - |s|\big| \le |s_n - s| \to 0.$$

(In the first expression the outer absolute value signs refer to the absolute value for $\mathbb{R}$, while the inner ones are the valuation on k.)

The field operations are all continuous, as we now show. Negation is the simplest. Suppose $s_n \to s$. Then

$$|(-s_n) - (-s)| = |-(s_n - s)| = |s_n - s| \to 0,$$

so $-s_n \to -s$.

To prove that inversion is continuous consider a sequence $\{s_n\}$ converging to some $s \neq 0$. If $|s_n - s| < |s|/2$, then the triangle inequality implies that $|s_n| \ge |s| - |s_n - s| > |s|/2$, so if $\varepsilon > 0$ and $|s_n - s| < \frac{1}{2}\min\{\varepsilon|s|^2, |s|\}$, then

$$\left|\frac{1}{s_n} - \frac{1}{s}\right| = \left|\frac{s - s_n}{s_n s}\right| = \frac{|s_n - s|}{|s_n|\,|s|} < \frac{2|s_n - s|}{|s|^2} < \varepsilon.$$

Therefore $1/s_n \to 1/s$.

Since we haven't done anything related to topology for the last two chapters, its probably a good idea to review some relevant definitions. If X and Y are topological spaces, the **product topology** on $X \times Y$ is the topology whose open sets are the unions of sets of the form $U \times V$ where

$U \subset X$ and $V \subset Y$ are open. A sequence $\{(x_n, y_n)\}$ converges to (x, y) if it is eventually inside each such "open rectangle," so it converges if and only if $x_n \to x$ and $y_n \to y$. In particular, a sequence $\{(s_n, t_n)\}$ converges to (s, t) in $k \times k$ if and only if $|s_n - s| \to 0$ and $|t_n - t| \to 0$.

To see that addition is continuous observe that for any $s, t \in k$ we have $s_n + t_n \to s + t$ as $(s_n, t_n) \to (s, t)$ because, for any $\varepsilon > 0$,

$$|(s_n + t_n) - (s + t)| \leq |s_n - s| + |t_n - t| < \varepsilon$$

whenever $|s_n - s| < \varepsilon/2$ and $|t_n - t| < \varepsilon/2$. Multiplication is a bit more complicated. If $|s_n - s| < \sqrt{\varepsilon/3}$ and $|t_n - t| < \sqrt{\varepsilon/3}$, then

$$|s_n - s|\,|t_n - t| < \varepsilon/3,$$

and if $|s_n - s| < \varepsilon/3|t|$ (or $t = 0$) and $|t_n - t| < \varepsilon/3|s|$ (or $s = 0$) then

$$|s|\,|t_n - t| < \varepsilon/3 \quad \text{and} \quad |s_n - s|\,|t| < \varepsilon/3.$$

Therefore $s_n t_n \to st$ as $(s_n, t_n) \to (s, t)$ because

$$\begin{aligned} |s_n t_n - st| &= |(s_n - s)(t_n - t) + s(t_n - t) + (s_n - s)t| \\ &\leq |(s_n - s)(t_n - t)| + |s(t_n - t)| + |(s_n - s)t| \\ &= |s_n - s|\,|t_n - t| + |s|\,|t_n - t| + |s_n - s|\,|t| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon \end{aligned}$$

whenever $|s_n - s| < \frac{1}{3}\min\{\sqrt{3\varepsilon}, \varepsilon/|t|\}$ and $|t_n - t| < \frac{1}{3}\min\{\sqrt{3\varepsilon}, \varepsilon/|s|\}$.

The topology induced by a metric is automatically Hausdorff, and this is certainly a condition we would like k to satisfy. In Chapter 2 we showed that, as a consequence of the Least Upper Bound Axiom, $\mathbb{R}$ is complete with respect to the metric induced by the absolute value: every Cauchy sequence has a limit. Since a Cauchy sequence (with respect to the modulus) in $\mathbb{C}$ can be thought of as a pair of Cauchy sequences in $\mathbb{R}$, $\mathbb{C}$ is also complete in this sense. So, we might guess that the critical properties of $\mathbb{R}$ and $\mathbb{C}$ that we need are summarized by saying that k has a valuation, and it is complete with respect to the derived metric.

Not so fast. For any k there is a **trivial valuation** given by setting $|0| = 0$ and $|s| = 1$ if $s \in k^*$. (Recall that for any field k, $k^* := k \setminus \{0\}$.) The associated metric induces the topology in which every set is open, so the induced topology is Hausdorff, the field operations are continuous, and in fact the metric is complete, but none of this is true in an interesting way. It is important that there be some $s \in k^*$ with $|s| \neq 1$. If $0 < |s| < 1$

(otherwise $0 < |1/s| < 1$) then, by choosing a sufficiently large interger r, we can make $|s^r| = |s|^r$ is arbitrarily small and $|s^{-r}| = |s|^{-r}$ arbitrarily large. Among other things, this implies that 0 is in the closure of k^*.

There is actually one more property we will need the field k to satisfy, namely that for all $R \geq 0$ the closed ball

$$J(R) := \{ t \in k : |t| \leq R \}$$

is compact. It is enough to assume that one such ball is compact.

Lemma 6.2. *If $J(1)$ is compact, then $J(R)$ is compact for all $R \geq 0$.*

Proof. Choose a $\beta \in k$ with $|\beta| \geq R$. Then the map $s \mapsto \beta s$ is a bijection between $J(1)$ and $J(|\beta|)$ (its inverse is $s \mapsto s/\beta$) so $J(|\beta|)$ is compact because (Theorem 3.47) it is the image of a compact set under a continuous function. Therefore $J(R)$ is a closed (as the preimage of the closed interval $[0, R]$ under the continuous function $|\cdot|$) subset of a compact set, hence compact by Theorem 3.38. □

Summarizing, we are assuming that k has a valuation, and that the topology induced by the metric derived from this valuation has the following properties:

- k is complete;

- 0 is in the closure of k^*, and the valuation is unbounded above;

- every closed ball $J(R)$ is compact.

Once again, we are really only interested in $\mathbb{R}$ and $\mathbb{C}$. The reason we work with an abstractly described k is not that we wish to extend the theory to other fields, although this is possible. The point is to call your attention to the properties of $\mathbb{R}$ and $\mathbb{C}$ that figure in our work.

6.2 A Weird Valuation

It is a digression, but the cost at this point is pretty low, so I'd like to mention another interesting valuation. (You may find it reassuring to learn that the material in this section is not a prerequisite for anything later in the book.) Let p be a prime number. Any $s \in \mathbb{Q}^*$ can be written in exactly one way as $s = \frac{a}{b} p^r$ where a, b, and r are integers with a nonzero, b positive,

a and b relatively prime[1], and neither a nor b divisible by p. The exponent r is called the **power** of p in s. Fixing an arbitrary α with $0 < \alpha < 1$, let

$$|s|_p := \alpha^r.$$

If $s' = \frac{a'}{b'}p^{r'}$, then

$$|ss'|_p = \alpha^{r+r'} = \alpha^r\alpha^{r'} = |s|_p|s'|_p.$$

If $r' \geq r$, then

$$s + s' = \Big(\frac{a}{b} + \frac{a'}{b'}p^{r'-r}\Big)p^r = \frac{ab' + a'bp^{r'-r}}{bb'}p^r.$$

Since b and b' aren't divisible by p, neither is bb'. If $r' > r$, then $ab'+a'bp^{r'-r}$ is not divisible by p, but if $r' = r$, then $ab' + a'b$ may be divisible by p, so $|\cdot|_p$ actually satisfies the **ultrametric inequality**

$$|s + s'|_p \leq \max\{|s|_p, |s'|_p\},$$

and $|s|_p = |s'|_p$ whenever the ultrametric inequality holds strictly.

The valuation $|\cdot|_p$ seems bizarre at first, and takes some getting used to. A sequence $s_1, s_2, \dots$ of rational numbers converges to s, relative to the metric induced by $|\cdot|_p$, if, for any positive integer R, there is an integer N such that the power of p in $s_n - s$ is *greater* than R whenever $n > N$. Among other things, a sequence of larger and larger (in the normal sense) integers can converge to a finite quantity. To take a quite simple example, the sequence $1, p, p^2, \dots$ converges to 0! The limit of a sequence of integers can even be a fraction:

$$p^n + \cdots + p + 1 = \frac{p^{n+1} - 1}{p - 1} \to -\frac{1}{p-1}.$$

The sequence

$$s_n = p + p^2 + p^4 + \cdots + p^{2^n}$$

can be used to show that $\mathbf{Q}$ is not complete with respect to the metric induced by $|\cdot|_p$. This sequence is Cauchy because the power of p in $s_m - s_n$ is $2^{\min\{m,n\}}$ when $m \neq n$, but it has no limit in $\mathbf{Q}$: the power of p in s_n is one, so zero is not a limit, and if $s = \frac{a}{b}p^r$ is a nonzero rational, then

$$s_n - s = \frac{b(p + \cdots + p^{2^n}) - ap^r}{b},$$

[1] Two integers are **relatively prime** if they have no common factor, so that their greatest common divisor is 1.

and the power of p in $b(p + \cdots + p^{2^n}) - ap^r$ eventually stops changing as n increases.

In the same way that we constructed the real numbers from the rationals, it is possible to embed $\mathbf{Q}$ (endowed with the metric derived from $|\cdot|_p$) in a larger field. In fact the method of construction is feasible, and important, for a general metric space (X, d). Two Cauchy sequences $\{x_n\}$ and $\{y_n\}$ in X are said to be **equivalent** if $\lim_{n\to\infty} d(x_n, y_n) = 0$. With obvious modifications, the argument given in Section 2.9 can be used to show that this is actually an equivalence relation, and we denote the equivalence class of $\{x_n\}$ by $[\{x_n\}]$. Let $\overline{X}$ be the set of equivalence classes of Cauchy sequences.

If $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences, then for any m and n the triangle inequality gives

$$d(x_m, y_m) \le d(x_m, x_n) + d(x_n, y_n) + d(y_n, y_m),$$

and the same inequality holds with m and n reversed, so

$$|d(x_m, y_m) - d(x_n, y_n)| \le d(x_m, x_n) + d(y_m, y_n)$$

and consequently $d(x_1, y_1), d(x_2, y_2), \ldots$ is a Cauchy sequence. Suppose that $\{x'_n\}$ is equivalent to $\{x_n\}$ and $\{y'_n\}$ is equivalent to $\{y_n\}$, and observe that

$$d(x'_n, y'_n) \le d(x'_n, x_n) + d(x_n, y_n) + d(y_n, y'_n).$$

Since $d(x'_n, x_n) \to 0$ and $d(y_n, y'_n) \to 0$, we have $\lim d(x'_n, y'_n) \le \lim d(x_n, y_n)$, and of course the reverse inequality also holds. Therefore we can define a function $\overline{d} : \overline{X} \times \overline{X} \to [0, \infty)$ by setting

$$\overline{d}\big([\{x_n\}], [\{y_n\}]\big) := \lim_{n\to\infty} d(x_n, y_n).$$

It is easy to see that $\overline{d}$ is a metric: symmetry and the triangle inequality follow from the fact that d has these properties, and the definition of equivalence insures that $\overline{d}\big([\{x_n\}], [\{y_n\}]\big) = 0$ precisely when $[\{x_n\}] = [\{y_n\}]$. There is a function $\iota : X \to \overline{X}$ taking each x to the equivalence class of the sequence $x, x, \ldots$, and this is an **isometry**: $\overline{d}(\iota(x), \iota(y)) = d(x, y)$ for all $x, y \in X$. (Usually it is simpler to think of obtaining $\overline{X}$ by "adding" points to X, so that X is a subset of $\overline{X}$.)

The metric space $(\overline{X}, \overline{d})$ is called the **completion** of (X, d). As the terminology suggests, we tend to think that all the points of $\overline{X}$ are present implicitly even if, for whatever reason, we are particularly interested in the points in X. If (X, d) is already a complete metric space, then $\overline{X}$ doesn't have any points that aren't already in X, so ι is a bijection. It's not that hard to show that $(\overline{X}, \overline{d})$ is complete, but a full explanation would be rather lengthy, so we'll leave it as an exercise if you like.

Now suppose that X is a field, and that the metric is derived from a valuation. We briefly review the main points of the discussion in Section 2.9 that do not depend on the order axioms, which are equally valid in this more general setting. Term-by-term negations, sums, and products of Cauchy sequences are Cauchy, and the sequence $\{x_n^{-1}\}$ is Cauchy, and inequivalent to $0, 0, \ldots$ whenever $\{x_n\}$ is a Cauchy sequence that has no zero terms and is not equivalent to $0, 0, \ldots$. If we define addition and multiplication of elements of $\overline{X}$ by setting

$$[\{x_n\}] + [\{y_n\}] := [\{x_n + y_n\}] \quad \text{and} \quad [\{x_n\}] \cdot [\{y_n\}] := [\{x_n y_n\}],$$

then these definitions do not depend on the choices of representatives, and the field axioms (F1)-(F9) are easily seen to be satisfied, simply by examining each one and asking whether it holds. (As we explained in Section 2.9, in truth (F7) is a bit trickier than the others.)

The field obtained in this way from $|\cdot|_p$ is called the field of ***p*-adic numbers**. It was introduced by Kurt Hensel (1861-1941) in 1897, it has played an increasingly important role in number theory since then, and calculus with respect to this field is an active topic of contemporary research in number theory. There are valuations on algebraic number fields, and on other sorts of fields as well, so the general idea has a larger significance than this particular example might suggest.

6.3 Normed Spaces

We now know that we will be working with vector spaces over a field k that is endowed with a certain type of valuation. Of course these vector spaces must also be endowed with topologies, but it turns out that the definition of the derivative actually requires that the topologies are of a rather special sort, namely that they are derived from norms. For vector spaces over $\mathbb{R}$ we already know what this means. To handle vector spaces over $\mathbb{C}$ we extend Definition 3.7 to vector spaces over a field with a valuation.

Definition 6.3. *Let k be a field with a valuation $|\cdot|$, and let V be a vector space over k. A **norm** on V is a function*

$$\|\cdot\| : V \to [0, \infty)$$

such that:

(i) for all $x \in V$, $\|x\| = 0$ if and only if $x = 0$;

(ii) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $x \in V$ and all $\alpha \in k$;

(iii) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in V$.

The discussion in Section 3.1 applies to this slightly extended definition with very minor modifications, and you might want to review it now. Actually, there was only one issue for the case $k = \mathbb{R}$ that was not straightforward, namely showing that the function

$$x \mapsto \|x\|_2 := \sqrt{x_1^2 + \cdots + x_n^2}$$

was a norm for $\mathbb{R}^n$. In the end this followed from the Cauchy-Schwartz inequality. For $\mathbb{C}^n$ it is clear that the function

$$\|z\|_2 := \sqrt{|z_1|^2 + \cdots + |z_n|^2}$$

satisfies (i) and (ii), and it satisfies (iii) because if $z_i = x_i + iy_i$, then $|z_i|^2 = x_i^2 + y_i^2$, so it reduces to $\| \cdot \|_2$ on $\mathbb{R}^{2n}$. The functions

$$\|z\|_1 := |z_1| + \cdots + |z_n| \quad \text{and} \quad \|z\|_\infty := \max\{|z_1|, \ldots, |z_n|\}$$

also satisfy (i) and (ii) obviously, and they satisfy (iii) by virtue of the calculations

$$\|z+w\|_1 = |z_1+w_1|+\cdots+|z_n+w_n| \leq |z_1|+|w_1|+\cdots+|z_n|+|w_n| = \|z\|_1+\|w\|_1$$

and

$$\|z+w\|_\infty = \max\{|z_1+w_1|, \ldots, |z_n+w_n|\} \leq \max\{|z_1|+|w_1|, \ldots, |z_n|+|w_n|\}$$
$$\leq \max\{|z_1|, \ldots, |z_n|\} + \max\{|w_1|, \ldots, |w_n|\} = \|z\|_\infty + \|w\|_\infty.$$

As was the case in Chapter 3, any norm determines a metric defined by the formula $d(x, y) = \|x - y\|$. For $\delta > 0$ let

$$B(\delta) := \{\, v \in V : \|v\| \leq \delta \,\}$$

be the closed unit ball of radius δ centered at the origin in V. Then a set $U \subset V$ is open if, for each $v \in U$, there is some $\delta > 0$ such that $v + B(\delta) \subset U$. Here are some basic properties that are applied very frequently:

Lemma 6.4. *Addition, scalar multiplication, and the norm itself, are continuous functions.*

Proof. Addition: For any $x, y, x', y' \in V$ we have

$$\|(x' + y') - (x + y)\| \leq \|x' - x\| + \|y' - y\|,$$

so $x' + y' \in x + y + B(\delta)$ whenever $x' \in x + B(\delta/2)$ and $y' \in y + B(\delta/2)$.

Scalar Multiplication: For any $x, x' \in V$ and $\alpha, \alpha' \in k$ we apply (iii), then (ii), obtaining

$$\|\alpha' x' - \alpha x\| = \|\alpha'(x' - x) + (\alpha' - \alpha)x\| \leq |\alpha'| \, \|x' - x\| + |\alpha' - \alpha| \, \|x\|.$$

If $|\alpha' - \alpha| < \min\{1, \delta/2\|x\|\}$, so that $|\alpha'| \leq 1 + |\alpha|$, and $x' \in x + B(\delta/2(1+|\alpha|))$ then $\alpha' x' \in \alpha x + B(\delta)$ because

$$|\alpha'| \, \|x' - x\| + |\alpha' - \alpha| \, \|x\| \leq (1 + |\alpha|) \cdot \delta/2(1 + |\alpha|) + (\delta/2\|x\|) \cdot \|x\| = \delta.$$

The Norm: For any $v, v' \in V$ the triangle inequality (iii) gives

$$\|v'\| \leq \|v\| + \|v' - v\| \quad \text{and} \quad \|v\| \leq \|v'\| + \|v - v'\|,$$

so for any v and $\delta > 0$ we have $\|v'\| \in (\|v\| - \delta, \|v\| + \delta)$ whenever $v' \in v + B(\delta)$ because $\big| \|v'\| - \|v\| \big| \leq \|v' - v\| \leq \delta$. □

Corollary 6.5. *Any linear function $\ell : k^n \to V$ is continuous if k^n has the product topology.*

Proof. For some $v_1, \dots, v_n$ we can regard ℓ as the composition

$$(\alpha_1, \dots, \alpha_n) \to (\alpha_1 v_1, \dots, \alpha_n v_n) \to \alpha_1 v_1 + \cdots + \alpha_n v_n.$$

The first map is continuous because each $\alpha_i \mapsto \alpha_i v_i$ is scalar multiplication and a cartesian product of continuous functions is continuous (Lemma 3.41). Of course the second map is continuous because addition is continuous. □

The differential calculus, as described here, can be generalized, in a fairly straightforward manner, to certain types of infinite dimensional spaces, but there are some technical details that would need to be addressed, and, in relation to the rest of what we do in this book, there wouldn't be that much reward. Perhaps even more important, attending to the infinite dimensional case would weaken the focus on what is, in the end, the central aspect of our discussion of the calculus, namely calculating stuff. Henceforth we assume that V is n-dimensional.

The following notation will appear in two of the proofs below: for $\varepsilon > 0$ let the n-fold cartesian product of $J(\varepsilon)$ be

$$J^n(\varepsilon) := \{\, (\alpha_1, \dots, \alpha_n) \in k^n : \max\{|\alpha_1|, \dots, |\alpha_n|\} \leq \varepsilon \,\}.$$

We also need to introduce a very important piece of topological terminology. If $f : X \to Y$ is a bijection, where X and Y are topological spaces, then we say that f is a **homeomorphism** if f and f^{-1} are both continuous. Homeomorphisms are the isomorphisms of the category of topological spaces and continuous functions.

Proposition 6.6. *If $\ell : k^n \to V$ is a linear isomorphism and k^n has the product topology, then ℓ is a homeomorphism.*

Proof. We know that ℓ is continuous, so we need to show that ℓ^{-1} is also continuous, which amounts to $\ell(U)$ being open in V whenever $U \subset k^n$ is open. If $U \subset k^n$ is open, then for each $x \in U$ there is some $\varepsilon > 0$ such that $x + J^n(\varepsilon) \subset U$, so to show that $\ell(U)$ is open it suffices to show that $\ell(x+J^n(\varepsilon))$ contains $\ell(x)+B(\delta)$ when $\delta > 0$ is sufficiently small. Therefore it suffices to show that for any $\varepsilon > 0$ there is $\delta > 0$ such that $B(\delta) \subset \ell(J^n(\varepsilon))$.

Fix $\varepsilon > 0$, and let

$$\partial J^n(\varepsilon) := \{\, (\alpha_1, \dots, \alpha_n) : \max\{|\alpha_1|, \dots, |\alpha_n|\} = \varepsilon \,\} \subset J^n(\varepsilon).$$

The function $(\alpha_1, \dots, \alpha_n) \mapsto \max\{|\alpha_1|, \dots, |\alpha_n|\}$ is continuous (the valuation is continuous, a cartesian product of continuous functions is continuous (Lemma 3.41), and the maximum operator is continuous) so $\partial J^n(\varepsilon)$ is a closed subset of $J^n(\varepsilon)$. Since $J^n(\varepsilon)$ is a cartesian product of compact sets, it is compact, so (Theorem 3.38) $\partial J^n(\varepsilon)$ is compact.

We are assuming that k contains scalars β with $|\beta|$ arbitrarily small, and it suffices to demonstrate the claim with ε replaced by a smaller number, so we may assume that $\varepsilon = |\beta|$ for some $\beta \in k$. This implies that $\partial J^n(\varepsilon)$ is nonempty. Let

$$\delta := \min_{\alpha \in \partial J^n(\varepsilon)} \|\ell(\alpha)\|.$$

Since $\|\ell(\alpha)\|$ is a continuous function of α (because ℓ and the norm are continuous) and $\partial J^n(\varepsilon)$ is nonempty and compact, Theorem 3.48 implies that $\delta = \|\ell(\alpha^*)\|$ for some $\alpha^* \in \partial J^n(\varepsilon)$. Since $\partial J^n(\varepsilon)$ does not contain the origin it follows that $\delta > 0$.

We still have to show that $B(\delta) \subset \ell(J^n(\varepsilon))$, so fix $x \in B(\delta)$. Since ℓ is a linear isomorphism, $x = \ell(\alpha)$ for some $\alpha \in k^n$. There is an index i such that $|\alpha_i| = \max\{|\alpha_1|, \dots, |\alpha_n|\}$. We want to show that $|\alpha_i| \leq \varepsilon$ because this implies that $x \in \ell(J^n(|\alpha_i|)) \subset \ell(J^n(\varepsilon))$. Of course $|\alpha_i| \le \varepsilon$ holds automatically if $\alpha_i = 0$, so we may assume that $\alpha_i \ne 0$.

The key point is that $(\beta/\alpha_i)\alpha \in \partial J^n(\varepsilon)$ because

$$\max\{|(\beta/\alpha_i)\alpha_1|, \dots, |(\beta/\alpha_i)\alpha_n|\} = |(\beta/\alpha_i)\alpha_i| = |\beta| = \varepsilon.$$

We now obtain $|\alpha_i| \le \varepsilon$ because the definition of δ gives

$$\delta \le \|\ell\big((\beta/\alpha_i)\alpha\big)\| = |\beta/\alpha_i|\, \|\ell(\alpha)\| = (\varepsilon/|\alpha_i|)\|x\| \le (\varepsilon/|\alpha_i|)\delta.$$

□

A linear isomorphism between two n-dimensional normed spaces can be written as a composition of linear isomorphisms from the first space to k^n and from k^n to the second space, so it too is a homeomorphism. In particular, the identity function from V endowed with one norm to V endowed with a second norm is a homeomorphism, which means that any two norms are **topologically equivalent** in the sense of inducing the same topology on V.

The most general and flexible expression of this idea is as follows:

Proposition 6.7. *If V and W are normed spaces with V finite dimensional, and $\ell : V \to W$ is linear, then ℓ is continuous.*

Proof. If $\iota : k^n \to V$ is a linear isomorphism, then (by the last two results) $\ell \circ \iota$ and ι^{-1} are continuous, and $\ell = (\ell \circ \iota) \circ \iota^{-1}$. □

We conclude this section with two important technical results.

Lemma 6.8. *For any $\delta > 0$, $B(\delta)$ is compact.*

Proof. Let $\ell : k^n \to V$ be a linear isomorphism. For any $R > 0$, $J^n(R)$ is compact because (Theorem 3.42) it is a finite cartesian product of compact sets, and consequently $\ell(J^n(R))$ is compact because (Theorem 3.47) ℓ is continuous. Since the norm is continuous, $B(\delta)$ is closed, so (Theorem 3.38) it is compact if it is contained in $\ell(J^n(R))$ for some R. Since ℓ^{-1} is continuous, $B(\varepsilon) \subset \ell(J^n(1))$ for some $\varepsilon > 0$. Let β be a scalar with $|\beta| \ge \delta/\varepsilon$. Then

$$B(\delta) \subset B(|\beta|\varepsilon) = \beta B(\varepsilon) \subset \beta\ell(J^n(1)) = \ell(\beta J^n(1)) = \ell(J^n(|\beta|)).$$

□

The definition of the derivative considers the ratio of the norms of two vectors, so for technical purposes, the sense in which different norms are equivalent is, at least on its surface, stronger than topological equivalence.

Proposition 6.9. *For any norms $\| \cdot \|$ and $\| \cdot \|_*$ on V there are numbers $M, m > 0$ such that $m\|v\| \leq \|v\|_* \leq M\|v\|$ for all $v \in V$.*

Proof. Let $M := \max_{\|v\| \leq 1} \|v\|_*$. Since $B(1)$ is nonempty and compact, and $\| \cdot \|_*$ is continuous, M is well defined and finite. If $M = 0$ (this happens when $V = \{0\}$) replace M with any positive number. Then for all v we have

$$\|v\|_* = \big\|v/\|v\|\big\|_* \cdot \|v\| \leq M\|v\|.$$

A symmetric argument gives $m > 0$ such that $\|v\| \leq (1/m)\|v\|_*$ for all v. □

6.4 Defining the Derivative

Finally we are in a position to define this chapter's central concept. Fix V and W, two finite dimensional vector spaces over k, and assume that each is endowed with a norm. The definition of a derivative will be applied to functions whose domains are subsets of V, and which have W as their range. We would like to to allow more general domains than all of V, but at the same time the definition will be a matter of placing restrictions on the behavior of the function in a neighborhood of a point, and it would have little force, and be hard to work with, if the domain didn't contain at least one neighborhood of the point in question, so it makes sense to insist that the domain be open. Therefore we fix an open $U \subset V$.

Definition 6.10. *Let $f : U \to W$ be a function, and let $\overline{x}$ be a point in U. We say that f is **differentiable** at $\overline{x}$ if there is a linear function $\ell : V \to W$ such that for any $\varepsilon > 0$ there is $\delta > 0$ such that*

$$\big\|f(x) - [f(\overline{x}) + \ell(x - \overline{x})]\big\| \leq \varepsilon\|x - \overline{x}\|$$

*for all $x \in U$ with $\|x - \overline{x}\| < \delta$. If this is the case we say that ℓ is the **derivative** of f at $\overline{x}$, and we denote it by $Df(\overline{x})$.*

That is, the affine function $x \mapsto f(\overline{x}) + \ell(x - \overline{x})$ is an asymptotically accurate approximation of f near $\overline{x}$ in the sense that its error, in proportion to $\|x - \overline{x}\|$, can be made arbitrarily small by restricting x to be sufficiently close to $\overline{x}$.

As we have seen on various occasions, it is sometimes necessary to show that a definition makes sense, and of course from a logical point of view this needs to be done before we can use the definition. Definition 6.10 defines the derivative of f at $\overline{x}$ to be a linear function with certain properties, but how do we know that there's only one linear function satisfying the specified conditions?

Let's suppose that $\ell' : V \to W$ is also a linear transformation such that for any $\varepsilon > 0$ we have

$$\big\|f(x) - [f(\overline{x}) + \ell'(x - \overline{x})]\big\| \le \varepsilon\|x - \overline{x}\|$$

when $\|x - \overline{x}\|$ is sufficiently small. Then

$$\begin{aligned}
&\|\ell'(x - \overline{x}) - \ell(x - \overline{x})\| \\
&\qquad = \big\|\big(f(x) - [f(\overline{x}) + \ell'(x - \overline{x})]\big) - \big(f(x) - [f(\overline{x}) + \ell'(x - \overline{x})]\big)\big\| \\
&\qquad \le \big\|f(x) - [f(\overline{x}) + \ell'(x - \overline{x})]\big\| + \big\|f(x) - [f(\overline{x}) + \ell'(x - \overline{x})]\big\| \\
&\qquad \le 2\varepsilon\|x - \overline{x}\|
\end{aligned}$$

when $\|x - \overline{x}\|$ is sufficiently small. Now consider any $v \in V$, and recall that we are assuming that there are nonzero scalars $\alpha \in k$ with $|\alpha|$ arbitrarily small, so that $\|\alpha v\|$ can be made small enough to imply that $\|\ell'(\alpha v) - \ell(\alpha v)\| \le 2\varepsilon\|\alpha v\|$, in which case

$$|\alpha|\,\|\ell'(v) - \ell(v)\| = \|\ell'(\alpha v) - \ell(\alpha v)\| < 2\varepsilon\|\alpha v\| = 2\varepsilon|\alpha|\,\|v\|.$$

Dividing by $|\alpha|$ gives $\|\ell'(v) - \ell(v)\| \le \varepsilon\|v\|$. This is true for all ε and v, so $\ell'(v) = \ell(v)$ for all v, which means that $\ell' = \ell$.

The next priority is to show that, as we have mentioned more than once, the definition of the derivative doesn't depend on the choice of norms for V and W. Once we've settled the question we'll be free to use whichever norm happens to be most convenient, and this will often simplify arguments and calculations.

Suppose that in addition to the given norms on V and W, both of which are denoted by $\|\cdot\|$, we have another norm for each of the spaces, which is denoted by $\|\cdot\|_*$. Proposition 6.9 gives numbers $M_V > m_V > 0$ and $M_W > m_W > 0$ such that

$$m_V\|v\| \le \|v\|_* \le M_V\|v\| \quad \text{and} \quad m_W\|w\| \le \|w\|_* \le M_W\|w\|$$

for all $v \in V$ and all $w \in W$. Assuming that f is differentiable at $\overline{x}$ with respect to the given norms, we want to show that $\ell = Df(\overline{x})$ is also the derivative of f at $\overline{x}$ with respect to the new norms. Fix $\varepsilon_* > 0$, and set $\varepsilon := (m_V/M_W)\varepsilon_*$. Choose $\delta > 0$ such that

$$\big\|f(x) - [f(\overline{x}) + \ell(x - \overline{x})]\big\| \le \varepsilon\|x - \overline{x}\|$$

for all $x \in U$ with $\|x - \overline{x}\| < \delta$, and set $\delta_* := m_V\delta$. If $\|x - \overline{x}\|_* < \delta_*$, then

$$\|x - \overline{x}\| \le \|x - \overline{x}\|_*/m_V < \delta_*/m_V = \delta,$$

so that

$$\big\|f(x) - [f(\overline{x}) + \ell(x - \overline{x})]\big\|_* \le M_W \big\|f(x) - [f(\overline{x}) + \ell(x - \overline{x})]\big\|$$
$$\le M_W \varepsilon \|x - \overline{x}\| \le M_W \varepsilon \|x - \overline{x}\|_* / m_V = \varepsilon_* \|x - \overline{x}\|_*,$$

which is just what we want.

Before going further we give two basic results that appear frequently in proofs. Since an affine function is a perfect approximation of itself, the definition of the derivative (Definition 6.10) has the following immediate and obvious consequence:

Proposition 6.11. *Suppose V and W are finite dimensional vector spaces over k, $U \subset V$ is open, and $a : U \to W$ is affine, so that there is $w_0 \in W$ and a linear transformation $\ell : V \to W$ such that $a(x) = w_0 + \ell(x)$ for all $x \in U$. Then for any $\overline{x} \in U$, a is differentiable at $\overline{x}$ and*

$$Da(\overline{x}) = \ell.$$

The following fact will soon seem so obvious as to be beneath mention.

Lemma 6.12. *If $f : U \to W$ is differentiable at $\overline{x}$, then it is continuous at $\overline{x}$.*

Proof. Consider a particular $\varepsilon > 0$. Since $Df(\overline{x})$ is linear, it is continuous (Proposition 6.7) so there is $\delta > 0$ such that $\|Df(\overline{x})v\| < \varepsilon/2$ whenever $\|v\| < \delta$. Replacing δ with a smaller number if need be, we can insist that the inequality in the definition of the derivative is satisfied whenever $\|x - \overline{x}\| < \delta$, and that $\delta < 1/2$. When $\|x - \overline{x}\| < \delta$ we have

$$\big\|f(x) - f(\overline{x})\big\| \le \big\|f(x) - [f(\overline{x}) + Df(\overline{x})(x - \overline{x})]\big\| + \big\|Df(\overline{x})(x - \overline{x})\big\|$$
$$< \varepsilon\|x - \overline{x}\| + \varepsilon/2 < \varepsilon \cdot \delta + \varepsilon/2 < \varepsilon.$$

□

6.5 The Derivative's Significance

Especially if you have never studied the derivative before, it's a good idea to pause for a bit and just *look* at it. Figure 6.1 is the picture that haunts the dreams of first year calculus students. The curve is the graph of a function $f : \mathbb{R} \to \mathbb{R}$. The straight line is the graph of the affine approximation

$$t \mapsto f(\overline{t}) + Df(\overline{t})(t - \overline{t}).$$

In the figure the magnitude of the error,

$$\big|f(t) - (f(\overline{t}) + Df(\overline{t})(t-\overline{t}))\big|,$$

is fairly large, but when $|t-\overline{t}|$ is small the error is small, not only absolutely but also in relation to $|t-\overline{t}|$. With a little imagination you should be able to "transport this into the third dimension" so that it applies to a function $f : \mathbb{R} \to \mathbb{R}^2$ whose graph is a curve in $\mathbb{R}^3$.

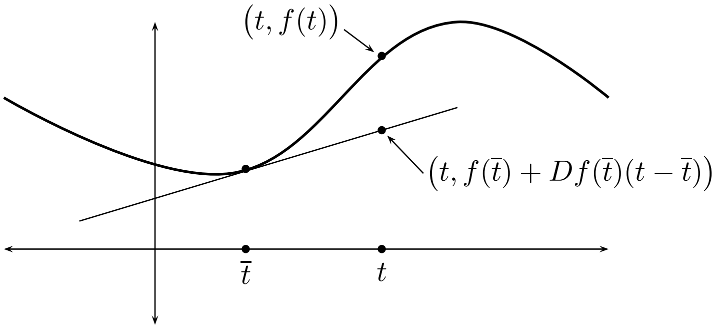

$\big(t, f(t)\big)$
$\big(t, f(\overline{t}) + Df(\overline{t})(t-\overline{t})\big)$
$\overline{t}$
t

Figure 6.1

Figure 6.2 shows the affine approximation

$$(x,y) \mapsto f(\overline{x},\overline{y}) + Df(\overline{x},\overline{y})(x-\overline{x}, y-\overline{y})$$

of a function $f : \mathbb{R}^2 \to \mathbb{R}$. Its graph is the plane tangent to the graph of f.

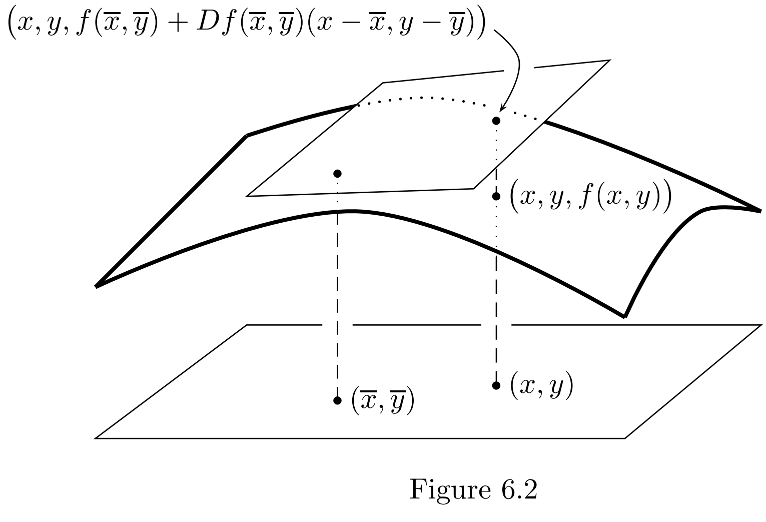

$\big(x, y, f(\overline{x},\overline{y}) + Df(\overline{x},\overline{y})(x-\overline{x}, y-\overline{y})\big)$
$\big(x, y, f(x,y)\big)$
$(\overline{x},\overline{y})$
(x,y)

Figure 6.2

Figure 6.3 illustrates the derivative of a function $f : \mathbb{R}^2 \to \mathbb{R}^2$. Now it would take four dimensions to show the graphs of f and its affine approximation, so we need a different method of visualizing the situation. We take a coordinate system near $(\overline{x}, \overline{y})$ and show the image of this coordinate system under f, and its image under the affine approximation.

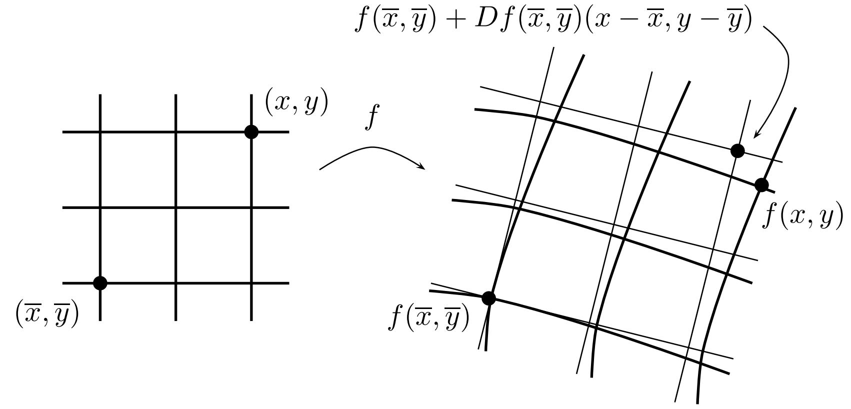

$f(\overline{x},\overline{y}) + Df(\overline{x},\overline{y})(x - \overline{x}, y - \overline{y})$
(x, y)
f
$f(x, y)$
$(\overline{x}, \overline{y})$
$f(\overline{x}, \overline{y})$

Figure 6.3

Why are differentiability and the derivative so important? The remainder of the book describes certain applications, but these don't really add up to more than a hint at an answer to this question. Here are very brief descriptions of some of the main themes in the overall significance of the concept.

(A) Let $L(V, W)$ be the set of linear functions from V to W. It is a finite dimensional vector space if addition and scalar multiplication are defined in the obvious "pointwise" manner: $(\ell + \ell')(v) := \ell(v) + \ell'(v)$; $(\alpha\ell)(v) := \alpha\ell(v)$. (In fact for any set S the space of functions from S to W is a vector space if the vector operations are defined pointwise, and $L(V, W)$ is a evidently a linear subspace of the space of functions from V to W.) Suppose that $f : U \to W$ is differentiable at every point in U. Then there is a derived function $Df : U \to L(V, W)$ which may have interesting properties and be useful in various ways. For example, it might be differentiable everywhere, in which case the second derivative

$$D(Df) : U \to L(V, L(V, W))$$

might be interesting and useful. And so forth.

(B) Requiring that a function f be differentiable at every point at U, and in particular requiring that Df be continuous, imposes conditions on f that are powerful and useful in analysis, and at the same time most functions we would typically be inclined to consider have these properties.

(C) If $V = W = k$ and f is differentiable at $\overline{u}$, then there is a scalar $f'(\overline{u}) \in k$ such that $Df(\overline{u})$ is the linear transformation $v \mapsto f'(\overline{u})v$. (Warning: in elementary calculus courses the derivative of f at $\overline{u}$ is *defined* to be $f'(\overline{u})$.) In the particular case $k = \mathbb{C}$, if f is differentiable at every point of U, then (quite remarkably!) the function $f' : U \to \mathbb{C}$ is also differentiable at every point of U, and in fact f is holomorphic. As we will see in Section 7.6, if U is connected, then a holomorphic $f : U \to \mathbb{C}$ is determined by its power series at any point $z_0 \in U$. Holomorphic functions are in this sense quite "rigid," and there are intricate relationships between their local and global properties that have been and continue to be one of mathematic's richest sources of subtle problems and deep theorems.

(D) If $k = \mathbb{R}$ and $f : U \to \mathbb{R}$ attains its maximum, or its minimum, at a point $\overline{x}$ where f is differentiable, then $Df(\overline{x}) = 0$. Hilltops and valley floors are flat. This fundamental insight affects the theory of optimization in numerous ways, and deserves a formal statement:

Theorem 6.13. *Suppose $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}$ is differentiable at $\overline{x}$. Then $Df(\overline{x}) = 0$ if either $f(\overline{x}) \geq f(x)$ for all $x \in U$ or $f(\overline{x}) \leq f(x)$ for all $x \in U$.*

Proof. Suppose that $f(\overline{x}) \geq f(x)$ for all $x \in U$. (The proof when $\overline{x}$ is a minimizer is similar.) Aiming at a contradiction, suppose that $Df(\overline{x})v \neq 0$ for some $v \in \mathbb{R}^n$. For any $\varepsilon > 0$ we have

$$f(\overline{x} + \alpha v) - [f(\overline{x}) + Df(\overline{x})(\alpha v)] > -\varepsilon \|\alpha v\|$$

when $|\alpha|$ is sufficiently small, so that

$$Df(\overline{x})(\alpha v) + \big(f(\overline{x} + \alpha v) - [f(\overline{x}) + Df(\overline{x})(\alpha v)]\big) \geq \alpha Df(\overline{x})v - \varepsilon|\alpha|\,\|v\|.$$

If $\varepsilon < |Df(\overline{x})v|/\|v\|$, then the right hand side is positive when $Df(\overline{x})v > 0$ and $\alpha > 0$, and also when $Df(\overline{x})v < 0$ and $\alpha < 0$, so that $f(\overline{x} + \alpha v) > f(\overline{x})$ for some α, contrary to assumption. □

(E) As we will see shortly, there are powerful methods for actually computing derivatives. The information about f extracted in this fashion is useful in all sorts of ways. Algorithms for computing derivatives contribute to many, many other algorithms including, obviously, algorithms related to optimization.

(F) Let the motion of a particle in space during the time interval (a, b) be described by $p : (a, b) \to \mathbb{R}^3$. The **velocity** of the particle at time t is

$$v(t) = (v_1(t), v_2(t), v_3(t))$$

where, for each $i = 1, 2, 3$, $v_i(t) = p_i'(t)$ is the derivative (in the sense described in (C)) of the corresponding component of p. In turn the **acceleration** of the particle at time t is

$$a(t) = (a_1(t), a_2(t), a_3(t))$$

where each $a_i(t) = v_i'(t)$ is the derivative of the corresponding component of v. Let m be the mass of the particle. The **force** acting on the particle at time t is

$$f(t) = ma(t).$$

Newton's theory of gravitation, as it affects a small particle in the gravitation field of a mass M located at the origin, is given by the equation

$$f = -G\frac{Mm}{\|p\|^3}p$$

where G is the universal gravitational constant. (Note that the acceleration does not depend on m, as was shown by Galileo.) A relationship between the various derivatives of a function is called a **differential equation**. The empirical substance of Newton's theory is that the trajectories that could be observed in this system are precisely those functions p that satisfy this differential equation at all times. Differential equations are also used to describe other fundamental physical theories, and a host of theories in other sciences, with enormous success.

(G) Suppose we are given a function $f : (a, b) \to \mathbb{R}$ and we wish to find a function $F : (a, b) \to \mathbb{R}$ such that $F'(t) = f(t)$ for all t. The process of going from f to some such F is called **integration**. Like the theory of the derivative, the theory of integration has evolved continuously since its initial development by both Leibniz and Newton, and it has roughly coequal status in terms of its importance to mathematics as a whole.

6.6 The Chain Rule

The chain rule characterizes the derivative of a composition of two differentiable functions. It is by far the most important theorem concerning the derivative, in part because it is the key to computing the derivative of almost every function you'll ever encounter, but also because it has a conceptual importance that is best appreciated when expressed in the language of category theory.

Suppose that, in addition to V and W, there is a third vector space X, and, in addition to $f : U \to W$, there is a function g whose domain is an open superset of the image of f and whose range is X. Assume that f is differentiable at $\overline{x} \in U$ and g is differentiable at $f(\overline{x})$. Now imagine that you are very small, or your world is very large, so that even with precise measurements it is difficult to tell whether life is governed by f and g or their affine approximations. It seems reasonable, and after reflection almost inevitable, that you should also have a hard time distinguishing between $g \circ f$ and the composition of the affine approximation of f with the affine approximation of g. That is, the composition of the affine approximations of f and g should be an asymptotically accurate approximation of $g \circ f$. This is what the chain rule says.

Based on this intuition, we "know" that the chain rule is true, but if you tried to read the proof below without appreciating this, it could easily seem like a rather messy and unmotivated calculation that just happened to work out. What makes for the mess is that several sorts of error need to be controlled. There is the difference between g and its affine approximation, and there is the difference between f and its affine approximation, as transmitted by composition with g. Finally there is a third term that reflects an interaction or compounding of the two sorts of error.

The argument involves a technical point that we deal with first. Recall that $L(V, W)$ is the vector space of linear transformations $\ell : V \to W$. Let

$$B_V := \{ v \in V : \|v\| \leq 1 \}.$$

(This set was denoted by $B(1)$ in the notational system of Section 6.3.) For $\ell \in L(V, W)$ let

$$\|\ell\| := \sup_{v \in B_V} \|\ell(v)\| \in [0, \infty].$$

We say that ℓ is **bounded** if $\|\ell\| < \infty$. (When V is finite dimensional this is automatic: applying Lemmas 6.8 and 6.4, Proposition 6.7, and Theorem 3.47, we see that B_V is compact and ℓ and the norm are continuous, so

$\{\, \|\ell(v)\| : v \in B_V \,\}$ is compact, hence bounded.) For any $\ell, \ell' \in L(V,W)$ we have

$$\begin{aligned}\|\ell + \ell'\| := \sup_{v \in B_V} \|\ell(v) + \ell'(v)\| &\leq \sup_{v \in B_V} (\|\ell(v)\| + \|\ell'(v)\|) \\ &\leq \sup_{v \in B_V} \|\ell(v)\| + \sup_{v \in B_V} \|\ell'(v)\| = \|\ell\| + \|\ell'\|,\end{aligned}$$

so $\ell + \ell'$ is bounded whenever ℓ and ℓ' are bounded. Similarly, for any $\alpha \in k$ we have

$$\|\alpha\ell\| = \sup_{v \in B_V} \|\alpha\ell(v)\| = \sup_{v \in B_V} |\alpha|\, \|\ell(v)\| = |\alpha| \sup_{v \in B_V} \|\ell(v)\| = |\alpha|\, \|\ell\|,$$

so $\alpha\ell$ is bounded whenever ℓ is bounded. Therefore the set of bounded elements of $L(V,W)$ is a linear subspace of $L(V,W)$.

The restriction of the function $\ell \mapsto \|\ell\|$ to this subspace is called the **operator norm** on the space of bounded linear functions from V to W. Clearly $\|\ell\| = 0$ if and only if $\ell = 0$, and the calculations above establish that properties (ii) and (iii) of the definition of a norm hold, so the operator norm is actually a norm. Aside from the definition itself, which gives a useful piece of notation for expressing the ideas in the proof below, the only property of the operator norm that we will use in this chapter (the operator norm will appear again in Chapter 7) is the **operator norm inequality**

$$\|\ell(v)\| = \big\|\ell(v/\|v\|)\|v\|\big\| = \|\ell(v/\|v\|)\| \cdot \|v\| \leq \|\ell\| \cdot \|v\|,$$

which is an immediate consequence of the definition.

Proceeding to the main event:

Theorem 6.14 (The Chain Rule)**.** *Suppose that V, W, and X are finite dimensional vector spaces over k, $U_V \subset V$ and $U_W \subset W$ are open, $f : U_V \to U_W$ is differentiable at $\overline{x}$, and $g : U_W \to X$ is differentiable at $f(\overline{x})$. Then $g \circ f$ is differentiable at $\overline{x}$, and*

$$D(g \circ f)(\overline{x}) = Dg(f(\overline{x})) \circ Df(\overline{x}).$$

Proof. To achieve more compact formulas we set

$$\ell := Df(\overline{x}) \quad \text{and} \quad m := Dg(f(\overline{x})).$$

The argument is a matter of analyzing the difference between $g(f(x))$ and the affine approximation $g(f(\overline{x})) + m(\ell(x - \overline{x}))$ given by $m \circ \ell$. We first decompose the error into two parts: for any $x \in U_V$,

$$\begin{aligned}&g(f(x)) - [g(f(\overline{x})) + m(\ell(x - \overline{x}))] \\ &\qquad = g(f(x)) - [g(f(\overline{x})) + m(f(x) - f(\overline{x}))] \\ &\qquad\qquad + m[f(x) - (f(\overline{x}) + \ell(x - \overline{x}))].\end{aligned}$$

Fix norms for V, W, and X. The triangle inequality gives

$$\begin{aligned}\big\|g(f(x)) - [g(f(\overline{x})) + m(\ell(x-\overline{x}))]\big\| \\ \le \big\|g(f(x)) - [g(f(\overline{x})) + m(f(x) - f(\overline{x}))]\big\| \\ + \big\|m[f(x) - (f(\overline{x}) + \ell(x-\overline{x}))]\big\|.\end{aligned}$$

For any $\varepsilon_g > 0$ there is $\delta_g > 0$ such that

$$\big\|g(f(x)) - [g(f(\overline{x})) + m(f(x) - f(\overline{x}))]\big\| \le \varepsilon_g \|f(x) - f(\overline{x})\|$$

whenever $\|f(x) - f(\overline{x})\| < \delta_g$. The operator norm inequality gives

$$\big\|m[f(x) - (f(\overline{x}) + \ell(x-\overline{x}))]\big\| \le \|m\| \cdot \|f(x) - (f(\overline{x}) + \ell(x-\overline{x}))\|,$$

and for any $\varepsilon_f > 0$ the definition of $\ell = Df(x)$ gives a $\delta_f > 0$ such that

$$\|f(x) - (f(\overline{x}) + \ell(x-\overline{x}))\| \le \varepsilon_f \|x - \overline{x}\|$$

whenever $\|x - \overline{x}\| < \delta_f$. Since (Lemma 6.12) f is continuous at $\overline{x}$, if δ_f is sufficiently small, then $\|f(x) - f(\overline{x})\| < \delta_g$ whenever $\|x - \overline{x}\| < \delta_f$. If all this is the case, then

$$\big\|g(f(x)) - [g(f(\overline{x})) + m(\ell(x-\overline{x}))]\big\| \le \varepsilon_g \|f(x) - f(\overline{x})\| + \varepsilon_f \|m\| \cdot \|x - \overline{x}\|$$

whenever $\|x - \overline{x}\| < \delta_f$. In order to bound the right hand side by a multiple of $\|x - \overline{x}\|$ we apply the operator norm inequality again, finding that

$$\begin{aligned}\|f(x) - f(\overline{x})\| &\le \|f(x) - (f(\overline{x}) + \ell(x-\overline{x}))\| + \|\ell(x-\overline{x})\| \\ &\le (\varepsilon_f + \|\ell\|)\|x - \overline{x}\|\end{aligned}$$

whenever $\|x - \overline{x}\| \le \delta_f$, in which case

$$\big\|g(f(x)) - [g(f(\overline{x})) + m(\ell(x-\overline{x}))]\big\| \le \big(\varepsilon_g(\varepsilon_f + \|\ell\|) + \varepsilon_f\|m\|\big)\|x - \overline{x}\|.$$

Since, for any given $\varepsilon > 0$, it is possible to choose $\varepsilon_f, \varepsilon_g > 0$ small enough that $\varepsilon_g(\varepsilon_f + \|\ell\|) + \varepsilon_f\|m\| \le \varepsilon$, this establishes the result. □

Soon we will study how the chain rule can be used to perform myriad concrete calculations, but first I want to show how the power of the chain rule is, in a certain sense, "explained" by a concept from category theory.

If $\mathcal{C}$ and $\mathcal{D}$ are categories, a **covariant**[2] **functor** $F : \mathcal{C} \to \mathcal{D}$ consists of an assignment of an object $F(X) \in \mathrm{Ob}(\mathcal{D})$ to each object $X \in \mathrm{Ob}(\mathcal{C})$, together with a system of functions

$$F : \mathcal{C}(X, Y) \to \mathcal{D}(F(X), F(Y))$$

for the various pairs of objects $X, Y \in \mathrm{Ob}(X)$, such that identities are preserved and F commutes with composition:

(a) for each $X \in \mathrm{Ob}(\mathcal{C})$,
$$F(\mathrm{Id}_X) = \mathrm{Id}_{F(X)};$$

(b) for all $X, Y, Z \in \mathrm{Ob}(\mathcal{C})$, $f \in \mathcal{C}(X, Y)$, and $g \in \mathcal{C}(Y, Z)$,
$$F(g \circ f) = F(g) \circ F(f).$$

Here are two examples:

Example 1:

- Let $\mathcal{C}$ be the category with the following objects and morphisms:
  - $\mathrm{Ob}(\mathcal{C})$ is the class of triples (V, U, x) in which V is a finite dimensional vector space over k, $U \subset V$ is open (relative to some norm, hence any norm because they are all equivalent) and $x \in U$;
  - for each pair of objects (V, U, x) and (V', U', x'),
  $$\mathcal{C}\big((V, U, x), (V', U', x')\big)$$
  is the set of functions $f : U \to U'$ with $f(x) = x'$ that are differentiable at x.

  The composition of any two morphisms is in turn a morphism because if $f \in \mathcal{C}\big((V, U, x), (V', U', x')\big)$ and $g \in \mathcal{C}\big((V', U', x'), (V'', U'', x'')\big)$, then the chain rule implies that $g \circ f$ is differentiable at x, so it is an element of $\mathcal{C}\big((V, U, x), (V'', U'', x'')\big)$.

- Let $\mathcal{D}$ be the category with the following objects and morphisms:

[2] The term 'covariant' serves the purpose of distinguishing the functors discussed here from "contravariant" functors. A **contravariant functor** $K : \mathcal{C} \to \mathcal{D}$ associates an object $K(X) \in \mathrm{Ob}(\mathcal{D})$ with each $X \in \mathrm{Ob}(\mathcal{C})$ and a map $K : \mathcal{C}(X, Y) \to \mathcal{D}(Y, X)$ (note the reversal of X and Y) to each pair $X, Y \in \mathrm{Ob}(X)$. Like covariant functors, contravariant functors must preserve identities and commute with composition. In general contravariant functors are about as important as covariant ones, but they won't figure in our work.

- $\mathrm{Ob}(\mathcal{D})$ is the class of finite dimensional vector spaces over k;
- for all pairs of objects V and V', $\mathcal{D}(V,V')$ is the space $L(V,V')$ of linear transformations from V to V'.

Then there is a functor $F : \mathcal{C} \to \mathcal{D}$ given by

$$F(V,U,x) = V \quad \text{and} \quad F(f) = Df(x)$$

when $(V,U,x) \in \mathrm{Ob}(\mathcal{C})$ and $f \in \mathcal{C}\big((V,U,x),(V',U',x')\big)$. Preservation of identities—that is, $D\mathrm{Id}_U(x) = \mathrm{Id}_V$—follows immediately from Proposition 6.11. The chain rule says precisely that F commutes with composition: if $f : (V,U,x) \to (V',U',x')$ and $g : (V',U',x') \to (V'',U'',x'')$ are morphisms, then

$$F(g \circ f) = D(g \circ f)(x) = Dg(x') \circ Df(x) = F(g) \circ F(f).$$

Example 2: Fix a finite dimensional vector space V over k. Let $\mathcal{D}_V$ be the category whose only object is V, with $\mathcal{D}(V,V) = \mathrm{End}(V)$. Let $\mathcal{E}_k$ be the category with a single object, denoted by $\mathcal{O}_k$, and $\mathcal{E}_k(\mathcal{O}_k,\mathcal{O}_k) = k$, with "composition" of morphisms defined to be multiplication. (In particular, $\mathrm{Id}_{\mathcal{O}_k} = 1 \in k$.) Then there is a functor $G_V : \mathcal{D}_V \to \mathcal{E}_k$ with $G_V(V) = \mathcal{O}_k$ (of course) and

$$G_V(\ell) = \det(\ell)$$

for $\ell \in \mathrm{End}(V)$. Preservation of identities was built into the theory of the determinant when we imposed the requirement that $\det(\mathrm{Id}_V) = 1$. The multiplicative property of the determinant (Theorem 5.13) implies that G_V commutes with composition:

$$G_V(\ell' \circ \ell) = \det(\ell' \circ \ell) = \det(\ell') \cdot \det(\ell) = G_V(\ell') \circ G_V(\ell).$$

There is a general principle of mathematics that is quite possibly just an empty tautology, at least if it is interpreted completely literally, namely that the only way to say something about a mathematical object is to compute the value of some function that has the object in its domain. Functors do this in a very systematic way, and consequently the derivative and the determinant have a relentlessness that goes a long way toward explaining why they are so useful.

Functors can be composed: if, in addition to $F : \mathcal{C} \to \mathcal{D}$, we have a second covariant functor $G : \mathcal{D} \to \mathcal{E}$, then there is a functor $G \circ F$ given by $X \mapsto G(F(X))$ for all $X \in \mathrm{Ob}(\mathcal{C})$ and $f \mapsto G(F(f))$ for all $X, Y \in \mathrm{Ob}(\mathcal{C})$

and $f \in \mathcal{C}(X,Y)$. The conditions defining a functor hold automatically: if $X \in \mathrm{Ob}(\mathcal{C})$, then

$$G(F(\mathrm{Id}_X)) = G(\mathrm{Id}_{F(X)}) = \mathrm{Id}_{G(F(X))},$$

and if $f \in \mathcal{C}(X,Y)$ and $g \in \mathcal{C}(Y,Z)$, then

$$G(F(g \circ f)) = G(F(g) \circ F(f)) = G(F(g)) \circ G(F(f)).$$

After all of my incessant harping on such points, you will, I am sure, be completely unsurprised to learn that there is a category whose objects are categories and whose morphisms are covariant functors.

To display a concrete example, fix a finite dimensional vector space V, let $\mathcal{C}_V$ be the restriction of the category $\mathcal{C}$ in Example 1 to those (V,U,x) with this first component, so that a morphism in $\mathcal{C}_V$ is essentially a function between open subsets of V, together with a particular point in the domain where the function is differentiable. Let F_V be the restriction of F to $\mathcal{C}_V$. If G_V is the functor given by Example 2, then the composition

$$G_V \circ F_V : \mathcal{C}_V \to \mathcal{E}_k$$

computes the determinant $G_V(F_V(f)) = \det(Df(x))$ of the derivative of a morphism $f \in \mathcal{C}\big((V,U,x),(V,U',x')\big)$.

Systems of interacting categories and functors seemingly present the possibility of encompassing a huge amount of mathematical information in structures that are tractable because of this fact. Indeed, during the last half century, particularly in topology and algebraic geometry, categories and functors have been the fundamental building blocks of some very elaborate and powerful machines. We will get a small taste of this in Section 9.5 when we study one of the most important topological functors.

6.7 Partial Derivatives

There is a one-to-one relationship between elements of k and linear functions from k to k. While thinking about a derivative as a linear function is always the "proper" thing to do, computations always boil down to arithmetical operations with numbers, and numbers are convenient in other ways as well. In recognition of this we introduce the following notation. Let W be a finite dimensional vector space over k, let $U \subset k$ be open, and let $g : U \to W$ be a function. If g is differentiable at a point $\overline{t} \in U$, then $g'(\overline{t})$ is the element of W such that

$$Dg(\overline{t})t = tg'(\overline{t}).$$

On the left hand side we are evaluating the linear transformation $Dg(\overline{t})$ at the "vector" $t \in k^1$, and on the right hand side we are multiplying the vector $g'(\overline{t})$ by the scalar t. In many applications, such as the one coming up next, W is one dimensional, and in these cases we regard $g'(\overline{t})$ as an element of k rather than a "vector" in k^1.

Now consider an open set $U \subset k^m$, a function $f : U \to k$, and a point $\overline{x} \in U$. For $i = 1, \dots, m$ let $\tau_i : k \to k^m$ be the linear function

$$\tau_i(t) = (\overline{x}_1, \dots, t, \dots, \overline{x}_m).$$

Then the **partial derivative** of f with respect to x_i at $\overline{x}$ is defined to be

$$\tfrac{\partial f}{\partial x_i}(\overline{x}) := (f \circ \tau_i)'(\overline{x}_i).$$

That is, $\frac{\partial f}{\partial x_i}(\overline{x})$ is a scalar measuring the rate at which f changes as we move away from $\overline{x}$ by changing the i^{th} coordinate.

If you've already had one or more calculus courses, you are certainly aware that the usual presentation of the subject begins with the calculation of the derivatives of simple univariate functions. Various techniques are developed, and the derivatives of more advanced functions are computed, all within the univariate world. This might lead you to expect that, when you finally arrive at the higher level, the notion of a partial derivative will be sophisticated and subtle, but from a computational point of view this isn't true at all: the way to compute a partial derivative is to treat the other variables, say y and z, in the same way that you dealt with constants like a and b when you were differentiating univariate functions. And it's not especially difficult to find simple, intuitive examples of multivariate functions one might differentiate, such as area being a function of height and width.

To the extent that there are subtleties, they largely revolve around the fact that partial derivatives are mostly used to represent linear functions, and linear algebra is frequently a course that comes *after* beginning calculus. In our approach this aspect is already built into the definition. Later we'll see that there actually are some technical subtleties arising from the fact that the existence of all partial derivatives isn't the same as differentiability, but at this point we give only a brief explanation of the central idea.

Consider a function $f : U \to k^n$. From a technical point of view it works perfectly well to think of f as an n-tuple $(f_1, \dots, f_n)$ of functions from U to k; in this point of view the partial $\frac{\partial f_j}{\partial x_i}(\overline{x})$ has already been defined. If all such partials exist, then the **Jacobian matrix** is

$$\begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\overline{x}) & \cdots & \frac{\partial f_1}{\partial x_m}(\overline{x}) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\overline{x}) & \cdots & \frac{\partial f_n}{\partial x_m}(\overline{x}) \end{pmatrix}.$$

At least for those with a certain minimum level of experience, it is intuitive and obvious that this is the matrix of $Df(\overline{x})$ with respect to the standard bases $\mathbf{e}_1, \ldots, \mathbf{e}_m$ and $\mathbf{f}_1, \ldots, \mathbf{f}_n$ for k^m and k^n, but we will now give a formal proof, partly in order to practice some manipulations involving the relationships between the various pieces of notation.

Let $\pi_j : (y_1, \ldots, y_n) \to y_j$ be the projection of k^n onto the j^{th} coordinate. Then $f_j = \pi_j \circ f$, so we can use the chain rule to compute that

$$D(\pi_j \circ f \circ \tau_i)(\overline{x}_i) = D\pi_j(f(\overline{x})) \circ Df(\overline{x}) \circ D\tau_i(\overline{x}_i) \in L(k,k).$$

Since τ_i is affine and π_j is linear, Proposition 6.11 gives $D\tau_i(\overline{x}_i) : t \mapsto t\mathbf{e}_i$ and $D\pi_j(y) = \pi_j$ for any $y \in k^n$. Therefore, for any $t \in k$,

$$D(\pi_j \circ f \circ \tau_i)(\overline{x}_i)t = \pi_j\bigl(Df(\overline{x})(t\mathbf{e}_i)\bigr) = t\pi_j(Df(\overline{x})\mathbf{e}_i).$$

In general the (j,i)-entry of the matrix of a linear transformation $\ell : k^m \to k^n$ with respect to the standard bases is *defined* to be $\pi_j(\ell(\mathbf{e}_i))$, so

$$\frac{\partial f_j}{\partial x_i}(\overline{x}) = (\pi_j \circ f \circ \tau_i)'(\overline{x}_i) = \pi_j(Df(\overline{x})\mathbf{e}_i)$$

is the (j,i)-entry of the matrix of $Df(\overline{x})$.

What else are partial derivatives good for? Well, lots of things, actually, and to get a representative sample, as well as to develop facility in computing with them, you will really need to take a course in multivariate calculus, or at least spend some time with a textbook for such a course. You won't need to understand all the details in order to understand the theoretical overview given here, but it will help to have some sense of the sorts of problems that motivate much of the material. While reading the rest of the chapter you might want to imagine the process of solving some particular chemistry problem, say determining how rapidly the viscosity of a liquid changes as you increase the concentration of some chemical. This might begin with certain theoretical relationship, which can be manipulated algebraicly. At some point certain partial derivatives have to be computed. Possibly after additional algebraic manipulation, the relevant expression has to be evaluated numerically. Although we won't do concrete calculations of this sort, the theory developed here is largely concerned with providing a useful toolkit for such tasks.

6.8 Computation of Derivatives

These days the word ‘calculus’ mostly refers to the processes of differentiation (which we are studying) and integration (which we are not, except for a small taste in Chapter 9) developed by Leibniz and Newton, but it also is used more generally to describe any “machine” for calculating things, especially if it involves processes that generate new instances from some collection of basic examples. In order to describe this sort of machine we need to specify the methods for generating new examples, and we need to give a collection of basic examples, which may be expanded later. This section presents the process of differentiation in this fashion, emphasizing the formal apparatus. As we’ll explain in the next section, it is somewhat more “technical” than the typical presentation in beginning calculus courses, but in exchange for this cost there is the benefit of a clearer and more systematic understanding of the subject.

The central components of our computational machine are methods of computing the derivatives of functions obtained by constructing new functions from functions whose derivatives we already know. There are three mains ways to construct new functions from given functions: composition; inversion; bundling. As we explained in Section 6.6, the chain rule computes the derivative of a composition of two functions, so we’ll now discuss the other two.

Assume that V and W are vector spaces of the same (finite) dimension, $U \subset V$ is open, $f : U \to W$ is an injection, and $f(U)$ is open. If we already knew that the inverse of an invertible differentiable function was differentiable at the image of each point in its domain where its derivative was nonsingular, we could use the chain rule to compute the derivative of the inverse:

$$\mathrm{Id}_V = D\mathrm{Id}_U(\overline{x}) = D(f^{-1} \circ f)(\overline{x}) = Df^{-1}(f(\overline{x})) \circ Df(\overline{x}),$$

so $Df^{-1}(f(\overline{x})) = Df(\overline{x})^{-1}$. But the proof that this formula is correct is complicated by the need to establish differentiability, which requires some rather messy concrete analysis based on the definition of the derivative.

Theorem 6.15. *If f is differentiable at $\overline{x} \in U$ with $Df(\overline{x})$ nonsingular, and f^{-1} is continuous at $\overline{y} := f(\overline{x})$, then f^{-1} is differentiable at $\overline{y}$ with*

$$Df^{-1}(\overline{y}) = Df(\overline{x})^{-1}.$$

One of the key ideas of the proof is that if ℓ is nonsingular and $\ell(x - \overline{x})$ is a sufficiently accurate approximation of $y - \overline{y}$, then $\|y - \overline{y}\|$ cannot be very small in comparison with $\|x - \overline{x}\|$. The calculations that lead to a quantitative expression of this intuition are best dealt with separately.

Lemma 6.16. *If $\ell : V \to W$ is a nonsingular linear transformation, $0 < \varepsilon < \|\ell^{-1}\|^{-1}$, $x, \overline{x} \in V$, $y, \overline{y} \in W$, and*

$$\|y - [\overline{y} + \ell(x - \overline{x})]\| \leq \varepsilon \|x - \overline{x}\|$$

then

$$\|x - \overline{x}\| \leq \frac{\|y - \overline{y}\|}{\|\ell^{-1}\|^{-1} - \varepsilon}.$$

Proof. For any $v \in V$ the operator norm inequality gives $\|v\| = \|\ell^{-1}(\ell(v))\| \leq \|\ell^{-1}\| \cdot \|\ell(v)\|$, so $\|\ell(v)\| \geq \|v\|/\|\ell^{-1}\|$. Since

$$y - \overline{y} = \ell(x - \overline{x}) + y - [\overline{y} + \ell(x - \overline{x})],$$

the triangle inequality for the norm gives

$$\|y - \overline{y}\| \geq \|\ell(x - \overline{x})\| - \|y - [\overline{y} + \ell(x - \overline{x})]\| \geq (1/\|\ell^{-1}\| - \varepsilon)\|x - \overline{x}\|.$$

□

We proceed to the heart of the argument.

Proof of 6.15. The proof will be explained in terms of a point $y \in f(U)$ and a number $\varepsilon > 0$ that will be fixed throughout. Our goal is to find $\delta > 0$ such that if $\|y - \overline{y}\| < \delta$, then

$$\big\|f^{-1}(y) - [f^{-1}(\overline{y}) - Df(x)^{-1}(y - \overline{y})]\big\| \leq \varepsilon \|y - \overline{y}\|. \qquad (*)$$

To simplify notation let $\ell := Df(\overline{x})$ and $x := f^{-1}(y)$, and note that

$$\begin{aligned} f^{-1}(y) - [f^{-1}(\overline{y}) - \ell^{-1}(y - \overline{y})] &= -\ell^{-1}\big(y - [\overline{y} + \ell(f^{-1}(y) - f^{-1}(\overline{y}))]\big) \\ &= -\ell^{-1}\big(y - [\overline{y} + \ell(x - \overline{x})]\big). \end{aligned}$$

The operator norm inequality gives

$$\big\| - \ell^{-1}\big(y - [\overline{y} + \ell(x - \overline{x})]\big)\big\| \leq \|\ell^{-1}\| \cdot \big\|y - [\overline{y} + \ell(x - \overline{x})]\big\|.$$

Choose $\tilde{\varepsilon} > 0$ small enough that $\tilde{\varepsilon} < \|\ell^{-1}\|^{-1}$ and $\tilde{\varepsilon}\|\ell^{-1}\|/(\|\ell^{-1}\|^{-1} - \tilde{\varepsilon}) \leq \varepsilon$. The definition of $Df(\overline{x})$ gives a $\tilde{\delta} > 0$ such that

$$\big\|y - [\overline{y} + \ell(x - \overline{x})]\big\| \leq \tilde{\varepsilon}\|x - \overline{x}\|$$

whenever $\|x - \overline{x}\| < \tilde{\delta}$, so that (by the result above)

$$\|x - \overline{x}\| \leq \frac{\|y - \overline{y}\|}{\|\ell^{-1}\|^{-1} - \tilde{\varepsilon}}.$$

By assumption f^{-1} is continuous at $\overline{y}$, so there is $\delta > 0$ such that $\|x - \overline{x}\| = \|f^{-1}(y) - f^{-1}(\overline{y})\| < \tilde{\delta}$ whenever $\|y - \overline{y}\| < \delta$, in which case these inequalities evidently combine to give $(*)$. □

Differentiation of new functions created by bundling together existing functions is simpler than inversion. We will only explicitly deal with the case of two functions, but it will be obvious that the method could be extended to any finite number of functions, and you should understand the discussion in this more general sense.

Suppose that W and W' are two vector spaces over k. Their **direct sum** $W \oplus W'$ is the cartesian product $W \times W'$ endowed with the obvious vector operations:

$$(w_1, w_1') + (w_2, w_2') := (w_1 + w_2, w_1' + w_2') \quad \text{and} \quad \alpha(w, w') := (\alpha w, \alpha w').$$

(You should quickly run through the conditions defining a vector space, verifying that they are satisfied by these operations.)

If V is another vector space over k, $U \subset V$ is open and $f : U \to W$ and $f' : U \to W'$ are functions, then we can define a function

$$f \oplus f' : U \to W \oplus W'$$

by setting

$$(f \oplus f')(x) := (f(x), f'(x)).$$

If $\ell : V \to W$ and $\ell' : V \to W'$ are linear, then $\ell \oplus \ell' : V \to W \oplus W'$ is also linear, obviously.

Theorem 6.17. *If $f : U \to W$ and $f' : U \to W'$ are differentiable at $\overline{x}$, then so is $f \oplus f'$, and*

$$D(f \oplus f')(\overline{x}) = Df(\overline{x}) \oplus Df'(\overline{x}).$$

Proof. Fix norms on V, W, and W'. The definition of the derivative is unaffected by the choice of norm, so we can impose any norm we like on $W \oplus W'$. For the relevant computations it is convenient to choose the norm

$$\|(w, w')\| := \|w\| + \|w'\|.$$

(You should verify that for any norms on W and W', this is a norm on $W \oplus W'$.) Let $\ell := Df(\overline{x})$ and $\ell' := Df'(\overline{x})$.

Fix $\varepsilon > 0$. There is $\delta > 0$ such that

$$\|f(x) - [f(\overline{x}) + \ell(x - \overline{x})]\| \leq (\varepsilon/2)\|x - \overline{x}\|$$

and

$$\|f'(x) - [f'(\overline{x}) + \ell'(x - \overline{x})]\| \leq (\varepsilon/2)\|x - \overline{x}\|$$

whenever $\|x - \overline{x}\| < \delta$. The definition of the direct sum now gives

$$\begin{aligned}\big(f \oplus f'\big)(x) - \big[(f \oplus f')(\overline{x}) + (\ell \oplus \ell')(x - \overline{x})\big] \\ &= \big(f(x), f'(x)\big) - \big[(f(\overline{x}), f'(\overline{x})) + (\ell(x - \overline{x}), \ell'(x - \overline{x}))\big] \\ &= \big(f(x) - [f(\overline{x}) + \ell(x - \overline{x})], f'(x) - [f'(\overline{x}) + \ell'(x - \overline{x})]\big).\end{aligned}$$

In view of our choice of norm for $W \oplus W'$ it is now clear that

$$\big\|(f \oplus f')(x) - [(f \oplus f')(\overline{x}) + (\ell \oplus \ell')(x - \overline{x})]\big\| \leq \varepsilon\|x - \overline{x}\|$$

whenever $\|x - \overline{x}\| < \delta$. □

There is a slightly different sort of bundling that also comes up frequently. Suppose that V' is another finite dimensional vector space, and $U' \subset V'$ is open. If $f : U \to W$ and $g : U' \to W'$ are functions, then there is the obvious derived function $(x, x') \mapsto (f(x), g(x'))$. To express this in terms of the bundling operation above we can introduce the projections $\pi : (x, x') \mapsto x$ and $\pi' : (x, x') \mapsto x'$ from $V \oplus V'$ to V and V' respectively. Then the function in question can be thought of as

$$(f \circ \pi) \oplus (g \circ \pi') : U \times U' \to W \oplus W'.$$

Since π and π' are linear, the result above, the basic facts about differentiating affine functions (Proposition 6.11) and the chain rule provide the tools we need to compute the derivative of this function.

Now that we understand how to compute the derivatives of compositions, inverses, and direct sums, we should start to gather some basic functions whose derivatives are known. We have already dealt with affine functions, and this is as good a point as any to mention a quite trivial but extremely important special case. If $U, U' \subset V$ are open with $U \subset U'$ and $i_{U,U'} : U \to U'$ is the inclusion, then Proposition 6.11 gives $Di_{U,U'}(x) := \mathrm{Id}_V$. If $f : U' \to W$ is a function, then the restriction $f|_U$ of f to U is the composition $f \circ i_{U,U'}$, and for any $x \in U$ we can (to be quite formal about things) use the chain rule to compute that

$$Df|_U(x) = Df(x) \circ Di_{U,U'}(x) = Df(x) \circ \mathrm{Id}_V = Df(x).$$

The inclusion functions are so trivial that it's easy to not even notice them, but they come up all the time.

Addition of scalars and negation are linear, so we have:

Proposition 6.18. *Let $A : k^2 \to k$ be addition—that is, $A(s,t) := s + t$. Then A is differentiable at each $(\overline{s}, \overline{t}) \in k^2$, and $DA(\overline{s}, \overline{t}) = A$.*

Proposition 6.19. *Let $N : k \to k$ be negation—that is, $N(s) := -s$. Then N is differentiable at each $\overline{s} \in k$, and $DN(\overline{s}) = N$.*

Multiplication is a bit more complicated, but the underlying intuition is simple, and will eventually become quite familiar: for any $\overline{s}$, $\overline{t}$, Δs, and Δt we have

$$(\overline{s} + \Delta s)(\overline{t} + \Delta t) = \overline{s}\overline{t} + \overline{s}\Delta t + \overline{t}\Delta s + \Delta s \Delta t,$$

and if Δs and Δt are both very small, then $\Delta s \Delta t$ is negligible.

Proposition 6.20. *Let $M : k^2 \to k$ be multiplication—that is, $M(s,t) := st$. Then M is differentiable at each $(\overline{s}, \overline{t}) \in k^2$, and $DM(\overline{s}, \overline{t})$ is the linear transformation $(s,t) \mapsto \overline{s}t + \overline{t}s$.*

Proof. We endow k^2 with the norm $\|(s,t)\| := |s| + |t|$. If $\varepsilon > 0$, $\delta \leq 2\varepsilon$, and $\|(s,t) - (\overline{s}, \overline{t})\| < \delta$, then $|s - \overline{s}|, |t - \overline{t}| < 2\varepsilon$, so

$$\begin{aligned}\left|st - [\overline{s}\overline{t} + (\overline{s}(t - \overline{t}) + \overline{t}(s - \overline{s}))]\right| &= |st - \overline{s}\overline{t} - \overline{s}t + \overline{s}\overline{t} - \overline{t}s + \overline{t}\overline{s}| \\ = |st - \overline{s}t - s\overline{t} + \overline{s}\overline{t}| &= |(s - \overline{s})(t - \overline{t})| = |s - \overline{s}|\,|t - \overline{t}| \\ &= |s - \overline{s}|(\tfrac{1}{2}|t - \overline{t}|) + (\tfrac{1}{2}|s - \overline{s}|)|t - \overline{t}| \\ &\leq \varepsilon(|s - \overline{s}| + |t - \overline{t}|) = \varepsilon\|(s,t) - (\overline{s}, \overline{t})\|.\end{aligned}$$

□

We can illustrate the application of the results above by computing the derivative of inversion. The formal details are ponderous, but again there is a simple calculation that suggests the right answer: if $I\overline{s} = 1$ and $(I + \Delta I)(\overline{s} + \Delta s) = 1$, then $I\Delta s + \overline{s}\Delta I + \Delta I \Delta s = 0$, so

$$\Delta I = -\frac{I}{\overline{s}}\Delta s - \frac{1}{\overline{s}}\Delta I \Delta s = -\frac{1}{\overline{s}^2}\Delta s - \frac{1}{\overline{s}}\Delta I \Delta s.$$

If Δs is very small, then so is ΔI, so that $\Delta I \Delta s$ is negligible.

Proposition 6.21. *Let $I : k^* \to k$ be inversion—that is, $I(s) := 1/s$. Then I is differentiable at each $\overline{s} \in k^*$, and $DI(\overline{s}) : s \mapsto -s/\overline{s}^2$.*

Proof. Since $M \circ (\mathrm{Id}_k \oplus I)$ is a constant function, its derivative is zero (by Proposition 6.11) so the chain rule gives

$$0 = D(M \circ (\mathrm{Id}_k \oplus I))(\overline{s}) = DM(\overline{s}, 1/\overline{s}) \circ D(\mathrm{Id}_k \oplus I)(\overline{s}).$$

Theorem 6.17 implies that $D(\mathrm{Id}_k \oplus I)(\overline{s}) = D\mathrm{Id}_k(\overline{s}) \oplus DI(\overline{s})$, and Proposition 6.11 gives $D\mathrm{Id}_k(\overline{s}) = \mathrm{Id}_k$, so for any $s \in k$ we have $(D\mathrm{Id}_k(\overline{s}) \oplus DI(\overline{s}))s = (s, DI(\overline{s})s)$. Applying the formula for the derivative of M gives

$$0 = DM(\overline{s}, 1/\overline{s})(s, DI(\overline{s})s) = \overline{s}DI(\overline{s})s + s/\overline{s},$$

and the desired result is obtained by solving for $DI(\overline{s})s$. □

Our formalism is a bit like a Swiss Army Knife. It can do pretty much anything, but it doesn't do any one thing in the most efficient or elegant way, and you wouldn't say it was "sleek." These aspects are all on display in the proof above, and in the proof of the next result, which gives formulas for addition and multiplication of univariate functions. Combining univariate functions in this way is quite common, so these formulas are quite useful.

Proposition 6.22. *Suppose that $U \subset k$ is open and $f, g : U \to k$ are differentiable at $\overline{t}$. Then:*

- $D(f + g)(\overline{t}) = Df(\overline{t}) + Dg(\overline{t})$;
- $D(fg)(\overline{t}) = f(\overline{t})Dg(\overline{t}) + g(\overline{t})Df(\overline{t})$.

Proof. Clearly, $f + g = A \circ (f \oplus g)$, so

$$D(f + g)(\overline{t}) = D(A \circ (f \oplus g))(\overline{t}) = DA((f \oplus g)(\overline{t})) \circ D(f \oplus g)(\overline{t})$$
$$= A \circ (Df(\overline{t}) \oplus Dg(\overline{t})) = Df(\overline{t}) + Dg(\overline{t}).$$

Here the second equality is the chain rule, the third applies the results concerning the derivative of A and the differentiation of direct sums, and the last follows from the definitions of A and the direct sum. Similarly, $fg = M \circ (f \oplus g)$, so

$$D(fg)(\overline{t}) = D(M \circ (f \oplus g))(\overline{t}) = DM((f \oplus g)(\overline{t})) \circ D(f \oplus g)(\overline{t})$$
$$= DM(f(\overline{t}), g(\overline{t})) \circ (Df(\overline{t}) \oplus Dg(\overline{t})) = f(\overline{t})Dg(\overline{t}) + g(\overline{t})Df(\overline{t}).$$

□

6.9 Practical Computation

It's interesting to compare what we've done so far with the standard presentation of differentiation in calculus courses. As we mentioned earlier, usually the derivative of a univariate function $f : U \to k$ on an open set $U \subset k$ at a point $\overline{t}$ is *defined* to be the slope $f'(\overline{t})$ of the line that is tangent to the graph of the function at $(\overline{t}, f(\overline{t}))$. Consider the formulas for differentiation of sums, products, and compositions of univariate functions in the two frameworks:

$$D(f+g)(\overline{t}) = Df(\overline{t}) + Dg(\overline{t});$$

$$(f+g)'(\overline{t}) = f'(\overline{t}) + g'(\overline{t});$$

$$D(fg)(\overline{t}) = f(\overline{t})Dg(\overline{t}) + g(\overline{t})Df(\overline{t});$$

$$(fg)'(\overline{t}) = f(\overline{t})g'(\overline{t}) + g(\overline{t})f'(\overline{t});$$

$$D(g \circ f)(\overline{t}) = Dg(f(\overline{t})) \circ Df(\overline{t});$$

$$(g \circ f)'(\overline{t}) = g'(f(\overline{t}))f'(\overline{t}).$$

On the left we have elements of $\text{End}(k) = L(k, k)$ which are added, multiplied by scalars, and composed. On the right we have elements of k which are added and multiplied. For any vector space V, $\text{End}(V)$ is a vector space over k, and a ring if we define multiplication to be functional composition. *The standard approach to calculus depends on the fact that* $\text{End}(k)$ *is isomorphic to* k*, with multiplication in* k *representing both composition of elements of* $\text{End}(k)$ *and multiplication of an element of* $\text{End}(k)$ *by a scalar.*

Is this a wonderful blessing or a confusing dirty trick? Probably a bit of both. In the explanation of differentiation given here there is a direct connection between what the derivative represents and what it is, and the passage from definitions to computational methods is overtly systematic, but we had to develop an overarching theoretical framework before we could work in this style. (One might argue that this is a benefit and not a cost!) Our approach is inherently multi-dimensional, but in the standard approach the transition from univariate functions to multivariate functions is difficult, in large part because the concepts from linear algebra that the standard approach finesses so cleverly can no longer be avoided.

When it comes to actual computations the standard approach has clear advantages. If $f'(u)$ is defined at every point in U and $f' : U \to k$ is differentiable at $\overline{u}$, then $f''(\overline{u}) := (f')'(\overline{u})$ denotes the **second derivative**. In general the r^{th} derivative (including when $r = 0, 1, 2$, if the context is suitable) is denoted by $f^{(r)}$ when it exists. Within our formalism just writing down the space $L(V, L(V, W))$ in which the second derivative lives is a nuisance. The advantages of the numerical approach can be illustrated by deriving some basic formulas. For $m = 0, 1, 2, \ldots$ let $g_m : k \to k$ be the function $g_m(t) := t^m$. Since g_0 and g_1 are affine we have $g_0'(t) = 0$ and $g_1'(t) = 1$ for all t. Using induction, we find that $g_m'(t) = mg_{m-1}(t)$ for all t, since the rule for differentiating products gives

$$g_m' = (g_{m-1}g_1)' = g_{m-1}g_1' + g_{m-1}'g_1 = t^{m-1} \cdot 1 + (m-1)t^{m-2} \cdot t = mt^{m-1}.$$

Differentiating g_m repeatedly gives

$$g_m^{(r)}(t) = m(m-1)\cdots(m-r+1)t^{m-r} = \frac{m!}{(m-r)!}t^{m-r} = r!\binom{m}{r}t^{m-r},$$

so $g_m^{(m)}(t) = m!$. The derivative of a constant function is zero (of course it is affine) so for any $c \in k$ and any $f : U \to k$ the product formula gives

$$(cf)' = 0 \cdot f + c \cdot f' = cf'.$$

For a univariate polynomial

$$f(t) = c_r t^r + \cdots + c_1 t + c_0$$

we obtain $c_h = \frac{1}{h!}f^{(h)}(0)$ for each $h = 0, \ldots, r$, so we have the interesting formula

$$f(t) = \frac{1}{r!}f^{(r)}(0)t^r + \frac{1}{(r-1)!}f^{(r-1)}(0)t^{r-1} + \cdots + f^{(1)}(0)t + f^{(0)}(0).$$

In our framework these computations would be extremely cumbersome due to all the distinctions between the spaces k, $L(k,k)$, $L(k, L(k,k))$, etc., that have no real bearing on the calculations. Still, at the end of a computation, when you are thinking about what it means and why it is or isn't interesting or to the point, the conceptual stuff can be important. "Think conceptually, act computationally" seems to be an appropriate motto.

In the one dimensional case there is a one-to-one relationship between our notion of derivative and the concept taken as definitional in the standard approach. The relationship between partial derivatives and our notion of differentiability for multivariate functions is more complex. It turns out that simply requiring that the partial derivatives of $f : U \to W$ be defined at each point is not very useful, because some very poorly behaved functions meet this condition. We'll illustrate the pathologies of partial derivatives, and practice the computational methods described above, by studying the function $\tilde{f} : \mathbb{R}^2 \to \mathbb{R}$ given by

$$\tilde{f}(x,y) := \begin{cases} \frac{x^3y^3}{x^6+y^6}, & x > 0 \text{ and } y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

First of all, observe that $\tilde{f}(z,z) = 1/2$ for all $z > 0$, so $\tilde{f}$ is not even continuous at $(0,0)$.

Before we can deal with $\tilde{f}$ we need to learn how to differentiate quotients. Suppose that $U \subset k$ is open and $g, h : U \to k$ with $g(t)h(t) = 1$ for all t, so that $g = 1/h$. Applying the product rule gives $g'h + h'g = 0$, so that

$$(1/h)'(t) = g'(t) = -h'(t)g(t)/h(t) = -h'(t)/h^2(t).$$

(One can also obtain this by applying Proposition 6.21 and the chain rule to $I \circ g$ where $I : s \mapsto 1/s$ is inversion.) Applying the formula for differentiation of products gives the general formula for the derivative of a quotient:

$$(g/h)' = g'(1/h) + g(1/h)' = g'/h - gh'/h^2 = \frac{g'h - gh'}{h^2}.$$

By the way, the best way to learn these formulas is to force yourself to *not* memorize them, instead remembering that they can be derived from the product formula. After you've been through the derivations yourself a couple times, they'll start to stick, and you'll have the added confidence of knowing that if you are ever the least bit uncertain about the details ("Is it $g'h - gh'$ or $gh' - g'h$ on top?") you can just work it out again quickly.

The way to compute a partial derivative, say with respect to the variable x, is to treat all the other variables as "parameters," which means that they are treated as (perhaps unknown) constants in the calculation, even though they might be thought of as varying in some larger context, or from the point of view of a different sort of question. So, when $x > 0$ and $y > 0$ we can apply the formula above and what we learned earlier about differentiating polynomials to $\tilde{f}$, viewed as a function of x with y treated as a parameter:

$$\frac{\partial \tilde{f}}{\partial x}(x,y) = \frac{(3x^2y^3)(x^6+y^6) - (x^3y^3)(6x^5)}{(x^6+y^6)^2} = \frac{3x^2y^3(y^6-x^6)}{(x^6+y^6)^2}$$

when $x > 0$ and $y > 0$.

If $y < 0$ or $x < 0$, then $\frac{\partial \tilde{f}}{\partial x}(x,y) = 0$ because $\tilde{f}(x',y) = 0$ for all x' in some interval $(x-\delta, x+\delta)$, and in fact this is true even when $x \geq 0$ and $y = 0$. The remaining case is $x = 0$ and $y > 0$. For any $\varepsilon > 0$, if $\delta \leq \sqrt{\varepsilon y^3}$ and $|x| < \delta$, then $|\tilde{f}(x,y) - \tilde{f}(0,y)| = 0$ if $x < 0$, and

$$|\tilde{f}(x,y) - \tilde{f}(0,y)| = \frac{x^3y^3}{x^6+y^6} \leq \frac{x^2}{y^3}x \leq \varepsilon|x|$$

when $x \geq 0$, so $\frac{\partial \tilde{f}}{\partial x}(0,y) = 0$. We have shown that

$$\frac{\partial \tilde{f}}{\partial x}(x,y) = \begin{cases} \frac{3x^2y^3(y^6-x^6)}{(x^6+y^6)^2}, & x > 0 \text{ and } y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The analysis of $\frac{\partial \tilde{f}}{\partial y}$ is symmetric. The point of this example is that $\tilde{f}$ has well defined partials at *every* point of $\mathbb{R}^2$ even though it is actually discontinuous at the origin.

Before continuing the discussion of this example we need to introduce the terminology and notation associated with higher order partial derivatives, which are partial derivatives of the partial derivative functions. Suppose that $U \subset k^n$ is open, $f : U \to k$ is a function, $\overline{x} \in U$, and $1 \leq i \leq n$. When it is defined the **second partial derivative** of f with respect to x_i is

$$\frac{\partial^2 f}{\partial x_i^2}(\overline{x}) := \frac{\partial(\frac{\partial f}{\partial x_i})}{\partial x_i}(\overline{x}).$$

If $1 \leq i, j \leq n$, then the so-called **mixed second partial** of f with respect to x_i and x_j is

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\overline{x}) := \frac{\partial(\frac{\partial f}{\partial x_i})}{\partial x_j}(\overline{x}).$$

More generally, if $1 \leq i_1, \ldots, i_r \leq n$, then $\frac{\partial^r f}{\partial x_{i_1} \cdots \partial x_{i_r}}(\overline{x})$ denotes the result of partial differentiation first with respect to x_{i_1}, then with respect to x_{i_2}, and so forth. We say that this is a partial derivative of **order** r. The **higher order partial derivatives** of f are those of order 2 and higher.

Calculations similar to those above, which are left to you, show that each second partial derivative of $\tilde{f}$ is defined at every point of $\mathbb{R}^2$. If we wanted all third partials to exist, we could replace the exponents '3' and '6' in the definition of $\tilde{f}$ with '4' and '8.' In fact extending this pattern shows that for any integer r there are functions from $\mathbb{R}^2$ to $\mathbb{R}$ that are discontinuous in spite of having all partial derivatives up to order r defined at every point of the domain.

The most popular assumption in analytic work is that all partial derivatives up to order r are defined *and continuous* everywhere in the domain. A function satisfying this condition is said to be C^r. (The pronunciation is "see-are.") If a function is C^r for every integer r, then it is C^∞. In this system of terminology 'C^0' is a synonym for 'continuous.'

As we will see in the next section, a real valued function whose domain is an open subset of a vector space over $\mathbb{R}$ is differentiable everywhere, in our sense, with a continuous derivative, if and only if it is C^1. Since the definition of Df does not depend on a choice of coordinate system, it follows that the choice of a coordinate system does not affect whether a function is C^1. In contrast, as the example above makes very clear, the existence of partial derivatives can easily depend on the coordinate system.

6.10 Rolle, Clairaut, Taylor

The guiding intuition of our work has been that differentiability *means* that an accurate affine approximation exists. It would be consistent with this outlook to say that the meaning of "differentiability of order r" is that the function can be well approximated by a polynomial of degree r. In the standard approach definitions of higher order differentiability are typically expressed in terms of partial derivatives, which at first sight might seem a bit to the side of the real point. However, Taylor's theorem (due to Brook Taylor (1685-1731)) states that a function can be approximated by a polynomial of degree r, with error that is "small of order r," if it is C^r. Thus the assumption that the function is C^r is strong enough to imply the property of conceptual interest, and in addition to this it is easy to work with and there is little interest in functions with partial derivatives that exist but are discontinuous. For all these reasons the assumption that the function is C^r has become part of the standard framework.

Up until this point we have been working over a rather general field k, but the next theorem due to Michel Rolle (1652-1719), which is at the heart of the proofs in this part of the theory, is specifically about the case $k = \mathbb{R}$. We'll use it in the rest of the section to prove Claurаut's theorem, which is about second order partial derivatives, and Taylor's theorem. These results will be extended to the case $k = \mathbb{C}$ in the next chapter, but the proofs there will not be straightforward extensions of the arguments below.

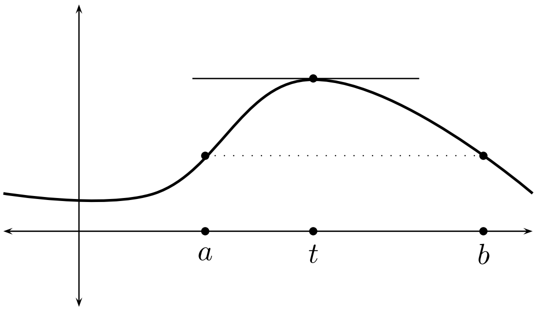


Figure 6.4

Theorem 6.23 (Rolle's Theorem)**.** *If $a < b$, $f : [a, b] \to \mathbb{R}$ is continuous, $f(a) = f(b)$, and f is differentiable at each $t \in (a, b)$, then $f'(t) = 0$ for some $t \in (a, b)$.*

Proof. Since $[a, b]$ is compact (Lemma 3.37) and f is continuous, f attains maximum at some point t^*. (Theorem 3.48.) If $a < t^* < b$, then $f'(t^*) = 0$, by Theorem 6.13 applied to $f|_{(a,b)}$. Similarly, f attains its minimum at some t_*, and $f'(t_*) = 0$ if $a < t_* < t$. The remaining possibility is that $t^*, t_* \in \{a, b\}$, but then $f(t) = f(a)$ for all t, so that $f'(t) = 0$ for all t. □

In proofs the following (very slight and obvious) generalization typically saves a step in the argument. For this reason it has its own name.

Theorem 6.24 (Mean Value Theorem)**.** *If $a < b$, $f : [a, b] \to \mathbb{R}$ is continuous, and f is differentiable at each $t \in (a, b)$, then for some $t \in (a, b)$ we have*

$$f'(t) = \frac{f(b) - f(a)}{b - a}.$$

Proof. Let $g : [a, b] \to \mathbb{R}$ be the function

$$g(t) := f(t) - \frac{f(b) - f(a)}{b - a}(t - a).$$

Then $g(a) = g(b)$, and Rolle's theorem gives a number $t \in (a, b)$ with

$$0 = g'(t) = f'(t) - \frac{f(b) - f(a)}{b - a}.$$

□

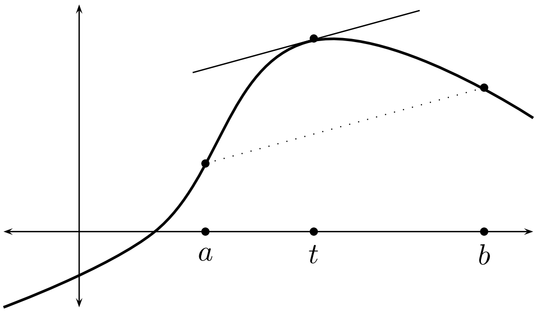


Figure 6.5

Earlier we saw that if f is a univariate polynomial of degree r, then

$$f(t) = \frac{1}{r!}f^{(r)}(0)t^r + \frac{1}{(r-1)!}f^{(r-1)}(0)t^{r-1} + \cdots + f^{(1)}(0)t + f^{(0)}(0).$$

The general idea of Taylor's theorem is that even if f isn't itself a polynomial, it will still (under suitable hypotheses) be well approximated by this sort of polynomial. We begin with the univariate version because the proof is a straightforward reflection of the underlying logic.

Theorem 6.25 (Univariate Taylor's Theorem). *If a and b are real numbers with $a < b$, $r \geq 0$ is an integer, and $f : (a,b) \to \mathbb{R}$ is C^r, then for each $\overline{t} \in (a,b)$ and $\varepsilon > 0$ there is $\delta > 0$ such that*

$$\Big| f(\overline{t} + t) - \sum_{i=0}^{r} \frac{1}{i!} f^{(i)}(\overline{t}) t^i \Big| \leq \varepsilon |t|^r$$

whenever $|t| < \delta$.

Proof. Fix $\overline{t}$ and ε. To reduce the amount of clutter we define a new function $g : (a - \overline{t}, b - \overline{t}) \to \mathbb{R}$ by setting

$$g(t) := f(t + \overline{t}) - \sum_{i=0}^{r} \frac{1}{i!} f^{(i)}(\overline{t}) t^i.$$

It is clear, in view of our rules for differentiating sums and products, that g is C^r and $g^{(i)}(0) = 0$ for each $1 \leq i \leq r$. Our goal is to find $\delta > 0$ such that $|g(t)| \leq \varepsilon |t|^r$ whenever $|t| < \delta$.

We will argue by induction on r. When $r = 0$ the assertion is simply that g is continuous, which is true because f is C^0 (i.e., continuous) by assumption, so suppose the claim has already been demonstrated with $r - 1$ in place of r. The hypotheses of that case are satisfied by g', so there is $\delta > 0$ such that $|g'(t)| \leq \varepsilon |t|^{r-1}$ whenever $|t| < \delta$. Consider such a t. If $t = 0$, then the claim holds simply because $g(0) = 0$. Otherwise the mean value theorem implies that there is t' strictly between 0 and t such that $g'(t') = g(t)/t$, so that

$$|g(t)| = |g'(t')| \cdot |t| \leq \varepsilon |t'|^{r-1} |t| < \varepsilon |t|^r.$$

□

The most important fact about higher order partial derivatives is that, provided that $k = \mathbb{R}$ and suitable continuity assumptions hold, the order of differentiation doesn't matter. This result is sometimes called Young's theorem, but it is properly attributed to Alexis Clairaut (1713-1765).

Theorem 6.26 (Clairaut's Theorem). *Suppose that $U \subset \mathbb{R}^n$ is open, $f : U \to \mathbb{R}$ is a function, and for some $1 \leq i < j \leq n$ the partial derivatives $\frac{\partial f}{\partial x_i}$, $\frac{\partial f}{\partial x_j}$, $\frac{\partial^2 f}{\partial x_i \partial x_j}$, and $\frac{\partial^2 f}{\partial x_j \partial x_i}$ are defined everywhere in U, with $\frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\frac{\partial^2 f}{\partial x_j \partial x_i}$ continuous at $\overline{x}$. Then*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\overline{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\overline{x}).$$

Proof. Let $\mathbf{e}_i$ and $\mathbf{e}_j$ be the standard unit basis vectors. Since U is open, for sufficiently small $\delta_i, \delta_j > 0$ the rectangle

$$\{\, \overline{x} + s_i\mathbf{e}_i + s_j\mathbf{e}_j : 0 \le s_i \le \delta_i, 0 \le s_j \le \delta_j \,\}$$

is contained in U. We will use the mean value theorem to express the quantity

$$A := f(\overline{x} + \delta_i\mathbf{e}_i + \delta_j\mathbf{e}_j) - f(\overline{x} + \delta_i\mathbf{e}_i) - f(\overline{x} + \delta_j\mathbf{e}_j) + f(\overline{x})$$

in terms of the relevant second partials.

For $0 \le s_i \le \delta_i$ let

$$g(s_i) := f(\overline{x} + s_i\mathbf{e}_i + \delta_j\mathbf{e}_j) - f(\overline{x} + s_i\mathbf{e}_i).$$

Then $A = g(\delta_i) - g(0)$. The mean value theorem implies that there is some s_i strictly between 0 and δ_i such that

$$A = g'(s_i)\delta_i = \Big[\frac{\partial f}{\partial x_i}(\overline{x} + s_i\mathbf{e}_i + \delta_j\mathbf{e}_j) - \frac{\partial f}{\partial x_i}(\overline{x} + s_i\mathbf{e}_i)\Big]\delta_i.$$

Applying the mean value theorem a second time, there is a number s_j strictly between 0 and δ_j such that

$$\frac{A}{\delta_i} = \frac{\partial f}{\partial x_i}(\overline{x} + s_i\mathbf{e}_i + \delta_j\mathbf{e}_j) - \frac{\partial f}{\partial x_i}(\overline{x} + s_i\mathbf{e}_i) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\overline{x} + s_i\mathbf{e}_i + s_j\mathbf{e}_j)\delta_j.$$

The same argument can be applied with i and j reversed to get $s'_i \in (0, \delta_i)$ and $s'_j \in (0, \delta_j)$ such that

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\overline{x} + s_i\mathbf{e}_i + s_j\mathbf{e}_j) = \frac{A}{\delta_i\delta_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}(\overline{x} + s'_i\mathbf{e}_i + s'_j\mathbf{e}_j).$$

Since the second partials are continuous at $\overline{x}$, by choosing δ_i and δ_j sufficiently small we can force the left hand side to belong to any neighborhood of $\frac{\partial^2 f}{\partial x_i \partial x_j}(\overline{x})$ and the right hand side to belong to any neighborhood of $\frac{\partial^2 f}{\partial x_j \partial x_i}(\overline{x})$, and of course this is impossible unless these two quantities are equal. □

The multivariate version of Taylor's theorem asserts that a certain polynomial provides a good approximation of a C^r real valued function. We now define and characterize this polynomial, not just for the case of the reals, but for a general field. (This generality will be useful in connection with the complex version of Taylor's theorem.) If $U \subset k^n$ is open, $r \ge 0$ is an integer, $\overline{x} \in U$, and the partial derivatives up to order r of the function $f : U \to k$ are defined at $\overline{x}$, then the r^{th} *order Taylor's series of f at $\overline{x}$* is the polynomial

$$T^r(f;\overline{x})(y) = \sum_{j=0}^{r} T_j(f;\overline{x})(y)$$

where $T_0(f;\overline{x})(y) = f(\overline{x})$ (regarded as a polynomial) and for $j = 1, \dots, r$,

$$T_j(f;\overline{x})(y) := \frac{1}{j!}\Big(\sum_{i_1,\dots,i_j=1}^{n} \frac{\partial^j f}{\partial x_{i_1} \cdots \partial x_{i_j}}(\overline{x}) y_{i_1} \cdots y_{i_j} \Big) \in k[y_1, \dots, y_n].$$

To simplify the notation we will often suppress reference to $\overline{x}$, since it will remain fixed throughout the analysis, for instance writing $T^r(f)(y)$ in place of $T^r(f;\overline{x})(y)$, and sometimes y will also be implicit.

We will need to check that $T^r(f)(y)$ has the same partial derivatives as f. The next result presents a single step in the calculation that proves this.

Lemma 6.27. *If Clairaut's theorem holds for the field k, so that higher order partials do not depend on the order of differentiation, then for each $k = 1, \dots, n$,*

$$\frac{\partial T^r(f)}{\partial y_k}(y) = T^{r-1}\Big(\frac{\partial f}{\partial x_k}\Big)(y).$$

Proof. The derivative of the constant term vanishes, of course, so the desired formula is obtained by summing the following computation (which is explained in detail below) over $j = 1, \dots, r$.

$$\begin{aligned}
\frac{\partial T_j(f)}{\partial y_k}(y) &= \frac{\partial}{\partial y_k}\Big(\frac{1}{j!} \sum_{i_1,\dots,i_j=1}^{n} \frac{\partial^j f}{\partial x_{i_1} \cdots \partial x_{i_j}} y_{i_1} \cdots y_{i_j} \Big) \\
&= \frac{1}{j!} \sum_{i_1,\dots,i_j=1}^{n} \sum_{h=1}^{j} \frac{\partial^j f}{\partial x_{i_1} \cdots \partial x_{i_j}} y_{i_1} \cdots y_{i_{h-1}} \frac{\partial y_{i_h}}{\partial y_k} y_{i_{h+1}} \cdots y_{i_j} \\
&= \frac{1}{j!} \sum_{\substack{i_1,\dots,i_j=1 \\ i_h=k}}^{n} \frac{\partial^{j-1}(\partial f/\partial x_k)}{\partial x_{i_1} \cdots \partial x_{i_{h-1}} \partial x_{i_{h+1}} \cdots \partial x_{i_j}} y_{i_1} \cdots y_{i_{h-1}} y_{i_{h+1}} \cdots y_{i_j} \\
&= \frac{1}{(j-1)!}\Big(\sum_{i_1,\dots,i_{j-1}=1}^{n} \frac{\partial^{j-1}(\partial f/\partial x_k)}{\partial x_{i_1} \cdots \partial x_{i_{j-1}}} y_{i_1} \cdots y_{i_{j-1}} \Big) \\
&= T_{j-1}\Big(\frac{\partial f}{\partial x_k}\Big)(y).
\end{aligned}$$

Here the first and last equality are just the definition. The second applies the rule for differentiating sums, then the rule for differentiating products, generalized to any number of factors. (The general idea of the inductive proof should be clear after you've seen the computation $(fgh)' = f'(gh) + f(gh)' = f'(gh) + f(g'h + gh') = f'gh + fg'h + fgh'$.) Note that $\frac{\partial y_{i_h}}{\partial y_k}$ is a constant function with value 1 or 0 according to whether $k = i_h$. The third equality first removes those terms with $k \neq i_h$, then uses Clairaut's theorem to interchange the order of differentiation, putting differentiation with respect to $x_{i_h} = x_k$ at the beginning. The fourth equality applies the fact that the function

$$(i_1, \ldots, i_j, h) \mapsto (i_1, \ldots, i_{h-1}, i_{h+1}, \ldots, i_j)$$

is a j-to-one (so that $1/j!$ is replaced with $1/(j-1)!$) map from

$$\{\, (i_1, \ldots, i_j, h) : 1 \leq i_1, \ldots, i_j \leq n,\ 1 \leq h \leq j, \text{ and } i_h = k \,\}$$

to

$$\{\, (i_1, \ldots, i_{j-1}) : 1 \leq i_1, \ldots, i_{j-1} \leq n \,\}.$$

□

We can apply the last result repeatedly to compute higher order partial derivatives of the Taylor's series.

Lemma 6.28. *If Clairaut's theorem holds for the field k, then for all $j = 1, \ldots, r$ and $1 \leq k_1, \ldots, k_j \leq n$,*

$$\frac{\partial^j T^r(f)}{\partial y_{k_1} \cdots \partial y_{k_j}} = T^{r-j}\Big(\frac{\partial f}{\partial x_{k_1} \cdots \partial x_{k_j}}\Big).$$

Proof. The last result is the case $j = 1$, so, by induction, we may assume that the result has already been established with j replaced by any $h = 1, \ldots, j-1$, and also by $j - h$, and these combine to give us what we want:

$$\begin{aligned}\frac{\partial^j T^r(f)}{\partial y_{k_1} \cdots \partial y_{k_j}} &= \frac{\partial^{j-h}}{\partial y_{k_{h+1}} \cdots \partial y_{k_j}}\Big[T^{r-h}\Big(\frac{\partial^h f}{\partial x_{k_1} \cdots \partial x_{k_h}}\Big)\Big] \\ &= T^{r-j}\Big(\frac{\partial^j f}{\partial x_{k_1} \cdots \partial x_{k_j}}\Big).\end{aligned}$$

□

In its overall outline the proof of the general version of Taylor's theorem follows the argument give for the univariate case, but naturally everything is a bit more complicated.

Theorem 6.29 (Multivariate Taylor's Theorem)**.** *Suppose that $U \subset \mathbb{R}^n$ is open, $r \geq 0$ is an integer, and $f : U \to \mathbb{R}$ is C^r. Then for each $\overline{x} \in U$ and $\varepsilon > 0$ there is $\delta > 0$ such that*

$$\big|f(\overline{x}+y) - T^r(f;\overline{x})(y)\big| \leq \varepsilon\|y\|^r$$

whenever $\|y\| < \delta$.

When $r = 0$ the assertion is simply that f is continuous. This is already present in the assumptions, but it will be a nice place to start an induction. In the case $r = 1$ the claim is just that $T^1(f)$ is an asymptotically accurate approximation of f near $\overline{x}$: that is, $Df(\overline{x})$ is defined and $x \mapsto T^1(f;\overline{x})(x - \overline{x}) = f(\overline{x}) + Df(\overline{x})(x - \overline{x})$ is the associated affine approximation of f.

Proof. Fix $\overline{x}$ and ε. As in the univariate case, we simplify the calculations by introducing a new function $g : \{\, x - \overline{x} : x \in U \,\} \to \mathbb{R}$ given by

$$g(y) := f(\overline{x}+y) - T^r(f)(y).$$

Our goal is to find $\delta > 0$ such that $|g(y)| \leq \varepsilon\|y\|^n$ whenever $\|y\| < \delta$. For all $j = 1,\ldots,r$ and $1 \leq k_1,\ldots,k_j \leq n$ the last result gives

$$\frac{\partial^j T^r(f)}{\partial y_{k_1}\cdots\partial y_{k_j}}(0) = T^{r-j}\Big(\frac{\partial f}{\partial x_{k_1}\cdots\partial x_{k_j}}\Big)(0) = \frac{\partial f}{\partial x_{k_1}\cdots\partial x_{k_j}}(\overline{x})$$

and thus

$$\frac{\partial^j g}{\partial y_{k_1}\cdots\partial y_{k_j}}(0) = 0.$$

In this sense the remainder of the argument can be understood as a matter of demonstrating that the theorem holds when all the partials of f up to order r vanish.

The case $r = 0$ follows from the continuity of g, so, by the principle of induction, we may suppose the claim has already been demonstrated with $r-1$ in place of r. The hypotheses of that case are satisfied by $\frac{\partial g}{\partial y_1},\ldots,\frac{\partial g}{\partial y_n}$, so there is $\delta > 0$ such that

$$\Big|\frac{\partial g}{\partial y_i}(y)\Big| \leq (\varepsilon/n)\|y\|^{r-1}$$

for all i whenever $\|y\| < \delta$. Fix such a y, and define $h : [0,1] \to \mathbb{R}$ by setting $h(t) := g(ty)$. The mean value theorem implies that there is some t between 0 and 1 such that

$$g(y) = g(y) - g(0) = h(1) - h(0) = h'(t).$$

The chain rule, and the rules for differentiating sums and products, give

$$h'(t) = \sum_{i=1}^{n} \frac{\partial g}{\partial y_i}(ty)y_i.$$

It is not hard to see that if the theorem holds for one norm on $\mathbb{R}^n$, then it holds for any other equivalent norm, and since all norms are equivalent we can specify that the norm is $\|y\| = \sqrt{y_1^2 + \cdots + y_n^2}$. For this norm we have $|y_i| \le \|y\|$ for all i, so

$$|g(y)| = |h'(t)| \le \sum_{i=1}^{n} \left|\frac{\partial g}{\partial y_i}(ty)\right| \cdot |y_i| \le \sum_{i=1}^{n} (\varepsilon/n)\|ty\|^{r-1}|y_i| \le \varepsilon\|y\|^r.$$

□

6.11 Derivatives of Sequences of Functions

If a function is defined by a power series, and is consequently the limit of a sequence of polynomials, one would naturally guess that the derivative of the limiting function is the limit of the sequence of derivatives of the polynomials. This is true, as we will see in the next chapter, but there are certain subtleties. It can happen that a sequence $\{f_k\}$ of C^1 functions converges to a C^1 function f, but the sequence of derivatives $\{f_k'\}$ doesn't converge. Properly speaking, we don't yet have the tools to complete the analysis, but readers who have any familiarity with the trigonometric functions will be able to recognize that if $f_k : \mathbb{R} \to \mathbb{R}$ is the function $f_k(t) = \sin(kt)/k$, then $\{f_k\}$ converges uniformly to the constant zero function, but $f_k'(t) = \cos(kt)$ oscillates in the interval $[-1,1]$, and in fact the sequence $\{f_k'\}$ is not even pointwise convergent.

The key result for this topic has quite strong hypotheses: we are given a sequence of C^1 functions that converges pointwise and whose sequence of derivative functions converges uniformly on compacta. (As stated, these hypotheses are, in a sense, artificially weak, since they imply that the sequence itself converges uniformly on compacta, but we won't bother to prove this.)

The only additional information we obtain is that the limit of the sequence of derivatives is the derivative of the limit of the sequence of functions. However, this conclusion is just what we need to understand the derivatives of functions defined by power series like $\exp(\cdot)$, $\sin(\cdot)$, and $\cos(\cdot)$.

We begin with the univariate case for $k = \mathbb{R}$ because we can apply Rolle's theorem. Actually, the proof in this case already contains the essential ideas, although perhaps not in a transparent form, and you might find it completely convincing and at the same time not much of an explanation. To get a better sense of the analysis it may help to just draw some figures illustrating the situation described in the hypotheses when f is a constant function.

Lemma 6.30. *Suppose $U \subset \mathbb{R}$ is open, $\{f_k\}$ is a sequence of functions from U to $\mathbb{R}$ that converges pointwise to f, f'_k is defined and continuous for each k, and $\{f'_k\}$ converges uniformly on compacta to $g : U \to \mathbb{R}$. Then f is C^1 with $f' = g$.*

Proof. Proposition 3.51 implies that g is continuous on some neighborhood of each of its points, which (Proposition 3.21) is the same as simply being continuous, so if we can show that $f' = g$, then f' is continuous and f is C^1. Fixing $\bar{t} \in U$, it suffices to show that $f'(\bar{t}) = g(\bar{t})$.

Fix $\varepsilon > 0$. Since U is open and g is continuous we can choose $\delta > 0$ such that $[\bar{t} - \delta, \bar{t} + \delta] \subset U$ and $\big|g(s) - g(\bar{t})\big| < \frac{1}{6}\varepsilon$ whenever $|s - \bar{t}| \leq \delta$. For sufficiently large k we have $|f'_k(s) - g(s)| < \frac{1}{6}\varepsilon$ for all $s \in [\bar{t} - \delta, \bar{t} + \delta]$ because $f'_k \to g$ uniformly on compacta, so that

$$\big|f'_k(s) - g(\bar{t})\big| \leq \big|f'_k(s) - g(s)\big| + \big|g(s) - g(\bar{t})\big| < \tfrac{1}{3}\varepsilon. \qquad (*)$$

Fix a particular $t \in [\bar{t} - \delta, \bar{t} + \delta]$ with $t \neq \bar{t}$. (The inequality we are trying to establish holds automatically when $t = \bar{t}$.) For sufficiently large k we have

$$\big|f(t) - f_k(t)\big| < \tfrac{1}{3}\varepsilon|t - \bar{t}| \quad \text{and} \quad \big|f(\bar{t}) - f_k(\bar{t})\big| < \tfrac{1}{3}\varepsilon|t - \bar{t}|$$

because $f_k \to f$ pointwise. (A key point here is that we are allowed to choose k *after* we have committed to a particular t.) For any k the mean value theorem gives a number s_k strictly between $\bar{t}$ and t such that $f_k(t) - f_k(\bar{t}) = f'_k(s_k)(t - \bar{t})$, so $(*)$ gives

$$\big|f_k(t) - [f_k(\bar{t}) + g(\bar{t})(t - \bar{t})]\big| = \big|(f'_k(s_k) - g(\bar{t}))(t - \bar{t})\big| \leq \tfrac{1}{3}\varepsilon|t - \bar{t}|.$$

For large k the inequalities above combine to give

$$\big|f(t) - [f(\bar{t}) + g(\bar{t})(t - \bar{t})]\big| \leq \big|f(t) - f_k(t)\big| + \big|f(\bar{t}) - f_k(\bar{t})\big|$$

$$+\big|f_k(t) - [f_k(\overline{t}) + g(\overline{t})(t - \overline{t})]\big| < \varepsilon|t - \overline{t}|.$$

□

Extending this result to the multivariate case is easy: we restrict to a one dimensional subset of the domain, then apply the univariate case. In the next chapter we'll extend this to a sequence of complex valued functions defined on an open subset of $\mathbb{C}^n$.

Theorem 6.31. *Suppose $U \subset \mathbb{R}^n$ is open, $\{f_k\}$ is a sequence of functions from U to $\mathbb{R}$ that converges pointwise to f, $\frac{\partial f_k}{\partial x_j}$ is defined and continuous for each k and each $j = 1, \ldots, n$, and each $\{\frac{\partial f_k}{\partial x_j}\}$ converges uniformly on compacta to a function $g_j : U \to \mathbb{R}$. Then f is C^1 with*

$$\frac{\partial f}{\partial x_1} = g_1, \ldots, \frac{\partial f}{\partial x_n} = g_n.$$

Proof. Proposition 3.51 implies that each g_j is continuous, so f is necessarily C^1 if we show that $\frac{\partial f}{\partial x_j} = g_j$ for all j. Fix a particular $\overline{x} \in U$ and $j = 1, \ldots, n$. Let $\mathbf{e}_j$ be the j^{th} standard unit basis vector of $\mathbb{R}^n$, and let

$$U_j := \{\, s \in \mathbb{R} : \overline{x} + s\mathbf{e}_j \in U \,\}.$$

Let $\varphi_k, \varphi, \gamma : U_j \to \mathbb{R}$ be the functions

$$\varphi_k(s) := f_k(\overline{x} + s\mathbf{e}_j), \quad \varphi(s) := f(\overline{x} + s\mathbf{e}_j), \quad \gamma(s) := g_j(\overline{x} + s\mathbf{e}_j).$$

Then $\{\varphi_k\}$ converges pointwise to φ. In addition, $\varphi_k'(s) = \frac{\partial f}{\partial x_j}(\overline{x} + s\mathbf{e}_j)$, so each φ_k' is defined and continuous, and $\{\varphi_k'\}$ converges uniformly on compacta to γ. The last result now implies that $\varphi'(0) = \gamma(0)$, which is the same as $\frac{\partial f}{\partial x_j}(\overline{x}) = g_j(\overline{x})$. □

Chapter 7

Complex Differentiation

The last chapter covered the main facts concerning differentiation of real valued functions, so we now turn to the special properties of differentiation for a complex valued function $f : U \to \mathbb{C}$ where $U \subset \mathbb{C}^n$ is open. These functions are special in several ways, and for this reason (and also because complex numbers occur less frequently in scientific modelling than real numbers) they are of somewhat less interest to science as a whole, but they are extremely important in pure mathematics. The study of their properties is one of the main entry points to "higher" mathematics.

We'll always assume that f is differentiable at each point of U, and all our arguments will be based on simple facts about differentiation. But it turns out that the slightly stronger assumption that f is C^1 implies quite a bit more, and there is consequently a certain potential for terminological confusion. So, we first explain how things can become muddled and what we will do about it.

For those who have achieved a certain level of education a function is **analytic** if each point in the domain has a power series centered at that point that agrees with the function in a neighborhood of the point. This notion is meaningful over the field of complex numbers, in the sense that the coefficients in the power series are complex numbers, and the variables are understood as taking complex values, but it is also meaningful over the real field. Thus one tends to speak of **complex analytic** functions and **real analytic** functions.

We've already encountered the notion of a C^r real valued function on an open subset of $\mathbb{R}^n$. The definition for the complex field is formally the same: all the partial derivatives up to order r (in the sense of partial differentiation given by the complex field) exist everywhere and are continuous. An open subset of $\mathbb{C}^n$ can be regarded as an open subset of $\mathbb{R}^{2n}$, and in this sense one may regard a complex valued function on an open subset of $\mathbb{C}^n$ as a pair of real valued functions defined on an open subset of $\mathbb{R}^{2n}$. If the complex function is C^r then, as we will see shortly, the associated real valued functions are C^r in the real sense, but the converse is very far from being true. One of the most remarkable facts of mathematics is that our function f is C^1 in the complex sense if and only if it is complex analytic, and if this is the case then each partial derivative function is also C^1 so that, by induction, f is C^∞. That is, *if f is C^1, then it is both C^∞ and complex analytic.* For those who have already mastered the basics, this happy state of affairs is described by saying that f is **holomorphic**.

The problem is that although, in the end, there is only one type of function in question, there are various states of knowledge concerning it, and any proof of a property of such functions begins in one such state and terminates in another. Some complex analysis texts use the word 'analytic' to describe a function from an open subset of $\mathbb{C}$ to $\mathbb{C}$ that is differentiable everywhere. This approach may make sense in the context of a large body of material focused on such functions, but I don't think it would work very well here. We'll sometimes use the word 'holomorphic' to describe a function satisfying any set of conditions implying all the properties described above, but we'll be careful to explain clearly just which properties are being applied in each argument. For the time being this means that we are assuming only that the given f is differentiable everywhere.

7.1 The Cauchy-Riemann Equations

We begin with the case $n = 1$, so f is a complex valued function defined on an open subset of $\mathbb{C}$. Under Argand's identification of $\mathbb{C}$ with the plane $\mathbb{R}^2$ we can also think of f as a map from an open subset of the plane to the plane. We wish to maintain a clear distinction between the two ways of looking at f, so we describe the situation as follows. Let

$$\iota : \mathbb{R}^2 \to \mathbb{C} \quad \text{be the function} \quad \iota(x, y) := x + iy.$$

Setting $\tilde{U} := \iota^{-1}(U)$, the function associated with f is

$$\tilde{f} = (u, v) := \iota^{-1} \circ f \circ \iota|_{\tilde{U}} : \tilde{U} \to \mathbb{R}^2.$$

Assuming that f is differentiable at a point z, we will aim at an understanding of this condition in terms of properties of $\tilde{f}$, u, and v at $(x, y) := \iota^{-1}(z)$.

The real and imaginary parts of the derivative of f at z can be expressed in terms of u and v:

$$\begin{aligned} f'(z) &= \lim_{\Delta x\to 0}\frac{f(z+\Delta x)-f(z)}{\Delta x} \\ &= \lim_{\Delta x\to 0}\frac{u(x+\Delta x,y)-u(x,y)}{\Delta x}+i\frac{v(x+\Delta x,y)-v(x,y)}{\Delta x} \\ &= \frac{\partial u}{\partial x}(x,y)+i\frac{\partial v}{\partial x}(x,y). \end{aligned}$$

But it is also the case that

$$\begin{aligned} f'(z) &= \lim_{\Delta y\to 0}\frac{f(z+i\Delta y)-f(z)}{i\Delta y} \\ &= \lim_{\Delta y\to 0}\frac{u(x,y+\Delta y)-u(x,y)}{i\Delta y}+i\frac{v(x,y+\Delta y)-v(x,y)}{i\Delta y} \\ &= -i\frac{\partial u}{\partial y}(x,y)+\frac{\partial v}{\partial y}(x,y). \end{aligned}$$

Equating the real and imaginary parts of these expressions for $f'(z)$ gives

$$\frac{\partial u}{\partial x}(x,y)=\frac{\partial v}{\partial y}(x,y) \quad\text{and}\quad \frac{\partial v}{\partial x}(x,y)=-\frac{\partial u}{\partial y}(x,y).$$

These are called the **Cauchy-Riemann** equations, in honor of the work of these two men in developing the theory of differentiable functions of a complex variable, even though they had appeared in works of d'Alembert in 1752 and were subsequently studied by Euler. They are very famous and very important. One simple and rather unexpected consequence is that complex conjugation $x+iy\mapsto x-iy$ is *not* differentiable in the complex sense because the first of the two equations does not hold.

It turns out that the Cauchy-Riemann equations are not just necessary conditions of complex differentiability but also (for functions that are C^1 in the real sense) sufficient.

Lemma 7.1. *Assume that u and v are C^1. Fixing $z=x+iy\in U$, let*

$$u_x:=\frac{\partial u}{\partial x}(x,y),\ u_y:=\frac{\partial u}{\partial y}(x,y),\ v_x:=\frac{\partial v}{\partial x}(x,y),\ v_y:=\frac{\partial v}{\partial y}(x,y).$$

If $u_x=v_y$ and $u_y=-v_x$, then $f'(z)$ is defined and

$$f'(z)=u_x+iv_x=v_y-iu_y.$$

Proof. For ξ, ψ such that $(x+\xi, y+\psi) \in \tilde{U}$ let

$$A(\xi,\psi) := u(x+\xi, y+\psi) - [u(x,y) + u_x\xi + u_y\psi]$$

and

$$B(\xi,\psi) := v(x+\xi, y+\psi) - [v(x,y) + v_x\xi + v_y\psi].$$

Taylor's theorem implies that there is $\delta > 0$ such that

$$\big|A(\xi,\psi)\big| \le \frac{\varepsilon}{\sqrt{2}}(\xi^2+\psi^2)^{1/2} \quad \text{and} \quad \big|B(\xi,\psi)\big| \le \frac{\varepsilon}{\sqrt{2}}(\xi^2+\psi^2)^{1/2}$$

whenever $(\xi^2+\psi^2)^{1/2} < \delta$. (Here we are using our freedom to work with whichever norm on $\mathbb{R}^2$ happens to be convenient.)

If $\zeta := \xi + i\psi$, then the equations $u_x = v_y$ and $u_y = -v_x$ give

$$\begin{aligned}(u_x + iv_x)\zeta &= (u_x\xi - v_x\psi) + i(v_x\xi + u_x\psi)\\ &= (u_x\xi + u_y\psi) + i(v_x\xi + v_y\psi).\end{aligned}$$

Since $f = u + iv$, these equations combine to give

$$f(z+\zeta) - [f(z) + (u_x + iv_x)\zeta] = A(\xi,\psi) + iB(\xi,\psi).$$

We conclude that

$$\big|f(z+\zeta) - [f(z) + (u_x+iv_x)\zeta]\big| = \sqrt{A^2+B^2} \le \varepsilon(\xi^2+\psi^2)^{1/2} = \varepsilon|\zeta|$$

whenever $\zeta = \xi + i\psi$ with $z + \zeta \in U$ and $|\zeta| < \delta$, which is just what we need. □

The extension of these results to the multivariate case is straightforward. Let $\tilde{U} \subset \mathbb{R}^{2n}$ be open, and let $u, v : \tilde{U} \to \mathbb{R}$ be functions. For $(x,y) = (x_1,\ldots,x_n,y_1,\ldots,y_n) \in \mathbb{R}^{2n}$ let

$$\iota(x,y) := x + iy = (x_1 + iy_1, \ldots, x_n + iy_n).$$

Let $U := \iota(\tilde{U})$, and let $f : U \to \mathbb{C}$ be the function

$$f(x+iy) := u(x,y) + iv(x,y).$$

The following proof (which you might want to skim quickly without studying in detail) simply applies our findings in the univariate case to the restrictions of f, u, and v to a one dimensional (in the complex sense) subset of U and the corresponding subset of $\tilde{U}$.

Theorem 7.2. *Assume that u and v are C^1. For any $z = x + iy \in U$ and $j = 1, \ldots, n$, if $\frac{\partial f}{\partial z_j}(z)$ is defined, then*

$$\frac{\partial f}{\partial z_j}(z) = \frac{\partial u}{\partial x_j}(x,y) + i\frac{\partial v}{\partial x_j}(x,y) = \frac{\partial v}{\partial y_j}(x,y) - i\frac{\partial u}{\partial y_j}(x,y). \qquad (*)$$

Conversely, if

$$\frac{\partial u}{\partial x_j}(x,y) = \frac{\partial v}{\partial y_j}(x,y) \quad \textit{and} \quad \frac{\partial u}{\partial y_j}(x,y) = -\frac{\partial v}{\partial x_j}(x,y),$$

then $\frac{\partial f}{\partial z_j}(z)$ is defined and $(*)$ holds.

Proof. Let $\mathbf{e}_j \in \mathbb{R}^n$ and $\mathbf{f}_j \in \mathbb{C}^n$ be the respective j^{th} standard unit basis vectors. (That is, $\mathbf{e}_j$ and $\mathbf{f}_j$ are both $(0, \ldots, 1, \ldots, 0)$ where the 1 is the j^{th} component.) Let

$$V := \{\, w \in \mathbb{C} : z + w\mathbf{f}_j \in U \,\},$$

and let $\varphi : V \to \mathbb{C}$ be the function $\varphi(w) := f(z + w\mathbf{f}_j)$. By virtue of the definition of partial derivatives, for all $w \in V$ the partial $\frac{\partial f}{\partial z_j}(z + w\mathbf{f}_j)$ is defined if and only if $\varphi'(w)$ is defined, in which case they are equal.

Let

$$\tilde{V} := \{\, (s,t) \in \mathbb{R}^2 : s + it \in V \,\},$$

and let $\mu, \nu : \tilde{V} \to \mathbb{R}$ be the functions

$$\mu(s,t) := u(x + s\mathbf{e}_j, y + t\mathbf{e}_j) \quad \text{and} \quad \nu(s,t) := v(x + s\mathbf{e}_j, y + t\mathbf{e}_j).$$

Then $\varphi(s + it) = \mu(x,t) + i\nu(s,t)$ for all $(s,t) \in \tilde{V}$. Again, the definition of partial derivatives gives

$$\frac{\partial u}{\partial x_j}(x + s\mathbf{e}_j, y + t\mathbf{e}_j) := \frac{\partial \mu}{\partial s}(s,t), \quad \frac{\partial v}{\partial x_j}(x + s\mathbf{e}_j, y + t\mathbf{e}_j) := \frac{\partial \nu}{\partial s}(s,t),$$
$$\frac{\partial u}{\partial y_j}(x + s\mathbf{e}_j, y + t\mathbf{e}_j) := \frac{\partial \mu}{\partial t}(s,t), \quad \frac{\partial v}{\partial y_j}(x + s\mathbf{e}_j, y + t\mathbf{e}_j) := \frac{\partial \nu}{\partial t}(s,t),$$

for all $(s,t) \in \tilde{V}$. The results for the univariate case now state that $\varphi'(0)$ is defined if and only if

$$\frac{\partial \mu}{\partial s}(s,t) = \frac{\partial \nu}{\partial t}(s,t) \quad \text{and} \quad \frac{\partial \mu}{\partial t}(s,t) = -\frac{\partial \nu}{\partial s}(s,t),$$

in which case

$$\varphi'(0) = \frac{\partial \mu}{\partial s}(0,0) + i\frac{\partial \nu}{\partial s}(0,0) = \frac{\partial \nu}{\partial t}(0,0) - i\frac{\partial \mu}{\partial t}(0,0),$$

which is exactly what we want. □

This section's final result considers a sequence of C^1 functions on U that converges uniformly on compacta, and whose associated sequence of derivative functions also converges uniformly on compacta. Extending Theorem 6.31 to the complex case, we will show that the derivative of the limiting function is the limit of the sequence of derivatives. The idea of the proof is to first apply the real version of the result to the real and complex parts of the sequence of functions, then observe that, by continuity, the Cauchy-Riemann equations hold in the limit.

Theorem 7.3. *Suppose $U \subset \mathbb{C}^n$ is open, $\{f_k\}$ is a sequence of functions from U to $\mathbb{C}$ that converges uniformly on compacta to f, $\frac{\partial f_k}{\partial z_j}$ is defined and continuous for each k and each $j = 1, \dots, n$, and each $\{\frac{\partial f_k}{\partial z_j}\}$ converges uniformly on compacta to a function $g_j : U \to \mathbb{C}$. Then f is C^1 with*

$$\frac{\partial f}{\partial z_1} = g_1, \dots, \frac{\partial f}{\partial z_n} = g_n.$$

Proof. Fix a particular j between 1 and n. As above let $f_k(z) = u_k(x, y) + iv_k(x, y)$ and $f(z) = u(x, y) + iv(x, y)$. Then $\{u_k\}$ and $\{v_k\}$ converge uniformly on compacta to u and v. For each k the Cauchy-Riemann equations are satisfied:

$$\frac{\partial f_k}{\partial z_j}(z) = \frac{\partial u_k}{\partial x_j}(x, y) + i\frac{\partial v_k}{\partial x_j}(x, y) = \frac{\partial v_k}{\partial y_j}(x, y) - i\frac{\partial u_k}{\partial y_j}(x, y).$$

Let $g_j(z) = s_j(x, y) + it_j(x, y)$. Then $\{\partial u_k/\partial x_j\}$ and $\{\partial v_k/\partial y_j\}$ converge uniformly on compacta to s_j and $\{\partial v_k/\partial x_j\}$ and $\{-\partial u_k/\partial y_j\}$ converge uniformly on compacta to t_j. In view of all this the real version of the result (Theorem 6.31) implies that

$$\frac{\partial u}{\partial x_j}(x, y) = s_j(x, y) = \frac{\partial v}{\partial y_j}(x, y), \quad \frac{\partial v}{\partial x_j}(x, y) = t_j(x, y) = -\frac{\partial u}{\partial y_j}(x, y).$$

In particular, the Cauchy-Riemann equations are satisfied, so Theorem 7.2 tells us that $\frac{\partial f}{\partial z_j} = \frac{\partial u}{\partial x_j} + i\frac{\partial v}{\partial x_j} = s_j + it_j = g_j$, as desired. □

7.2 Conformal Mappings

Let's go back to the univariate framework, so $U \subset \mathbb{C}$ is open, $\tilde{U} = \iota^{-1}(U)$, $\tilde{f} = (u, v) : \tilde{U} \to \mathbb{R}^2$ is C^1, and $f := \iota \circ (u, v) \circ \iota^{-1} : U \to \mathbb{C}$. We would like a better geometric understanding of complex differentiation. The basic idea of the derivative is that near a point $z \in U$, $f(z + w)$ is well approximated by $f(z) + f'(z)w$. In Section 3.9 we showed that any complex number can be written in the form $re^{i\theta}$ for some $r \geq 0$ and some $\theta \in \mathbb{R}$, which may be taken in the interval $[0, 2\pi)$ if we like. If $f'(z) = re^{i\theta}$ and $w = se^{i\phi}$, then $f'(z)w = rse^{i(\theta+\phi)}$. Multiplication induces an action of the group $\mathbb{C}^*$ (with multiplication as the group operation) on $\mathbb{C}$; in terms of the geometry of the Argand plane the action of $re^{i\theta}$ is counterclockwise rotation through the angle θ, followed by compression or dilation by the factor r.

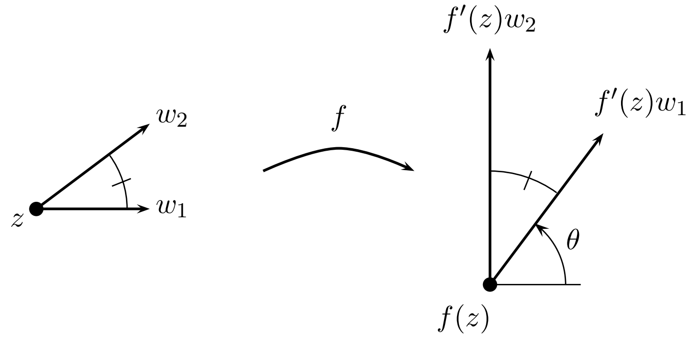


Figure 7.1

Suppose now that $f'(z) = a + ib$ and $w = s + it$. Then $f'(z)w = (as - bt) + i(at + bs)$, so we may also think of the action of $f'(z)$ as a linear transformation from $\mathbb{R}^2$ to itself with matrix

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

Note that the determinant $a^2 + b^2$ of this matrix is nonnegative.

We would like to achieve a clear understanding of the relationship between this algebraic description of $f'(z)$ and the geometric properties identified above. To this end, we will now lay out the basic properties of linear transformations that preserve distance. Suppose $\ell : \mathbb{R}^n \to \mathbb{R}^n$ is such a linear transformation, so that $\|\ell(v) - \ell(w)\| = \|v - w\|$ for all $v, w \in \mathbb{R}^n$. Since $\ell(v) - \ell(w) = \ell(v - w)$, this is the case if and only if $\|\ell(v)\| = \|v\|$ for any $v \in \mathbb{R}^n$. Substituting the definition of the norm, then squaring both sides of this equation, gives

$$\langle \ell(v), \ell(v) \rangle = \langle v, v \rangle.$$

In order to get a better idea of what this means we replace v with $v - w$ in this equation, then expand both sides using basic properties of the inner product, obtaining

$$\big\langle \ell(v), \ell(v)\big\rangle - 2\big\langle \ell(v), \ell(w)\big\rangle + \big\langle \ell(w), \ell(w)\big\rangle = \big\langle v, v\big\rangle - 2\big\langle v, w\big\rangle + \big\langle w, w\big\rangle.$$

The leftmost terms on the two sides of this equation are equal, as are the rightmost terms. Subtracting these and dividing by -2 yields

$$\big\langle \ell(v), \ell(w)\big\rangle = \big\langle v, w\big\rangle.$$

If this is true for all v and w, then it holds when $w = v$, of course, so we have shown that ℓ preserves distances if and only if this last equation holds for all $v, w \in \mathbb{R}^n$. We say that ℓ is an **orthogonal transformation** if it has these properties.

There is a remarkably simple algebraic characterization of this condition. Recall that the inner product $\big\langle v, w\big\rangle$ can be written as the matrix product $v^T w$ if we ignore the distinction between the resulting 1×1 matrix and its entry. In general, if A and B are conformable matrices, then $(AB)^T = B^T A^T$: the (j, i)-entry of $(AB)^T$ is the (i, j)-entry of AB, namely the inner product of the i^{th} row of A and the j^{th} column of B, and of course this is the inner product of j^{th} row of B^T and the i^{th} column of A^T. In particular, $(Av)^T = v^T A^T$, so if A is the matrix of an orthogonal transformation ℓ, then

$$v^T (A^T A) w = (Av)^T (Aw) = \big\langle \ell(v), \ell(w)\big\rangle = \big\langle v, w\big\rangle = v^T w$$

for all $v, w \in \mathbb{R}^n$. Letting v and w vary over the standard unit basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ shows that $A^T A$ is the $n \times n$ identity matrix. The converse is true as well: if $A^T A$ is the identity matrix, then

$$\big\langle \ell(v), \ell(w)\big\rangle = (Av)^T (Aw) = v^T A^T A w = v^T w = \big\langle v, w\big\rangle$$

for all $v, w \in \mathbb{R}^n$. We have shown that:

Proposition 7.4. *An $n \times n$ matrix A with real entries is the matrix of an orthogonal transformation if and only if it is invertible with $A^{-1} = A^T$.*

What does this mean concretely for a 2×2 matrix? Suppose that

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix} = \begin{bmatrix} a^2 + b^2 & ac + bd \\ ac + bd & c^2 + d^2 \end{bmatrix}.$$

Then $ac = -bd$ because the off diagonal entries are zero, so $a^2 c^2 = b^2 d^2$ and

$$a^2 = a^2(c^2 + d^2) = (b^2 + a^2) d^2 = d^2,$$

from which it follows that

$$b^2 = 1 - a^2 = 1 - d^2 = c^2.$$

Therefore $c = \pm b$, $d = \pm a$, and $ac = -bd$, so the matrix of an orthogonal transformation $\ell : \mathbb{R}^2 \to \mathbb{R}^2$ is either

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} a & b \\ b & -a \end{bmatrix}$$

for numbers a and b such that $a^2 + b^2 = 1$. (It's easy to check that, in fact, the product of either of these matrices with its transpose is the identity.) Note that the determinant of the first matrix is positive and the determinant of the second matrix is negative.

Definition 7.5. *If $\tilde{U} \subset \mathbb{R}^2$ is open, a function $\tilde{f} : \tilde{U} \to \mathbb{R}^2$ is a **conformal mapping** if it is C^1 and, for each $(x, y) \in \tilde{U}$, $D\tilde{f}(x, y)$ is a nonnegative multiple of an orthogonal transformation with positive determinant.*

We have seen that if $f : U \to \mathbb{C}$ is C^1, then $\tilde{f} : \tilde{U} \to \mathbb{R}^2$ is conformal. But we have also developed the tools that prove the converse. If $\tilde{f}$ is conformal, the analysis above implies that for each $(x, y) \in \tilde{U}$ we have

$$\begin{pmatrix} \frac{\partial u}{\partial x}(x, y) & \frac{\partial u}{\partial y}(x, y) \\ \frac{\partial v}{\partial x}(x, y) & \frac{\partial v}{\partial y}(x, y) \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

for some $a, b \in \mathbb{R}$, which means that the Cauchy-Riemann equations hold everywhere, implying that f is C^1 in the complex sense. Summarizing:

Theorem 7.6. *Suppose $U \subset \mathbb{C}$ is open, $f : U \to \mathbb{C}$, $\tilde{U} = \iota^{-1}(U)$, and $\tilde{f} := \iota^{-1} \circ f \circ \iota|_{\tilde{U}}$. Then f is differentiable if and only if $\tilde{f}$ is conformal.*

7.3 Complex Clairaut and Taylor

Our next task is to show that the main results concerning higher order partial derivatives in the real case—mixed partials do not depend on the order of differentiation, and Taylor's series approximations are asymptotically accurate—extend to the complex case. We begin with a precise statement of the desired results:

Theorem 7.7. *Suppose that $U \subset \mathbb{C}^n$ is open, $f : U \to \mathbb{C}$ is C^r for some $r \geq 1$, and $\overline{z} \in U$. Then for each $\varepsilon > 0$ there is $\delta > 0$ such that whenever $\|w\| < \delta$,*

$$\left| f(\overline{z} + w) - T^r(f; \overline{z})(w) \right| \leq \varepsilon \|w\|^r.$$

If, for some $1 \le j,k \le n$ with $j \neq k$, the second order partial derivatives $\frac{\partial^2 f}{\partial z_j \partial z_k}$ and $\frac{\partial^2 f}{\partial z_k \partial z_j}$ are defined everywhere in U and continuous at $\overline{z} \in U$ (of course this is automatic when $r \ge 2$) then

$$\frac{\partial^2 f}{\partial z_j \partial z_k}(\overline{z}) = \frac{\partial^2 f}{\partial z_k \partial z_j}(\overline{z}).$$

The rest of the section is devoted to the proofs of these claims. We will use the same framework as before: $\iota : \mathbb{R}^{2n} \to \mathbb{C}^n$ is the function $\iota(x,y) := x + iy$, $\tilde{U} := \iota^{-1}(U)$, and $f = (u \circ \iota^{-1}) + i(v \circ \iota^{-1})$. Explicit reference to ι will often be suppressed, in the sense that in equations involving x, y, and z, or s, t, and w, it will be understood that $z = \iota(x,y)$ and $w = \iota(s,t)$.

The main tool underlying the arguments in the real case, namely the mean value theorem, does not apply to the complex case. Instead, as in the proof of Theorem 7.3, we will obtain the complex versions of these results by combining the real versions with what we already know about complex differentiation. This style of argument, in which a theorem is used to prove a generalization, variant, or extension of itself, is called a *bootstrap*, after the phrase "pull yourself up by your bootstraps."

For the complex case of Clairaut's theorem this is quite simple: when $1 \le j < k \le n$ we have the following calculation:

$$\begin{aligned}\frac{\partial^2 f}{\partial z_j \partial z_k} &= \frac{\partial}{\partial z_k}\big(\frac{\partial u}{\partial x_j} + i\frac{\partial v}{\partial x_j}\big) = \frac{\partial^2 u}{\partial x_j \partial x_k} + i\frac{\partial^2 v}{\partial x_j \partial x_k} \\ &= \frac{\partial^2 u}{\partial x_k \partial x_j} + i\frac{\partial^2 v}{\partial x_k \partial x_j} = \frac{\partial}{\partial z_j}\big(\frac{\partial u}{\partial x_k} + i\frac{\partial v}{\partial x_k}\big) = \frac{\partial^2 f}{\partial z_k \partial z_j}.\end{aligned}$$

Here the third equality is the real case of Clairaut's theorem, and all other equalities follow from the fact that the (complex) partial derivative with respect to a variable z_h can be computed by taking the (real) partial derivative with respect to z_h's real part.

You should take note of how, in this calculation, we left out the arguments of the partial derivative functions because including them would make a mess, and it is very clear that all the partial derivatives are to be evaluated at $\overline{z}$ or $(\overline{x}, \overline{y})$. This sort of abbreviation is essential to readability in all but the simplest discussions of theories expressed in terms of multivariate calculus, and we will see several more examples below. More generally, the readability of mathematics is enhanced by notation that is spare and light, even if that involves quite a bit in the way of requiring the reader to fill in the (hopefully obvious) missing information.

We now turn to the complex version of Taylor's theorem. The first order partial derivatives of u and v are the real and complex parts of the first order partial derivatives of f, and by applying this principal repeatedly we see that the partials of u and v up to order r are the real and complex parts of various partials of f of the same order. The reason for mentioning this is that it implies that u and v are C^r, so the real version of Taylor's theorem tells us that u and v are well approximated in a neighborhood of $(\overline{x}, \overline{y})$ by their r^{th} order Taylor series. Thus, for any $\varepsilon > 0$ there is $\delta > 0$ such that

$$\left|u(\overline{x}+s,\overline{y}+t) - T^r(u)(s,t)\right| \le \frac{\varepsilon}{\sqrt{2}}\|(s,t)\|^r$$

and

$$\left|v(\overline{x}+s,\overline{y}+t) - T^r(v)(s,t)\right| \le \frac{\varepsilon}{\sqrt{2}}\|(s,t)\|^r$$

whenever $\|(s,t)\| < \delta$, in which case

$$\left|f\big((\overline{x}+s)+i(\overline{y}+t)\big) - T^r(u)(s,t) - iT^r(v)(s,t)\right| \le \varepsilon\|(s,t)\|^r.$$

Since we are using the Euclidean norms, if $w = \iota(s,t)$, then

$$\|w\| = \sqrt{\|w_1\|^2 + \cdots + \|w_n\|^2} = \sqrt{s_1^2 + t_1^2 + \cdots + s_n^2 + t_n^2} = \|(s,t)\|,$$

so it follows immediately[1] that if $\|w\| < \delta$, then

$$\left|f(\overline{z}+w) - T^r(u)(\iota^{-1}(w)) - iT^r(v)(\iota^{-1}(w))\right| \le \varepsilon\|w\|^r.$$

Therefore the complex version of Taylor's theorem holds if $T^r(f) = (T^r(u) + iT^r(v)) \circ \iota^{-1}$, which is the same thing as

$$T^r(f) \circ \iota = T^r(u) + iT^r(v) \in \mathbb{C}[s_1, \dots, s_n, t_1, \dots, t_n]. \qquad (*)$$

The rest of the section is devoted to the proof of this formula.

At first glance it somehow "feels" strange that there is even anything here in need of proof. The "hard" work of taking limits has already been done, and what remains is "just algebra." How could it be that deep or complex? However, a typical term

$$\frac{\partial f}{\partial z_{i_1} \cdots \partial z_{i_j}}(\overline{z})w_{i_1} \cdots w_{i_j}$$

[1] In view of Proposition 6.9, Taylor's theorem is actually valid for any norms, but the proof for arbitrary norms has a few additional details.

of $T^r(f)$ breaks into 2^j terms when expanded in terms of $s_{i_1}, \ldots, s_{i_j}$ and $t_{i_1}, \ldots, t_{i_j}$, and while there are various ways to express this expansion, they are cumbersome and hard to manipulate, or at least so it seems. Instead of making a frontal assault, our argument will employ a rather clever induction over both n and r.

We first explain the general nature of our methods. Consider a polynomial $g = a_m X^m + \cdots + a_1 X + a_0 \in R[X]$ where R may be any commutative ring with unit. We can define a **formal derivative** by applying the rules for differentiating polynomials:

$$Dg := m a_m X^{m-1} + \cdots + 2a_2 X + a_1 \in R[X].$$

The formal derivative will agree with the analytic derivative of a function defined by g whenever there is such a thing, but even if there is no analytic sense in which Dg is the derivative of a function, the purely formal properties of this "derivative" may still be interesting. For example, purely algebraic computations can be used to prove (give it a try!) that, for all $g, h \in R[X]$,

$$D(gh) = Dg \cdot h + g \cdot Dh.$$

It is obvious that $D(g + h) = Dg + Dh$, so computation of formal derivatives is governed by the same rules that we used to compute derivatives of polynomial functions.

This idea extends to partial derivatives: if $g = \sum_{0 \le i+j \le r} a_{i,j} X^i Y^j \in R[X, Y]$, then we can define the formal partial derivative

$$\frac{\partial g}{\partial X} := \sum_{1 \le i+j \le r, 1 \le i} i a_{i,j} X^{i-1} Y^j,$$

and of course $\partial g/\partial Y$ is defined similarly. The specific result we will use is:

Lemma 7.8. *Assume that R is an integral domain, and that the homomorphism from $\mathbb{Z}$ to R taking 1 to 1 is injective. If*

$$g = \sum_{0 \le i+j \le r} a_{i,j} X^i Y^j \quad \textit{and} \quad h = \sum_{0 \le i+j \le r} b_{i,j} X^i Y^j$$

are elements of $R[X, Y]$ with $\partial g/\partial X = \partial h/\partial X$, $\partial g/\partial Y = \partial h/\partial Y$, and $a_{0,0} = b_{0,0}$, then $g = h$.

Proof. We need to show that $a_{i,j} = b_{i,j}$ for all relevant i and j. By assumption $a_{0,0} = b_{0,0}$. Since $\partial g/\partial X = \partial h/\partial X$, we have $i a_{i,j} = i b_{i,j}$ whenever $i \ge 1$. Since i (that is, the image in R of $i \in \mathbb{Z}$) is nonzero and an integral domain has no zero divisors, it follows that $a_{i,j} = b_{i,j}$ whenever $i \ge 1$. Similarly, $a_{i,j} = b_{i,j}$ whenever $j \ge 1$ because $\partial g/\partial Y = \partial h/\partial Y$. □

We will prove $(*)$ by applying this result to $T^r(f)\circ\iota$ and $T^r(u)+iT^r(v)$, where we regard these as polynomials in the variable s_n and t_n with coefficients in $\mathbb{C}[s_1,\dots,s_{n-1},t_1,\dots,t_{n-1}]$. The remainder of the argument verifies the hypotheses of the lemma, namely that

$$\begin{aligned}(T^r(f)\circ\iota)&(s_1,\dots,s_{n-1},0,t_1,\dots,t_{n-1},0) =\\ &= T^r(u)(s_1,\dots,s_{n-1},0,t_1,\dots,t_{n-1},0)\\ &\qquad + iT^r(v)(s_1,\dots,s_{n-1},0,t_1,\dots,t_{n-1},0),\end{aligned}$$

and that

$$\frac{\partial(T^r(f)\circ\iota)}{\partial s_n}=\frac{\partial T^r(u)}{\partial s_n}+i\frac{\partial T^r(v)}{\partial s_n} \text{ and } \frac{\partial(T^r(f)\circ\iota)}{\partial t_n}=\frac{\partial T^r(u)}{\partial t_n}+i\frac{\partial T^r(v)}{\partial t_n}.$$

If either $n=0$ or $r=0$, then $(*)$ holds trivially, so, by the principle of induction, we may assume that $(*)$ has already been established when the pair (n,r) is replaced by either $(n-1,r)$ or $(n,r-1)$. The first hypothesis of the lemma is simply $(*)$ evaluated at

$$(s_1,\dots,s_{n-1},0,t_1,\dots,t_{n-1},0),$$

which is $(*)$ for the pair $(n-1,r)$. (A bit more precisely, it is $(*)$ for the function $\varphi:\{\,\zeta\in\mathbb{C}^{n-1}:(\zeta,\overline{z}_n)\in U\,\}\to\mathbb{C}$ given by $\varphi(\zeta):=f(\zeta,\overline{z}_n)$.)

In the following computation the first equality is from the chain rule, Lemma 6.27 gives the second and final equalities, and the third equality is the case $(n,r-1)$ applied to $\frac{\partial f}{\partial z_n}=\frac{\partial u}{\partial x_n}+i\frac{\partial v}{\partial x_n}$:

$$\begin{aligned}\frac{\partial(T^r(f)\circ\iota)}{\partial s_n}&=\frac{\partial T^r(f)}{\partial w_n}\circ\iota=T^{r-1}(\tfrac{\partial f}{\partial z_n})\circ\iota=T^{r-1}(\tfrac{\partial u}{\partial x_n})+iT^{r-1}(\tfrac{\partial v}{\partial x_n})\\ &=\frac{\partial T^r(u)}{\partial s_n}+i\frac{\partial T^r(v)}{\partial s_n}.\end{aligned}$$

(It is a good idea to slow down here and really make sure you understand each equality!)

The computation for the partial with respect to t_n follows the same pattern, but with one additional complication, namely the fourth equality below applies the Cauchy-Riemann equations:

$$\begin{aligned}\frac{\partial(T^r(f)\circ\iota)}{\partial t_n}&=i\big(\frac{\partial T^r(f)}{\partial w_n}\circ\iota\big)=i\big(T^{r-1}(\tfrac{\partial f}{\partial z_n})\circ\iota\big)=i\big(T^{r-1}(\tfrac{\partial u}{\partial x_n})+iT^{r-1}(\tfrac{\partial v}{\partial x_n})\big)\\ &=T^{r-1}(\tfrac{\partial u}{\partial y_n})+iT^{r-1}(\tfrac{\partial v}{\partial y_n})=\frac{\partial T^r(u)}{\partial t_n}+i\frac{\partial T^r(v)}{\partial t_n}.\end{aligned}$$

We have now established the hypotheses of the lemma, thereby proving $(*)$ and completing the proof of the complex version of Taylor's theorem.

7.4 Functions Defined by Power Series

In this section, and the two that come after it, we study functions, such as the exponential and trigonometric functions, that are defined by power series. We will show that the function defined by the series in the disk given by the radius of convergence is C^1, and that it's derivative (or any partial derivative in the multivariate case) coincides with the function defined by the power series obtained by term-by-term differentiation. Since this analysis can be repeated, it follows that a power series defines a C^∞ function.

Recall that a function is **analytic** if, for each point in its domain, there is a power series centered at that point that converges to the function in some neighborhood of that point. Of course a function defined by a power series satisfies this condition at the point where the series is centered, and we will show that it is analytic by giving an explicit formula for its power series at a nearby point.

The current section develops the results described above in the univariate case, and the multivariate case is considered in the next section. The underlying ideas in the two cases are really the same, and if you have a clear understanding of the univariate case you should be well equipped to deal with the technical complexities of the general case.

Consider a power series

$$\sum_{k=0}^{\infty} c_k (z-a)^k$$

centered at $a \in \mathbb{C}$ whose coefficients $c_0, c_1, c_2, \ldots$ are in $\mathbb{C}$. In Section 3.9 we defined the radius of convergence to be

$$R := \liminf_{k\to\infty} 1/\sqrt[k]{|c_k|}.$$

Assume that R is positive, and let

$$D = \{\, z \in \mathbb{C} : |z-a| < R \,\}.$$

We showed (Lemma 3.54) that if $0 < r < R$, then the series converges uniformly and absolutely on the closed disk $\{\, z \in \mathbb{C} : |z-a| \le r \,\}$. Any compact subset of D is contained in some such disk (the interiors of such disks are an open cover that must have a finite subcover) so the series converges uniformly on compacta to a function

$$f : D \to \mathbb{C}.$$

Now consider the power series $\sum_{k=1}^{\infty} kc_k(z-a)^{k-1}$ obtained by applying the rules for differentiating sums and products. Since

$$\sqrt[k-1]{|kc_k|} = \sqrt[k-1]{k}(\sqrt[k]{|c_k|})^{\frac{k}{k-1}}$$

and $\sqrt[k-1]{k} \to 1$, the radius of convergence of this power series is also R, and it also converges uniformly on compacta in D. Theorem 7.3 now implies that the function defined by this series is f' because, for each $K = 1, 2, \dots$,

$$\sum_{k=1}^{K} kc_k(z-a)^{k-1} \quad \text{is the derivative of} \quad \sum_{k=0}^{K} c_k(z-a)^k.$$

In particular, f is C^1, but of course this argument applies equally to the power series for f', so, by induction, f is C^∞.

For the exponential and trigonometric functions

$$\exp(z) := \sum_{k=0}^{\infty} \frac{z^k}{k!}, \quad \cos(z) := \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k}}{(2k)!}, \quad \sin(z) := \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k+1}}{(2k+1)!}$$

differentiating the power series term-by-term gives

$$\exp'(z) = \exp(z), \quad \cos'(z) = -\sin(z), \quad \sin'(z) = \cos(z).$$

(To get the second equation first replace k with $k+1$, then differentiate.)

If $U \subset \mathbb{C}$ is open, $f : U \to \mathbb{C}$ is differentiable at $t \in U \cap \mathbb{R}$, and $f(U \cap \mathbb{R}) \subset \mathbb{R}$, then the derivative of $f|_{U\cap\mathbb{R}}$ with respect to the field $\mathbb{R}$ is the same as the derivative of f at t with respect to the field $\mathbb{C}$. (Make sure you see that this is an automatic consequence of the definitions of the two derivatives.) Therefore the same formulas characterize the derivatives of the exponential and trigonometric functions when these are regarded as functions from $\mathbb{R}$ to $\mathbb{R}$. Now that we know how to differentiate these functions and polynomials, and we have the rules for differentiating sums, products, quotients, and compositions of functions, we have purely mechanical procedures that allow us to differentiate an enormous variety of functions. Of course those of you who have already taken first year calculus know this *very* well.

We now turn to this section's second issue, which is to prove that f is analytic. For a given point $b \in D$ we will show that there is a power series $\sum_{k=0}^{\infty} c'_k(z-b)^k$ centered at b that converges to f absolutely in a neighborhood of b. In view of the relationship between the coefficients of a power series and the derivatives of the function it defines, necessarily

$$c'_k = \frac{1}{k!} f^{(k)}(b) = \frac{1}{k!} \sum_{\ell=k}^{\infty} \ell(\ell-1)\cdots(\ell-k+1)c_\ell(b-a)^{\ell-k}$$

$$= \frac{1}{k!}\sum_{\ell=k}^{\infty}\frac{\ell!}{(k-\ell)!}c_\ell(b-a)^{\ell-k} = \sum_{\ell=k}^{\infty}\binom{\ell}{k}c_\ell(b-a)^{\ell-k}.$$

Let R' be the radius of convergence of the power series $\sum_{k=0}^{\infty} c'_k(z-b)^k$. One would hope that $R' \geq R - |b-a|$, and this is in fact the case, but it's not exactly clear how one might prove this by working directly with the definition of R'. Instead recall that, by Proposition 3.55, if $|z-b| < R'$, then the series converges absolutely, and if $|z-b| > R'$, then the series does not converge absolutely. Consequently the inequality $R' \geq R - |b-a|$ follows if we can show that the series$\sum_{k=0}^{\infty} c'_k(z-b)^k$ converges absolutely when $|z-b| < R - |b-a|$. The following computation does this by first changing the order of summation, then applying basic facts about the absolute value, after which the binomial theorem can be invoked:

$$\sum_{k=0}^{\infty}\sum_{\ell=k}^{\infty}\left|c_\ell\binom{\ell}{k}(b-a)^{\ell-k}(z-b)^k\right| = \sum_{\ell=0}^{\infty}\sum_{k=0}^{\ell}\left|c_\ell\binom{\ell}{k}(b-a)^{\ell-k}(z-b)^k\right|$$

$$= \sum_{\ell=0}^{\infty}|c_\ell|\Big(\sum_{k=0}^{\ell}\binom{\ell}{k}|b-a|^{\ell-k}|z-b|^k\Big) = \sum_{\ell=0}^{\infty}|c_\ell|\big(|z-b|+|b-a|\big)^\ell < \infty.$$

The final inequality follows from the fact that $|z-b|+|b-a| < R$, as was explained in detail in Section 3.9.

We still need to show that the function defined by the power series $\sum_{k=0}^{\infty} c'_k(z-b)^k$ agrees with f near b. The absolute convergence established above implies that the various terms can be summed in any order. This fact validates the following computation applying the binomial theorem:

$$\begin{aligned}\sum_{k=0}^{\infty} c'_k(z-b)^k &= \sum_{k=0}^{\infty}\Big(\sum_{\ell=k}^{\infty}\binom{\ell}{k}c_\ell(b-a)^{\ell-k}\Big)(z-b)^k \\ &= \sum_{\ell=0}^{\infty}c_\ell\Big(\sum_{k=0}^{\ell}\binom{\ell}{k}(b-a)^{\ell-k}(z-b)^k\Big) \\ &= \sum_{\ell=0}^{\infty}c_\ell\big((z-b)+(b-a)\big)^\ell \\ &= \sum_{\ell=0}^{\infty}c_\ell(z-a)^\ell = f(z).\end{aligned}$$

7.5 Multivariate Power Series

We now generalize the analysis above to multivariate power series. Aside from the fact that we have to do it before we can say we've done it, the main point of interest here is actually the system of notation, which allows lots of stuff to be compressed into tight little packages. This tends to increase the mental distance between the symbols and what they represent, and the calculations are still a bit messy, so you may find this rather difficult and tedious reading. (The details won't reappear later, so if things start to get difficult you can skip ahead to the last few paragraphs of this section, intending to return some morning after you've had a good night's sleep and a strong cup of coffee.) At the same time the excellent match between the notation and the analysis is, in my opinion, extremely elegant and aesthetically pleasing.

We'll work with the variables $z = (z_1, \ldots, z_n)$. An **exponent vector** for this system of variables is an n-tuple $\alpha = (\alpha_1, \ldots, \alpha_n)$ whose components are nonnegative integers. Let E_n be the set of exponent vectors. We achieve a compact notation for **monomials** by setting

$$z^\alpha := z_1^{\alpha_1} \cdots z_n^{\alpha_n}.$$

A power series in z centered at $a \in \mathbb{C}^n$ is then an infinite sum

$$\sum_{\alpha \in E_n} c_\alpha (z - a)^\alpha \tag{$*$}$$

where the coefficients c_α are in $\mathbb{C}$.

The **total degree** of an exponent vector α is

$$|\alpha| := \alpha_1 + \cdots + \alpha_n.$$

For each $k = 0, 1, 2, \ldots$ let E_n^k be the set of exponent vectors of total degree k. We can now define the **radius of convergence** of the series to be

$$R := \liminf_{k \to \infty} \frac{1}{\max_{\alpha \in E_n^k} \sqrt[k]{|c_\alpha|}}.$$

The analysis of uniform absolute convergence is only slightly more complicated than in the univariate case, the one new idea being that the number of elements in E_n^k, which we will denote by $|E_n^k|$, grows slowly enough as k increases that it doesn't affect the conclusion. The number $|E_n^k|$ can be analyzed with considerable sophistication and precision, but for us a quite crude bound is good enough, so we will simply show that $|E_n^k| \le (k+1)^{n-1}$ whenever $n \ge 1$ and $k \ge 0$. When $n = 1$ the unique exponent of total degree k corresponds to z_1^k, so $|E_1^k| = 1$, and (by induction) we may assume that the inequality holds with n replaced by $n-1$. An element of E_n^k may be thought of as a pair whose first element is $\alpha_1 \in \{0, \dots, k\}$ and whose second element is an element of $E_{n-1}^{k-\alpha_1}$, so $|E_n^k|$ can be written as a sum $|E_{n-1}^0| + \cdots + |E_{n-1}^k|$ of $k+1$ terms, each of which is no greater than $(k+1)^{n-2}$, so the sum is no greater than $(k+1)^{n-1}$.

When $0 < r < R$ let

$$B(r) := \{\, z \in \mathbb{C}^n : \|z - a\|_\infty \le r \,\}$$

be the closed disk of radius r centered at a. (Recall that the norm $\|\cdot\|_\infty$ on $\mathbb{C}^n$ is given by $\|w\|_\infty := \max\{|w_1|, \dots, |w_n|\}$.) We claim that the series $(*)$ converges absolutely and uniformly on each such $B(r)$. Fixing r, choose numbers r_1 and r_2 with $r < r_1 < r_2 < R$. If K is large enough, then for all $k > K$ we have $(k+1)^{n-1} < (r_1/r)^k$ for all $k > K$ (because exponential growth dominates the growth of any polynomial) and $\max_{\alpha \in E_n^k} \sqrt[k]{|c_\alpha|} < 1/r_2$. For any $w \in \mathbb{C}^n$ we have $|w^\alpha| = \prod_i |w_i|^{\alpha_i} \le \|w\|_\infty^{|\alpha|}$. In view of all this, if $\|z - a\|_\infty \le r$, then

$$\Big|\sum_{k=K+1}^{\infty} \sum_{\alpha \in E_n^k} c_\alpha (z-a)^\alpha\Big| \le \sum_{k=K+1}^{\infty} \sum_{\alpha \in E_n^k} |c_\alpha| r^{|\alpha|} \le \sum_{k=K+1}^{\infty} |E_n^k| (1/r_2)^k r^k$$

$$\le \sum_{k=K+1}^{\infty} (r_1/r)^k (r/r_2)^k = \sum_{k=K+1}^{\infty} (r_1/r_2)^k = \frac{(r_1/r_2)^K}{1 - r_1/r_2}.$$

The final expression goes to 0 as $K \to \infty$, so the series converges absolutely, and it converges uniformly on $B(r)$.

Fix an index j with $1 \le j \le n$ and let $\mathbf{e}_j = (0, \dots, 1, \dots, 0)$ be the j^{th} standard unit basis vector. Applying the rules for computing partial derivatives shows that the partial derivative of the given power series with respect to z_j is

$$\sum_{k=0}^{\infty} \sum_{\alpha \in E_n^k, \alpha_j > 0} \alpha_j c_\alpha (z-a)^{\alpha - \mathbf{e}_j} = \sum_{k=0}^{\infty} \sum_{\alpha \in E_n^k} (\alpha_j + 1) c_{\alpha + \mathbf{e}_j} (z-a)^\alpha.$$

Of course $\{\, \alpha + \mathbf{e}_j : \alpha \in E_n^k \,\} \subset E_n^{k+1}$ and $\alpha_j \le k$ when $\alpha \in E_n^k$, so

$$\liminf_{k \to \infty} \frac{1}{\max_{\alpha \in E_n^k} \sqrt[k]{\big|(\alpha_j + 1) c_{\alpha + \mathbf{e}_j}\big|}} \ge R.$$

Therefore the series above also converges absolutely and uniformly on $B(r)$. Now Theorem 7.3 implies that this limit is the partial with respect to z_j of the function defined by the given power series. Since this argument can be applied iteratively, it follows that the function defined by the given power series is C^∞. Summarizing the analysis to this point:

Theorem 7.9. *Let $\sum_{\alpha \in E_n} c_\alpha (z-a)^\alpha$ be a power series whose radius of convergence*

$$R := \liminf_{k\to\infty} \frac{1}{\max_{\alpha \in E_n^k} \sqrt[k]{|c_\alpha|}}$$

is positive, and let $D = \{ z \in \mathbb{C}^n : \|z-a\|_\infty < R\}$. Then the series converges absolutely and uniformly on compacta to a function $f : D \to \mathbb{C}$ that is C^∞, and its partial derivatives are the functions defined by the power series obtained from term-by-term differentiation.

We need some more notation to handle the calculations related to analyticity of f. For $\alpha \in E_n$ we define

$$\alpha! := \alpha_1! \times \cdots \times \alpha_n!.$$

For $\alpha, \beta \in E_n$ with $\beta \le \alpha$ let

$$\binom{\alpha}{\beta} := \frac{\alpha!}{\beta!(\alpha-\beta)!}.$$

There is now the following multivariate extension of the binomial theorem.

Lemma 7.10. *For all $z, w \in \mathbb{C}^n$ and all $\alpha \in E_n$,*

$$(z+w)^\alpha = \sum_{0 \le \beta \le \alpha} \binom{\alpha}{\beta} z^{\alpha-\beta} w^\beta.$$

Proof. This is a big calculation employing the univariate binomial theorem and the distributive law:

$$\begin{aligned}
(z+w)^\alpha &= (z_1+w_1)^{\alpha_1} \cdots (z_n+w_n)^{\alpha_n} \\
&= \Big(\sum_{\beta_1=0}^{\alpha_1} \binom{\alpha_1}{\beta_1} z_1^{\alpha_1-\beta_1} w^{\beta_1} \Big) \cdots \Big(\sum_{\beta_n=0}^{\alpha_n} \binom{\alpha_n}{\beta_n} z_n^{\alpha_n-\beta_n} w^{\beta_n} \Big) \\
&= \sum_{0 \le \beta \le \alpha} \binom{\alpha_1}{\beta_1} \cdots \binom{\alpha_n}{\beta_n} z_1^{\alpha_1-\beta_1} \cdots z_n^{\alpha_n-\beta_n} w_1^{\beta_1} \cdots w_n^{\beta_n} \\
&= \sum_{0 \le \beta \le \alpha} \binom{\alpha}{\beta} z^{\alpha-\beta} w^\beta.
\end{aligned}$$

□

If $U \subset \mathbb{C}^n$ is open, $f : U \to \mathbb{C}^n$ is C^r, and $\alpha \in E_n$ with $|\alpha| \le r$, then, in order to have more compact notation, we write $\partial^\alpha f$ in place of

$$\frac{\partial^{|\alpha|} f}{\partial z_1^{\alpha_1} \cdots \partial z_n^{\alpha_n}}.$$

When $0 \le \beta \le \alpha$ the rules for differentiating polynomials give

$$\frac{\partial^{\beta_i} z^\alpha}{\partial z_i^{\beta_i}} = \alpha_i(\alpha_i - 1) \cdots (\alpha_i - \beta_i + 1) z_1^{\alpha_1} \cdots z_{i-1}^{\alpha_{i-1}} z_i^{\alpha_i - \beta_i} z_{i+1}^{\alpha_{i+1}} \cdots z_n^{\alpha_n}$$

$$= \frac{\alpha_i!}{(\alpha_i - \beta_i)!} z_1^{\alpha_1} \cdots z_{i-1}^{\alpha_{i-1}} z_i^{\alpha_i - \beta_i} z_{i+1}^{\alpha_{i+1}} \cdots z_n^{\alpha_n}.$$

Since we have established Clairhaut's theorem for functions on $\mathbb{C}^n$, we can do this repeatedly for $i = 1, \ldots, n$ without worrying about the order of differentiation, and the end result boils down to the tight little formula

$$\partial^\beta z^\alpha = \frac{\alpha!}{(\alpha - \beta)!} z^{\alpha - \beta}.$$

Under the hypotheses of Theorem 7.9, when $\|b - a\|_\infty < R$ we can compute $\partial^\beta f(b)$ by term-by-term partial differentiation:

$$\partial^\beta f(b) = \sum_{\alpha \in E_n, \alpha \ge \beta} c_\alpha \frac{\alpha!}{(\alpha - \beta)!} (b - a)^{\alpha - \beta}.$$

In particular, when $b = a$ every term in this sum other than the constant term vanishes, so $\partial^\beta f(a) = \beta! c_\beta$.

Theorem 7.11. *Under the hypotheses of Theorem 7.9, fix $b \in D$, and for $\beta \in E_n$ let*

$$c'_\beta := \frac{1}{\beta!} \partial^\beta f(b) = \sum_{\alpha \in E_n, \alpha \ge \beta} c_\alpha \binom{\alpha}{\beta} (b - a)^{\alpha - \beta}.$$

Then the series $\sum_{\beta \in E_n} c'_\beta (z - b)^\beta$ converges absolutely to f on

$$\{\, z \in \mathbb{C}^n : \|z - b\|_\infty < R - \|b - a\|_\infty \,\},$$

and consequently (since b was arbitrary) f is analytic.

Proof. Absolute convergence of $\sum_{\beta\in E_n} c'_\beta(z-b)^\beta$ follows if we establish that

$$S := \sum_{\beta\in E_n} \sum_{\alpha\in E_n,\alpha\geq\beta} c_\alpha \binom{\alpha}{\beta} (b-a)^{\alpha-\beta}(z-b)^\beta$$

converges absolutely. We adopt the following notation: if $z \in \mathbb{C}^n$ then $|z| := (|z_1|,\dots,|z_n|)$. For any exponent vector α we have $|z^\alpha| = |z|^\alpha$ where the left hand side is the usual modulus for $\mathbb{C}$. Therefore

$$|S| \leq \sum_{\beta\in E_n} \sum_{\alpha\in E_n,\alpha\geq\beta} |c_\alpha| \binom{\alpha}{\beta} |b-a|^{\alpha-\beta}|z-b|^\beta$$

$$= \sum_{\alpha\in E_n} |c_\alpha| \Big[\sum_{0\leq\beta\leq\alpha} \binom{\alpha}{\beta} |b-a|^{\alpha-\beta}|z-b|^\beta \Big].$$

(The second relation reorders the series, even though we haven't yet shown absolute convergence, but this is permitted because all terms are nonnegative real numbers, so the sum under one ordering has the same, possibly infinite, limit as the sum under any other ordering.) The multivariate binomial theorem now gives

$$|S| \leq \sum_{\alpha\in E_n} |c_\alpha|(|b-a| + |z-b|)^\alpha.$$

We claim that

$$(|b-a| + |z-b|)^\alpha \leq \|\,|b-a| + |z-b|\,\|_\infty^{|\alpha|} \leq (\|b-a\|_\infty + \|z-b\|_\infty)^{|\alpha|}$$

for all $\alpha \in E_n$. The first inequality comes from the fact that $x^\alpha \leq \|x\|_\infty^{|\alpha|}$ for any $x \in \mathbb{R}^n$ and $\alpha \in E_n$. The second inequality is justified by the observation that

$$\|\,|z| + |w|\,\|_\infty = \max\{|z_1| + |w_1|,\dots,|z_n| + |w_n|\}$$

$$\leq \max\{|z_1|,\dots,|z_n|\} + \max\{|w_1|,\dots,|w_n|\} = \|z\|_\infty + \|w\|_\infty$$

for all $z, w \in \mathbb{C}^n$. We now have $|S| \leq \sum_{\alpha\in E_n} |c_\alpha| r^{|\alpha|}$ for some $r < R$, and in our analysis of the absolute convergence of $\sum_{\alpha\in E_n} c_\alpha(z-a)^\alpha$ on D we showed that $\sum_{\alpha\in E_n} |c_\alpha| r^{|\alpha|}$ is finite.

Now that we know that S converges absolutely we are free to rearrange the terms in the summation. This (together with the multivariate binomial formula) justifies the computation

$$\begin{aligned}\sum_{\beta\in E_n} c'_\beta(z-b)^\beta &= \sum_{\beta\in E_n}\Big[\sum_{\alpha\in E_n,\alpha\ge\beta}\binom{\alpha}{\beta}c_\alpha(b-a)^{\alpha-\beta}\Big](z-b)^\beta\\ &= \sum_{\alpha\in E_n} c_\alpha\Big[\sum_{0\le\beta\le\alpha}\binom{\alpha}{\beta}(b-a)^{\alpha-\beta}(z-b)^\beta\Big]\\ &= \sum_{\alpha\in E_n} c_\alpha(z-a)^\alpha = f(z).\end{aligned}$$

□

We've shown that functions defined by power series are analytic and C^∞. There is one more element of the overall picture:

Theorem 7.12. *If $U \subset \mathbb{C}^n$ is open and $f : U \to \mathbb{C}$ is C^1, then it is analytic.*

This is certainly one of the most remarkable results in all of mathematics, and also one of the most important. Of course it seems magical that C^1 functions are automatically C^∞, but analyticity is actually much stronger still. As we'll explain in detail in the next section, a holomorphic function on a connected domain is completely determined by its power series at any point in the domain, and in particular the entire function can be recovered from the restriction of the function to an arbitrarily small neighborhood of that point. In contrast, in Section 7.7 we'll see that real valued C^∞ functions defined on open subsets of $\mathbb{R}^n$ are, in a certain sense, completely flexible.

Although we have mentioned various famous results that are not proven here, the treatment to this point has been rigorous in the sense of proving all the results used in our own analysis. As we explained in the first chapter, from a psychological point of view, real skill and facility with mathematics is impossible without this sort of rigorous understanding of foundations. Due to the use of advanced concepts and results not covered here, and its overall length and complexity, the proof of Theorem 7.12 is at a quite different level from anything else in this book, and it cannot be included. Although we won't make much use of it to prove other things (it says that certain types of objects don't exist, and mostly we simply won't consider them) Theorem 7.12 plays an important role in the next two chapters because it informs our expectations concerning the properties of the objects studied there, and the failure to prove it is our primary violation of this standard of self-contained, exact understanding.

7.6 Analytic Continuation

Suppose that X and Y are topological spaces, and $x_0 \in X$. We say that two functions $f : X \to Y$ and $f' : X \to Y$ *have the same germ at* x_0 if there is a neighborhood W of x_0 such that $f|_W = f'|_W$. This is an equivalence relation. (Reflexivity and symmetry are immediate, and transitivity is easy, but you should think through the details for yourself.) The equivalence class containing f is called the *germ* of f at x_0.

Our goal in this section is:

Theorem 7.13. *If $U \subset \mathbb{C}^n$ is open and connected, $f, g : U \to \mathbb{C}$ are analytic functions, and $a_0 \in U$ is a point at which f and g have the same power series, or the same germ, then $f = g$.*

In this sense analytic functions are "rigid."

If $f : U \to \mathbb{C}$ is analytic and $a_0 \in U$, then the definition of analyticity implies that the power series of f at a_0 determines the germ of f at a_0. But Theorem 7.11 implies that the germ determines the power series: at each point in the domain of an analytic function the power series is determined by the partial derivatives (of all orders) of the function. To prove Theorem 7.13 it is enough to show that the conclusion holds when f and g have the same power series at a_0.

Recall that if $z \in \mathbb{C}^n$ and $r > 0$, then the ball of radius r centered at z, with respect to the norm $\|\cdot\|_\infty$, is

$$\mathbf{U}_r(z) := \{\, z' \in \mathbb{C}^n : \|z' - z\|_\infty < r \,\}.$$

For $z \in U$ let r_z be the supremum of the set of r such that:

(a) $\mathbf{U}_r(z) \subset U$;

(b) the radii of convergence of the power series of f and g centered at z are greater than r;

(c) $f|_{\mathbf{U}_r(z)}$ and $g|_{\mathbf{U}_r(z)}$ agree with the functions defined by the power series of f and g centered at z.

The definition of an analytic function implies that $r_z > 0$.

Consider any two points $z, z' \in U$ with $z' \in \mathbf{U}_{r_z}(z)$. To show that

$$r_{z'} \geq r_z - \|z' - z\|_\infty \tag{$*$}$$

we check the three conditions above. The triangle inequality gives

$$\mathbf{U}_{r_z - \|z'-z\|_\infty}(z') \subset \mathbf{U}_{r_z}(z) \subset U.$$

Since f and g agree on $\mathbf{U}_{r_z}(z)$ with the functions defined by their power series at z, Theorem 7.11 implies the radii of convergence of the power series of f and g at z' are at least $r_z - \|z' - z\|_\infty$, and that f and g agree with the functions defined by these power series on $\mathbf{U}_{r_z - \|z'-z\|_\infty}(z')$.

In order to prove Theorem 7.13 it suffices to show that $f(a_1) = g(a_1)$ for any $a_1 \in U$. Fix such an a_1. The proof uses a process called **analytic continuation** which develops a quite simple idea. Consider a point $z_1 \in \mathbf{U}_{r_{a_0}}(a_0)$. The definition of r_{a_0} implies that f and g agree in $\mathbf{U}_{r_{a_0}}(a_0)$ with the function defined by their common power series, so f and g have the same power series at z_1. If z_2 is a point in $\mathbf{U}_{r_{z_1}}(z_1)$, then f and g agree on a neighborhood of z_2, so they have the same power series there, we can choose a point $z_3 \in \mathbf{U}_{r_{z_2}}(z_2)$, and so forth. If we can continue this process until $a_1 \in \mathbf{U}_{r_{z_k}}(z_k)$, then $f(a_1) = g(a_1)$ as desired.

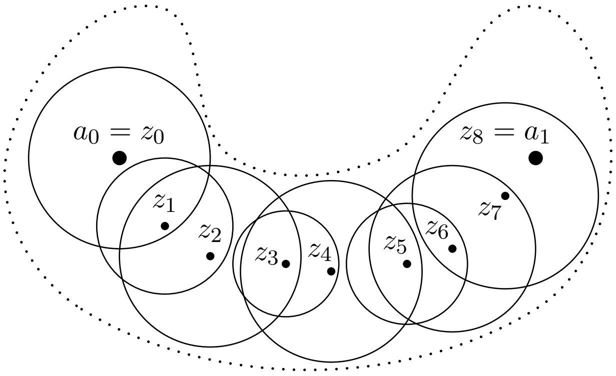


Figure 7.3

The remaining task is to show that there are points

$$a_0 = z_0, z_1, \dots, z_{k-1}, z_k = a_1$$

in U such that $\|z_{h+1} - z_h\|_\infty < r_{z_h}$ for all $h = 0, \dots, k-1$. Below we will show that there is a continuous $\gamma : [0,1] \to U$ with $\gamma(0) = a_0$ and $\gamma(1) = a_1$. Assuming that we have such a γ, let S be the set of t such that there exist

$$0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = t$$

such that $\|\gamma(t_{h+1}) - \gamma(t_h)\|_\infty < r_{\gamma(t_h)}$ for all $h = 0, \dots, k-1$. We would like to show that $1 \in S$. Of course S is nonempty because it contains 0, and it is bounded above by 1; let $\bar{t}$ be its least upper bound.

Now observe that $(*)$ implies that $z \in \mathbf{U}_{r_{z'}}(z')$ whenever $z' \in \mathbf{U}_{r_z/2}(z)$, and in particular $\gamma(\overline{t}) \in \mathbf{U}_{r_{\gamma(t_k)}}(\gamma(t_k))$ whenever $t_0, \ldots, t_k$ is a sequence as above with $\|\gamma(t_k) - \gamma(\overline{t})\| < \frac{1}{2} r_{\gamma(\overline{t})}$. The definition of $\overline{t}$ implies that such a sequence exists, and we can extend it by setting $t_{k+1} := \overline{t}$. Moreover, if $\overline{t} < 1$ we can extend it again by letting t_{k+2} be some number greater than $\overline{t}$ with $\|\gamma(t_{k+2}) - \gamma(\overline{t})\| < r_{\gamma(t)}$. This shows both that $\overline{t} \in S$ and that $\overline{t} < 1$ implies a contradiction of the definition of $\overline{t}$. Therefore $1 = \overline{t} \in S$.

All that remains now is to show that a suitable γ exists. A **path** in a topological space X is a continuous function $\gamma : [0,1] \to X$, and we say that X is **path connected** if, for any $a_0, a_1 \in X$, there is a path γ with $\gamma(0) = a_0$ and $\gamma(1) = a_1$. So, we need to show that any connected open $U \subset \mathbb{C}^n$ is path connected. This may seem obvious, but there is a bit more to say about it than one might expect, and we will see a famous example that "everyone" knows, so you should know it too.

A path connected space is necessarily connected. To show this suppose that X is disconnected, so that $X = U_0 \cup U_1$, where U_0 and U_1 are nonempty, open, and disjoint. Then X is not path connected because if $\gamma : [0,1] \to X$ was a path with $\gamma(0) \in U_0$ and $\gamma(1) \in U_1$, then $\gamma^{-1}(U_0)$ and $\gamma^{-1}(U_1)$ would be nonempty, open, disjoint sets whose union was all of $[0,1]$, and this is impossible because $[0,1]$ is connected (Lemma 3.65).

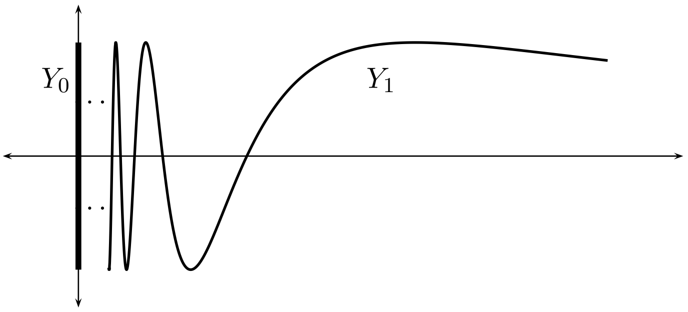


Figure 7.4

The **topologist's sine curve** is a very well known example of a connected space that is not path connected. This is the set $Y = Y_0 \cup Y_1 \subset \mathbb{R}^2$ where

$$Y_0 := \{\, (0,s) : -1 \leq s \leq 1 \,\} \quad \text{and} \quad Y_1 := \{\, (s, \sin \tfrac{1}{s}) : 0 < s \leq 1 \,\}.$$

As shown in Figure 7.4, Y_1 goes up and down infinitely many times as $s \to 0$ from above. Since Y_0 and Y_1 are each the image of a continuous function from an interval in $\mathbb{R}$ to Y, they are path connected, hence connected. If $Y = U_1 \cup U_2$, with U_1 and U_2 disjoint and open, then for $i = 0, 1$ we have $Y_i = (U_1 \cap Y_i) \cup (U_2 \cap Y_i)$, so either $Y_i \subset U_1$ or $Y_i \subset U_2$. This means that the only way that Y might be disconnected is if Y_0 and Y_1 are both open subsets, but this isn't the case: every neighborhood of each point of Y_0 intersects Y_1.

It might seem obvious that there can't be a continuous path with $\gamma(0) \in Y_0$ and $\gamma(1) \in Y_1$, but let's work through a rigorous proof anyway. Since Y_0 is a closed subset of Y, $\gamma^{-1}(Y_0)$ is a closed subset of $[0, 1]$, so it contains its least upper bound b. The definition of continuity gives $\delta > 0$ such that $\|\gamma(t) - \gamma(b)\| < 1$ whenever $b - \delta < t < b + \delta$. Since $[b, b + \delta)$ is connected, so is $\pi(\gamma([b, b + \delta)))$ where $\pi : \mathbb{R}^2 \to \mathbb{R}$ is the projection onto the first component. Now observe that $\pi(\gamma(b)) = 0$, and $\pi(\gamma(t)) > 0$ for all $t \in (b, b + \delta)$, so there are t in this interval such that the second component of $\gamma(t)$ is 1 and other t such that the second component of $\gamma(t)$ is -1, which contradicts the definition of δ.

Returning to the general framework, the **path component** $X(a_0)$ of a point $a_0 \in X$ is the set of all points that are connected to a_0 by some path in X. That is, $a_1 \in X(a_0)$ if and only if $a_1 = \gamma(1)$ for some path $\gamma : [0, 1] \to X$ with $\gamma(0) = a_0$. Clearly $X(a_0)$ contains every path connected subset of X containing a_0, and it is itself path connected, so it is the largest path connected subset of X containing a_0. Two path components of X are the same if they have a point in common, so distinct path components are disjoint, and the path components are the elements of a partition of X. Since $X(a_0)$ contains the image of a path from a_0 to each of its points, $X(a_0)$ is connected, so $X(a_0)$ is a subset of the connected component of X containing a_0. In general, if we have two partitions of a set and every element of the first partition is a subset of some element of the second partition, then we say that the first partition is a **refinement** of the second partition and the second partition is a **coarsening** of the first. The partition of X into path components is a refinement of the partition into connected components.

The space X is **locally path connected** if every point in X has a path connected neighborhood. The path components of a locally path connected space are open, obviously. Consequently the path components contained in any given connected component constitute a partition of that connected component into disjoint nonempty open sets, so there cannot be more than one such path component: in a locally path connected space each path component is a connected component. In particular, if there is a single connected component (i.e., the space is connected) then there is a single path component (i.e., the space is path connected). That is, *a space is path connected if it is both connected and locally path connected.* Of course an open subset of $\mathbb{R}^n$ or $\mathbb{C}^n$ is locally path connected because each of its points is contained in an open ball that is in turn contained in the set, and the open ball contains the line segment between the center and any other point. Therefore a connected open subset of $\mathbb{C}^n$ is path connected.

The proof of Theorem 7.13 is now complete.

7.7 Smooth Functions

A real valued function on an open set $U \subset \mathbb{R}^n$ is often said to be **smooth** if it is C^∞, though you should be warned that 'smooth' is a word that different authors use in different ways, according to convenience. (There will be an example of this later on.) Almost all the functions we deal with on a regular basis are smooth, so this may seem like a fairly normal state of affairs. On the other hand, most of the functions we're familiar with, like the exponential and trigonometric functions, are actually defined by power series, and, as we saw above, such functions are "rigid" in the sense that if the domain is connected, then the function is completely determined by its restriction to any open subset of the domain, no matter how small. The point of this section is to show that, in contrast, C^∞ functions can flop around in any way whatsoever. The restriction of the function to a small neighborhood of one point doesn't tell you anything at all about the restriction to a small neighborhood of some other point.

The key to this topic is a particular example of a smooth function that is not given by a power series in any neighborhood of a certain point in its domain. The construction depends on certain basic properties of the exponential function. Let

$$P(t) = a_0 + a_1 t + \cdots + a_r t^r \in \mathbb{R}[t]$$

be a polynomial function of t. In the power series $\exp(t) = \sum_{j=0}^{\infty} \frac{1}{j!} t^j$ the terms $t^k/k!$ for $k > r$ will be much larger than any $a_i t^i$ when t is sufficiently large, so $\exp(t)/|P(t)| \to \infty$ as $t \to \infty$. The multiplicative property of the exponential function gives $\exp(-t) = 1/\exp(t)$, so we have $P(t)\exp(-t) \to 0$ as $t \to \infty$. The particular implication of this used in the following is that

$$\lim_{t \to 0, t > 0} P(\tfrac{1}{t}) \exp(-\tfrac{1}{t}) = 0. \qquad (*)$$

Let $\beta : \mathbb{R} \to \mathbb{R}$ be the function

$$\beta(t) := \begin{cases} 0, & t \leq 0, \\ \exp(-\frac{1}{t}), & t > 0. \end{cases}$$

Earlier in this chapter we showed that when the exponential function is regarded as a function from $\mathbb{C}$ to $\mathbb{C}$, it is its own derivative. Just by combining this fact with the definition of the derivative, we can see that the exponential function is also its own derivative when it is regarded as a function from $\mathbb{R}$ to $\mathbb{R}$. Applying the chain rule, the product rule, the quotient rule, and the rules for differentiating polynomials, we can compute that

$$\beta'(t) = \tfrac{1}{t^2}\exp(-\tfrac{1}{t}) \quad \text{and} \quad \beta''(t) = \Big(\tfrac{-2}{t^3} + \tfrac{1}{t^4}\Big)\exp(-\tfrac{1}{t})$$

when $t > 0$.

There is a pattern emerging here. Using induction, we will show that for each $r = 0, 1, 2, \ldots$ there is a polynomial $P_r \in \mathbb{R}[t]$ such that

$$\beta^{(r)}(t) := \begin{cases} 0, & t \le 0, \\ P_r(\frac{1}{t})\exp(-\frac{1}{t}), & t > 0. \end{cases}$$

This is true when $r = 0$, so, by induction, we may assume that it has already been established with $r-1$ in place of r. It is, of course, clear that $\beta^{(r)}(t) = 0$ when $t < 0$. For $t > 0$ the chain rule and the formulas for the derivatives of products and quotients give

$$\beta^{(r)}(t) = P'_{r-1}(\tfrac{1}{t}) \cdot \tfrac{-1}{t^2} \cdot \exp(-\tfrac{1}{t}) + P_{r-1}(\tfrac{1}{t}) \cdot \exp(-\tfrac{1}{t}) \cdot \tfrac{1}{t^2},$$

so we can verify the claim for $t > 0$ by setting

$$P_r(s) := s^2\big(-P'_{r-1}(s) + P_{r-1}(s)\big).$$

We still need to show that $\beta^{(r)}(0) = 0$. Of course

$$\frac{\beta^{(r-1)}(t) - \beta^{(r-1)}(0)}{t} = 0$$

whenever $t < 0$. Since

$$-|\beta^{(r-1)}(t)| \le \beta^{(r-1)}(t) \le |\beta^{(r-1)}(t)|$$

we have

$$-|\tfrac{1}{t}P_{r-1}(\tfrac{1}{t})|\exp(-\tfrac{1}{t}) \le \frac{\beta^{(r-1)}(t) - \beta^{(r-1)}(0)}{t} \le |\tfrac{1}{t}P_{r-1}(\tfrac{1}{t})|\exp(-\tfrac{1}{t})$$

for all $t > 0$, and $(*)$ implies that $|\frac{1}{t}P_{r-1}(\frac{1}{t})|\exp(-\frac{1}{t}) \to 0$ as $t \to 0$ from above. This verifies that 0 satisfies the definition of the derivative of $\beta^{(r-1)}$ at 0. Thus β is C^∞.

There is no neighborhood of 0 on which the function β is the limit of a power series $\sum_{r=0}^{\infty} a_r t^r$ with a positive radius of convergence. To see this observe that if all the coefficients a_r are zero, then the function defined by this power series is identically zero, but $\beta(t) > 0$ for all $t > 0$. On the other hand suppose that K is the smallest integer such that $a_K \neq 0$. Using the definition of the radius of convergence, one can easily show that

$$|a_K t^K| > \sum_{k=K+1}^{\infty} |a_k t^k| \geq \Big| \sum_{k=K+1}^{\infty} a_k t^k \Big|$$

when $|t|$ is sufficiently small. (The idea is a variant of the one used to prove the maximum modulus principle.) Therefore the function defined by the power series is nonzero for small negative values of t.

The function β can be used to construct many different kinds of smooth functions. For example, to get a C^∞ "bump" function $B : \mathbb{R}^n \to \mathbb{R}^n$ that is positive on the interior of the unit cube and vanishes everywhere else, set

$$B(x) := \prod_{i=1}^{n} \beta(x_i)\beta(1 - x_i).$$

We can reduce the diameter of this bump by rescaling the domain, we can translate it, we can multiple it by any scalar, and we can add such bump functions to each other. There are various precise senses in which such constructions can be used to show that any continuous function can be approximated by smooth functions, which means that the possibilities for smooth functions are not much different from the possible behaviors of continuous functions.

The contrasting properties of analytic and smooth functions are reflected in their roles in mathematics. Because there are almost no restrictions on what C^∞ functions can do, they are not so interesting in and of themselves, but their flexibility makes them versatile and powerful tools in various constructions that are important in topology. On the other hand, the subtle relationship between local information and the global properties of complex analytic functions makes them a source of profound problems that have been, and continue to be, the subject of some of the deepest research in mathematics.

There is one other type of function that should be mentioned. If $U \subset \mathbb{R}^n$ is open, a function $f : U \to \mathbb{R}$ is **real analytic** if, for each $a \in U$, there is a power series centered at a that converges to f absolutely in some neighborhood of a. Although a great many important functions are real analytic, the theory of such functions is not very prominent, in large part because many properties of real analytic functions can easily be derived from corresponding properties of complex analytic functions. If $\tilde{U} \subset \mathbb{C}^n$ is open, $\tilde{f} : \tilde{U} \to \mathbb{C}$ is complex analytic, $U \subset \tilde{U} \cap \mathbb{R}^n$ is open, and $\tilde{f}(U) \subset \mathbb{R}$, then $\tilde{f}|_U$ is real analytic. It turns out that every real analytic function is of this sort, as we'll explain in detail below. That is, a real analytic function is just a complex analytic function that happens to take on real values at the real points in its domain, and this characterization will imply everything we need to know about such functions.

Let $f : U \to \mathbb{R}$ be real analytic. For each $x \in U$ let $R_x > 0$ be a number small enough that $\mathbf{U}_{R_x}(x) \cap \mathbb{R}^n \subset U$, and also small enough that there is a function $\tilde{f}_x : \mathbf{U}_{R_x}(x) \to \mathbb{C}$ that agrees with f on $\mathbf{U}_{R_x}(x) \cap \mathbb{R}^n$ and is defined by a power series centered at x whose radius of convergence is at least R_x. The coefficients of this power series are determined by the partial derivatives of f, so Theorem 7.13 implies that for the chosen R_x there is a unique such $\tilde{f}_x$. Let

$$\tilde{U} := \bigcup_{x \in U} \mathbf{U}_{R_x/2}(x).$$

We would like to define $\tilde{f} : \tilde{U} \to \mathbb{C}$ by requiring that

$$\tilde{f}(z) = \tilde{f}_x(z)$$

whenever $z \in \mathbf{U}_{R_x/2}(x)$. In order for this to make sense it has to be the case that $\tilde{f}_x(z) = \tilde{f}_{x'}(z)$ whenever $z \in \mathbf{U}_{R_x/2}(x) \cap \mathbf{U}_{R_{x'}/2}(x')$, as we will show below. Provided we can do this, $\tilde{f}$ is complex analytic because it agrees with $\tilde{f}_x$ on each $\mathbf{U}_{R_x/2}(x)$, and it agrees with f on U because each $\tilde{f}_x$ agrees with f on $\mathbf{U}_{R_x/2}(x) \cap U$.

So suppose that $z \in \mathbf{U}_{R_x/2}(x) \cap \mathbf{U}_{R_{x'}/2}(x')$. Since we can interchange x and x', we may assume that $R_x > R_{x'}$, in which case the triangle inequality implies that $x' \in \mathbf{U}_{R_x}(x)$. Since $\tilde{f}_x$ and $\tilde{f}_{x'}$ are analytic functions that have the same power series as f at x' (the coefficients are determined by the partial derivatives of f at x') and $\mathbf{U}_{R_x}(x) \cap \mathbf{U}_{R_{x'}}(x')$ is open, the principle of analytic continuation (Theorem 7.13) implies that they agree on the connected component of $\mathbf{U}_{R_x}(x) \cap \mathbf{U}_{R_{x'}}(x')$ that contains x', so we would like to show that z is an element of this connected component. But z is an arbitrary element of $\mathbf{U}_{R_x}(x) \cap \mathbf{U}_{R_{x'}}(x')$, so we need to show that this set is connected.

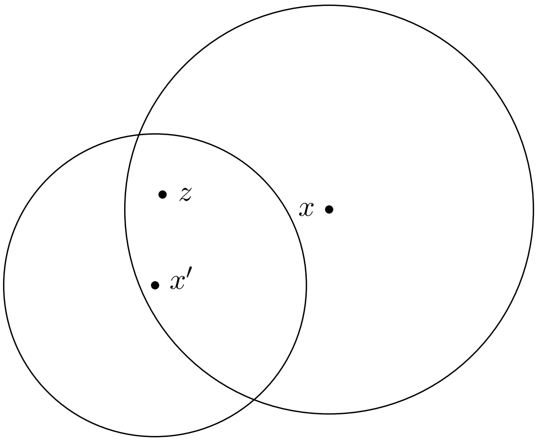


Figure 7.5

It would be quick and easy to show that $\mathbf{U}_{R_x}(x) \cap \mathbf{U}_{R_{x'}}(x')$ is path connected, hence connected, because it contains the line segment between x' and any other element, and in fact the bottom line of our discussion will be exactly this, but this is a particular instance of a very significant idea, so we take the time to explain things in more general terms. A subset C of $\mathbb{R}^m$ (or of $\mathbb{C}^m$, thought of as $\mathbb{R}^{2m}$) is **convex** if it contains the line segment between any two of its points. That is, for all $y, y' \in C$,

$$\{\, (1-t)y + ty' : 0 \le t \le 1 \,\} \subset C.$$

Visualizing this concept is easy: cubes, rectangles, disks, the interior of an ellipse, and so on, are all convex; a kidney shaped swimming pool isn't.

We will apply three simple facts about convex sets. First, an immediate consequence of the definition of convexity is that any convex set is path connected.

Second, the open balls defined by any norm are convex. To see this we begin by observing that for any $r > 0$, the open ball of radius r centered at the origin is convex: if $\|y\|, \|y'\| < r$ and $0 \le t \le 1$, then

$$\|(1-t)y + ty'\| \le \|(1-t)y\| + \|ty'\| = (1-t)\|y\| + t\|y'\| < r.$$

The convexity of all open balls follows from the fact that convexity is translation invariant: $C \subset \mathbb{R}^m$ is convex if and only if $a + C = \{\, a + y : y \in C \,\}$ is convex for any $a \in \mathbb{R}^m$.

The final fact is simply that the intersection of any two convex sets (actually the intersection of any, possibly infinite, collection of convex sets) is convex: if $y, y' \in C \cap C'$ and C and C' both contain the line segment between y and y', then so does their intersection. Obviously these facts combine to imply that $\mathbf{U}_{R_x}(x) \cap \mathbf{U}_{R_{x'}}(x')$ is convex, hence path connected.

I am afraid that this brief and passing mention of convexity might give a drastically understated impression of the overall significance of this concept, which is fundamental to many aspects of geometry and analysis. Because the concept is simple (at least on the surface) and fairly obvious, that it is important probably shouldn't be too surprising. What seems more remarkable is that convexity is very much a 20th century concept. This seems to be another gift of the set theory revolution: unless you are accustomed to talking about sets, there is simply no way to talk about convex sets.

Since we have shown that $\mathbf{U}_{R_x}(x) \cap \mathbf{U}_{R_{x'}}(x')$ is connected, we have completed the demonstration that any real analytic function is the restriction of a complex analytic function. Here are some simple consequences. If f is real analytic, then it is C^∞ because (Theorem 7.9) $\tilde{f}$ is C^∞. If $g : U \to \mathbb{R}$ is also real analytic, and f and g have the same power series at a point x, or agree in a neighborhood of x, then g is also the restriction of some $\tilde{g} : \tilde{U}' \to \mathbb{C}$, and (Theorem 7.11) $\tilde{f}$ and $\tilde{g}$ agree on the path component of $\tilde{U} \cap \tilde{U}'$ containing x, so f and g agree on the path component of U containing x. In particular, if a real analytic function on $\mathbb{R}$ is identically zero on an open interval, then it must be identically zero everywhere, which again shows that β is not real analytic.

7.8 The Inverse Function Theorem

Continuous functions $f : U \to \mathbb{R}^2$, where $U \subset \mathbb{R}^2$ is open, occur frequently in any course on multivariate calculus, and there are familiar physical analogs such as what happens when you wrap a melon in cellophane. Suppose that f is C^1, and consider a particular point $x \in U$. Rather complicated and messy "folds" or multiple wrappings (e.g., the complex function $z \mapsto z^2$ at the origin in $\mathbb{C}$) can happen when $Df(x)$ is singular, but experience leads us to expect that if $Df(x)$ is nonsingular, then f has a simple structure near x, with a neighborhood of x mapped "nicely" onto a neighborhood of $f(x)$. This section's result gives a precise rendering of this intuition.

If $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}^m$ is C^1, a point $x \in U$ is a **regular point** of f if the image of $Df(x)$ is all of $\mathbb{R}^m$. This can't happen when

$n < m$. When $n = m$ it is the same as $Df(x)$ being nonsingular: $Df(x)$ has a nonzero determinant, and is a linear isomorphism. We now fix an order of differentiability $1 \leq r \leq \infty$. A C^r **diffeomorphism** is a bijection $f : U \to V$, where U and V are open subsets of $\mathbb{R}^n$ for some n, such that f and f^{-1} are C^r.

Theorem 7.14 (Inverse Function Theorem)**.** *If $f : U \to \mathbb{R}^n$ is C^r, where $U \subset \mathbb{R}^n$ is open, and x is a regular point of f, then there is an open neighborhood $W \subset U$ of x such that $f|_W$ is a C^r diffeomorphism between W and $f(W)$.*

Hopefully this seems plausible, perhaps almost obvious, if you look again at Figure 6.3. But the proof will be a minor adventure, and we should say a bit about where the difficulty lies. Once we have an open neighborhood W of x that is mapped bijectively onto a neighborhood of $f(x)$, it's not that hard to show that $(f|_W)^{-1}$ is C^r:

Lemma 7.15. *Suppose that $U \subset \mathbb{R}^n$ is open, $f : U \to \mathbb{R}^n$ is C^r and maps U bijectively onto $f(U)$, $f(U)$ is open, and every point of U is a regular point of f. Then f^{-1} is C^r.*

Proof. Fix $y \in f(U)$. Theorem 6.15 tells us that f^{-1} is differentiable at y, and that $Df^{-1}(y) = Df(f^{-1}(y))^{-1}$. In view of Cramer's rule (Theorem 5.19) the entries of the matrix of $Df^{-1}(y)$ are rational functions[2] of the entries of the matrix of $Df(f^{-1}(y))$ (namely the partials $\frac{\partial f_j}{\partial x_j}(f^{-1}(y))$) and are consequently continuous functions of y, so f^{-1} is C^1. When $r \geq 2$, the chain rule and the basic facts concerning differentiation of sums, products, and quotients allow you to compute the second order partials of f^{-1} as (*extremely* complicated) rational functions of the first order partials of f^{-1} and the partials, up to order 2, of f, and again these functions are continuous. This reasoning can be repeated inductively up to order r. □

We will also give a complex version of the implicit function theorem whose proof requires the complex version of this result.

Lemma 7.16. *Suppose that $U \subset \mathbb{C}^n$ is open, $f : U \to \mathbb{C}^n$ is holomorphic and maps U bijectively onto $f(U)$, $f(U)$ is open, and every point of U is a regular point of f. Then f^{-1} is holomorphic.*

The proof that f^{-1} is C^1 is exactly the same, word for word; this is possible because Theorem 6.15 and Cramer's rule were proved for sufficiently general

[2] A **rational function** is a quotient of polynomials.

fields. (Of course Theorem 7.12 says that f^{-1} is C^∞ and analytic, but these properties play no role in the logic of this section's argument.)

It is also not terribly difficult to show that, in the setting of the inverse function theorem, the restriction of the function to some neighborhood of the point in question is injective. From a technical perspective the key idea is given by the following multidimensional generalization of the mean value theorem. Let $U \subset \mathbb{R}^n$ be an open set containing the image of the path

$$\gamma : s \mapsto (1-s)x_0 + sx_1$$

between two of its points x_0 and x_1, and let $g : U \to \mathbb{R}^m$ be a C^1 function. Suppose that as you travel from x_0 to x_1 on the path γ, your shadow travels from $g(x_0)$ to $g(x_1)$ along the path $g \circ \gamma$. Your path is direct while your shadow's path may meander, so there must be some point during this journey when the ratio of your shadow's speed to your speed is at least as large as the ratio $\|g(x_1) - g(x_0)\|/\|x_1 - x_0\|$ of distances traveled. In addition, by the operator norm inequality, at each time s the speed of $g \circ \gamma$ cannot be greater than $\|Dg(\gamma(s))\|$ times the speed of γ, so:

Proposition 7.17. *In the setting described above,*

$$\|g(x_1) - g(x_0)\| \le \|x_1 - x_0\| \cdot \sup_{0 \le s \le 1} \|Dg((1-s)x_0 + sx_1)\|.$$

Proof. Define $\phi : [0,1] \to \mathbb{R}$ by letting

$$\phi(s) := \big\langle g((1-s)x_0 + sx_1) - g(x_0), g(x_1) - g(x_0)\big\rangle.$$

The chain rule implies that ϕ is C^1, so the mean value theorem gives a t strictly between 0 and 1 such that

$$\phi'(t) = \phi(1) - \phi(0) = \|g(x_1) - g(x_0)\|^2.$$

The function $w \mapsto \langle w, g(x_1) - g(x_0)\rangle$ is linear, so it is its own derivative, and in view of this the chain rule yields

$$\phi'(t) = \big\langle Dg((1-t)x_0 + tx_1)(x_1 - x_0), g(x_1) - g(x_0)\big\rangle.$$

Combining these, then applying the Cauchy-Schwartz inequality, yields

$$\|g(x_1) - g(x_0)\|^2 \le \|Dg((1-t)x_0 + tx_1)(x_1 - x_0)\| \cdot \|g(x_1) - g(x_0)\|,$$

after which we can divide by $\|g(x_1) - g(x_0)\|$ and apply the operator norm inequality. □

Suppose $U \subset \mathbb{R}^n$ is open, $f : U \to \mathbb{R}^n$ is C^1, and x is a regular point of f. We wish to use this result to show that $f|_V$ is injective when $V \subset U$ is a small convex neighborhood of x (e.g., the open ball of radius ε for some $\varepsilon > 0$). Let $g : U \to \mathbb{R}^n$ be the function $g(y) := f(y) - Df(x)y$. Then

$$f(x_1) - f(x_0) = Df(x)(x_1 - x_0) + g(x_1) - g(x_0)$$

for any $x_0, x_1 \in V$, and the triangle inequality gives

$$\|f(x_1) - f(x_0)\| \geq \|Df(x)(x_1 - x_0)\| - \|g(x_1) - g(x_0)\|.$$

To bound the first term on the right hand side we observe that

$$\|Df(x)(x_1 - x_0)\| = \|x_1 - x_0\| \cdot \left\| Df(x) \frac{x_1 - x_0}{\|x_1 - x_0\|} \right\|$$

$$\geq \|x_1 - x_0\| \cdot \min_{v \in S^{n-1}} \|Df(x)v\|,$$

where $S^{n-1} := \{ v \in \mathbb{R}^n : \|v\| = 1 \}$ is the unit sphere centered at the origin of $\mathbb{R}^n$. Of course S^{n-1} is closed and bounded, hence compact, and the function $v \mapsto \|Df(x)v\|$ is continuous, so (Theorem 3.48) it has a minimizer, and it is positive at any minimizer because $Df(x)$ is nonsingular. The result above implies that there is an s between 0 and 1 such that

$$\|g(x_1) - g(x_0)\| \leq \|Dg((1-s)x_0 + sx_1)\| \cdot \|x_1 - x_0\|,$$

and Dg is continuous, so when V is sufficiently small we have

$$\|Dg((1-s)x_0 + sx_1)\| = \|Df((1-s)x_0 + sx_1) - Df(x)\| < \min_{v \in S^{n-1}} \|Df(x)v\|,$$

in which case $\|f(x_1) - f(x_0)\| > 0$ and $f(x_1) \neq f(x_0)$.

In proving the inverse function theorem, the really tough nut is showing that some open neighborhood of x is mapped *onto* a neighborhood of $f(x)$. Our proof will take advantage of the assumption that f is C^1, but it turns out that this hypothesis is unnecessary, by virtue of a famous result called **invariance of domain** due to L. E. J. Brouwer (1881-1966) that is good to know about, even though we won't be able to prove it here. This result asserts that if $U \subset \mathbb{R}^n$ is open and $f : U \to \mathbb{R}^n$ is continuous and injective, then $f(U)$ is open and f^{-1} is continuous, so that f is a homeomorphism onto its image.

We'll use a fixed point theorem to deal with this aspect of the proof. If $f : X \to X$ is a function from a set X to itself, a **fixed point** of f is a point x^* that is mapped to itself—that is, $f(x^*) = x^*$—and a **fixed point theorem** is a result asserting that, under certain hypotheses, a fixed point necessarily exists.

If (X, d) is a metric space, a function $c : X \to X$ is a **contraction** if there is a number α strictly between 0 and 1 such that

$$d(c(x), c(x')) \leq \alpha d(x, x') \quad \text{for all } x, x' \in X.$$

The **modulus of contraction** of c is the greatest lower bound of the set of α such that this inequality holds for all x, x'. If, for some particular x, x', this inequality fails for some α, then it also fails for slightly larger α, so it must hold for all x, x' when α is equal to the modulus of contraction.

Theorem 7.18 (Contraction Mapping Theorem)**.** *If (X, d) is a metric space and $c : X \to X$ is a contraction, then c has at most one fixed point. If, in addition, (X, d) is nonempty and complete, then a fixed point exists.*

Proof. Let α be the modulus of contraction of c. We first prove uniqueness. If x^* and x^{**} are both fixed points, then

$$d(x^*, x^{**}) = d(c(x^*), c(x^{**})) \leq \alpha d(x^*, x^{**}).$$

Since $\alpha < 1$, this inequality implies that $d(x^*, x^{**}) = 0$, so $x^* = x^{**}$.

To prove existence we construct a sequence $x_0, x_1, x_2, \dots$ inductively by letting x_0 be any point of X and setting $x_{i+1} := c(x_i)$ for each i. Then

$$d(x_i, x_{i+1}) \leq \alpha d(x_{i-1}, x_i) \leq \alpha^2 d(x_{i-2}, x_{i-1}) \leq \cdots \leq \alpha^i d(x_0, x_1).$$

It follows that the sequence is Cauchy, because if $j > i$, then

$$\begin{aligned} d(x_i, x_j) &\leq d(x_i, x_{i+1}) + \cdots + d(x_{j-1}, x_j) \\ &\leq (\alpha^i + \cdots + \alpha^{j-1}) d(x_0, x_1) = \frac{\alpha^i - \alpha^j}{1 - \alpha} d(x_0, x_1). \end{aligned}$$

Since X is complete, the sequence has a limit x^*. For each i we have

$$d(c(x^*), x^*) \leq d(c(x^*), x_i) + d(x_i, x^*) \leq \alpha d(x^*, x_{i-1}) + d(x_i, x^*).$$

Both terms goes to zero as $i \to \infty$, so $d(c(x^*), x^*) = 0$ and $c(x^*) = x^*$. □

While we are discussing fixed points we should mention the following very famous result.

Theorem 7.19 (Brouwer Fixed Point Theorem)**.** *If $D = \{\, x \in \mathbb{R}^n : \|x\| \leq 1 \,\}$ and $f : D \to D$ is continuous, then f has a fixed point.*

Originally proved in 1910 by Brouwer, this is typically regarded as one of the most important results coming out of the subfield of topology called algebraic topology. Among other things, once Brouwer's fixed point theorem is known it is not so hard to prove invariance of domain. Nowadays various methods of proof are known, but unfortunately none of them is simple enough to present here.

We're now ready to prove the inverse function theorem. Suppose that x is a regular point of a C^r function $f : U \to \mathbb{R}^n$ where $U \subset \mathbb{R}^n$ is open. Of course our goal is to find an open neighborhood W of x such that $f|_W$ is a C^r diffeomorphism onto its image. If we know how to do this when x and $f(x)$ are both the origin, then we can obtain the general case by applying this special case to the function $v \mapsto f(x+v) - f(x)$, so we can assume that $x = 0 = f(x)$.

If $L : \mathbb{R}^n \to \mathbb{R}^n$ is linear and nonsingular, $W \subset U$ is an open neighborhood of the origin, and $(L \circ f)|_W$ is a C^r diffeomorphism onto its image, then $f|_W = L^{-1} \circ (L \circ f)|_W$ itself is a C^r diffeomorphism onto its image because it is a composition of C^r diffeomorphisms. This means that we are free to prove the result with f replaced by $L \circ f$ for any such L, and $L = Df(0)^{-1}$ is the choice that works well. The chain rule gives

$$D(Df(0)^{-1} \circ f)(0) = Df(0)^{-1} \circ Df(0) = \mathrm{Id}^{\mathbb{R}^n},$$

so the upshot of this line of reasoning is that it suffices to prove the result when $Df(0) = \mathrm{Id}_{\mathbb{R}^n}$.

Since f is C^r, the entries of the matrix of $Df(x)$ are continuous functions of x, and the determinant is a continuous (in fact polynomial) function of its entries. Therefore the set of regular points is an open subset of U containing x, and the theorem will follow if we can show that it holds when U is replaced with this set.

In view of all this, to establish the inverse function theorem in full generality it suffices to prove the following special case.

Proposition 7.20. *If $U \subset \mathbb{R}^n$ is an open set containing the origin and $f : U \to \mathbb{R}^n$ is a C^r function such that $f(0) = 0$ and $Df(0) = \mathrm{Id}_{\mathbb{R}^n}$, and every point of U is a regular point of f, then there is an open $W \subset U$ containing the origin such that $f(W)$ is open and $f|_W : W \to f(W)$ is a C^r diffeomorphism.*

Proof. For $y \in \mathbb{R}^n$ let $A^y : U \to \mathbb{R}^n$ be the function

$$A^y(x) := y + x - f(x).$$

Note that x is a fixed point of A^y if and only if $f(x) = y$. Since

$$DA^y(x) = \mathrm{Id}_{\mathbb{R}^n} - Df(x),$$

and in particular $DA^y(0) = 0$, it seems reasonable to hope that if y is close to the origin, then the restriction of A^y to some neighborhood of the origin is a contraction mapping this neighborhood into itself.

The continuity of Df allows us to choose $r > 0$ with $\mathbf{U}_r(0) \subset U$ and $\|DA^y(x)\| < \frac{1}{2}$ for all $x \in \mathbf{U}_r(0)$ and all $y \in \mathbb{R}^n$. Consider a particular $y \in \mathbf{U}_{r/2}(0)$ and $x, x' \in \mathbf{U}_r(0)$. Since $\mathbf{U}_r(0)$ is convex, it contains the line segment between x and x', so Proposition 7.17 gives a number t strictly between 0 and 1 such that

$$\|A^y(x) - A^y(x')\| \le \|DA^y((1-t)x + tx')\| \cdot \|x - x'\| \le \tfrac{1}{2}\|x - x'\|.$$

Thus $A^y|_{\mathbf{U}_r(0)}$ is a contraction. It maps $\mathbf{U}_r(0)$ to itself because for each x in this set we can apply the last inequality, obtaining

$$\|A^y(x)\| \le \|A^y(x) - A^y(0)\| + \|A^y(0)\| \le \tfrac{1}{2}\|x\| + \|y\| < \tfrac{1}{2}r + \tfrac{1}{2}r = r.$$

The contraction mapping theorem implies that $A^y|_{\mathbf{U}_r(0)}$ has a unique fixed point.

Thus for each $y \in \mathbf{U}_{r/2}(0)$ there is a unique $x \in \mathbf{U}_r(0)$ such that $f(x) = y$. Let $W := \mathbf{U}_r(0) \cap f^{-1}(\mathbf{U}_{r/2}(0))$. Then $f|_W$ is a bijection between W and $\mathbf{U}_{r/2}(0)$. Of course W contains the origin because $f(0) = 0$, and it is open because f is continuous. Since W contains only regular points of f, Lemma 7.15 implies that $f|_W$ is a C^r diffeomorphism. □

The inverse function theorem is valid for holomorphic functions, and for real analytic functions. Instead of proving these variants from scratch we use bootstrap arguments. Extending our terminology to the complex case, if $U \subset \mathbb{C}^n$ is open and $f : U \to \mathbb{C}^m$ is a function, we will say that $x \in U$ is a **regular point** of f if f is differentiable (in the complex sense) at x and the image of $Df(x)$ is all of $\mathbb{C}^m$. A **holomorphic diffeomorphism** is a bijection $f : U \to V$, where U and V are open subsets of some $\mathbb{C}^n$, such that f and f^{-1} are holomorphic.

Theorem 7.21. *If $U \subset \mathbb{C}^n$ is open, $f : U \to \mathbb{C}^n$ is holomorphic, and $z \in U$ is a regular point of f, then there is an open neighborhood $W \subset U$ of z such that $f|_W$ is a holomorphic diffeomorphism onto its image.*

Proof. As we emphasized throughout this chapter, if we think of f as a function mapping an open subset of $\mathbb{R}^{2m}$ into $\mathbb{R}^{2m}$, then it is C^1 because each of its partial derivatives (in the real sense) is the real or complex part of one of the partials in the complex sense. For each $w \in U$ we can reinterpret $Df(w)$ as a linear function from $\mathbb{R}^{2n}$ to itself, and since it is surjective when viewed as a complex function, it must also be surjective when viewed as a real function. In particular $Df(z)$ is nonsingular in the real sense. We can now apply the real version of the inverse function theorem, obtaining an open $W \subset U$ containing z such that $f(W)$ is open, $f|_W$ is injective, and for each $w \in W$, $Df(w)$ is nonsingular in the real sense and therefore also in the complex sense. Lemma 7.16 says precisely that in this circumstance $f|_W$ and $(f|_W)^{-1}$ are inverse holomorphic diffeomorphisms. □

Proposition 7.22. *If $U \subset \mathbb{R}^n$ is open, $f : U \to \mathbb{C}^n$ is real analytic, and $x \in U$ is a regular point of f, then there is an open neighborhood $W \subset U$ of x such that $f|_W$ is a real analytic diffeomorphism onto its image.*

Proof. In the last section we showed that for each $i = 1, \dots, n$ there is an open $\tilde{U}_i \subset \mathbb{C}$ and a holomorphic function $\tilde{f}_i : \tilde{U}_i \to \mathbb{C}$ such that $U \subset \tilde{U}_i \cap \mathbb{R}^n$ and $f_i = \tilde{f}_i|_U$. Let $\tilde{U} := \tilde{U}_1 \cap \ldots \cap \tilde{U}_n$, and let

$$\tilde{f} := \tilde{f}_1|_{\tilde{U}} \times \cdots \times \tilde{f}_n|_{\tilde{U}} : \tilde{U} \to \mathbb{C}^n.$$

Applying the last result to $\tilde{f}$ gives an open neighborhood $\tilde{W} \subset \tilde{U}$ of x such that $\tilde{f}|_{\tilde{W}}$ is a holomorphic diffeomorphism onto its image. Let $W := \tilde{W} \cap \mathbb{R}^n$. Then $f|_W$ is real analytic, so it is C^∞, and the first version of the inverse function theorem allows us to replace W with a smaller open neighborhood of x such that $f|_W$ is a C^∞ diffeomorphism onto its image. Since $(\tilde{f}|_{\tilde{W}})^{-1}$ is complex analytic, $(f|_W)^{-1}$ is real analytic. □

Chapter 8

Curved Space

> *I have therefore first set myself the task of constructing the concept of a multiply extended quantity from general notions of quantity. It will be shown that a multiply extended quantity is susceptible of various metric relations, so that Space constitutes only a special case of a triply extended quantity. From this however it is a necessary consequence that the theorems of geometry cannot be deduced from general notions of quantity, but that those properties which distinguish Space from other conceivable triply extended quantities can only be deduced from experience. Thus arises the problem of seeking out the simplest data from which the metric relations of Space can be determined, a problem which by its very nature is not completely determined, for there may be several systems of simple data which suffice to determine the metric relations of Space; for the present purposes, the most important system is that laid down as a foundation of geometry by Euclid. These data are—like all data—not logically necessary, but only of empirical certainty, they are hypotheses; one can therefore investigate their likelihood, which is certainly very great within the bounds of observation, and afterwards decide upon the legitimacy of extending them beyond the bounds of observation, both in the direction of the immeasurably large, and in the direction of the immeasurably small.*
>
> –Bernhard Riemann

Riemann is my favorite mathematician. Although he proved quite a few major theorems, his most important contributions were foundational concepts that have been the focus of a great deal of research, both in mathematics and physics, ever since. This chapter introduces the notion of a manifold, what he calls a "multiply extended quantity" in this passage from his 1854 Habilitationsschrift. (The translation above is by Michael Spivak.) Manifolds are objects like the circle, spheres in various dimensions, the torus, and so forth, that resemble Euclidean space on a small scale. Understanding the possible global structures of a manifold is a fundamental issue in topology that is well understood in the two dimensional case, in large part due to concepts introduced by Riemann. Extending the most basic aspect of this analysis to higher dimensions is the topic of a famous conjecture that was proved very recently. As he points out above, it is possible for a manifold to be curved, in which case Euclidean geometry will be increasingly inaccurate as one moves to larger scales. Generalizing Gauss' work on the two dimensional case, he developed the concepts that quantify curvature in terms of measurements that can be made within the manifold itself. Sixty years later this machinery was used by Einstein to express the general theory of relativity. The structure of space and time on "immeasurably small" scales is a central issue of contemporary research in theoretical physics.

Before delving into all that, however, I'd like to mention one other very famous aspect of Riemann's work. The story begins with a beautiful proof that there are infinitely many primes due to Euler. If there were only finitely many primes the product

$$\prod_{p \text{ a prime}} \frac{1}{1-p^{-1}} = \prod_{p \text{ a prime}} \Big(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\Big)$$

would be finite. Each integer n factors as a product $p_1^{e_1} \cdots p_k^{e_k}$ of powers of primes, and $\frac{1}{n} = \frac{1}{p_1^{e_1}} \cdots \frac{1}{p_k^{e_k}}$ appears exactly once in the distributive expansion of the right hand side, so we could conclude that

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots < \infty,$$

but this is false[1]. As Euler observed, this argument actually proves something much stronger and more interesting, namely that the sum

$$\sum_{p \text{ a prime}} \tfrac{1}{p} = \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{5} + \tfrac{1}{7} + \tfrac{1}{11} + \cdots$$

[1] To prove that the **harmonic series** $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ diverges observe that

$$\sum_n \frac{1}{n} = \sum_{i=0}^{\infty} \Big(\sum_{2^{i-1} < n \leq 2^i} \frac{1}{n} \Big) > \sum_{i=0}^{\infty} \Big(\sum_{2^{i-1} < n \leq 2^i} \frac{1}{2^i} \Big) = \sum_{i=0}^{\infty} \frac{1}{2} = \infty.$$

diverges. To see this observe that if $0 < a \le \frac{1}{2}$, then $1 \le 1 + a - 2a^2 = (1-a)(1+2a)$, so that

$$(1-a)^{-1} \le 1 + 2a < \sum_{j=0}^{\infty} (2a)^j / j! = \exp(2a).$$

If $0 < a_1, a_2, a_3, \ldots \le \frac{1}{2}$ and $\prod_{n=1}^{\infty} (1-a_n)^{-1} = \infty$, then $\sum_{n=1}^{\infty} a_n = \infty$ because for any integer N we have

$$\prod_{n=1}^{N} (1-a_n)^{-1} \le \prod_{n=1}^{N} (1+2a_n) < \prod_{n=1}^{N} \exp(2a_n) = \exp\big(2 \sum_{n=1}^{N} a_n\big).$$

Euler went on to make some clever guesses about the rate of divergence of the sequence $\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11} + \cdots$ by "taking the logarithm" of both sides of the "equation"

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots = \prod_p \frac{1}{1-p^{-1}}.$$

Of course this isn't rigorous, but he also pointed out that for any $s > 1$ the equation

$$1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots = \prod_p \frac{1}{1-p^{-s}} = \prod_p \Big(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots\Big)$$

is valid, so one can try to develop these ideas precisely by studying the asymptotic behavior of the function

$$\zeta(s) := \prod_p \frac{1}{1-p^{-s}}$$

as $s \to 1$ from above.

This is called the **Riemann zeta-function** because, even though Euler originated it, Riemann had the idea of treating ζ as a function of a complex variable. Riemann was able to show that the function defined by the formula above is the restriction to $\{\, s \in \mathbb{R} : s > 1 \,\}$ of an analytic function defined on all of $\mathbb{C} \setminus \{1\}$. (Of course the principle of analytic continuation (Theorem 7.13) implies that there is at most one such analytic function.) It turns out that issues concerning the distribution of prime numbers are intimately related to the location of the zeros of ζ. Riemann showed that aside from zeros at $-2, -4, -6, \ldots$, all the zeros lie in the **critical strip**

$\{ s = x + iy : 0 \leq x \leq 1 \}$, and he conjectured that they actually all lie on the line $x = \frac{1}{2}$. This conjecture is now called the **Riemann hypothesis**. It was included by Hilbert in his list of problems, and it is one of the Clay Mathematics Institute's seven Millenium Prize problems. There is something of a consensus that it is currently the most important open problem in mathematics.

8.1 Manifolds

We start with the idea that a "multiply extended quantity" should look like a Euclidean space in a neighborhood of each of its points. Recall that a homeomorphism is a bijection between two topological spaces that is continuous and whose inverse is continuous. At first glance "has a neighborhood homeomorphic to an open subset of a Euclidean space" seems like a reasonable rendering of "looks like a Euclidean space locally," but there will be a bit more to it.

Definition 8.1. *Let n be a nonnegative integer. An n-dimensional* ***manifold*** *is a Hausdorff topological space M such that for each $p \in M$ there is a open set $U \subset M$ containing p and a homeomorphism $\varphi : U \to V$ between U and an open set $V \subset \mathbb{R}^n$.*

The function φ is called a **coordinate chart**, and its inverse is called a **parameterization**. An **atlas** for M is a collection of coordinate charts $\{ \varphi_i : U_i \to V_i \}_{i \in I}$ (here I is an arbitrary index set) whose domains cover M, i.e., $\bigcup_{i \in I} U_i = M$.

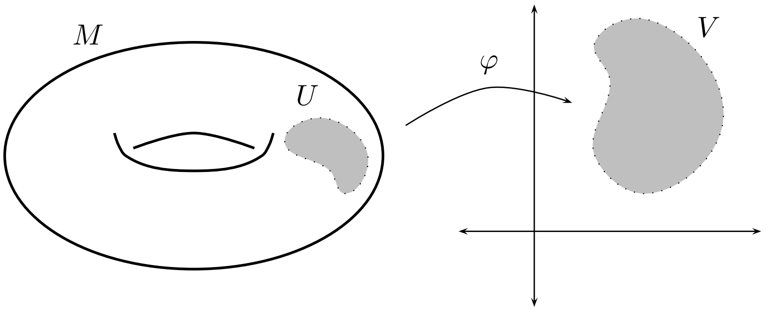


Figure 8.1

Hopefully you are developing the habit of looking at definitions critically, in which case there is a question that should be bugging you. *Why do we want M to be Hausdorff?* Consider the following topological space, which is known as the **line with two origins**. Let

$$X = \mathbb{R} \cup \{0^*\}$$

where 0^* is some object that is not an element of $\mathbb{R}$. Impose a topology on X by specifying that the open sets of X are the open (in the usual sense) subsets of $\mathbb{R}$ together with all the sets $\{0^*\} \cup U$ where $U \subset \mathbb{R}$ is open (in the usual sense) and $(-\varepsilon, 0) \cup (0, \varepsilon) \subset U$ for some $\varepsilon > 0$. (Please check that $\emptyset$ and X are open, and that finite intersections and arbitrary unions of open sets are open.) Note that X is not Hausdorff because 0 and 0^* do not have disjoint neighborhoods. The map $t \mapsto t$ is a homeomorphism between $X \setminus \{0^*\}$ and $\mathbb{R}$, as is the map from $X \setminus \{0\}$ to $\mathbb{R}$ that takes each $t \neq 0$ to itself and takes 0^* to 0. (Please check that the latter map and its inverse both take open sets to open sets.) Since every point in X is in the domain of one of these homeomorphisms, the definition above would have been satisfied by this rather obnoxious space, and others of its ilk, if the Hausdorff requirement had been omitted. Although we do not wish to study such spaces, during proofs we sometimes encounter situations in which we do not (yet) know that the space in question is Hausdorff, so the following terminology is useful: a topological space X is an n-dimensional **quasimanifold** if every point of X has a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^n$.

What this example points out is that we want every point in M to have a neighborhood that "looks like" an open set in Euclidean space not only because it is homeomorphic to such a set, but also because its closure doesn't contain any points that shouldn't be there. Specifically, if $\varphi : U \to V$ is as in the definition, and $C \subset V$ is a compact neighborhood of $\varphi(p)$, then $\varphi^{-1}(C)$ is compact because (Theorem 3.47) φ^{-1} is continuous. In addition, C is closed in V because (Theorem 3.40) V is Hausdorff, so $\varphi^{-1}(C)$ is closed in U because φ is continuous. We want $\varphi^{-1}(C)$ to be "clean" in the sense that its closure in M doesn't contain any points in $M \setminus U$. That is, we want $\varphi^{-1}(C)$ to be closed, not just in U, but also in M. Requiring M to be Hausdorff accomplishes this because (Theorem 3.40) a compact subset of a Hausdorff space is closed.

There are other senses in which one can ask whether Definition 8.1 is really what we want. For example, some rather advanced results imply that the topology of M can be derived from a metric, but if this wasn't automatic we would probably want to include "metrizability" as a requirement of our definition. Sometimes a definition is "obviously" correct in the sense that there really isn't any choice about how to express the concept in question, but there are other definitions that are, in the end, seemingly quite simple, but which are actually the result of years of research. As you might have already guessed, Riemann never actually managed to give a fully satisfactory definition of a "multiply extended quantity," and Definition 8.1 is in fact one of the crown jewels of 20th century topology.

The rest of this section enumerates some methods of constructing new manifolds from given manifolds, and some basic examples. Although this might help firm up your understanding of what the definition means, that's not really the main point. The examples here are the simplest, the most symmetric, and the most obvious manifolds one could think of. They come up "naturally" for various reasons, and they are what anyone looking for an example of something will think of first. "Everyone" knows them, so you should too.

First of all, any set X can be endowed with the **discrete topology**, which is the topology in which all subsets of X are open. Any set with the discrete topology is a 0-dimensional manifold, and any 0-dimensional manifold has the discrete topology. The main idea is simply that for any $p \in X$, $\{p\}$ is an open set containing p that is homeomorphic to $\mathbb{R}^0$.

Let M be a manifold with atlas $\{\varphi_i : U_i \to V_i\}_{i\in I}$. Any open subset W of M is itself a manifold with atlas $\{\varphi_i|_{U_i\cap W}\}_{i\in I}$. In particular, any open subset of $\mathbb{R}^n$ is a manifold. If $f : M \to \mathbb{R}^k$ is continuous, then its graph

$$\mathrm{Gr}(f) = \{\, (p, f(p)) : p \in M \,\}$$

is a manifold. To construct an atlas, for each $i \in I$ let $\tilde{U}_i := \mathrm{Gr}(f|_{U_i})$ and let $\tilde{\varphi}_i : \tilde{U}_i \to V_i$ be the function

$$\tilde{\varphi}_i(p, f(p)) := \varphi_i(p).$$

Graphs of functions from open subsets of $\mathbb{R}^2$ to $\mathbb{R}$ are easy to visualize, and they can be analyzed using the partial derivatives of the function. Historically, the first major study of manifolds as geometric objects, Gauss' *Disquisitiones Generales Circa Superficies Curvas*, was devoted to them.

Let M' be a second manifold with atlas $\{\varphi'_j : U'_j \to V'_j\}_{j\in J}$. Then $M \times M'$ (endowed with the product topology) is a manifold. In detail, a cartesian product of two Hausdorff spaces is Hausdorff, a cartesian product (in the obvious sense) of two homeomorphisms is a homeomorphism, and $\{\, U_i \times U'_j : i \in I, j \in J \,\}$ is a cover of $M \times M'$ so

$$\{\varphi_i \times \varphi'_j : U_i \times U'_j \to V_i \times V'_j\}_{(i,j)\in I\times J}$$

(with $\varphi_i \times \varphi'_j$ defined in the "obvious" way) is an atlas for $M \times M'$.

For any field k the **n-sphere** is

$$S^n(k) := \{ (p_0, \dots, p_n) \in k^{n+1} : p_0^2 + \cdots + p_n^2 = 1 \}.$$

Let

$$U_N := \{ p \in S^n(k) : p_0 \neq 1 \} \quad \text{and} \quad U_S := \{ p \in S^n(k) : p_0 \neq -1 \}.$$

When $k = \mathbb{R}$ we think of these as the Northern and Southern Hemispheres. Let $\varphi_N : U_N \to k^n$ and $\varphi_S : U_S \to k^n$ be the functions

$$\varphi_N(p) := \left(\tfrac{p_1}{1-p_0}, \dots, \tfrac{p_n}{1-p_0}\right) \quad \text{and} \quad \varphi_S(p) := \left(\tfrac{p_1}{1+p_0}, \dots, \tfrac{p_n}{1+p_0}\right).$$

These functions are called **stereographic projections**.

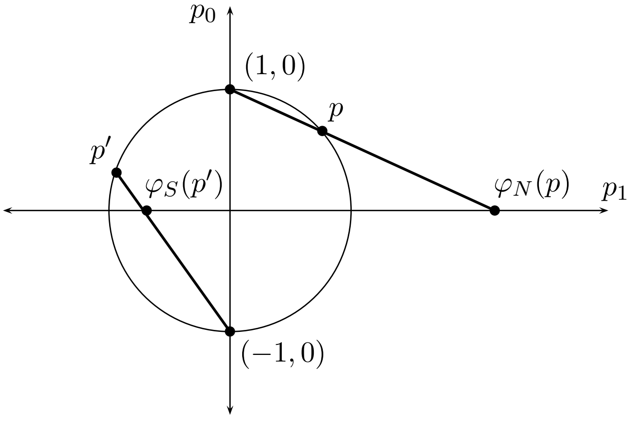


Figure 8.2

To understand φ_N geometrically, observe that

$$p = p_0(1, 0, \dots, 0) + (1 - p_0)(0, \varphi_N(p)),$$

so that $\varphi_N(p)$ consists of the last n coordinates of the point where the ray emanating from $(1, 0, \dots, 0)$ and passing through p intersects the plane $\{ y \in k^{n+1} : y_0 = 0 \}$. The description of φ_S is similar, with the rays emanating from $(-1, 0, \dots, 0)$.

For $x \in k^n$ let $\sigma(x) := x_1^2 + \cdots + x_n^2$, and let $V := \{ x \in k^n : \sigma(x) \neq -1 \}$. If $\varphi_N(p) = x$, then summing the squares of the components on both sides of the equation above gives $1 = p_0^2 + (1 - p_0)^2 \sigma(x)$. Subtracting p_0^2 from both sides and dividing by $1 - p_0$ results in $1 + p_0 = (1 - p_0)\sigma(x)$, and this reduces to $1 = -1$ if $\sigma(x) = -1$, so it must be the case that $x \in V$. When this is the case we can solve this equation for p_0 and substitute into the equation above, arriving at

$$\varphi_N^{-1}(x) = \tfrac{(\sigma(x)-1, 2x_1, \ldots, 2x_n)}{\sigma(x)+1}.$$

Symmetrical analysis shows that φ_S is a bijection between U_S and V, and that

$$\varphi_S^{-1}(x) = \tfrac{(1-\sigma(x), 2x_1, \ldots, 2x_n)}{1+\sigma(x)}.$$

If k has a topology with respect to which addition, multiplication, negation, and inversion are continuous, then φ_N and φ_S are homeomorphisms. Of course $S^n(\mathbb{R})$ and $S^n(\mathbb{C})$ are Hausdorff spaces, so we have shown that they are manifolds.

Substituting the formula for φ_N^{-1} into the formula for φ_S, and substituting the formula for φ_S^{-1} into the formula for φ_N, leads to

$$\varphi_S(\varphi_N^{-1}(x_1, \ldots, x_n)) = \tfrac{1}{\sigma(x)}(x_1, \ldots, x_n) = \varphi_N(\varphi_S^{-1}(x_1, \ldots, x_n)). \qquad (*)$$

In Section 8.4 we'll explain the relevance of this formula.

The complex n-sphere $S^n(\mathbb{C})$ actually doesn't look very much like what you probably think a sphere should look like. Among other things, it is not compact when $n > 0$ because there are points $z \in S^n(\mathbb{C})$ with $|z_0|$ arbitrarily large. Possibly for this reason, $S^n(\mathbb{C})$ doesn't come up much, and in fact most authors write S^n rather than the more cumbersome $S^n(\mathbb{R})$. We'll do the same unless confusion seems likely.

Note that we are already in a position to construct a large collection of examples by taking cartesian products. The most famous of these is the **torus** $S^1 \times S^1$ which we've already seen illustrated in Figure 8.1.

The next set of examples is probably less familiar because it doesn't figure in the secondary school curriculum, at least where I went to high school, but at a higher level it is very important. For any field k, n-dimensional **projective space** $P^n(k)$ is the set of one dimensional linear subspaces of k^{n+1}. Each line through the origin in $\mathbb{R}^{n+1}$ intersects the unit sphere at two antipodal points, and we can think of constructing $P^n(\mathbb{R})$ by identifying antipodal points of S^n. Thus $P^1(\mathbb{R})$ is a circle that has the larger circle S^1 wrapped around itself twice. I must confess that I have never seen an illustration of $P^2(\mathbb{R})$ that gave me a good idea of the space. I try to think of starting with the Northern hemisphere and sewing opposite equatorial points to each other, but my visual imagination just gets tangled up[2].

[2] In 1901 Hilbert assigned his student Werner Boy (1879-1914) the problem of showing that there is no immersion of $P^2(\mathbb{R})$ in $\mathbb{R}^3$. If M and N are C^1 manifolds and M is n-dimensional, a C^1 function (as defined in Section 8.4) $f : M \to N$ is an **immersion**

The formal description of $P^n(k)$ goes as follows. For $x \in k^{n+1} \setminus \{0\}$, the line spanned by x is

$$[x] := \{\, \alpha x : \alpha \in k \,\}.$$

For each $i = 0, \dots, n$ let

$$U_i := \{\, [x] \in P^n(k) : x_i \neq 0 \,\},$$

let $V_i := k^n$, and let $\varphi_i : U_i \to V_i$ be the function

$$\varphi_i([x]) = (\tfrac{x_0}{x_i}, \dots, \tfrac{x_{i-1}}{x_i}, \tfrac{x_{i+1}}{x_i}, \dots, \tfrac{x_n}{x_i}).$$

These definitions make sense because the truth value of the condition $x_i \neq 0$ and the formula defining φ_i are unaffected if we replace x with αx for any nonzero α.

Because S^n is a subset of $\mathbb{R}^{n+1}$, it inherits a subspace topology, which is Hausdorff, so all we had to do to show that $\{\varphi_N, \varphi_S\}$ is an atlas for S^n was to observe that the domains of these two functions cover the sphere, and each is a homeomorphism. For projective space we are doing something a bit different insofar as the maps $\varphi_0, \dots, \varphi_n$ are used to *define* the topology of $P^n(k)$ when $k = \mathbb{R}$ or $k = \mathbb{C}$. We need to think carefully about how this is done and what conditions need to be satisfied in order for the resulting space to be a manifold.

Most obviously, we need it to be the case that

$$U_0 \cup \dots \cup U_n = P^n(k).$$

For each $[x]$ there is at least one i such that $x_i \neq 0$, so this is the case.

Each φ_i must be a bijection. In fact each

$$(x_0, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \in V_i$$

is the image of

$$[x_0, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n]$$

under φ_i, so φ_i is surjective. To see that φ_i is injective observe that if $\varphi_i([x]) = \varphi_i([x'])$, then $x_i \neq 0 \neq x'_i$ and

$$(\tfrac{x_0}{x_i}, \dots, \tfrac{x_{i-1}}{x_i}, \tfrac{x_{i+1}}{x_i}, \dots, \tfrac{x_n}{x_i}) = (\tfrac{x'_0}{x'_i}, \dots, \tfrac{x'_{i-1}}{x'_i}, \tfrac{x'_{i+1}}{x'_i}, \dots, \tfrac{x'_n}{x'_i}),$$

if, for each $p \in M$, the derivative (as defined in Section 8.6) $Df(p)$ has an n-dimensional image. Boy came up with an example, known as Boy's surface, showing that Hilbert's conjecture was wrong. There are many images of Boy's surface on the internet.

so $x'_j = (x'_i/x_i)x_j$ for all $j = 0, \ldots, n$, which means that $[x] = [x']$.

For the next part of the discussion let's generalize the framework. Suppose we are given a set M, a collection of subsets $\{U_i\}_{i \in I}$ with $\bigcup_{i \in I} U_i = M$, where I is an arbitrary index set, and a bijection $\varphi_i : U_i \to V_i$ for each $i \in I$ between U_i and an open $V_i \subset \mathbb{R}^n$. We want to know what conditions have to hold in order for it to be possible to impose a topology on M that makes it a manifold with atlas $\{\varphi_i\}_{i \in I}$. And if it is possible, what is the topology?

There are some rather obvious conditions that must hold whenever M is a manifold, or even a quasimanifold, with atlas $\{\varphi_i\}_{i \in I}$. For any i and j, $U_i \cap U_j$ must be open, so (because φ_i is a homeomorphism) $\varphi_i(U_i \cap U_j)$ must be an open subset of V_i. In addition, if $W_i \subset \varphi_i(U_i \cap U_j)$ is open, then $\varphi_i^{-1}(W_i)$ is open (again, because φ_i is a homeomorphism) and[3] $(\varphi_i \circ \varphi_j^{-1})^{-1}(W_i) = \varphi_j(\varphi_i^{-1}(W_i))$ is open. That is, each $\varphi_i \circ \varphi_j^{-1}$ is continuous, and in fact a homeomorphism because its inverse is $\varphi_j \circ \varphi_i^{-1}$. It turns out that these conditions are not just "necessary," in the sense that they hold whenever M is a quasimanifold with atlas $\{\varphi_i\}_{i \in I}$, but also "sufficient" in the sense that M is a quasimanifold with atlas $\{\varphi_i\}_{i \in I}$ whenever they hold.

Proposition 8.2. *If M is a set, $\{U_i\}_{i \in I}$ a collection of subsets of M whose union is all of M, and, for each i, $\varphi_i : U_i \to V_i$ is a bijection between U_i and an open $V_i \subset \mathbb{R}^n$, then the following are equivalent:*

(a) for each $i, j \in I$, $\varphi_i(U_i \cap U_j)$ is open and

$$\varphi_i \circ \varphi_j^{-1} : \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$$

is a homeomorphism;

(b) the collection τ of all sets of the form $\bigcup_{i \in I} \varphi_i^{-1}(W_i)$, where each $W_i \subset V_i$ is open, is a topology, and M endowed with this topology is an n-dimensional quasimanifold.

Proof. We have already seen that (b) implies (a), so all we have to do is show that (a) implies (b). Our first task to show that τ actually is the collection of open sets of a topology. Obviously τ includes $\emptyset$ and M itself. If A is an index set and, for each $\alpha \in A$ and $i \in I$, $W_{i,\alpha}$ is an open subset of V_i, then

$$\bigcup_{\alpha \in A} \Big(\bigcup_{i \in I} \varphi_i^{-1}(W_{i,\alpha}) \Big) = \bigcup_{i \in I} \varphi_i^{-1} \Big(\bigcup_{\alpha \in A} W_{i,\alpha} \Big)$$

[3] Here, in order to reduce clutter, we are "abusing notation" by writing φ_j^{-1} in place of $\varphi_j^{-1}|_{\varphi_j(U_i \cap U_j)}$. There will be similar abuses throughout the remainder of the book when confusion seems unlikely.

is an element of τ because each $\bigcup_{\alpha\in A} W_{i,\alpha}$ is open. Therefore τ contains arbitrary unions of its elements. If, for each i, W_i and W_i' are open subsets of V_i, then

$$\Big(\bigcup_{i\in I}\varphi_i^{-1}(W_i)\Big)\cap\Big(\bigcup_{j\in I}\varphi_j^{-1}(W_j')\Big)=\bigcup_{i\in I}\bigcup_{j\in I}\big(\varphi_i^{-1}(W_i)\cap\varphi_j^{-1}(W_j')\big)$$

$$=\bigcup_{i\in I}\Big(\bigcup_{j\in I}\varphi_i^{-1}\big(W_i\cap\varphi_i(\varphi_j^{-1}(W_j'))\big)\Big)$$

is an element of τ because each $\varphi_i(\varphi_j^{-1}(W_j'))$ is open and τ contains unions of its elements. Therefore τ contains the intersection of any two of its elements.

It remains to show that each φ_i is a homeomorphism. For each i,

$$\varphi_i^{-1}(W_i)=\varphi_i^{-1}(W_i)\cup\Big(\bigcup_{j\neq i}\varphi_j^{-1}(\emptyset)\Big)\in\tau$$

whenever $W_i\subset V_i$ is open, so φ_i is continuous. If, for each j, $W_j\subset V_j$ is open, then

$$\varphi_i\Big(\bigcup_{j\in I}\varphi_j^{-1}(W_j)\Big)=\bigcup_{j\in I}\varphi_i(\varphi_j^{-1}(W_j))$$

is open because each $\varphi_j\circ\varphi_i^{-1}$ is continuous, so φ_i^{-1} is continuous. □

It is worth pointing out that when (a) holds, τ is the only topology such that each φ_i is a homeomorphism. The continuity of the various φ_i implies that each element of τ must be open in M. Conversely, any open set $U\subset M$ must be in τ because each $U\cap U_i$ and $\varphi_i(U\cap U_i)$ are open, and

$$U=\bigcup_{i\in I}\varphi_i^{-1}(\varphi_i(U\cap U_i)).$$

When is M a Hausdorff space? The issue is clarified by the following characterization of Hausdorff spaces in terms of the product topology.

Proposition 8.3. *If X is a topological space, then X is Hausdorff if and only if the **diagonal***

$$\Delta:=\{\,(x,x):x\in X\,\}$$

is a closed subset of $X\times X$ (endowed with the product topology).

Proof. In the product topology $(X\times X)\setminus\Delta$ is open if and only if for any $(x,x')\in(X\times X)\setminus\Delta$ there are open neighborhoods U and U' of x and x' with $(U\times U')\cap\Delta=\emptyset$, which is the same as saying that $U\cap U'=\emptyset$. □

Suppose $\{U_i\}_{i\in I}$ is an open cover of X. Then $\{U_i \times U_j\}_{i,j\in I}$ is an open cover of $X \times X$, and since closedness is a local property (Proposition 3.24) X is Hausdorff if and only if each $(U_i \times U_j) \cap \Delta$ is relatively closed in $U_i \times U_j$. Under the conditions given by Proposition 8.4 above, each $\varphi_i \times \varphi_j : U_i \times U_j \to V_i \times V_j$ is a homeomorphism, so:

Proposition 8.4. *If M is an n-dimensional quasimanifold, $\{U_i\}_{i\in I}$ is an open cover, and for each i, $\varphi_i : U_i \to V_i$ is a homeomorphism, where $V_i \subset \mathbb{R}^n$ is open, then the following are equivalent:*

(a) for each $i, j \in I$,

$$C_{ij} := \{\, (\varphi_i(p), \varphi_j(p)) : p \in U_i \cap U_j \,\}$$

is closed in $V_i \times V_j$;

(b) M is a Hausdorff space, hence an n-dimensional manifold.

To get a concrete image of how (a) can fail to hold, recall our two coordinate charts for the line with two origins.

We need to check that the conditions in the last two results hold in the specific case of $P^n(\mathbb{R})$ and $P^n(\mathbb{C})$. Consider i and j between 0 and n. Then

$$\varphi_i(U_i \cap U_j) = \{\, (x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) : x_j \neq 0 \,\}$$

is open, and $\varphi_i \circ \varphi_j^{-1}$ is continuous because

$$\varphi_i(\varphi_j^{-1}(y_0, \ldots, y_{j-1}, y_{j+1}, \ldots, y_n)) = \varphi_i([y_0, \ldots, y_{j-1}, 1, y_{j+1}, \ldots, y_n])$$

$$= (\tfrac{y_0}{y_i}, \ldots, \tfrac{y_{i-1}}{y_i}, \tfrac{y_{i+1}}{y_i}, \ldots, \tfrac{y_{j-1}}{y_i}, \tfrac{1}{y_i}, \tfrac{y_{j+1}}{y_i}, \ldots, \tfrac{y_n}{y_i}). \qquad (**)$$

To show that C_{ij} is closed, suppose that $[x^1], [x^2], \ldots$ is a sequence in $U_i \cap U_j$ with

$$\big(\varphi_i([x^k]), \varphi_j([x^k])\big) \to (y^i, y^j) \in V_i \times V_j.$$

We need to find $[x] \in U_i \cap U_j$ with $(\varphi_i([x]), \varphi_j([x])) = (y^i, y^j)$. Suppose

$$y^i = (a_0, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \text{ and } y^j = (b_0, \ldots, b_{j-1}, b_{j+1}, \ldots, b_n).$$

Since φ_i and φ_j are homeomorphisms, $[x]$ must be the limit of the sequence $\{[x^k]\}$. Each $[x^k]$ is in $U_i \cap U_j$, so there are numbers a_h^k and b_h^k such that

$$[a_0^k, \ldots, a_{i-1}^k, 1, a_{i+1}^k, \ldots, a_n^k] = [x^k] = [b_0^k, \ldots, b_{j-1}^k, 1, b_{j+1}^k, \ldots, b_n^k],$$

and we have

$$[a_0^k,\dots,a_{i-1}^k,1,a_{i+1}^k,\dots,a_n^k] \to \varphi_i^{-1}(y^i) = [a_0,\dots,a_{i-1},1,a_{i+1},\dots,a_n]$$

and

$$[b_0^k,\dots,b_{j-1}^k,1,b_{j+1}^k,\dots,b_n^k] \to \varphi_j^{-1}(y^j) = [b_0,\dots,b_{j-1},1,b_{j+1},\dots,b_n].$$

The key point is that $a_j \neq 0$. To see this observe that

$$a_j^k = a_j^k/a_i^k = b_j^k/b_i^k = 1/b_i^k,$$

so if $a_j = 0$, then $b_i^k \to \infty$, and of course this is impossible. Therefore $a_j^k \neq 0$ when k is sufficiently large, in which case $b_h^k = b_h^k/b_j^k = a_h^k/a_j^k$ for all $h = 0,\dots,n$, and

$$b_h = \lim_{k\to\infty} b_h^k = \lim_{k\to\infty} a_h^k/a_j^k = a_h/a_j$$

so

$$[a_0,\dots,a_{i-1},1,a_{i+1},\dots,a_n] = [b_0,\dots,b_{j-1},1,b_{j+1},\dots,b_n].$$

That is, $\varphi_i^{-1}(y^i) = \varphi_j^{-1}(y^j)$, and consequently $(y^i,y^j) \in C_{ij}$.

In case you're wondering, the ideas used to construct projective space aren't restricted to one dimensional subspaces. For $0 < m < n$ the set of m-dimensional linear subspaces of k^n is called the **Grassmannian** of m-planes in k^n. There are various ways of defining an atlas on the Grassmannian, but it seems that each of them involves a lot of work. Since the conclusion in question—that the Grassmannian is a manifold—is quite intuitive (more precisely, it would be shocking if it was false) we won't explore the matter further.

8.2 Differentiable Manifolds

There are various sorts of manifolds. The definition in the last section describes what is sometimes called a **topological manifold**, in contrast with more highly structured types. The variants defined in this section can be usefully understood as fitting into a larger conceptual framework that we now describe.

During the 19th century the understanding of the fundamental nature of geometry shifted in important ways. Prior to that era, the geometry of the universe had been thought to be largely a matter of logical necessity, to the point where many people felt that Euclid's Parallel Postulate should be a logical consequence of the other axioms. The discovery of spaces satisfying the other axioms, but not the Parallel Postulate (we'll see one in Section 9.2) refuted that view, and opened the door to the investigation of geometry in a wide variety of spaces, using a variety of axiomatic frameworks.

There then arose the question of how to organize this knowledge. In 1872 Felix Klein (1849-1925) proposed a system for classifying geometric theories that came to be known as the Erlangen Program. A fundamental concept in this approach is a collection (often a group) of allowed coordinate transformations. Given such a collection, the meaningful geometrical concepts are those that are preserved by the coordinate transformations. A large collection of transformations entails a small set of very general concepts, while a small collection allows a rich geometry.

To illustrate this concretely consider $\mathbb{R}^n$. The inverse of a homeomorphism is a homeomorphism, and a composition of two homeomorphisms is a homeomorphism, so, for any space, the homeomorphisms from the space to itself are a group with composition as the group operation. (This is a particular instance of Theorem 1.7.) The group of all homeomorphisms $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ is the largest group of symmetries of $\mathbb{R}^n$ that one would usually care to consider. The meaningful geometric concepts in this framework are those that are preserved by homeomorphisms, for example connectedness.

An **affine transformation** of $\mathbb{R}^n$ is a map

$$x \mapsto \ell(x) + x_0$$

in which $\ell : \mathbb{R}^n \to \mathbb{R}^n$ is a nonsingular linear transformation and x_0 is some point in $\mathbb{R}^n$. The inverse

$$y \mapsto \ell^{-1}(y) - \ell^{-1}(x_0)$$

is also an affine transformation, and it is easy to see that compositions of affine transformations are affine transformations, so the affine transformations are a subgroup of the group of homeomorphisms from $\mathbb{R}^n$ to itself. An affine transformation maps a d-dimensional affine subspace to another d-dimensional affine subspace, so concepts like "line," "plane," and "parallel" are meaningful in the associated geometry, but distances and angles are not preserved by affine transformations.

A **Euclidean motion** is an affine transformation $x \mapsto \ell(x)+x_0$ in which ℓ is an orthogonal transformation. It is easy to show that inverses and compositions of orthogonal transformations are orthogonal transformations, and in turn it is easy to pass from this result to the conclusion that inverses and compositions of Euclidean motions are Euclidean motions, so the Euclidean motions constitute a subgroup of the group of affine transformations. The quantities preserved by Euclidean motions, and concepts derived from them, are what we usually think of as the subject matter of geometry.

At least since Einstein, an important method in physics is to investigate the group of symmetries that preserve the structure of physical laws. (Earlier physicists were, of course, aware that Newtonian physics exhibited symmetries, but systematic exploitation of this fact only became possible with the development of group theory.) In our discussion of the transformations expressing these symmetries, and the relativistic transformations that replaced them, we will consider one dimensional space. (This simplification does not do violence to any of the key ideas.) One dimensional Newtonian physics is invariant under all transformations $(x,t) \mapsto (x',t')$ of space and time of the form

$$x' = x_0 + x + tv \quad \text{and} \quad t' = t_0 + t.$$

Here x_0 and t_0 are parameters reflecting translations of the origins of space and time respectively, and v gives the relative velocities of the two coordinate systems. An **inertial trajectory** is the graph of an affine function mapping time into space. Any physically valid coordinate system should respect the principle of inertia—in the absence of an external force, a particle maintains a constant velocity—so it should map inertial trajectories to inertial trajectories, and this is reflected in the affine character of the transformation above. The other physical ideas expressed by this group of transformations are that distance and time are distinct concepts, and each of them has an absolute character that is independent of the observer's frame of reference. That is, for any two points (x,t) and $(\tilde{x},\tilde{t})$ the distance and the time elapsed between them are the same in the two coordinate systems because

$$\tilde{x} - x = \tilde{x}' - x' \quad \text{and} \quad \tilde{t} - t = \tilde{t}' - t'.$$

(A physicist would, quite properly, regard this explanation as incomplete because it does not specify procedures for measuring distances and time intervals.)

Until the late 19th century it hadn't occurred to physicists that other interesting groups of transformations might be "nearby" in the sense of giving a similar description of the physics of low velocities. A **Lorentz transformation** is a linear change of coordinates of the form

$$(x',t') = \Big(\frac{x + vt}{\sqrt{1 - v^2/c^2}}, \frac{t + vx/c^2}{\sqrt{1 - v^2/c^2}}\Big)$$

where c is a constant that is the speed of light in physical applications. (A completely general description would include translation parameters x_0 and t_0, as above. How to include them is obvious, and these parameters would tend to complicate the notation and computations without adding interesting or challenging concepts to the discussion, so typically one restricts attention to those transformation in which these parameters are zero.) The inverse of this transformation is

$$(x,t) = \Big(\frac{x' - vt'}{\sqrt{1 - v^2/c^2}}, \frac{t' - vx'/c^2}{\sqrt{1 - v^2/c^2}}\Big),$$

which is also a Lorentz transformation. Proving that the composition of two Lorentz transformations is, in turn, a Lorentz transformation, is a straightforward but bulky calculation that we omit. Taking this as given, we see that the Lorentz transformations constitute a group.

Several properties of Lorentz transformations should be noted. A Lorentz transformation is affine, so it respects the principle of inertia. If we look at the limit as $c \to \infty$ we recover a Newtonian transformation. When v is much smaller than c, $\sqrt{1 - v^2/c^2}$ is very close to one, so Newtonian physics is accurate with respect to velocities that are much smaller than the speed of light. The experimental observation that prompted the development of the theory of relativity was the Michelson-Morley experiment, which showed that the speed of light is independent of the frame of reference of the observer. The Lorentz transformation above respects this insofar as $x' = ct'$ whenever $x = ct$ and $x' = -ct'$ whenever $x = -ct$, as you can check for yourself.

There is a precise sense in which the Lorentz transformation can be derived from the principle of inertia and the invariance of the speed of light. Consequently there is really no good way to avoid the counterintuitive aspects of the special theory of relativity, namely that time and space combine to form a single entity called **space-time**, and are no longer absolute insofar as perceived distances and time intervals depend on the velocity of the observer.

In the general theory of relativity the group of linear transformations described above is approximately valid locally. It characterizes the physics of phenomena that are restricted to regions of space-time that are "small" or "approximately flat" in the sense that the curvature induced by the influence of gravity is slight and can be ignored. Mathematically, the theory is expressed in terms of differential equations, which might be thought of as a matter of restricting attention to "infinitesimal" regions. Roughly speaking, general relativity describes the influence of gravity as a matter of the larger scale structure of space-time being curved. In general relativity there is no a priori assumption that space-time is homeomorphic to $\mathbb{R}^4$, so the theory must be described by imposing structure on a general four dimensional manifold.

How does one "impose structure" on a manifold? In the discussion above we saw various examples in which certain coordinate systems were regarded as "physically" or "geometrically" valid. There were certain types of transformations that passed from one valid coordinate system to another. The meaningful concepts were those that were preserved by the transformations, and in a sense the structures of interest were really embedded in the collections of transformations. Roughly, an atlas of coordinate charts $\{\varphi_i\}_{i \in I}$ is OK if each $\varphi_i \circ \varphi_j^{-1}$ is one of these transformations.

We are now going to impose a structure on a manifold that makes concepts like "differentiable function" meaningful. Our work will follow the pattern described above: there will be an atlas of "allowed" or "valid" coordinate charts, and the concept of differentiability we have in mind will be expressed by the requirement that the changes of coordinates induced by going from one coordinate chart to another are differentiable. As we will see, if a function, say from the manifold to $\mathbb{R}$, is differentiable at a point p in the perspective afforded by one of these coordinate charts, then it is differentiable at p from the point of view of every other allowed coordinate chart that has p in its domain.

A **degree of differentiability** is an element of the set

$$\{\,1, 2, 3, \dots\,\} \cup \{\infty\}.$$

The expression "$1 \le r \le \infty$" is a notational shorthand indicating that r is a degree of differentiability, and similar expressions such as $2 \le r \le \infty$ have the obvious meanings. As we mentioned earlier, we sometimes treat 0 as a degree of differentiability by writing expressions such as $0 \le r < \infty$, where C^0 means "continuous."

Let M be an n-dimensional manifold. Fix a degree of differentiability r. A C^r **atlas** for M is an atlas $\{\varphi_i : U_i \to V_i\}_{i \in I}$ such that for all $i, j \in I$, the map

$$\varphi_j \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$$

is C^r. There is, at this point, a temptation to naively define a C^r manifold to be a manifold endowed with a C^r atlas of coordinate charts. The problem with this would be that two such objects could be "different" because they had different atlases, even though the two atlases induced the same class of C^r objects. Let us say that two C^r atlases for M are C^r **equivalent** if their union is in turn a C^r atlas. This notion of equivalence is reflexive and symmetric, obviously. Is it transitive? This question is resolved by the following technical result which underpins the whole subject.

Lemma 8.5. *If $f : U \to V$ and $g : V \to W$ are C^r, where $U \subset \mathbb{R}^m$, $V \subset \mathbb{R}^n$, and $W \subset \mathbb{R}^p$ are open, then*

$$g \circ f : U \to W$$

is also C^r.

Proof. We may assume that r is finite, because the result for $r = \infty$ follows once it is established for all finite r. We have to show that for each s with $1 \leq s \leq r$, each partial derivative function

$$\frac{\partial^s (g \circ f)}{\partial x_{i_1} \cdots \partial x_{i_s}} : U \to \mathbb{R}$$

is defined and continuous on U. This will follow if we can show that its value at a point $x \in U$ can be written as a polynomial function of the values at x of the partials of f up to order s and the values at $f(x)$ of the partials of g up to order s. This is clearly true when $s = 1$ by virtue of the chain rule, and if it is true with $s - 1$ in place of s, then it is true for the partial in question by virtue of the chain rule and the rules for differentiating sums and products. □

Suppose we have three C^r atlases for M, with the first C^r equivalent to the second and the second C^r equivalent to the third. Let $\varphi_i : U_i \to V_i$ and $\varphi_k : U_k \to V_k$ be elements of the first and third atlas respectively. The union of the first and third atlas will be a C^r atlas if, in this situation, it is always the case that

$$\varphi_k \circ \varphi_i^{-1} : \varphi_i(U_i \cap U_k) \to \varphi_k(U_i \cap U_k)$$

is a C^r diffeomorphism. For any $p \in U_i \cap U_k$ there is a chart $\varphi_j : U_j \to V_j$ in the second atlas with $p \in U_j$. Then

$$\varphi_k \circ \varphi_i^{-1} = (\varphi_k \circ \varphi_j^{-1}) \circ (\varphi_j \circ \varphi_i^{-1})$$

on $\varphi_i(U_i \cap U_j \cap U_k)$, and $\varphi_k \circ \varphi_j^{-1}$ and $\varphi_j \circ \varphi_i^{-1}$ are C^r because of the assumed equivalences, so the result above implies that $\varphi_k \circ \varphi_i^{-1}$ is C^r on this set, and consequently (because p was arbitrary) on all of $\varphi_i(U_i \cap U_k)$. Of course this line of reasoning shows that $\varphi_i \circ \varphi_k^{-1}$ is also C^r.

We have shown that "is C^r equivalent to" is transitive, so, in fact, it is an equivalence relation on the set of C^r atlases of M. It would make sense to define a C^r manifold to be a manifold together with an equivalence class of C^r atlases, but it turns out that there is a slightly simpler definition. If two atlases are C^r equivalent, then their union is a C^r atlas that is equivalent to each of them. Actually, the union of *any* collection of C^r equivalent atlases is another C^r atlas that is C^r equivalent to each atlas in the collection. Going a step further in this direction, consider the union of *all* C^r atlases that are C^r equivalent to a given C^r atlas. This atlas is maximal among all atlases that are C^r equivalent to the given C^r atlas, since any atlas that is C^r equivalent to it is C^r equivalent to the given atlas and consequently contained in it.

Definition 8.6. *A C^r* ***differentiable structure*** *for M is a C^r atlas that is maximal in the sense it is not a proper subset of any other C^r atlas. A C^r* ***manifold*** *is a manifold M endowed with a C^r differentiable structure. The elements of the differentiable structure are called C^r* ***coordinate charts*** *for M.*

As we explained above, if a C^r atlas is contained in a C^r differentiable structure, then that differentiable structure is the union of all the atlases that are C^r equivalent to the given atlas. In particular, in order to specify or construct a C^r manifold it suffices to describe a single C^r atlas.

8.3 Orientation

Let's talk about turning a left shoe into a right shoe. Try as you might, no matter how you rotate it you can't do this by moving it around in the confines of your bedroom, but it is at least conceivable that if you put a left shoe onto a rocket and shot it into the intergalactic void, after a few quintillion years it might return from some different direction as a right shoe.

The **Möbius strip** is one of the standard two dimensional illustrations of this phenomenon. To construct a Möbius strip you glue one of the short edges of a rectangle of paper to the other short edge after twisting it 180°. A bit more mathematically, think of starting with $[-2,2] \times (-1,1)$ and identifying the two vertical edges $\{-2\} \times (-1,1)$ and $\{2\} \times (-1,1)$ with a twist, so that $(-2,-t)$ and $(2,t)$ are the same point. Figure 8.3 illustrates how we can "reverse the orientation" of a two dimensional object by sliding it around the Möbius strip. You should compare this with Figure 5.1.

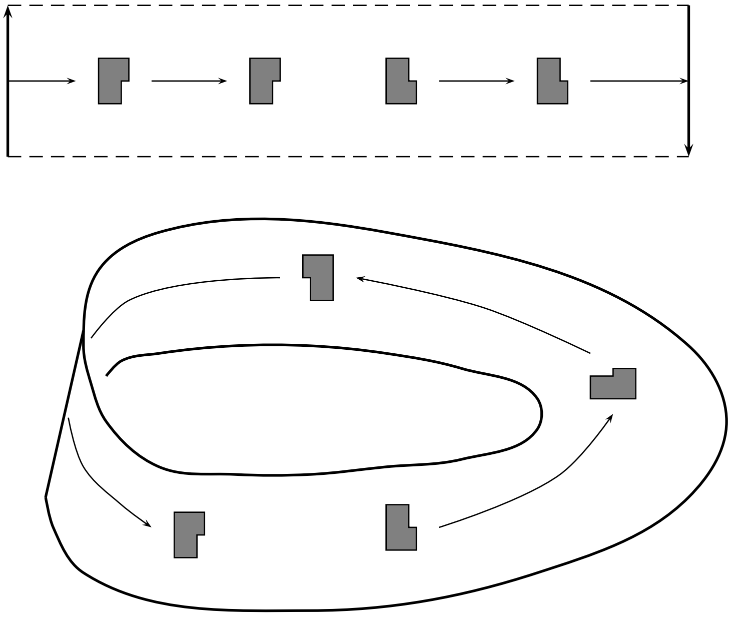


Figure 8.3

It isn't possible to reverse orientation by sliding an object around in the sphere S^2 or in the torus, because these manifolds admit a way of assigning an orientation to each coordinate chart in some atlas that is consistent in the sense that the assigned orientations of any two coordinate charts agree on the intersection of their domains. An **oriented atlas** for a manifold M is a C^1 atlas $\{\varphi_i : U_i \to V_i\}_{i\in I}$ such that for all $i, j \in I$ and all $p \in U_i \cap U_j$, the determinant of $D(\varphi_j \circ \varphi_i^{-1})(\varphi_i(p))$ is positive. We say that M is **orientable** if such an atlas exists, and otherwise M is **unorientable**.

Two dimensional projective space $P^2(\mathbb{R})$ is perhaps the simplest example of an unorientable manifold that (unlike the Möbius strip) is compact, but it is very hard to visualize. For this reason the **Klein bottle**, which was first described in 1882 by Felix Klein, is a bit better known. To construct a Klein bottle we glue two opposite edges of a square (the edges marked A in Figure 8.4) together, obtaining a tube, then glue the two circles bounding the tube to each other, but instead of doing this in the way that results in a torus we reverse the sense in which the two circles are identified with each other. In $\mathbb{R}^3$ you can't do this without having the tube intersect itself. Equivalently, but perhaps less easy to visualize, we can glue two opposite edges of the square to each other with a twist (the edges labeled B in Figure 8.4) obtaining a Möbius strip that now contains the points along its edge, then "zip up" the Möbius strip by gluing points that are across the strip from each other.

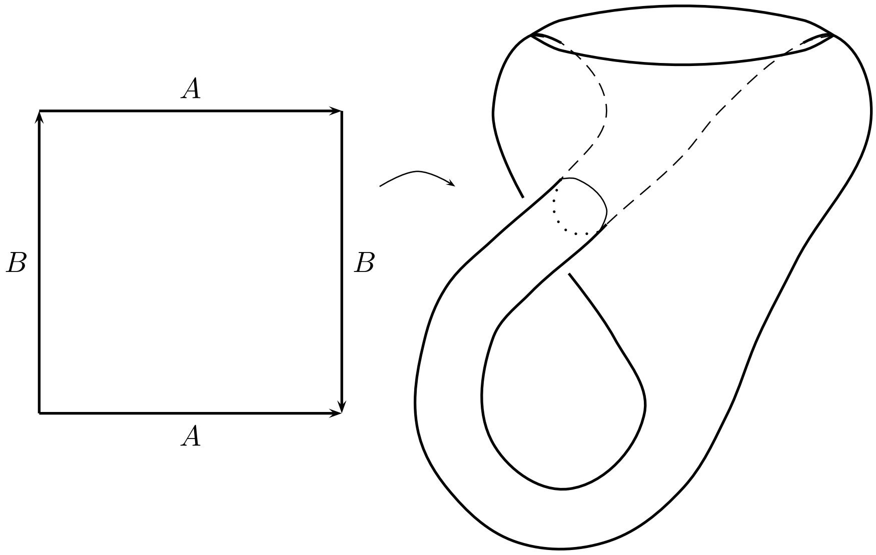


Figure 8.4

8.4 Differentiable Functions

Being able to talk about orientation is nice, but the main point of introducing C^r differentiable structures is to be able to say when a function from one C^r manifold to another is C^r, and eventually to define a notion of differentiation for such functions. After defining C^r functions we will show that C^r manifolds and C^r functions constitute a category. (Defining a category of topological manifolds is easy: we simply let the morphisms from one manifold to another be the continuous functions.)

Let M and N be C^r manifolds, where M is m-dimensional and N is n-dimensional. A C^r **function** from M to N is a function $f : M \to N$ such that

$$\varphi' \circ f \circ \varphi^{-1} : \varphi(U \cap f^{-1}(U')) \to V'$$

is C^r whenever $\varphi : U \to V$ is a C^r coordinate chart for M and $\varphi' : U' \to V'$ is a C^r coordinate chart for N. In a proof that a function is C^r one has to use the C^r coordinate charts that are given, so the following technical result is frequently invoked.

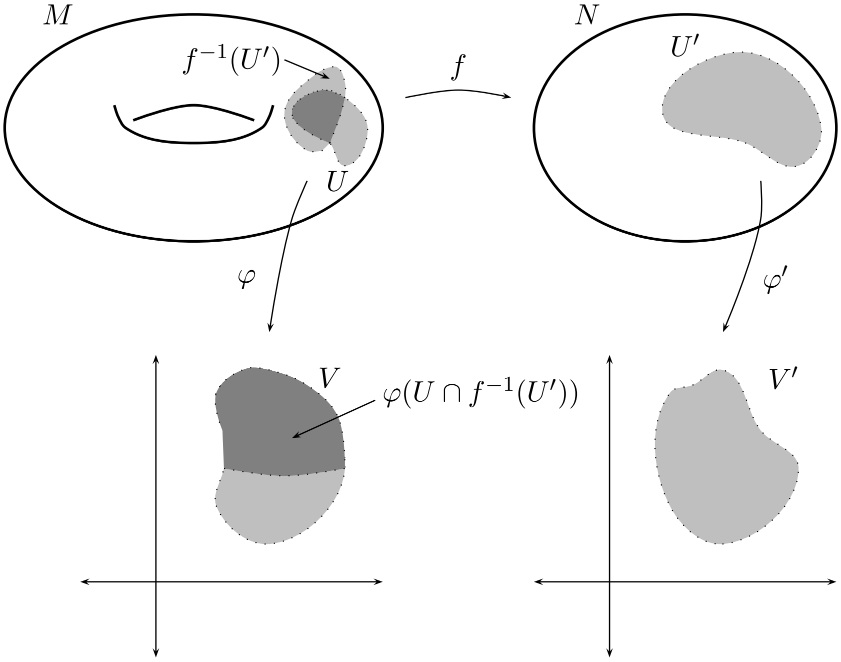


Figure 8.5

Lemma 8.7. *A function $f : M \to N$ is C^r if, for each $p \in M$, there exist C^r coordinate charts $\tilde{\varphi} : \tilde{U} \to \tilde{V}$ and $\tilde{\varphi}' : \tilde{U}' \to \tilde{V}'$ with $p \in \tilde{U}$ and $f(p) \in \tilde{U}'$ such that $\tilde{\varphi}' \circ f \circ \tilde{\varphi}^{-1}$ is C^r.*

Proof. We need to show that, in the situation described in the definition above, $\varphi' \circ f \circ \varphi^{-1}$ is C^r, and of course this amounts to showing that it is C^r in a neighborhood of each $p \in U \cap f^{-1}(U')$. Let $\tilde{\varphi} : \tilde{U} \to \tilde{V}$ and $\tilde{\varphi}' : \tilde{U}' \to \tilde{V}'$ be C^r coordinate charts with $p \in \tilde{U}$ and $f(p) \in \tilde{U}'$ such that $\tilde{\varphi}' \circ f \circ \tilde{\varphi}^{-1}$ is C^r. Since compositions of C^r functions between open subsets of Euclidean spaces are C^r (Lemma 8.5)

$$(\varphi' \circ \tilde{\varphi}'^{-1}) \circ (\tilde{\varphi}' \circ f \circ \tilde{\varphi}^{-1}) \circ (\tilde{\varphi} \circ \varphi^{-1}) = \varphi' \circ f \circ \varphi^{-1}$$

is C^r on the domain of definition of the left hand side. □

Incidentally, although no one would ever doubt that a C^r function is continuous, there's a bit more to the proof than one might expect: in the situation described in the definition ${\varphi'}^{-1} \circ (\varphi' \circ f \circ \varphi^{-1}) \circ \varphi$ is a composition of the continuous (by Lemma 6.12) function $\varphi' \circ f \circ \varphi^{-1}$ and two homeomorphisms, so f is continuous in a neighborhood of each point of M, and of course (Proposition 3.21) continuity is a local property.

We now verify the categorical properties.

Lemma 8.8. *If M, N, and P are C^r manifolds and $f : M \to N$ and $g : N \to P$ are C^r functions, then $g \circ f : M \to P$ is a C^r function.*

Proof. Suppose that $\varphi : U \to V$ and $\varphi'' : U'' \to V''$ are C^r coordinate charts for M and P respectively, and consider a point $p \in U \cap (g \circ f)^{-1}(U'')$. There is a C^r coordinate chart $\varphi' : U' \to V'$ for N whose domain U' contains $f(p)$. Lemma 8.5 implies that

$$\varphi'' \circ (g \circ f) \circ \varphi^{-1} = (\varphi'' \circ g \circ {\varphi'}^{-1}) \circ (\varphi' \circ f \circ \varphi^{-1})$$

is C^r on

$$U \cap f^{-1}(U') \cap (g \circ f)^{-1}(U'').$$

Since p was arbitrary, $g \circ f$ is C^r everywhere on $U \cap (g \circ f)^{-1}(U'')$. □

If you review the definition of a C^r atlas you will see that the assertion "Id_M is a C^r function" reduces to the condition that our atlas for M is, in fact, a C^r atlas. The verification that C^r manifolds and C^r functions constitute a category is completed by the observation that composition of C^r functions is associative, and

$$\mathrm{Id}_N \circ f = f = f \circ \mathrm{Id}_M$$

whenever $f : M \to N$ is C^r, simply because these are true for functions.

Mathematicians say that the category of C^r manifolds and C^r functions between them is "modelled on" C^r functions between open subsets of Euclidean spaces. This terminology can be made precise, but we won't do so, since we expect that even without a precise definition, we will be able to use it effectively in the following discussion. Specifically, we would like to have a category of manifolds that is modelled on complex analytic functions between open subsets of $\mathbb{C}^n$, for various n, and we would like to have a category of manifolds that is modelled on real analytic functions between open subsets of Euclidean spaces. (Theorem 7.12 states that a C^1 function from an open subset of $\mathbb{C}^n$ to $\mathbb{C}$ is analytic, so there is no distinct category of manifolds modelled on C^r functions between open subsets of $\mathbb{C}^n$.)

We won't go through the definitions explicitly, since in each case the discussion is, in every detail, "modelled on" what we did above. If you review this it will be apparent that the only difficulties concern the respective analogues of Lemma 8.5, which are the following two results. Recall that a function from an open subset of $\mathbb{C}^m$ to $\mathbb{C}^n$ is said to be **holomorphic** if it is C^1, and that this is the case if (Theorem 7.9) and only if (Theorem 7.12) the function is analytic and consequently C^∞.

Lemma 8.9. *If $f : U \to V$ and $g : V \to W$ are holomorphic, where $U \subset \mathbb{C}^m$, $V \subset \mathbb{C}^n$, and $W \subset \mathbb{C}^p$ are open, then $g \circ f : U \to W$ is also holomorphic.*

Proof. The chain rule implies that $g \circ f$ is differentiable at each point of U, and the argument used in the real case (that is, in the proof of Lemma 8.5) works equally well here to show that each first order partial derivative of $g \circ f$ is continuous. Therefore $g \circ f$ is C^1 in the complex sense, so (Theorem 7.12) it is holomorphic. □

Lemma 8.10. *If $f : U \to V$ and $g : V \to W$ are real analytic, where $U \subset \mathbb{R}^m$, $V \subset \mathbb{R}^n$, and $W \subset \mathbb{R}^p$ are open, then $g \circ f : U \to W$ is also real analytic.*

Proof. There are open sets $\tilde{U} \subset \mathbb{C}^m$ and $\tilde{V} \subset \mathbb{C}^n$ with $U \subset \tilde{U} \cap \mathbb{R}^m$ and $V \subset \tilde{V} \cap \mathbb{R}^n$ and complex analytic functions $\tilde{f} : \tilde{U} \to \mathbb{C}^n$ and $\tilde{g} : \tilde{V} \to \mathbb{C}^p$ such that $f = \tilde{f}|_U$ and $g = \tilde{g}|_V$. Since we can replace $\tilde{U}$ with $\tilde{f}^{-1}(\tilde{V})$, we may assume that $\tilde{f}(\tilde{U}) \subset \tilde{V}$. The result above implies that $\tilde{g} \circ \tilde{f}$ is complex analytic, so $g \circ f$ is real analytic because it is the restriction of $\tilde{g} \circ \tilde{f}$ to U. □

Now look again at equations $(*)$ and $(**)$ in Section 8.1. In view of the results above, these equations define analytic functions because they are compositions of basic arithmetic operations, and the basic arithmetic operations are analytic: addition, negation, and multiplication are polynomial functions, and for inversion we observe that for any $z_0 \in \mathbb{C}^*$ the power series

$$\frac{1}{z} = \frac{1/z_0}{1 + (z - z_0)/z_0} = \frac{1}{z_0}\Big(1 - \frac{(z - z_0)}{z_0} + \frac{(z - z_0)^2}{z_0^2} - \cdots\Big)$$

centered at z_0 converges absolutely in the open ball of radius $|z_0|$ around z_0. Therefore S^n and $P^n(\mathbb{R})$ are real analytic manifolds, and $S^n(\mathbb{C})$ and $P^n(\mathbb{C})$ are complex analytic manifolds.

Classification of isomorphism types is probably the most important theme in the study of manifolds. The morphisms in the category of topological manifolds are just the continuous functions, so two topological manifolds are isomorphic if they are homeomorphic. If M and N are C^r manifolds, and $f : M \to N$ is a C^r function, then f is a C^r **diffeomorphism** if it is a bijection and $f^{-1} : N \to M$ is also a C^r function. This is the notion of isomorphism for the category of C^r manifolds and C^r functions. Complex analytic diffeomorphisms (also known as holomorphic diffeomorphisms) and real analytic diffeomorphisms are defined analogously. For each of our categories a solution of the classification problem would be a list[4] of manifolds such that any manifold was diffeomorphic to exactly one element of the list.

Any manifold is the union of its connected components, each of which is a connected manifold, and given any collection of manifolds $\{M_\alpha\}_{\alpha \in A}$ of a certain dimension, we can create a new manifold by taking the so-called **disjoint union**: $M := \{\, (\alpha, p) : \alpha \in A, p \in M_\alpha \,\}$ endowed with the topology in which $U \subset M$ is open if and only if each $\{\, p \in M_\alpha : (\alpha, p) \in U \,\}$ is open in M_α. This is all quite trivial, so when we talk about classifying manifolds, what we really mean is classifying *connected* manifolds. Connected manifolds that are not compact can have "infinitely complexity" that is (with a few exceptions) beyond any hope of a humanly comprehensible enumeration, so attention is primarily focused on the classification of compact manifolds. It is also natural to consider the different dimensions separately. In each of the C^r categories, and in the real analytic category, there is precisely one (connected) compact one dimensional manifold, namely the circle.

As we will explain in Section 9.6, the classification of compact two dimensional topological manifolds is well understood. This classification "agrees" with the classification of two dimensional C^r manifolds for any $1 \leq r \leq \infty$ in the following sense: any compact two dimensional topological manifold is homeomorphic to a C^r manifold, and if two compact two dimensional C^r manifolds are homeomorphic, then they are C^r diffeomorphic. A one dimensional (in the sense of one *complex* dimension) holomorphic manifold is called a **Riemann surface**; as we will see in Section 9.4, the problem of classifying compact connected Riemann surfaces up to holomorphic diffeomorphism is much more complex than the classification of two dimensional topological manifolds. This is a reflection of the "rigidity" of analytic functions that we described in Section 7.6.

For a long time a very famous problem stood at the very beginning of the path to a classification of compact three dimensional topological manifolds, but this problem has recently been solved. Compact four dimensional topological manifolds are known to be complex in ways that preclude any

[4] The word "list" should be understood in a general sense that allows, for example, a function from some space of parameters. To constitute a solution of the classification problem the list must contain each isomorphism class exactly once, and it must be "easily computable." The latter requirement is inherently a bit vague.

hope for a classification that is, in a certain sense, "computable."

In dimension three the category of topological manifolds and the category of C^∞ manifolds are again "the same" in the sense described above: any compact three dimensional topological manifold is homeomorphic to a C^∞ manifold, and if two compact three dimensional C^∞ manifolds are homeomorphic, then they are C^∞ diffeomorphic. In 1960 Michel Kervaire (1927-2007) showed that in higher dimensions there are topological manifolds that are not homeomorphic to C^1 manifolds. However, in all higher dimensions it is still the case that any C^1 manifold is C^1 diffeomorphic to a C^∞ manifold, and if two C^∞ manifolds are C^1 diffeomorphic, then they are C^∞ diffeomorphic. These results are manifestations of the flexibility of C^∞ functions, as described in Section 7.7.

The comparison of the different categories of manifolds is the topic of some of the most surprising and famous results of the last several decades. In 1956 John Milnor (b. 1931) showed that there exist what came to be known as exotic 7-spheres. An **exotic 7-sphere** is a C^1 manifold that is homeomorphic to S^7, but not C^1 diffeomorphic to it. Even more remarkable results were proved by Simon Donaldson (b. 1957) in the early 1980's. An **exotic $\mathbb{R}^4$** is a C^1 manifold that is homeomorphic to $\mathbb{R}^4$ but not C^1 diffeomorphic to it. Donaldson produced an uncountable family of exotic $\mathbb{R}^4$'s, no two of which are C^1 diffeomorphic, and he produced a large collection of 4-dimensional topological manifolds that are not homeomorphic to C^1 manifolds. In this respect dimension four is quite anomalous: in 1963 Kervaire and Milnor showed that in each dimension $n \geq 5$ the number of diffeomorphism types of manifolds homeomorphic to S^n is finite, and they showed how to compute this number, but it is still unknown whether there exist exotic four spheres, or even whether the number of diffeomorphism types is necessarily finite, or at worst countable.

The Fields Medal is, perhaps, the closest analogue to a Nobel Prize in mathematics. It is awarded every four years in recognition of specific accomplishments to between two and four mathematicians, who must be under forty years of age, at the International Congress of the International Mathematical Union. Largely in honor of the work described above, Milnor was awarded the Fields Medal in 1962, and Donaldson received it in 1986.

8.5 The Tangent Space

Pursuant to our overall philosophy of differentiation, we would like to define the derivative of a C^1 function between C^1 manifolds, at a point, to be a linear function that provides an asymptotically accurate approximation of the function at that point. When the domain and range of the function were open subsets of Euclidean spaces, we could use those Euclidean spaces as the domain and range of the derivative, but with general manifolds this is not possible. Our first task, then, is to attach a vector space to each point of a C^1 manifold, aiming at using these spaces to define derivatives. There are various ways of doing this, all of which essentially reduce to the following idea: we know what differentiation means for any given coordinate systems for the domain and range, but we do not want to single out particular coordinate systems, so we build a structure that treats all coordinate systems equally by simultaneously including them all. This means equivalence classes, with a number of rather boring verifications that choices of representatives don't matter.

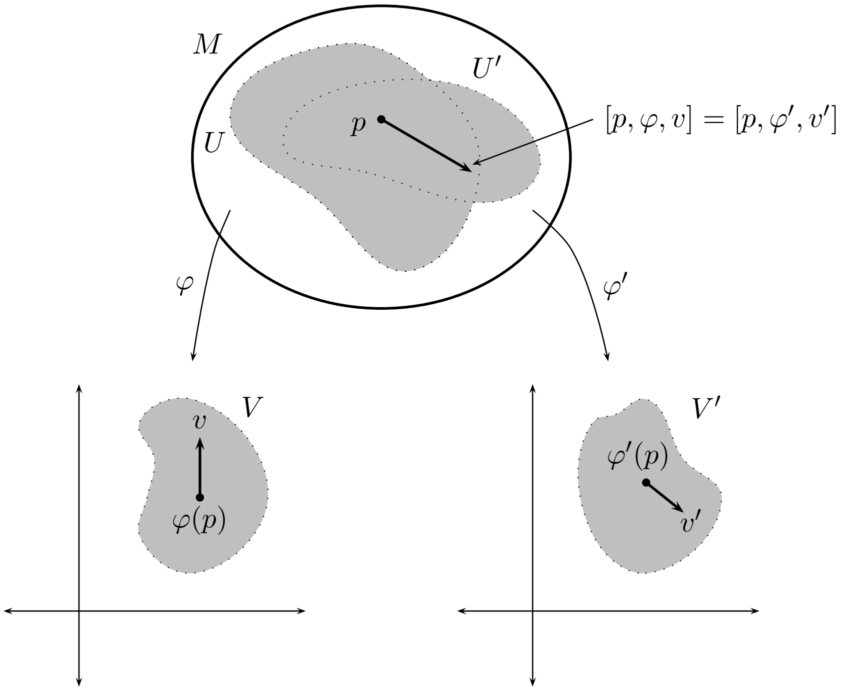


Figure 8.6

The procedure is essentially the same in all of the relevant categories. Let k be either $\mathbb{R}$ or $\mathbb{C}$, and fix an order of differentiation $1 \le r \le \infty$. (Since C^r objects are automatically C^1, the additional generality resulting from allowing r to be greater than 1 is largely spurious, but it is also customary and harmless.) The explicit description will pertain only to C^r objects relative to the field k, but you should understand it as applying equally to the real analytic and complex analytic cases. Sometimes the symbol C^ω is used to denote analyticity, so if you like you can think of ω being a possible value of r. Of course when $k = \mathbb{C}$, C^1 objects are automatically complex analytic.

Fix an m-dimensional C^r manifold M, and consider a point $p \in M$. If M was "nicely" embedded in k^ℓ for some ℓ, we could think of a tangent vector as a vector $v \in k^\ell$ located at a point p that was tangent to M in the normal geometric sense. We aren't given such an embedding, but if $\varphi : U \to V$ is a C^r coordinate chart with $p \in U$, a vector $v \in k^m$ (thought of as emanating from $\varphi(p)$) can be used to represent such a tangent vector. In this way we are led to define a **tangent vector** to be an equivalence class $[p, \varphi, v]$ of triples such as (p, φ, v), where (p, φ, v) and (p', φ', v') are *equivalent* if $p = p'$ and

$$D(\varphi' \circ \varphi^{-1})(\varphi(p))v = v'.$$

You can think of $[p, \varphi, v]$ as the velocity of a curve $c : (-\varepsilon, \varepsilon) \to M$ with $c(0) = p$ and $(\varphi \circ c)'(0) = v$, and in fact there is a different approach in which a tangent vector is defined to be an equivalence class of curves.

Let's check that "equivalence" is, in fact, an equivalence relation. To see that (p, φ, v) is equivalent to itself we compute that

$$D(\varphi \circ \varphi^{-1})(\varphi(p))v = D(\mathrm{Id}_V)(\varphi(p))v = \mathrm{Id}_{k^m}(v) = v.$$

Suppose (p, φ, v) is equivalent to (p', φ', v'). Let $x := \varphi(p)$ and $x' := \varphi'(p')$. Then (p', φ', v') is equivalent to (p, φ, v) because the derivative of the inverse of an invertible map is the inverse of its derivative, so that

$$\begin{aligned} D(\varphi \circ \varphi'^{-1})(x')v' &= D((\varphi' \circ \varphi^{-1})^{-1})(x')v' \\ &= \big(D(\varphi' \circ \varphi^{-1})(x)\big)^{-1} v' = v \end{aligned}$$

We have shown that equivalence is reflexive and symmetric. To check transitivity suppose that (p', φ', v') is also equivalent to (p'', φ'', v''). The chain rule gives

$$\begin{aligned} D(\varphi'' \circ \varphi^{-1})(x)v &= D\big((\varphi'' \circ \varphi'^{-1}) \circ (\varphi' \circ \varphi^{-1})\big)(x)v \\ &= \big(D(\varphi'' \circ \varphi'^{-1})(x') \circ D(\varphi' \circ \varphi^{-1})(x)\big)v \\ &= D(\varphi'' \circ \varphi'^{-1})(x')v' = v'', \end{aligned}$$

so (p, φ, v) is equivalent to (p'', φ'', v'').

The **tangent space** of M, denoted by TM, is the set of all the tangent vectors like $[p, \varphi, v]$. There is a map

$$\pi : TM \to M \quad \text{given by} \quad \pi[p, \varphi, v] = p$$

called the **projection**. For each $p \in M$ the **tangent space of M at p** is

$$T_pM := \pi^{-1}(p).$$

Then

$$TM = \bigcup_{p \in M} T_pM.$$

We wish to treat each T_pM as a vector space with vector operations defined by the formulas

$$[p, \varphi, v] + [p, \varphi, w] := [p, \varphi, v + w] \quad \text{and} \quad \alpha[p, \varphi, v] := [p, \varphi, \alpha v].$$

Since we are working with equivalence classes, we need to check that this makes sense. If $\varphi' : U' \to V'$ is another C^r coordinate chart with $p \in U'$, then $D(\varphi' \circ \varphi^{-1})(\varphi(p))$ and $D(\varphi \circ \varphi'^{-1})(\varphi'(p))$ are inverse linear isomorphisms, so:

(a) any particular coordinate chart φ whose domain contains p can be used to define the vector operations on T_pM because any $[p, \varphi', v'] \in T_pM$ is $[p, \varphi, v]$ for some v;

(b) since $D(\varphi' \circ \varphi^{-1})(\varphi(p))$ and $D(\varphi \circ \varphi'^{-1})(\varphi'(p))$ are linear isomorphisms, the definitions of the vector operations don't depend on whether they are expressed in terms of φ or φ'.

It's easy to check that T_pM satisfies the axioms for a vector space over k because they're satisfied by k^m.

For many purposes we only need the individual tangent spaces, but it is also interesting to look at TM as a manifold. For each C^r coordinate chart $\varphi : U \to V$ there is a derived function

$$T_\varphi : \pi^{-1}(U) \to V \times k^m \quad \text{given by} \quad T_\varphi([p, \varphi, v]) := (\varphi(p), v).$$

Proposition 8.11. *If $\{\varphi_i\}_{i \in I}$ is a C^r atlas for M, then TM (with the topology induced by $\{T_{\varphi_i}\}_{i \in I}$, as per Proposition 8.2) is a $2n$-dimensional manifold, and $\{T_{\varphi_i}\}_{i \in I}$ is a C^{r-1} atlas for TM.*

Proof. The domains $\pi^{-1}(U_i)$ of the T_{φ_i} cover TM because $\{U_i\}$ is a cover of M. We claim that for any $i, j \in I$, $T_{\varphi_j} \circ T_{\varphi_i}^{-1}$ is C^{r-1}. Fixing $p \in U_i \cap U_j$, let $[p, \varphi_i, v_i] = [p, \varphi_j, v_j]$ be an element of $\pi^{-1}(U_i) \cap \pi^{-1}(U_j)$, and let $x_i := \varphi_i(p)$. Our definitions give

$$T_{\varphi_j}\big(T_{\varphi_i}^{-1}(x_i, v_i)\big) = \big(\varphi_j(\varphi_i^{-1}(x_i)), D(\varphi_j \circ \varphi_i^{-1})(x_i)v_i\big).$$

Each component of the vector $D(\varphi_j \circ \varphi_i^{-1})(x_i)v_i$ is a polynomial function of the partial derivatives of $\varphi_j \circ \varphi_i^{-1}$ and the components of v_i, so $D(\varphi_j \circ \varphi_i^{-1})(x_i)v_i$ is a C^{r-1} function of (x_i, v_i), and of course $\varphi_j \circ \varphi_i^{-1}$ is C^r.

Therefore (Proposition 8.2) TM is a $2n$-dimensional quasimanifold because each $T_{\varphi_j} \circ T_{\varphi_i}^{-1}$ is a homeomorphism (of course its inverse is $T_{\varphi_i} \circ T_{\varphi_j}^{-1}$). Among other things, in the induced topology of TM each $\pi^{-1}(U_i)$ is an open set, and each T_{φ_i} is a homeomorphism. Observe that each $\pi|_{\pi^{-1}(U_i)}$ is continuous because it is the composition of T_{φ_i}, the projection $V_i \times k^n \to V_i$, and φ^{-1}. Consequently (because continuity is a local property) π is continuous.

In order to show that TM is Hausdorff we need to show that distinct points ξ and ξ' have disjoint neighborhoods. Since π is continuous and M is Hausdorff, if $\pi(\xi) \neq \pi(\xi')$ we can take $\pi^{-1}(U)$ and $\pi^{-1}(U')$ where U and U' are disjoint neighborhoods of $\pi(\xi)$ and $\pi(\xi')$, so we may assume that $\pi(\xi) = \pi(\xi')$. But if U_i contains this point, then T_{φ_i} is a homeomorphism between $\pi^{-1}(U_i)$, which contains both ξ and ξ', and $V_i \times k^n$, which is a Hausdorff space.

We can now conclude that TM is a manifold. It is easy to check that our argument has verified that $\{T_{\varphi_i}\}_{i \in I}$ satisfies each element of the definition of a C^{r-1} atlas. □

Lemma 8.12. *Under the hypotheses of the last result, $\pi : TM \to M$ is a C^{r-1} function.*

Proof. Since the diagram

$$\begin{tikzcd}
\pi^{-1}(U_i) \arrow[r, "\pi"] \arrow[d, "T_{\varphi_i}"'] & U_i \arrow[d, "\varphi_i"] \\
V_i \times k^n \arrow[r] & V_i
\end{tikzcd}$$

commutes, $\varphi_i \circ \pi \circ T_{\varphi_i}^{-1}$ is C^{r-1} because it coincides with the natural projection $V_i \times k^n \to V_i$. In view of Lemma 8.7, it follows that π is C^{r-1}. □

8.6 A Coordinate-Free Derivative

We continue to work with a field k, which may be either $\mathbb{C}$ or $\mathbb{R}$, and a fixed order of differentiability r between 1 and ∞. (It will continue to be the case that pretty much everything we do here also makes sense for real analytic manifolds and maps.) Let M and N be m- and n-dimensional C^r manifolds, let $f : M \to N$ be a C^r function, and fix a particular point $p \in M$ and a tangent vector $\xi \in T_pM$. When $k = \mathbb{R}$ our guiding intuition is that if ξ is the velocity at time 0 of a curve $c : (-\varepsilon, \varepsilon) \to M$ with $c(0) = p$, then the derivative $Df(p)$ of f at p should take ξ to the vector

$$Df(p)\xi \in T_{f(p)}N$$

that is the velocity of the curve $f \circ c$ at time 0. (When $k = \mathbb{C}$ we might imagine a function c whose domain is a small neighborhood of $0 \in \mathbb{C}$.)

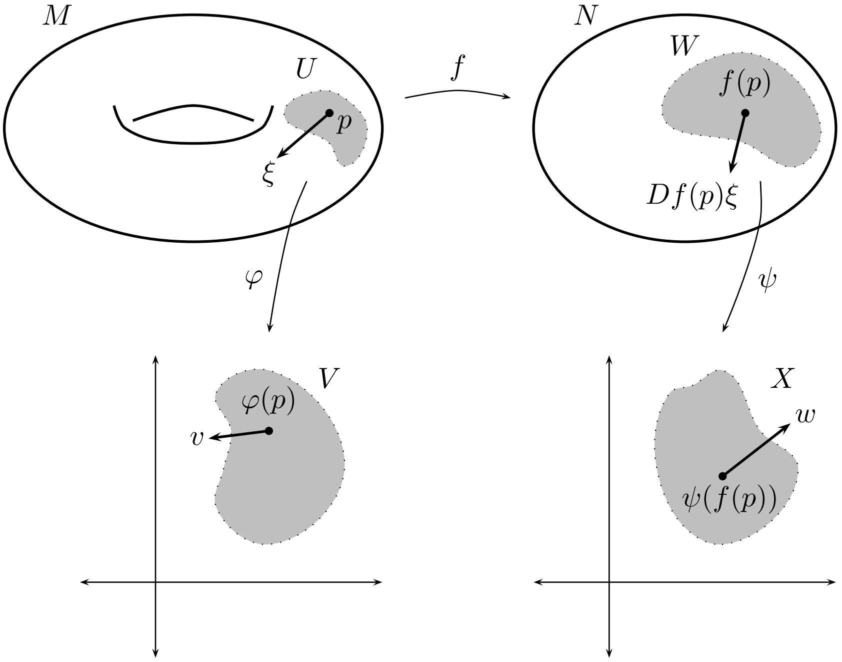


Figure 8.7

To ground the definition in concrete calculations we need to introduce coordinate charts. Let $\varphi : U \to V$ be a C^r coordinate chart for M with

$p \in U$, and let $\psi : W \to X$ be a C^r coordinate chart for N with $f(p) \in W$. We define

$$Df(p) : T_pM \to T_{f(p)}N$$

by setting

$$Df(p)[p, \varphi, v] := [f(p), \psi, w] \quad \text{where} \quad w = D(\psi \circ f \circ \varphi^{-1})(\varphi(p))v.$$

In view of the definition of the vector operations on T_pM and $T_{f(p)}N$ and the linearity of $D(\psi \circ f \circ \varphi^{-1})(\varphi(p))$, once we've shown that $Df(p)$ is well defined it will be obvious that it is linear.

The rest of the section develops the most basic properties of this definition: 1) independence of choice of representatives of equivalence classes; 2) the chain rule. We then combine the derivatives at the various points of M into a single map $Tf : TM \to TN$ that is shown to be C^{r-1}. Finally, we'll see how T can be viewed as a functor. There will be some bulky and forbidding looking formulas, but for those with prior experience with these sorts of structures this material is unsurprising and mechanical. From any point of view the symbolic mass of the calculations dwarfs the conceptual content, except, perhaps, in one sense: the uneventful quality of the exposition is a reflection of the fact that this is, in some sense, the "right" system of definitions.

To show that the definition of $Df(p)$ doesn't depend on the choices of coordinate charts suppose that $\varphi' : U' \to V'$ and $\psi' : W' \to X'$ are C^r coordinate charts for M and N with $p \in U'$ and $f(p) \in W'$ respectively. Let $x := \varphi(p)$, $x' := \varphi'(p)$, and $y := \psi(f(p))$. Suppose also that $[p, \varphi, v] = [p, \varphi', v']$, and let

$$w' := D(\psi' \circ f \circ {\varphi'}^{-1})(x')v'.$$

With φ' and ψ' in place of φ and ψ, the definition above gives

$$Df(p)[p, \varphi', v'] := [f(p), \psi', w'],$$

so our goal is to show that $[f(p), \psi, w] = [f(p), \psi', w']$. Since (p, φ, v) and (p, φ', v') are equivalent, $v = D(\varphi \circ {\varphi'}^{-1})(x')v'$, so that

$$D(\psi' \circ \psi^{-1})(y)w = D(\psi' \circ \psi^{-1})(y)\big(D(\psi \circ f \circ \varphi^{-1})(x)v\big)$$

$$= \Big(D(\psi' \circ \psi^{-1})(y) \circ D(\psi \circ f \circ \varphi^{-1})(x) \circ D(\varphi \circ {\varphi'}^{-1})(x')\Big)v'$$

$$= D(\psi' \circ f \circ {\varphi'}^{-1})(x')v' = w'.$$

(The penultimate equality here is, of course, the chain rule for functions between open subsets of Euclidean spaces.)

We turn to the chain rule. Let P be a p-dimensional C^r manifold, let $g : N \to P$ be another C^r function, and let $\chi : Y \to Z$ be a C^r coordinate chart for P with $g(f(p)) \in Y$. The definition of $D(g \circ f)$ gives

$$D(g \circ f)(p)[p, \varphi, v] = [g(f(p)), \chi, z] \text{ where } z = D(\chi \circ g \circ f \circ \varphi^{-1})(x)v.$$

There is now another application of the chain rule for functions between open subsets of Euclidean spaces:

$$\begin{aligned} z &= D(\chi \circ g \circ \psi^{-1} \circ \psi \circ f \circ \varphi^{-1})(x)v \\ &= D(\chi \circ g \circ \psi^{-1})(y)\big(D(\psi \circ f \circ \varphi^{-1})(x)v\big) \\ &= D(\chi \circ g \circ \psi^{-1})(y)w, \end{aligned}$$

so $[g(f(p)), \chi, z] = Dg(f(p))[f(p), \psi, w]$ and

$$D(g \circ f)(p)[p, \varphi, v] = Dg(f(p))[f(p), \psi, w] = Dg(f(p))\big(Df(p)[p, \varphi, v]\big).$$

We have shown that

$$D(g \circ f)(p) = Dg(f(p)) \circ Df(p).$$

The derivatives of f at the various points of M can be combined into a single function $Tf : TM \to TN$ defined by

$$Tf([p, \varphi, v]) := Df(p)[p, \varphi, v].$$

The chain rule passes up to this level of aggregation: if $g : N \to P$ is a second C^r function, then $T(g \circ f) = Tg \circ Tf$ because for any $p \in M$ and $[p, \varphi, v] \in T_pM$ we have

$$Tg(Tf([p, \varphi, v])) = Tg(Df(p)[p, \varphi, v]) = Dg(f(p))\big(Df(p)[p, \varphi, v])\big)$$

$$= D(g \circ f)(p)[p, \varphi, v] = T(g \circ f)([p, \varphi, v]).$$

We claim that Tf is C^{r-1}. Since (Proposition 8.11) $\{T_{\varphi_i}\}_{i \in I}$ is a C^{r-1} atlas for TM whenever $\{\varphi_i\}_{i \in I}$ is a C^r atlas for M, what this amounts to is that $T_\psi \circ Tf \circ T_\varphi^{-1}$ is a C^{r-1} function whenever $\varphi : U \to V$ and $\psi : W \to X$ are C^r coordinate charts for M and N with $f(U) \subset W$. Consider $p \in U$ and $[p, \varphi, v] \in T_pM$, and let $x := \varphi(p)$ and $y := \psi(f(p))$. Then

$$Df(p)[p, \varphi, v] := [f(p), \psi, w] \quad \text{where} \quad w = D(\psi \circ f \circ \varphi^{-1})(x)v.$$

Since $T_\varphi([p,\varphi,v]) = (x,v)$ and $T_\psi([f(p),\psi,w] = (y,w)$ we have

$$\big(T_\psi \circ Tf \circ T_\varphi^{-1}\big)(x,v) = \big(y, D(\psi \circ f \circ \varphi^{-1})(x)v\big).$$

Of course $\psi \circ f \circ \varphi^{-1}$ is C^r because f is C^r, so $y = \psi(f(\varphi^{-1}(x)))$ is a C^r function of x. The entries of the matrix of $D(\psi \circ f \circ \varphi^{-1})$ are the partial derivatives of the component functions of $\psi \circ f \circ \varphi^{-1}$, so $D(\psi \circ f \circ \varphi^{-1})(x)v$ is a C^{r-1} function of (x,v).

We've associated a C^{r-1} manifold TM with each C^r manifold M, and we've associated a C^{r-1} function $Tf : TM \to TN$ with each C^r function $f : M \to N$ between C^r manifolds. *We claim that T is a functor.* We have already seen that T commutes with composition: $T(g \circ f) = Tg \circ Tf$ whenever $f : M \to N$ and $g : N \to P$ are C^r functions. The only remaining verification is that $T\mathrm{Id}_M = \mathrm{Id}_{TM}$, and to establish this we need to show that $D\mathrm{Id}_M(p) = \mathrm{Id}_{T_pM}$ for any $p \in M$. Let $\varphi : U \to V$ and $\varphi' : U' \to V'$ be C^r coordinate charts with $p \in U \cap U'$, and consider a tangent vector $[p,\varphi,v] \in T_pM$. The definition of $D\mathrm{Id}_M(p)$ gives

$$D\mathrm{Id}_M(p)[p,\varphi,v] := [p,\varphi',v'] \quad \text{where} \quad v' = D(\varphi' \circ \mathrm{Id}_M \circ \varphi^{-1})(\varphi(p))v.$$

Of course $\varphi' \circ \mathrm{Id}_M \circ \varphi^{-1} = \varphi' \circ \varphi^{-1}$, so (p,φ,v) and (p,φ',v') are equivalent. That is, $[p,\varphi',v'] = [p,\varphi,v]$, so $D\mathrm{Id}_M(p)[p,\varphi,v] = [p,\varphi,v]$.

Above we talked about T as "a" functor, but of course in the case $k = \mathbb{R}$ we really have a functor for each order of differentiability r, and there are also functors for the complex holomorphic category and the real analytic category. You should take note of the usefulness of the category concept here: by making the fundamental structural elements explicit, it organizes our thinking in ways that allow us to perceive how proofs and constructions in related contexts are really the same, to the point where we can write a single argument that is simultaneously applicable to all the different categories we are studying. There are important instances of this in other parts of mathematics, and in some cases it has been possible to express the common features of an argument or construction in terms of properties of a very general type of category, thereby achieving an extreme, and extremely powerful, level of abstraction.

There is one additional point worth mentioning. It looks like the C^∞ category is a bit more convenient than the C^r category for $r < \infty$, since if M is C^r, then TM is only C^{r-1}, but if M is C^∞, then so is TM. This turns out to be a rather significant simplification with wide applicability. Many of the functions arising in physics and other sciences are C^∞. In "softer" sciences like economics it can happen that the underlying conceptual structure doesn't single out specific equations, so that technical assumptions can be imposed if they are convenient and they don't impose undue constraints on the qualitative properties of the model. This is often the case because, essentially as a result of the construction described in Section 7.7, it is possible to approximate C^r objects, or even continuous objects, with C^∞ objects. (In contrast, the rigidity of analytic objects is unrealistic in most scientific applications.) Differential topology—the subfield of mathematics studying C^r manifolds and C^r maps between them from the point of view of topological issues—takes great advantage of this phenomenon.

There is something tremendously satisfying about our work here. We have created a notion of differentiation that is, in every respect, free of *a priori* commitment to particular coordinate systems or any restriction on the large scale structure of the domain and range manifolds. Its key properties are expressed in its functorial nature. Our initial description of differentiation as a functor, in Section 6.6, was a bit strained and artificial, but that is no longer the case because the objects and morphisms in the domain and range categories of the various functors T are, in fact, the things we are really interested in. Achieving all this involved a certain amount of work, in that many definitions and verifications were required. If you are new to this sort of thing it might have been rather difficult reading, not least because the notation is quite bulky. As you become more experienced, you'll find that such formalities become less difficult and a bit tedious: when reading mathematics, it usually doesn't work to skip parts of the text, and it is important to make all of the logic of the argument explicit, but verifications like the ones in this section are quite predictable, with no conceptual revelations. Having done them all in a fully general framework, at least we won't have to do them again.

At the same time, if one were to ask how to compute the version of the derivative developed here, the only answer we have at this point is that you have to pass to coordinate systems for the domain and range, then apply the various rules developed in the Chapter 6. In this sense we haven't yet accomplished anything of concrete significance. However, this doesn't mean that our work to this point is useless. Like a blank piece of paper, in itself the framework we have developed says nothing, but it can be used to express many interesting theories.

8.7 The Regular Value Theorem

The simplest form of the general principle studied in this section is possibly familiar to you from elementary calculus, where it occurs in connection with a technique for computing the derivative of a function that is defined implicitly. Fix an order of differentiability $1 \leq r \leq \infty$. Let $f : U \to \mathbb{R}$ be a C^r function, where $U \subset \mathbb{R}^2$ is open, and let (x_0, y_0) be a point of U at which $\frac{\partial f}{\partial y}(x_0, y_0) \neq 0$. The **implicit function theorem** asserts that if $\delta > 0$ is sufficiently small, then there is a unique C^r function

$$g : (x_0 - \delta, x_0 + \delta) \to \mathbb{R}$$

such that $g(x_0) = y_0$ and $f(x, g(x)) = f(x_0, y_0)$ for all x in the domain of g. Moreover, the the graph of g completely characterizes $f^{-1}(f(x_0, y_0))$ near (x_0, y_0): there is a neighborhood $W \subset U$ of (x_0, y_0) such that

$$f^{-1}(f(x_0, y_0)) \cap W = \mathrm{Gr}(g) := \{\, (x, g(x)) : x \in (x_0 - \delta, x_0 + \delta) \,\}.$$

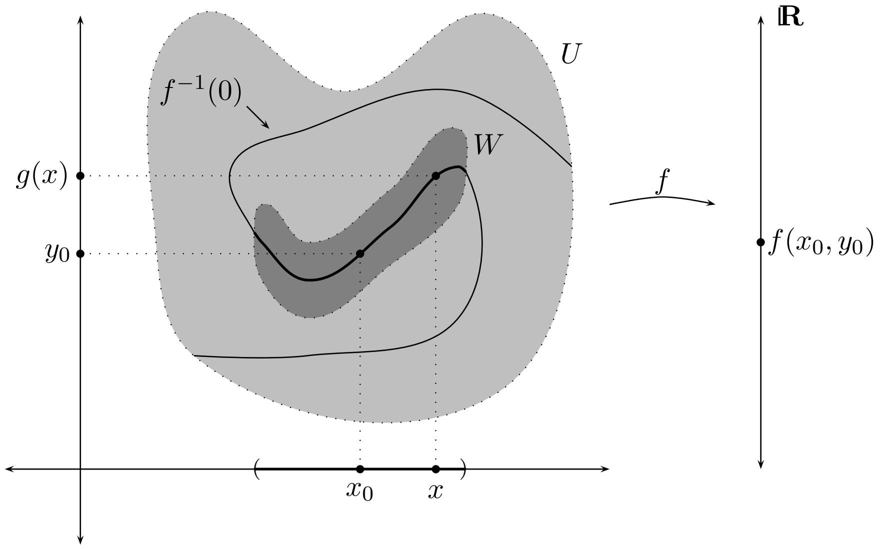


Figure 8.8

Given this setup, what you learn to do in elementary calculus is use the chain rule to compute that

$$0 = \frac{\partial f}{\partial x}(x, g(x)) + \frac{\partial f}{\partial y}(x, g(x)) \cdot g'(x),$$

then rearrange to get

$$g'(x) = -\frac{\partial f}{\partial x}(x, g(x)) \Big/ \frac{\partial f}{\partial y}(x, g(x)).$$

This can sometimes give a closed form expression for $g'(x_0)$ even when it is impossible to find a closed form expression for g.

Since we will prove a more general version, we won't dwell on the proof of this case of the implicit function theorem, but it is worth pointing out that the formula for $g'(x)$ conveys the intuition underlying the procedure for finding g. The change in f resulting from going from (x_0, y_0) to $(x_0+\Delta x, y_0)$ is

$$f(x_0 + \Delta x, y_0) - f(x_0, y_0),$$

which is well approximated by $\frac{\partial f}{\partial x}(x_0, y_0)\Delta x$, and one can use $\frac{\partial f}{\partial y}(x_0, y_0)$ to compute a change

$$\Delta y := -\Big(\frac{\partial f}{\partial x}(x_0, y_0) \Big/ \frac{\partial f}{\partial y}(x_0, y_0)\Big)\Delta x$$

such that $f(x_0 + \Delta x, y_0 + \Delta y)$ is approximately $f(x_0, y_0)$. It typically won't be exactly $f(x_0, y_0)$, but we can compute a further adjustment

$$-\big(f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0)\big) \Big/ \frac{\partial f}{\partial y}(x_0, y_0)$$

to the y-coordinate that will (in a suitably small neighborhood of (x_0, y_0), because $\frac{\partial f}{\partial y}$ is continuous) result in a better approximation. If we begin with x sufficiently close to x_0 the sequence of values of the y-coordinate computed by iterating this procedure converges to a point y such that $f(x, y) = f(x_0, y_0)$. The actual proof below will be an appeal to the inverse function theorem, so this process of successive approximations is hidden, but the proof of the inverse function theorem has a similar idea (the contraction mapping theorem) at its heart.

The discussion throughout the rest of the section applies equally to the three cases covered by the three versions of the inverse function theorem. In order to avoid saying everything three times we adopt the following framework: k is either $\mathbb{R}$ or $\mathbb{C}$, and the term 'smooth' means 'C^r' (where $1 \le r \le \infty$) or 'real analytic' if $k = \mathbb{R}$, or 'holomorphic' if $k = \mathbb{C}$.

Suppose we are given an open $U \subset k^m$, a smooth function $f : U \to k^n$, and a point $w_0 \in U$ such that the image of $Df(w_0)$ is all of k^n. (When this is not the case the structure of $f^{-1}(f(w_0))$ near w_0 can be extremely complicated, even when f is a polynomial function.) Let L be an

n-dimensional linear subspace of k^m that is complementary to $\ker Df(w_0)$, so that $L \cap \ker Df(w_0) = \{0\}$ and $L + \ker Df(w_0) = k^m$. Let $z_0 := f(w_0)$.

If f was affine, say $f(w) = a + \ell(w)$ where $\ell : k^m \to k^n$ is linear, then $f^{-1}(z_0) = w_0 + \ker \ell$ would be an $(m-n)$-dimensional affine subspace of k^n. When f is not affine, but merely differentiable, it should still be the case that if you were very small you would have a hard time distinguishing f from its affine approximation, so we should expect $f^{-1}(z_0)$ to "look like" $w_0 + \ker Df(w_0)$ near w_0. We can measure the resemblance between $f^{-1}(z_0)$ and $w_0 + \ker Df(w_0)$ near w_0 by looking at the intersections $(y+L) \cap f^{-1}(z_0)$ for y in a neighborhood $V \subset w_0 + \ker Df(w_0)$ of w_0. If the resemblance is close, then for each $y \in V$ there should be a single $g(y) \in L$ near 0 such that $f(y + g(y)) = z_0$, and the function $g : V \to L$ should be difficult to distinguish from the zero function (if you are sufficiently small) in the sense that $g(w_0) = 0$ and $Dg(w_0) = 0$.

In the customary presentation of the implicit function theorem, there is a given linear subspace L that is assumed to be complementary to $\ker Df(w_0)$, but the implicit function is not necessarily defined on a neighborhood of w_0 in $w_0 + \ker Df(w_0)$, but instead on an open subset of an $(m-n)$-dimensional subspace that is complementary to L. In this context it makes sense to let this subspace and L be coordinate subspaces of k^m, so we denote points in k^m by (x, y) where $x = (x_1, \dots, x_{m-n})$ and $y = (y_1, \dots, y_n)$. If $U \subset k^m$ is open and $f : U \to k^n$ is differentiable at (x, y), let $D_x f(x, y)$ and $D_y f(x, y)$ denote the "partial" derivatives given by varying x and y respectively. That is, $D_x f(x, y)$ is the derivative of the function $f(\cdot, y)$ at x, and $D_y f(x, y)$ is the derivative of the function $f(x, \cdot)$ at y.

Theorem 8.13 (Implicit Function Theorem). *Suppose that $U \subset k^m$ is open, $f : U \to k^n$ is smooth, and (x_0, y_0) is a point in U such that $D_y f(x_0, y_0)$ is nonsingular. Let $z_0 := f(x_0, y_0)$. Then there is a neighborhood $V \subset k^{m-n}$ of x_0, a neighborhood $W \subset U$ of (x_0, y_0), and a smooth function $g : V \to k^n$, such that*

$$f^{-1}(z_0) \cap W = \mathrm{Gr}(g) := \{\, (x, g(x)) : x \in V \,\}.$$

Proof. We will apply the inverse function theorem to the function $F : U \to k^m$ given by $F(x, y) := (x, f(x, y))$. It is easy to see that $DF(x_0, y_0)$ is nonsingular: the image of $D_y F(x_0, y_0)$ is the coordinate subspace of the last n coordinates, and the composition of $D_x F(x_0, y_0)$ with the projection onto the coordinate subspace of the first $m-n$ coordinates is the projection $(v, w) \mapsto v$. The inverse function theorem (whichever version pertains) gives an open neighborhood $W \subset U$ of (x_0, y_0) such that $F|_W$ is a smooth diffeomorphism onto its image. Let $V := \{\, x \in k^{m-n} : (x, z_0) \in F(W) \,\}$.

Since $x \mapsto (x, z_0)$ is continuous and $F(W)$ is open, this set is open, and it contains x_0 because $(x_0, y_0) \in W$. Let $\pi : k^m \to k^n$ is the projection $\pi(x, y) := y$, and let $g : V \to k^n$ be the function

$$g(x) := \pi\big((F|_W)^{-1}(x, z_0)\big).$$

Since it is a composition of smooth functions, g is smooth (Lemma 8.5, 8.9, or 8.10). In view of the definitions of F, V, and g,

$$\{\, (x, y) \in W : f(x, y) = z_0 \,\} = \{\, (x, y) \in W : F(x, y) = (x, z_0) \,\}$$

$$= (F|_W)^{-1}(V \times \{z_0\}) = \{\, (x, g(x)) : x \in V \,\}.$$

□

There is a closed form expression for the derivative of g:

Proposition 8.14. *In the situation described by the hypotheses and conclusion of the preceeding result,*

$$Dg(x_0) = -D_y f(x_0, y_0)^{-1} \circ D_x f(x_0, y_0).$$

Proof. We apply the chain rule to the identity $0 = f(x, g(x))$, obtaining

$$0 = D_x f(x_0, y_0) + D_y f(x_0, y_0) \circ Dg(x_0).$$

Since $D_y f(x_0, y_0)$ is nonsingular by assumption, we can take the composition of both sides with its inverse. □

The inverse function theorem and the implicit function theorem are really two expressions of a single underlying principle. In the proof above we saw how to use the inverse function theorem to prove the implicit function theorem, and if the implicit function theorem is already established it is equally easy to use it to prove the inverse function theorem, as we now explain.

Let $f : U \to k^m$ be a smooth function where $U \subset k^m$ is open, let x_0 be a regular point of f, and let $y_0 := f(x_0)$. Let $F : U \times k^m \to k^m$ be the function

$$F(x, y) := f(x) - y.$$

By assumption $D_x F(x_0, y_0) = Df(x_0)$ is nonsingular, so if we've already established the implicit function theorem we can apply it (with the roles of the variables x and y reversed) to get an open neighborhood $V \subset k^m$ of y_0 and a smooth function $g : V \to U$ such that

$$F^{-1}(0) \cap Z = \{\, (g(y), y) : y \in V \,\}$$

for some open neighborhood $Z \subset U \times k^m$ of (x_0, y_0). Then $g(V)$ is open because it is the set of x mapped to the open set Z by the continuous function $x \mapsto (x, f(x))$, and $f|_{g(V)}$ and g are inverse functions because for $(x, y) \in U \times k^m$ the following conditions are equivalent:

(a) $y \in V$ and $x = g(y)$;

(b) $(x, y) \in Z$ and $F(x, y) = 0$;

(c) $x \in g(V)$ and $y = f(x)$.

The version of the implicit function theorem given above (with $k = \mathbb{R}$ and 'smooth' meaning C^r) is frequently a "capstone" result in a multivariable calculus course, but in my opinion the proper formulation of the result, and an appreciation of its significance, are impossible without the manifold concept. In preparation for the explanation, let's expand our vocabulary a bit. Fix an m-dimensional smooth manifold M.

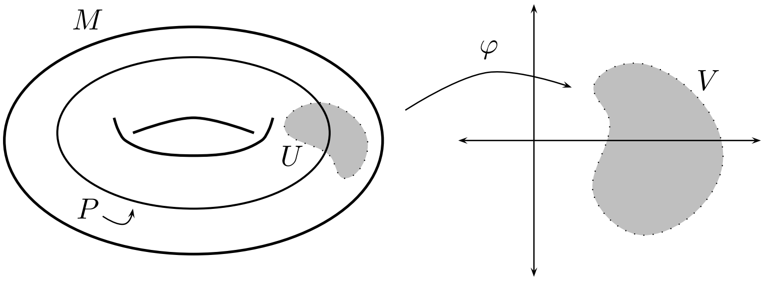


Figure 8.9

Definition 8.15. *A set $P \subset M$ is a p-dimensional*[5] *smooth* ***submanifold*** *of M if, for each $p \in P$, there is a smooth coordinate chart $\varphi : U \to V$ with $p \in U$ and $\varphi(U \cap P) = V \cap k^p$. The difference $m - p$ is called the* ***codimension*** *of P.*

[5] Here we are using p to denote either the dimension of P or a typical element of M, but the appropriate interpretation of the symbol will always be obvious. In more advanced books this sort of "overloading" of notation is more frequent, and often not mentioned explicitly.

(In this definition, and at appropriate points below, we identify k^p with $\{\, x \in k^m : x_{p+1} = \cdots = x_m = 0 \,\}$.)

The intuitive picture is pretty simple: a smooth submanifold of M is just a subset that happens to be a smooth manifold itself, in a way that's compatible with the differentiable structure of M. The following details flesh this out. If φ is as in the definition, then its restriction to $U \cap P$ is a coordinate chart for P, because the restriction of a homeomorphism to a subset of its domain is a homeomorphism. If $\varphi' : U' \to V' \subset k^m$ is another such coordinate chart, then $\varphi' \circ (\varphi|_{U \cap U' \cap P})^{-1}$ is smooth because it is the composition of

$$(x_1, \dots, x_p) \mapsto (x_1, \dots, x_p, 0, \dots, 0) \in k^m$$

with $\varphi' \circ (\varphi|_{U \cap U'})^{-1}$. Therefore the restrictions $\varphi|_{U \cap P}$ of coordinate charts of the sort given by the definition constitute a smooth atlas for P.

It happens very frequently in science that we are interested in a subset of a Euclidean space, or a manifold, given by the vanishing of some differentiable function. For example, if a physical system conserves energy and momentum, then its motion is confined to the set of configurations that have the initial values of these quantities. The conceptual significance of the implicit function theorem is that it gives conditions under which such a subset is necessarily a submanifold.

A bit more terminology helps with the explanation of this. Fix a second smooth manifold N, which we assume to be n-dimensional, and a smooth function $f : M \to N$.

Definition 8.16. *A point $p \in M$ is a **regular point** of f if*

$$Df(p) : T_pM \to T_{f(p)}N$$

*is surjective, and otherwise it is a **singular point** of f. A point $q \in N$ is a **singular value** of f if $f^{-1}(q)$ contains a singular point, and otherwise it is a **regular value** of f.*

Note that if $m < n$, then every point of M is automatically a singular point of f. Also, this system of terminology has the following paradoxical aspect: if $f^{-1}(q) = \emptyset$, then q is automatically a regular value of f, even though it isn't a "value" of f.

Theorem 8.17 (Regular Value Theorem)**.** *If q is a regular value of f, then $f^{-1}(q)$ is a codimension n smooth submanifold of M.*

Proof. Fix an arbitrary $p \in f^{-1}(q)$. Let $\varphi : U \to k^m$ and $\psi : V \to k^n$ be smooth coordinate charts for open neighborhoods $U \subset M$ and $V \subset N$ of p and q respectively. Since we can replace U with a smaller open neighborhood of p, we may assume that $f(U) \subset V$. We can also easily arrange for it to be the case that $\varphi(p) = 0$ and $\psi(q) = 0$.

By assumption p is a regular point, so $Df(p)$ is surjective. Of course $D\varphi^{-1}(0)$ and $D\psi(q)$ are linear isomorphisms, so (in view of the chain rule) $D(\psi \circ f \circ \varphi^{-1})(0)$ is surjective. Let $\mathbf{e}_1, \dots, \mathbf{e}_m$ be the standard unit basis of k^m. A maximal linearly independent subset of

$$\left\{D(\psi \circ f \circ \varphi^{-1})(0)\mathbf{e}_1, \dots, D(\psi \circ f \circ \varphi^{-1})(0)\mathbf{e}_m\right\}$$

has n elements because this set spans k^n, so, by reindexing, we can arrange for the last n elements of this set to be linearly independent. This means that $D_y(\psi \circ f \circ \varphi^{-1})(0)$ is nonsingular where, as before, we denote points in k^m by (x, y) with $x \in k^{m-n}$ and $y \in k^n$.

The implicit function theorem now gives a neighborhood $W \subset \varphi(U)$ of 0, a neighborhood $Z \subset k^{m-n}$ of 0, and a smooth $g : Z \to k^n$ such that

$$(\psi \circ f \circ \varphi^{-1})^{-1}(0) \cap W = \mathrm{Gr}(g) = \{\, (x, g(x)) : x \in Z \,\}.$$

All of the conditions given above continue to hold if we replace W with $W \cap \pi^{-1}(Z)$, so we may assume that $\pi(W) = Z$. Also, we can replace U with $\varphi^{-1}(W)$, so we may assume that $\varphi(U) = W$.

Let $\varphi = (\varphi_x, \varphi_y)$. The idea now is to modify φ_y in order to create a new coordinate chart in which the preimage of q is contained in the coordinate subspace $k^{m-n} \subset k^m$. Define $\tilde{\varphi} = (\tilde{\varphi}_x, \tilde{\varphi}_y) : U \to k^m$ by setting $\tilde{\varphi}_x := \varphi_x$ and $\tilde{\varphi}_y := \varphi_y - g \circ \varphi_x$. Then $\tilde{\varphi}$ is a smooth coordinate chart because $\tilde{\varphi}$ is smooth and $\tilde{\varphi}^{-1}(x, y) = \varphi^{-1}(x, y + g(x))$, so $\tilde{\varphi}^{-1}$ is also smooth. It displays $f^{-1}(q)$ as a smooth submanifold near p because for $p' \in U$ we have

$$f(p') = q \iff \varphi_y(p') = g(\varphi_x(p')) \iff \tilde{\varphi}_y(p') = 0.$$

□

There is a sense in which a "typical" element of N is a regular value of f. Imagine the graph of a C^∞ function from an open subset $U \subset \mathbb{R}^2$ to $\mathbb{R}$. The critical points of this function are the hilltops, hollows, places where a hillside happens to level out, and so forth, and the critical values are the values of the function at these points. Probably in your imagination there are only finitely many critical points and consequently only finitely many critical values. By having the function be constant in some connected region, one can have a continuum of critical points, but they all map to a single critical value. It is easy enough to create a function that has countably many critical values because it oscillates countably many times, but it is difficult to see how to create an uncountable set of critical values.

A crude intuition suggests why it might be difficult to have a large set of critical values: when two critical points map to different critical values, the region between those critical points has to be largely filled with regular points because the value of the function has to change as you go along any path from one of the critical points to the other. During the 1930's this intuition was made precise: in the setting of the regular value theorem when "smooth" means C^r (that is, M and N are m and n-dimensional C^r manifolds over the field $\mathbb{R}$ and $f : M \to N$ is C^r) a fundamental result called **Sard's theorem** states that if $r > m - n$ and $r \geq 1$, then "almost all" elements of N are regular values of f. We don't have the tools required to give a precise description of what "almost all" means, so you will have to be content with the assertion that it is quite a strong property. Among other things, it implies that the set of regular values is a dense subset of N. (Recall that a subset of a topological space is **dense** if its closure is the entire space.) This greatly enhances the power and applicability of the regular value theorem.

Although in my way of thinking about things, the regular value theorem is the "conceptually correct" formulation of the implicit function theorem, it should be admitted that others might feel that the implicit function theorem expresses computationally useful facts that are lost in the passage to a manifold-theoretic framework. This is less relevant to the comparison of the standard presentation of the inverse function theorem (using open subsets of k^m) with the "conceptually correct" version.

Theorem 8.18 (Inverse Function Theorem)**.** *If M and N are smooth m-dimensional manifolds, $f : M \to N$ is a smooth function, and p is a regular point of f, then p has an open neighborhood W such that $f|_W$ is a smooth diffeomorphism onto its image.*

Proof. Let $\varphi : U \to k^m$ and $\psi : V \to k^m$ be smooth coordinate charts for open sets $U \subset M$ and $V \subset N$ containing p and $f(p)$ respectively. If need be we can replace U with $U \cap f^{-1}(V)$, so we may assume that $f(U) \subset V$. Since φ and ψ are diffeomorphisms, $\varphi(p)$ is a regular point of $\psi \circ f \circ \varphi^{-1}$. The version of the inverse function theorem from Section 7.8 gives a neighborhood $\tilde{W}$ of $\varphi(p)$ such that $\psi \circ f \circ \varphi^{-1}|_{\tilde{W}}$ is a smooth diffeomorphism onto its image, which is an open subset of $\psi(V)$. Let $W := \varphi^{-1}(\tilde{W})$. Then, due to Lemma

8.5, Lemma 8.9, or Lemma 8.10, according to the meaning of "smooth,"

$$f|_W = \psi^{-1} \circ (\psi \circ f \circ \varphi^{-1}) \circ \varphi|_W$$

and

$$f^{-1}|_{f(W)} = \varphi^{-1} \circ (\varphi \circ f^{-1} \circ \psi^{-1}) \circ \psi|_{f(W)}$$

are smooth inverse diffeomorphisms. □

Chapter 9

Going Higher

We've now completed the book's main agenda of giving a conceptual description of and perspective on the material covered in the early college mathematics curriculum, through linear algebra and advanced calculus. In the remainder we'll discuss a few topics involving manifolds that were chosen because they apply what we have done, because they point to the concerns of advanced and contemporary mathematics, because they are simply quite beautiful and interesting, and because they were either originated by Riemann or have some relationship to his thought. If what came before was cake, the rest is icing.

This chapter is a bit different, and in some ways harder, than what has come before. Up to this point we've mainly been concerned with establishing a system of definitions, and although (with a couple exceptions) each chapter has featured one or two topics that are a bit more advanced and complex, most of the results we've proved have served to show that what we were doing was coherent, and had certain basic properties. The six essays that constitute this chapter each introduce concepts that are basic in the context of subsequent developments, but they are more strongly motivated by concrete mathematical questions and consequently feature more analysis, and denser argumentation.

9.1 Differential Geometry

All of the geometry of Euclidean space—distance, angle, shape—flows out of the standard inner product. The starting point of differential geometry is to endow a manifold with local geometry of this sort, then study the manifold's geometric properties, either at a somewhat larger scale where the manifold's curvature is apparent, or at a global scale. Gauss worked out the main ideas for two dimensional submanifolds of $\mathbb{R}^3$, and Riemann generalized the key concepts to the case of general dimension.

To keep things a bit simpler, in this section we'll work in the C^∞ category. It will be fairly easy to see that the definitions and analysis generalize to the C^r category for finite r, but it turns out that the additional generality doesn't allow qualitatively different phenomena: a precise explanation would be too complicated for inclusion here, but the general idea is that systematic elaboration of the consequences of the construction described in Section 7.7 leads eventually to the conclusion that any C^r phenomenon has a C^∞ approximation.

Everything we do makes sense in the real analytic category, but the qualitative or conceptual consequences of real analyticity seem not to have been studied extensively, and won't be discussed here. (At the same time the majority of concrete examples are real analytic.) For the category of holomorphic manifolds there are analogous constructions, but the subject has quite different qualitative properties, and is motivated by applications that are quite distant from those that give rise to the central concerns of real differential geometry.

So, let M be an n-dimensional C^∞ manifold over $\mathbb{R}$. Roughly, we would like to specify an inner product for each tangent space T_pM, and we want these inner products to "vary smoothly" as we move through the manifold. The precise description involves a new manifold. For each $p \in M$ let

$$T_p^2M := T_pM \times T_pM.$$

The union of these spaces is

$$T^2M := \bigcup_{p \in M} T_p^2M,$$

and $\pi_2 : T^2M \to M$ is the projection

$$\pi_2([p, \varphi, v_1], [p, \varphi, v_2]) := p.$$

If $\varphi : U \to V$ is a C^∞ coordinate chart for M, let

$$T_\varphi^2 : \pi_2^{-1}(U) \to V \times \mathbb{R}^n \times \mathbb{R}^n$$

be the function

$$T_\varphi^2([p, \varphi, v_1], [p, \varphi, v_2]) := (\varphi(p), v_1, v_2).$$

The discussion in Section 8.5 (with obvious adjustments) shows that $\{T^2_{\varphi_i}\}_{i\in I}$ is a C^∞ atlas for T^2M whenever $\{\varphi_i\}_{i\in I}$ is a C^∞ atlas for M.

The explicit but clunky notation $[p,\varphi,v]$ for tangent vectors is useful in connection with elementary foundational issues, but is unappealing in most other contexts. In the following definition elements of TM will be denoted by η,ζ,ξ, etc.

Definition 9.1. *A* ***Riemannian metric*** *for M is a C^∞ function*

$$\langle\cdot,\cdot\rangle : T^2M \to \mathbb{R}$$

such that for each $p \in M$ the restriction $\langle\cdot,\cdot\rangle_p$ of $\langle\cdot,\cdot\rangle$ to T^2_pM is an inner product. That is, for all $\eta,\zeta,\xi \in T_pM$ and all $\alpha \in \mathbb{R}$:

(a) $\langle\eta,\zeta\rangle_p = \langle\zeta,\eta\rangle_p$*;*

(b) $\langle\eta+\zeta,\xi\rangle_p = \langle\eta,\xi\rangle_p + \langle\zeta,\xi\rangle_p$*;*

(c) $\langle\alpha\eta,\zeta\rangle_p = \alpha\langle\eta,\zeta\rangle_p$*;*

(d) $\langle\eta,\eta\rangle_p \ge 0$ *with equality if and only if* $\eta = 0$.

A ***Riemannian manifold*** *is a C^∞ manifold endowed with a Riemannian metric. The inner product $\langle\cdot,\cdot\rangle_p$ induces a norm $\|\cdot\|_p$ on T_pM defined by the formula*

$$\|\zeta\|_p := \sqrt{\langle\zeta,\zeta\rangle_p}.$$

The large scale agenda is to use the Riemannian metric to define and study geometric concepts, and perhaps the most fundamental of these is distance. Before discussing distance in M, there are some generalities that pertain to any pathwise connected metric space (X,d). Recall that for $x_0,x_1 \in X$, a **path** or **curve** from x_0 to x_1 is a continuous function $\gamma : [a,b] \to X$ with $\gamma(0) = x_0$ and $\gamma(1) = x_1$. The **length** of γ is the supremum of the set of sums of the form

$$\sum_{i=1}^{k} d(\gamma(t_{i-1}),\gamma(t_i))$$

where $a \le t_0 < \ldots < t_k \le b$. The length of γ is always at least $d(x_0,x_1)$, and this supremum may easily be infinite. Let $d^*(x_0,x_1)$ be the infimum, over all paths γ from x_0 to x_1, of the length of γ. If $d^*(x,y) < \infty$ for all $x,y \in X$

(this can easily fail to be the case[1]) then d^* is a metric for X. (Make sure you understand why!) Clearly $d^*(x_0, x_1) \geq d(x_0, x_1)$, so every d-open set is d^*-open, but there can be d^*-open sets that aren't d-open. For someone who can only move around by following paths in X, d^* is the "real" metric, and its induced topology is the "real" topology of X.

In connection with M we wish to use the Riemannian metric to define a notion of curve length for C^1 curves, after which we can define the "real" metric of M using the procedure described above. Consider a C^1 function $\gamma : [a, b] \to M$. For $t \in \mathbb{R}$ we may take $[t, \mathrm{Id}_{\mathbb{R}}, 1]$ as the "standard" unit basis vector of the one dimensional vector space $T_t\mathbb{R}$. When t is in the domain of the curve γ we think of

$$\gamma'(t) := D\gamma(t)[t, \mathrm{Id}_{\mathbb{R}}, 1] \in T_{\gamma(t)}M$$

as the **velocity** of γ at time t, and the **speed** of γ at time t is $\|\gamma'(t)\|_{\gamma(t)}$. If, for example, $\|\gamma'(t)\|_{\gamma(t)} = s$ for all t, then the length of γ is $s(b - a)$. Roughly, we will define the length of γ to be the limit, in a certain sense, of the sums

$$\sum_{i=1}^{k} \|\gamma'(t_i)\|_{\gamma(t_i)}(t_i - t_{i-1})$$

where $a \leq t_0 < \ldots < t_k \leq b$.

In preparation for the precise definition of curve length for curves like γ we now discuss integration of continuous real valued functions on compact intervals. This is done with some regret, for the following reason. Any mathematical document needs to set and obey bounds on its scope in order to avoid growing to a length that defeats its purpose. Even though integration is coequal to differentiation as a component of the calculus, and is typically studied in conjunction with the topics discussed in this book, avoiding it has done much to prevent things from becoming even more bloated, and for this reason introducing it at this point feels wrong. In practice things aren't so bad: we will be able to keep the discussion brief by considering only the simplest case.

[1] For a concrete example let X be the image of the curve $\gamma : [0, 1] \to \mathbb{R}^2$ where $\gamma(t) := (t, t \sin 1/t)$. The distance between two points in $\mathbb{R}^2$ is at least as large as the absolute value of the difference between their respective second components, so for each $k = 1, 2, \ldots$ we have

$$\|\gamma(\tfrac{2}{(4k+1)\pi}) - \gamma(\tfrac{2}{(4k+3)\pi})\| \geq \tfrac{2}{(4k+1)\pi} + \tfrac{2}{(4k+3)\pi} > \tfrac{1}{(k+1)\pi}.$$

Since the harmonic series diverges, the length of any path from $(0, 0)$ to any other point in X is infinite.

There is another reason to be unhappy with our discussion of integration. Like differentiation, integration as it was understood by Newton and Leibniz could not be defined with complete rigor prior to the late 19^{th} century. The theory as it was formulated then, which is what is taught in introductory calculus courses, actually suffers from severe limitations on the class of functions it considers. There were attempts, including one by Riemann, to develop more general definitions, but a fully satisfactory theory emerged only in the 1920's. It is one of the greatest success stories of the set theory revolution, and a magnificent piece of abstraction, giving a solid foundation for probability and statistics, among many other things. Since the discussion of integration below avoids this material, it fails to live up to our attitude of enthusiastically embracing abstraction.

Suppose that $a \leq b$, and let $f : [a, b] \to \mathbb{R}$ be a continuous function. We adopt the following notation: if S is a set, D_S is the set of finite subsets of S. A typical element of $D_{[a,b]}$ is $\iota = \{t_0, \dots, t_k\}$; in this circumstance we always assume that $t_0 < \cdots < t_k$. For such an ι let

$$I_\iota(f) := \sum_{j=1}^{k} f(t_k)(t_k - t_{k-1}).$$

If $\gamma : [a, b] \to M$ is a C^1 curve and $f(t) := \|\gamma'(t)\|_{\gamma(t)}$ is the speed of γ at time t, then $f(t_k)(t_k - t_{k-1})$ is an approximation of the distance traversed by γ between t_{k-1} and t_k, so $I_\iota(f)$ is an approximation of the total distance travelled between times a and b. When f is everywhere nonnegative, we can think of $f(t_k)(t_k - t_{k-1})$ as the area of the rectangle $[t_{k-1}, t_k] \times [0, f(t_k)]$, so that $I_\iota(f)$ is an approximation of the area under the graph of f.

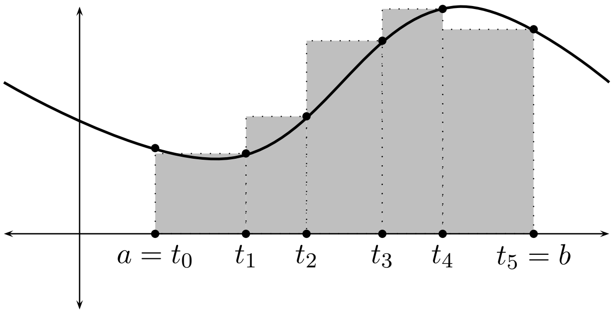


Figure 9.2

This approximation of area or distance travelled becomes more accurate as the set ι becomes larger, and we would like to define the integral of f to be the "limit" of $I_\iota(f)$ as the set $\{t_0, \dots, t_k\}$ "converges" to the interval $[a, b]$. One way to do this is to take a particular sequence of sets. For example, the k^{th} set could be

$$\{a, \tfrac{k-1}{k}a + \tfrac{1}{k}b, \dots, \tfrac{1}{k}a + \tfrac{k-1}{k}b, b\},$$

and we could define the integral of f to be

$$\lim_{k\to\infty} \sum_{i=1}^{k} f\big(\tfrac{k-i}{k}a + \tfrac{i}{k}b\big)\tfrac{b-a}{k}.$$

This works, and possibly most authors would develop the subject in this way, but we will take a more abstract approach that leads to a more flexible setup. Developing an initial understanding of this definition will take a bit more effort, but its pliability will make it easier to prove things.

Let R be a binary relation on a set D. We say that R is **antisymmetric** if there are no two elements $x, y \in D$ such that both xRy and yRx. That is, whenever xRy it is not also the case that yRx. As you probably recall, the relation R is **transitive** if xRz whenever xRy and yRz. A binary relation R on D is a **partial order** if it is antisymmetric and transitive. When R is a partial order and xRy, we say that x **precedes** y and that y **succeeds** x. The relation R is **irreflexive** if there is no $x \in D$ such that xRx. Of course antisymmetry implies irreflexivity, so in a partial ordering no element precedes or succeeds itself.

The real numbers (and the rational numbers, and the integers) are partially order by "is less than." In fact for any n, there is a partial order R on $\mathbb{R}^n$ in which xRy if and only if $x_i < y_i$ for all $i = 1, \dots, n$. The subsets of any set are partially ordered by "is a proper subset of," and the open sets of a topological space (and also the closed sets) are partially ordered by the restriction of this relation. I am afraid these examples do not begin to do justice to the significance and generality of partial orders. Partial orders arise naturally in all branches of mathematics, and it would not surprise me if you could list quite a few more examples without working up a sweat.

The pair $(D, <)$ is a **directed set** if D is a nonempty set, $<$ is a partial ordering of D, and for any two elements ι_1 and ι_2 of D there is third element ι such that $\iota_1 < \iota$ and $\iota_2 < \iota$. Taking $\iota_1 = \iota_2$, we find that every element of a directed set has a successor, that successor has a successor, and so forth, so a directed set is necessarily infinite. If S is an infinite set and $<$ is the "is a proper subset of" relation on D_S, then for any $\iota_1, \iota_2 \in D_S$ we can (if necessary) append some new element of S to $\iota_1 \cup \iota_2$ to create an ι with $\iota_1 < \iota$ and $\iota_2 < \iota$, so D_S is a directed set. Of course another example of a directed set is given by the real numbers (or the rational numbers, or the integers) with the partial ordering "is less than."

Let X be a topological space. A **net** in X is a function $\iota \mapsto x_\iota$ from D to X where $(D, <)$ is a directed set. Such a net is said to **converge** to a point $x \in X$ if, for every open set U containing x, there is some $\iota_U \in D$ such that $x_\iota \in U$ whenever ι is a successor of ι_U. When this is the case we write

$$x = \varinjlim x_\iota.$$

Since the integers, with the usual ordering, are a directed set, this convergence concept has convergence of sequences as a special case, and it actually plays a key role in the foundations of topology. Among other things, it can happen that a point x in a topological space X is a limit of a net in $X \setminus \{x\}$ even though there are no sequences in $X \setminus \{x\}$ that converge to x.

We can now define the **integral** of f to be 0 if $a = b$, and otherwise it is

$$\int_a^b f(t)\,dt := \varinjlim I_\iota(f)$$

where the limit is over $\iota \in D_{[a,b]}$. Of course the first agenda item is to show that this limit always exists, so that the integral is well defined. This will take a bit of work, in part because we take advantage of the nice opportunity it presents to explain an interesting and useful fact (Proposition 9.3 below) about functions between metric spaces.

Definition 9.2. *A function $g : X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is* ***uniformly continuous*** *if, for every $\varepsilon > 0$, there is some $\delta > 0$ such that*

$$d_Y(g(x), g(x')) < \varepsilon \quad \textit{whenever} \quad d_X(x, x') < \delta.$$

Comparing this with the definition of continuity, we see that there is a propositional function

$$P(\varepsilon, x, \delta, x') := \text{`}d_X(x, x') < \delta \ \Rightarrow\ d_Y(g(x), g(x')) < \varepsilon\text{'}$$

such that g is continuous if $(\forall\varepsilon)(\forall x)(\exists\delta)(\forall x')P(\varepsilon, x, \delta, x')$ and uniformly continuous if $(\forall\varepsilon)(\exists\delta)(\forall x)(\forall x')P(\varepsilon, x, \delta, x')$. In words, the definition of uniform continuity insists that you commit to a particular δ before you know x, while continuity allows you to select a different δ for each x. Therefore uniform continuity is, in general, a more demanding concept than continuity, but there is an important case in which the two concepts are equivalent.

Proposition 9.3. *Suppose (X, d_X) and (Y, d_Y) are metric spaces and $g : X \to Y$ is continuous. If X is compact, then g is uniformly continuous.*

Proof. Aiming at a contradiction, suppose that g is not uniformly continuous. As we explained at the beginning of Chapter 3, the negation of

$$(\forall\varepsilon)(\exists\delta)(\forall x)(\forall x')P(\varepsilon, x, \delta, x') \quad \text{is} \quad (\exists\varepsilon)(\forall\delta)(\exists x)(\exists x')\neg P(\varepsilon, x, \delta, x'),$$

so there is an $\varepsilon > 0$ such that for each $\delta_n := 1/n$ there are $x_n, x'_n \in X$ with $d_X(x_n, x'_n) < \delta_n$ and $d_Y(g(x_n), g(x'_n)) \geq \varepsilon$. Since X is compact (Theorem 3.44) the sequence $\{x_n\}$ has a subsequence $\{x_{n_i}\}_{i=1,2,\ldots}$ that converges to some point, say x. Clearly $\{x'_{n_i}\}_{i=1,2,\ldots}$ also converges to x, and the continuity of g implies that

$$\lim_{i\to\infty} g(x_{n_i}) = g(x) = \lim_{i\to\infty} g(x'_{n_i}).$$

In turn this implies that $d_Y(g(x_{n_i}), g(x'_{n_i})) < \varepsilon$ for sufficiently large i, contradicting our supposition. □

So, our given $f : [a, b] \to \mathbb{R}$ is uniformly continuous. The specific consequence of interest is:

Lemma 9.4. *Suppose that $\iota, \iota' \in D_{[a,b]}$ with $\iota < \iota'$, $\iota = \{t_0, \ldots, t_k\}$, $t_0 = a$, $t_k = b$, and $t_j - t_{j-1} < \delta$ for all $j = 1, \ldots, k$. If $|f(t) - f(t')| < \varepsilon$ whenever $|t - t'| < \delta$, then*

$$|I_{\iota'}(f) - I_\iota(f)| \leq \varepsilon(b - a).$$

Proof. Suppose that $\iota' = \{u_0, \ldots, u_\ell\}$. Then for each $j = 0, \ldots, k$ there is some h_j such that $u_{h_j} = t_j$, and $t_j - t_{j-1} = \sum_{h=h_{j-1}+1}^{h_j} u_h - u_{h-1}$. We have

$$I_\iota(f) = \sum_{j=1}^{k} f(t_j)(t_j - t_{j-1}) = \sum_{j=1}^{k} \sum_{h=h_{j-1}+1}^{h_j} f(t_j)(u_h - u_{h-1})$$

and

$$I_{\iota'}(f) = \sum_{h=1}^{\ell} f(u_h)(u_h - u_{h-1}) = \sum_{j=1}^{k} \sum_{h=h_{j-1}+1}^{h_j} f(u_h)(u_h - u_{h-1}).$$

If $h_{j-1} + 1 \leq h \leq h_j$, then $t_j - u_h < t_j - t_{j-1} < \delta$ and $|f(u_h) - f(t_j)| < \varepsilon$, so

$$\left|I_{\iota'}(f) - I_\iota(f)\right| \leq \sum_{j=1}^{k} \sum_{h=h_{j-1}+1}^{h_j} |f(u_h) - f(t_j)|(u_h - u_{h-1})$$

$$< \varepsilon \sum_{j=1}^{k} \sum_{h=h_{j-1}+1}^{h_j} (u_h - u_{h-1}) = \varepsilon(b-a).$$

□

Aiming at showing that $\int_a^b f(t)\,dt = \varinjlim I_\iota(f)$ is well defined, suppose that for each $n = 1, 2, \ldots$ we have chosen $\delta_{1/n} > 0$ small enough that $|f(t) - f(t')| < 1/n$ whenever $|t - t'| < \delta_{1/n}$, and we have also chosen $\iota_n = \{t_0, \ldots, t_k\} \in D_{[a,b]}$ with $t_0 = a$, $t_k = b$, and $t_j - t_{j-1} < \delta_{1/n}$ for all $j = 1, \ldots, k$. Then $\{I_{\iota_n}(f)\}$ is a Cauchy sequence: for any $m, m' > n$ we can choose $\iota \in D_{[a,b]}$ such that $\iota_m, \iota_{m'} < \iota$, in which case the last result gives

$$|I_{\iota_m}(f) - I_{\iota_{m'}}(f)| \le |I_{\iota_m}(f) - I_\iota(f)| + |I_\iota(f) - I_{\iota_{m'}}(f)|$$

$$\le (\tfrac{1}{m} + \tfrac{1}{m'})(b-a) < \tfrac{2}{n}(b-a).$$

Since $\mathbb{R}$ is complete, the sequence $\{I_{\iota_n}(f)\}$ has a limit. Moreover, since $|I_\iota(f) - I_{\iota_n}(f)| \le (b-a)/n$ whenever $\iota_n < \iota$, this limit satisfies the condition defining $\varinjlim I_\iota(f)$.

Now that we know that integration is a well defined operation, there are, of course, a great many things we could say about it. Since we can't hope to be comprehensive, it seems best, on the whole, to say as little as possible, but we need three basic properties that appear in almost every argument involving an integral.

Lemma 9.5. *If $m \le f(t) \le M$ for all t, then*

$$m(b-a) \le \int_a^b f(t)\,dt \le M(b-a).$$

Proof. This follows from the inequality $m(b-a) \le I_\iota(f) \le M(b-a)$, which is in turn an immediate consequence of the definition of $I_\iota(f)$. □

Lemma 9.6. *If $a = t_0 \le t_1 \le \cdots \le t_{k-1} \le t_k = b$, then*

$$\int_a^b f(t)\,dt = \int_{t_0}^{t_1} f(t)\,dt + \cdots + \int_{t_{k-1}}^{t_k} f(t)\,dt.$$

Proof. Since the integral over a degenerate interval consisting of a single point is 0, this holds automatically when $a = b$, and when $a < b$ we may assume that $t_0 < t_1 < \cdots < t_{k-1} < t_k$ because we can remove trivial terms from the right hand side of the asserted equation. Let $\iota := \{t_0, \dots, t_k\}$. If $\iota < \iota' = \{u_0, \dots, u_\ell\}$ with $u_{h_j} = t_j$ for each $j = 0, \dots, k$, then

$$\sum_{h=h_{j-1}+1}^{h_j} f(u_h)(u_h - u_{h-1}) = I_{\iota' \cap [t_{j-1}, t_j]}(f|_{[t_{j-1}, t_j]})$$

for each $j = 1, \dots, k$, and consequently

$$I_{\iota'}(f) = I_{\iota' \cap [t_0, t_1]}(f|_{[t_0, t_1]}) + \cdots + I_{\iota' \cap [t_{k-1}, t_k]}(f|_{[t_{k-1}, t_k]}).$$

The left hand side converges to $\int_a^b f(t)\, dt$. For any $\iota_1 \in D_{[t_0, t_1]}, \dots, \iota_k \in D_{[t_{k-1}, t_k]}$ we can easily construct an ι' with $\iota_1 < \iota' \cap [t_0, t_1], \dots, \iota_k < \iota' \cap [t_{k-1}, t_k]$. Therefore the right hand side of this equation can be made arbitrarily close to the right hand side of the asserted equation. □

Lemma 9.7. *The integral is a linear function from the space of continuous real valued functions on $[a, b]$ to $\mathbb{R}$: if $f, g : [a, b] \to \mathbb{R}$ are continuous and $\alpha \in \mathbb{R}$, then*

$$\int_a^b (\alpha f + g)(t)\, dt = \alpha \int_a^b f(t)\, dt + \int_a^b g(t)\, dt.$$

Proof. Once again, this is automatic when $a = b$, so suppose that $a < b$. For any $\iota \in D_{[a,b]}$ the equation $I_\iota(\alpha f + g) = \alpha I_\iota(f) + I_\iota(g)$ follows directly from the definition of $I_\iota(\cdot)$, and the claim obviously follows from this equation. □

In addition, there is a famous theorem that provides useful insights into the definition of curve length below, as well as being extremely important for many other reasons. Essentially it asserts that *integration and differentiation are inverse operations*. In our applications the derivative of the curve length function at time t is the speed at time t, and the integral of the speed from a to t is the distance traveled up to time t.

Theorem 9.8 (The Fundamental Theorem of Calculus)**.** *For continuous functions $f : [a, b] \to \mathbb{R}$ and $F : [a, b] \to \mathbb{R}$ the following conditions are equivalent:*

(a) $F(t) = F(a) + \int_a^t f(s)\, ds$ *for all* $t \in [a, b]$.

(b) F *is* C^1 *with* $F'(t) = f(t)$ *for all* $t \in (a, b)$.

Proof. Both (a) and (b) are true automatically when $a = b$, so assume that $a < b$.

First suppose (a) holds. Fixing $t \in (a, b)$, we will show that $F'(t) = f(t)$, after which it follows that F is C^1 because f is continuous by assumption. Consider $\varepsilon > 0$. Since f is continuous, there is $\delta > 0$ such that $|f(t') - f(t)| < \varepsilon$ for all $t' \in (t - \delta, t + \delta)$. Fix such a t' with $t' \geq t$. (The case $t' \leq t$ is similar.) We have

$$\begin{aligned} F(t') - F(t) &= \big(F(t') - F(a)\big) - \big(F(t) - F(a)\big) \\ &= \int_a^{t'} f(s)\,ds - \int_a^t f(s)\,ds = \int_t^{t'} f(s)\,ds, \end{aligned}$$

where the final equality comes from Lemma 9.6, and Lemma 9.5 implies that

$$(f(t) - \varepsilon)(t' - t) < \int_t^{t'} f(s)\,ds < (f(t) + \varepsilon)(t' - t).$$

Therefore

$$\Big|F(t') - [F(t) + f(t)(t' - t)]\Big| = \Big|\int_t^{t'} f(s)\,ds - f(t)(t' - t)\Big| < \varepsilon|t' - t|.$$

Now suppose that (b) holds. Clearly (a) will follow if we show that

$$\Big|\int_a^t f(s)\,ds - (F(t) - F(a))\Big| < \varepsilon(t - a)$$

for any $t \in (a, b)$ and $\varepsilon > 0$, so fix t and ε. Since f is uniformly continuous (Proposition 9.3) there is a $\delta > 0$ such that $|f(s) - f(s')| < \varepsilon$ whenever $|s - s'| < \delta$. Choose $t_0 < \cdots < t_k$ with $t_0 = a$, $t_k = t$, and $t_h - t_{h-1} < \delta$ for all $h = 1, \ldots, k$. For each such h the mean value theorem implies the existence of $\tilde{t}_h \in (t_{h-1}, t_h)$ such that

$$F(t_h) - F(t_{h-1}) = f(\tilde{t}_h)(t_h - t_{h-1}),$$

and Lemma 9.5 implies that

$$\Big|\int_{t_{h-1}}^{t_h} f(s)\,ds - f(\tilde{t}_h)(t_h - t_{h-1})\Big| < \varepsilon(t_h - t_{h-1}).$$

Applying Lemma 9.6, we compute that

$$\Big|\int_a^t f(s)\,ds - (F(t) - F(a))\Big| = \Big|\sum_{h=1}^k \Big(\int_{t_{h-1}}^{t_h} f(s)\,ds - (F(t_h) - F(t_{h-1}))\Big)\Big|$$

$$\leq \sum_{h=1}^{k} \Big| \int_{t_{h-1}}^{t_h} f(s)\,ds - f(\tilde{t}_h)(t_h - t_{h-1})) \Big| < \sum_{h=1}^{k} \varepsilon(t_h - t_{h-1}) = \varepsilon(t-a).$$

□

We haven't defined the derivative at a or the derivative at b of a function whose domain is $[a,b]$. There are sensible definitions using "one sided limits," and once such definitions are in place it is not hard to strengthen (b) to require also that $F'(a) = f(a)$ and $F'(b) = f(b)$. In the future we usually won't worry about this little detail. For instance, in the result below we assume that a function on $[c,d]$ is C^1, and this should be understood as meaning that the function is differentiable everywhere including the endpoints, and the derivative is continuous at every point in $[c,d]$.

Returning to the Riemannian setting, let M be a Riemannian manifold, and let $\gamma : [a,b] \to M$ be a C^1 path. With the theory of the integral under our belts, we can now define the **length** of γ to be the total distance travelled:

$$L(\gamma) := \int_a^b \|\gamma'(t)\|_{\gamma(t)}\,dt.$$

As everyone knows, the total distance travelled in going from $\gamma(a)$ to $\gamma(b)$ depends only on the image of γ, in the following sense. Suppose $\tilde{\gamma} : [c,d] \to M$ is a different curve that covers the same ground according to a different schedule without "backtracking," by which we mean that there is an increasing C^1 function $\tau : [c,d] \to [a,b]$ such that $\tau(c) = a$, $\tau(d) = b$, and $\tilde{\gamma} = \gamma \circ \tau$. In this circumstance we say that $\tilde{\gamma}$ is a **reparameterization** of γ. If you go from $\gamma(a)$ to $\gamma(b)$ according to the schedule specified by γ, while your friend's itinerary is $\tilde{\gamma}$, the two of you will cover the same total distance.

The following result is a more general formulation of the underlying principle.

Proposition 9.9 (Change of Variables Formula)**.** *If $f : [a,b] \to \mathbb{R}$ is continuous, $\lambda : [c,d] \to [a,b]$ is C^1, and $\lambda(c) \leq \lambda(d)$, then*

$$\int_c^d f(\lambda(\sigma))\lambda'(\sigma)\,d\sigma = \int_{\lambda(c)}^{\lambda(d)} f(s)\,ds.$$

This formula has a simple intuition: if you replace d with $d + \Delta d$, the right hand side increases by approximately $f(\lambda(d))\lambda'(d)\Delta d$. As you learn more advanced theories of integration you will see other change of variables formulas, each based on some variant of this insight. Note that, among other things, there is no need to require that $\lambda'(t) \ge 0$ for all t. Also, the only reason we require $\lambda(c) \le \lambda(d)$ is that we have not defined $\int_a^b f(t)\,dt$ when $b < a$. Setting $\int_a^b f(t)\,dt := -\int_b^a f(t)\,dt$ in this circumstance works perfectly well, even if it might be preferable, conceptually and aesthetically, to revise the discussion above so that the two cases are treated symmetrically.

Proof. For $a \le t \le b$ let $F(t) := \int_a^t f(s)\,ds$, and for $c \le \tau \le d$ let $G(\tau) := F(\lambda(\tau))$. The fundamental theorem of calculus and the chain rule imply that $G'(\tau) = f(\lambda(\tau))\lambda'(\tau)$, so

$$\int_c^d f(\lambda(\sigma))\lambda'(\sigma)\,d\sigma = G(d) - G(c) = F(\lambda(d)) - F(\lambda(c)) = \int_{\lambda(c)}^{\lambda(d)} f(s)\,ds,$$

where the first inequality is another application of the fundamental theorem of calculus, and the last is derived from Lemma 9.6. □

As promised above, we now show that two curves have the same length if they cover the same ground at different speeds. Suppose that $c < d$ and $\tau : [c,d] \to [a,b]$ is a C^1 function with $\tau(c) = a$, $\tau(d) = b$, and $\tau'(\sigma) \ge 0$ for all $\sigma \in [c,d]$, so that

$$\tilde{\gamma} := \gamma \circ \tau : [c,d] \to M$$

is a reparameterization of γ. The chain rule gives

$$\|\tilde{\gamma}'(\sigma)\|_{\tilde{\gamma}(\sigma)} = \|\gamma'(\tau(\sigma))\tau'(\sigma)\|_{\gamma(\tau(\sigma))} = \|\gamma'(\tau(\sigma))\|_{\gamma(\tau(\sigma))} \cdot |\tau'(\sigma)|,$$

so

$$L(\tilde{\gamma}) = \int_c^d \|\tilde{\gamma}'(\sigma)\|_{\tilde{\gamma}(\sigma)}\,d\sigma = \int_c^d \|\gamma'(\tau(\sigma))\|_{\gamma(\tau(\sigma))} \cdot |\tau'(\sigma)|\,d\sigma.$$

Since $|\tau'(\sigma)| = \tau'(\sigma)$ we can apply the change of variables formula:

$$L(\tilde{\gamma}) = \int_c^d \|\gamma'(\tau(\sigma))\|_{\gamma(\tau(\sigma))} \cdot \tau'(\sigma)\,d\sigma = \int_a^b \|\gamma'(s)\|_{\gamma(s)}\,ds = L(\gamma).$$

The curves of greatest geometric interest are those that are "as straight as possible." The curve γ is **distance minimizing** if there is no other curve $\tilde{\gamma} : [c,d] \to M$ with $\tilde{\gamma}(c) = \gamma(a)$, $\tilde{\gamma}(d) = \gamma(b)$, and $L(\tilde{\gamma}) < L(\gamma)$. (To visualize this imagine holding a piece of string against two points of a football and reducing its length until it it taut.) It is **locally distance minimizing** if, for each $t \in (a,b)$, there are a' and b' with $a \le a' < t < b' \le b$ such that $\gamma|_{[a',b']}$ is distance minimizing, and an analogous condition holds at the endpoints: $\gamma|_{[a,a+\varepsilon]}$ and $\gamma|_{[b-\varepsilon,b]}$ are distance minimizing for some $\varepsilon > 0$. A locally distance minimizing curve of constant speed seems like the natural generalization of the Newtonian notion of an inertial trajectory, and in fact this is true not only "in principle," but also in physical reality as described by the general theory of relativity.

Naturally we expect a distance minimizing curve to be locally distance minimizing. This is true, and the basic idea is simple and obvious: basic facts about integration (specifically, Lemma 9.6) give

$$L(\gamma) = L(\gamma|_{[a,a']}) + L(\gamma|_{[a',b']}) + L(\gamma|_{[b',b]})$$

when $a \le a' < b' \le b$. If γ is distance minimizing, then $\gamma|_{[a',b']}$ should be distance minimizing because otherwise a shorter curve from $\gamma(a')$ to $\gamma(b')$ could be reparameterized to the interval $[a',b']$, then combined with the rest of γ to give a shorter path from $\gamma(a)$ to $\gamma(b)$. However, there is actually a nasty technical detail here, insofar as the curve resulting from this gluing procedure need not be C^1. With a bit of work one can show that there is a nearby C^1 curve with approximately the same length, but the easier approach is to modify our definitions to allow the curves in our definition of distance minimization to be "piecewise" C^1, where a curve is piecewise C^1 if it is C^1 on a each of a finite collection of intervals that cover the domain.

Roughly, the image of a locally distance minimizing curve is the analogue in Riemannian geometry of the notion of a line in Euclidean geometry. However, we need to be careful in formulating this definition. In Euclidean geometry lines extend indefinitely, but if we try to extend a locally distance minimizing curve it can intersect itself. A great circle on the surface of the Earth does this in a well behaved fashion, in that if you keep following it you just go round and round, but by deforming the surface of the Earth we could actually arrange for the curve to intersect itself at an angle. This consideration suggests that our definition should have a local character, describing sets that are images of locally distance minimizing curves in a neighborhood of each point. Also, we do not want sets like the union of two parallel lines to satisfy the definition, so it makes sense to insist that the set be connected.

Definition 9.10. *A* ***geodesic*** *in M is a nonempty connected set $g \subset M$ such that for any $p \in g$ there is an open set $U \subset M$ containing p and a distance minimizing curve $\gamma : [-\varepsilon, \varepsilon] \to M$, for some $\varepsilon > 0$, such that $\gamma'(t) \neq 0$ for all t, $\gamma(0) = p$, and*

$$g \cap U = \operatorname{image}(\gamma) \cap U.$$

A geodesic is ***complete*** *if it is not a proper subset of another geodesic.*

We are now confronted with a number of interesting foundational issues. In Euclidean geometry any two distinct points are contained in exactly one line. Two points in the sphere that are **antipodal** (that is, diametrically opposite each other) are both contained in a continuum of distinct complete geodesics, so we know that this property of Euclidean geometry doesn't extend to Riemannian geometry at a global scale. Nevertheless, there is a local generalization: each $p \in M$ has an open neighborhood $U \subset M$ such that for each $q \in U$ there is a unique distance minimizing curve $\gamma : [0,1] \to U$ of constant speed with $\gamma(0) = p$ and $\gamma(1) = q$.

A closely related issue is that Newtonian physics is a **deterministic** theory, as is the description given by general relativity of a single particle moving in a force field. The rough idea is that if we know the "state" of the system at a point in time, then we can predict its state at any time in the future or infer its state at any time in the past. In the Riemannian context a constant speed distance minimizing curve need not continue forever, so this principle has a local character. Mathematically, what determinism boils down to in this particular setting is that for any $p \in M$, any $\zeta \in T_pM$, and any sufficiently small $\varepsilon > 0$, there is exactly one constant speed distance minimizing curve $\gamma : [-\varepsilon, \varepsilon] \to M$ with $\gamma(0) = p$ and $\gamma'(0) = \zeta$.

A precise development of the results described in the last two paragraphs would lead to quite a bit of interesting and important mathematics, but it would take many pages, and involve techniques that are somewhat more advanced than those described in this book. In addition, even after we had done all this, we would still be at the very beginning of a huge body of mathematics with many more foundational issues to consider. Instead, an in-depth exploration of the geometry of a single concrete example seems like a more effective way to introduce some of the ideas of Riemannian geometry.

9.2 Hyperbolic Space

Two dimensional hyperbolic space is a very special two dimensional Riemannian manifold. It is one of the two examples discovered in the 19^{th} century that show that the first four axioms of Euclid do not imply the parallel postulate. Our goal is to explain this rigorously.

The particular description of hyperbolic space studied here is called the **Poincaré disk model** after its originator Henri Poincaré (1854-1912). The Poincaré disk has a very simple definition. Let

$$H = \{\, (x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1 \,\}$$

be the usual open unit disk in $\mathbb{R}^2$. We impose the following Riemannian metric on T^2H: if $(x, y) \in H$ and $\zeta, \eta \in T_{(x,y)}H$, let

$$\langle \zeta, \eta \rangle^H := \frac{\langle \zeta, \eta \rangle}{(1 - x^2 - y^2)^2},$$

where the inner product on the right hand side is the standard one. Roughly this means that the distance between two points near (x, y) is magnified by a factor of approximately $(1 - x^2 - y^2)^{-1}$ in comparison with the distance between them when the unit disk has its usual metric. On its surface this definition doesn't tell us much, and our study of H will be a circuitous affair, with various interesting twists and turns.

The complete geodesics of H will play the role of lines in our comparison of the geometry of H with Euclidean geometry. Let

$$g_0 := \{ (s, 0) : -1 < s < 1 \}$$

be the intersection of the x-axis with H. In a journey between two points in g_0, going away from the x-axis increases the distance in H that one covers in order to achieve a certain amount of left-to-right progress as measured in the usual metric of $\mathbb{R}^2$, so we should expect that a distance minimizing path between two points of g_0 will not stray from g_0. The first step in our analysis is give a precise quantitative argument showing that this is the case. We should also expect that g_0 is a geodesic. Of course it's connected, so what we need to show is that any point has a neighborhood that is contained in the image of a distance minimizing curve.

Fixing s_a, s_b with $-1 < s_a < s_b < 1$, let $\gamma = (\gamma_x, \gamma_y) : [a, b] \to H$ be any C^1 curve with $\gamma(a) = (s_a, 0)$ and $\gamma(b) = (s_b, 0)$. We first compare the length of γ with the length of the curve $\tilde{\gamma} : (a, b) \to H$ given by $\tilde{\gamma}(s) := (\gamma_x(t), 0)$. We have $\gamma'(t) = (\gamma'_x(t), \gamma'_y(t))$, $\tilde{\gamma}'(t) = (\gamma'_x(t), 0)$, and

$$\begin{aligned} \|\gamma'(t)\|^H_{\gamma(t)} &= \sqrt{\frac{\gamma'_x(t)^2 + \gamma'_y(t)^2}{(1 - \gamma_x(t)^2 - \gamma_y(t)^2)^2}} \geq \sqrt{\frac{\gamma'_x(t)^2}{(1 - \gamma_x(t)^2)^2}} \\ &= \frac{|\gamma'_x(t)|}{1 - \gamma_x(t)^2} = \|\tilde{\gamma}'(t)\|^H_{\tilde{\gamma}(t)}. \end{aligned}$$

The definition of curve length, this formula, and monotonicity of integration (Lemma 9.5) give:

$$L(\gamma) = \int_a^b \|\gamma'(t)\|^H_{\gamma(t)}\, dt \geq \int_a^b \|\tilde{\gamma}'(t)\|^H_{\tilde{\gamma}(t)}\, dt = L(\tilde{\gamma}).$$

When is it the case that $L(\gamma) = L(\tilde{\gamma})$? A simple argument based on Lemmas 9.5 and 9.6 shows that this inequality is strict if there is even a single t such that $\|\gamma'(t)\|^H_{\gamma(t)} > \|\tilde{\gamma}'(t)\|^H_{\tilde{\gamma}(t)}$, since then (due to continuity) there is an interval of positive length along which it holds strictly. On the other hand, if $\|\gamma'(t)\|^H_{\gamma(t)} = \|\tilde{\gamma}'(t)\|^H_{\tilde{\gamma}(t)}$ for all t, then $L(\gamma) = L(\tilde{\gamma})$. We always have $\gamma'_x(t)^2 + \gamma'_y(t)^2 \geq \gamma'_x(t)^2$ and $1 - \gamma_x(t)^2 - \gamma_y(t)^2 \leq 1 - \gamma_x(t)^2$, so $\|\gamma'(t)\|^H_{\gamma(t)} > \|\tilde{\gamma}'(t)\|^H_{\tilde{\gamma}(t)}$ if $\gamma'_y(t) \neq 0$, or if $\gamma'_x(t) \neq 0$ and $\gamma_y(t) \neq 0$. It is intuitively obvious that $\gamma'_y(t) = 0$ for all t if and only if $\gamma_y(t) = 0$ for all t. (There is an easy argument based on the mean value theorem, which you might try to construct, that gives a formal proof.) Therefore $L(\gamma) = L(\tilde{\gamma})$ if and only if $\gamma_y(t) = 0$ for all t.

Next we compare the length of $\tilde{\gamma} = (\gamma_x, 0)$ with the length of $\eta|_{[s_a,s_b]}$ where $\eta = (\eta_x, 0) : (-1, 1) \to H$ is the function $\eta(s) := (s, 0)$. In preparation for the change of variables formula we note that for each t we have

$$\frac{\gamma'_x(t)}{1 - \gamma_x(t)^2} = \frac{{\eta_x}'(\gamma_x(t))}{1 - \eta_x(\gamma_x(t))^2} \cdot \gamma'_x(t)$$

because $\eta_x(\gamma_x(t)) = \gamma_x(t)$ and ${\eta_x}'(\gamma_x(t)) = 1$. Monotonicity of integration (Lemma 9.5) and the change of variables formula (Proposition 9.9) give:

$$L(\tilde{\gamma}) = \int_a^b \|\tilde{\gamma}'(t)\|^H_{\tilde{\gamma}(t)}\,dt = \int_a^b \frac{|\gamma'_x(t)|}{1 - \gamma_x(t)^2}\,dt \geq \int_a^b \frac{\gamma'_x(t)}{1 - \gamma_x(t)^2}\,dt$$

$$= \int_a^b \frac{{\eta_x}'(\gamma_x(t))}{1 - \eta_x(\gamma_x(t))^2} \cdot \gamma'_x(t)\,dt = \int_{s_a}^{s_b} \frac{{\eta_x}'(s)}{1 - \eta_x(s)^2}\,ds = L(\eta|_{[s_a,s_b]}),$$

with strict inequality if and only if $\gamma'_x(t) = \gamma'_x(t) < 0$ for some t.

We have shown that $L(\gamma) \geq L(\tilde{\gamma}) \geq L(\eta|_{[s_a,s_b]})$. Since γ could be any path between $(s_a, 0)$ and $(s_b, 0)$, $\eta|_{[s_a,s_b]}$ is distance minimizing, so g_0 is a geodesic. We have $L(\gamma) = L(\tilde{\gamma})$ if and only if $\gamma_y(t) = 0$ for all t, and we have $L(\tilde{\gamma}) = L(\eta|_{[s_a,s_b]})$ if and only if $\gamma'_x(t) \geq 0$ for all t. Throughout the analysis above we assumed that $s_a < s_b$, but of course the argument, with obvious modifications, works equally well when $s_a > s_b$. The bottom line is that γ is distance minimizing if and only if its image is contained in g_0 and it either always goes from left to right or always goes from right to left. The important geometric consequences of this are:

Lemma 9.11. *The geodesics contained in g_0 are precisely the open connected subsets. A geodesic contained in g_0 contains the image of every distance minimizing path between any two of its points, and any two points of g_0 are the endpoints of a distance minimizing path.*

This is a good start, but it's just a single complete geodesic. (That g_0 is complete is obvious, and will be proved eventually.) The main idea in what follows is to study the symmetries of H, so that information about g_0 can understood as applying to all geodesics.

We begin with some general considerations. Suppose that $f : M \to N$ is a C^∞ diffeomorphism where M and N are Riemannian manifolds with Riemannian metrics $\big\langle \cdot, \cdot \big\rangle^M$ and $\big\langle \cdot, \cdot \big\rangle^N$, and let p be a point in M. Of course $Df(p)$ and $Df^{-1}(f(p))$ are inverse linear isomorphisms because the chain rule gives

$$\mathrm{Id}_{T_pM} = D\mathrm{Id}_M(p) = D(f^{-1} \circ f)(p) = Df^{-1}(f(p)) \circ Df(p)$$

and

$$\mathrm{Id}_{T_{f(p)}N} = D\mathrm{Id}_N(f(p)) = D(f \circ f^{-1})(f(p)) = Df(p) \circ Df^{-1}(f(p)).$$

The point p is an **isometry point** of f if

$$\big\langle Df(p)\zeta, Df(p)\eta \big\rangle^N = \big\langle \zeta, \eta \big\rangle^M \quad \text{for all } \zeta, \eta \in T_pM, \tag{$\dagger$}$$

and if every $p \in M$ is an isometry point of f, then f is an **isometry**. If such an f exists we say that M and N are **isometric**.

We will make use of two basic facts about isometry points.

Lemma 9.12. *If p is an isometry point of f, then $f(p)$ is an isometry point of f^{-1}.*

Proof. Consider any $\zeta', \eta' \in T_{f(p)}N$. Since f is a diffeomorphism,

$$\zeta' = Df(p)\zeta \quad \text{and} \quad \eta' = Df(p)\eta$$

for some $\zeta, \eta \in T_pM$. Then

$$\zeta = Df^{-1}(f(p))\zeta' \quad \text{and} \quad \eta = Df^{-1}(f(p))\eta'$$

because $Df(p)^{-1} = Df^{-1}(f(p))$, so equation ($\dagger$) can be rewritten as

$$\big\langle \zeta', \eta' \big\rangle^N = \big\langle Df^{-1}(f(p))\zeta', Df^{-1}(f(p))\eta' \big\rangle^M.$$

□

Corollary 9.13. *If f is an isometry, then so is f^{-1}.*

Let P be a third Riemannian manifold with Riemannian metric $\big\langle \cdot, \cdot \big\rangle^P$, and let $g : N \to P$ be a second C^∞ diffeomorphism.

Lemma 9.14. *If p is an isometry point of f and $f(p)$ is an isometry point of g, then p is an isometry point of $g \circ f$.*

Proof. For any $\zeta, \eta \in T_pM$ the chain rule and the hypotheses allow us to compute that

$$\begin{aligned}\big\langle D(g\circ f)(p)\zeta, D(g\circ f)(p)\eta\big\rangle^P &= \big\langle Dg(f(p))(Df(p)\zeta), Dg(f(p))(Df(p)\eta)\big\rangle^P \\ &= \big\langle Df(p)\zeta, Df(p)\eta\big\rangle^N = \big\langle \zeta,\eta\big\rangle^M.\end{aligned}$$

□

Corollary 9.15. *If f and g are isometries, then so is $g \circ f$.*

Insofar as two isometric manifolds are really "the same," one should expect that every aspect of geometric structure is preserved. This principle borders on being automatic, but nonetheless we will present a formal verification that curve length is preserved. Let $\gamma : [a,b] \to M$ be a C^1 curve. Then for each t the chain rule gives $(f\circ\gamma)'(t) = Df(\gamma(t))\gamma'(t)$, and the definition of an isometry implies that

$$\big\langle (f\circ\gamma)'(t), (f\circ\gamma)'(t)\big\rangle^N = \big\langle Df(\gamma(t))\gamma'(t), Df(\gamma(t))\gamma'(t)\big\rangle^N = \big\langle \gamma'(t),\gamma'(t)\big\rangle^M.$$

For $p \in M$ and $q \in N$ we let $\|\cdot\|_p^M$ and $\|\cdot\|_q^N$ denote the norms derived from $\langle\cdot,\cdot\rangle_p^M$ and $\langle\cdot,\cdot\rangle_q^N$ respectively. Then

$$\begin{aligned}\|(f\circ\gamma)'(t)\|_{f(\gamma(t))}^N &= \sqrt{\big\langle (f\circ\gamma)'(t), (f\circ\gamma)'(t)\big\rangle^N} \\ &= \sqrt{\big\langle \gamma'(t),\gamma'(t)\big\rangle^M} = \|\gamma'(t)\|_{\gamma(t)}^M.\end{aligned}$$

Therefore

$$L(f\circ\gamma) = \int_a^b \|(f\circ\gamma)'(t)\|_{f(\gamma(t))}^N\,dt = \int_a^b \|\gamma'(t)\|_{\gamma(t)}^M\,dt = L(\gamma).$$

An important consequence of this is that γ is distance minimizing (or locally distance minimizing) if and only if $f\circ\gamma$ is distance minimizing (or locally ditance minimizing) so $g \subset M$ is a geodesic, or a complete geodesic, in M if and only if $f(g)$ is a geodesic, or a complete geodesic, in N.

For us the most interesting isometries will be those between a Riemannian and itself. These are called **symmetries**. For any Riemannian manifold the symmetries constitute a group with composition as the group operation because compositions and inverses of symmetries are symmetries. In line with the Erlangen program, we will investigate the geometry of H by studying its group of symmetries.

Concretely, what does it mean for (x,y) to be an isometry point of $f : H \to H$? According to the definition, (x,y) is an isometry point of f if and only if

$$\big\langle Df(x,y)\zeta, Df(x,y)\eta\big\rangle^H_{f(x,y)} = \big\langle \zeta, \eta\big\rangle^H_{(x,y)}$$

for all $\zeta, \eta \in T_{(x,y)}H$. Setting $(x',y') := f(x,y)$ and substituting the definition of $\big\langle \cdot, \cdot\big\rangle^H$, we find that this is the case if and only if

$$\frac{\big\langle Df(x,y)\zeta, Df(x,y)\eta\big\rangle_{(x',y')}}{(1-x'^2-y'^2)^2} = \frac{\big\langle \zeta, \eta\big\rangle_{(x,y)}}{(1-x^2-y^2)^2}$$

for all $\zeta, \eta \in T_{(x,y)}H$. That is,

$$\frac{1-x^2-y^2}{1-x'^2-y'^2}Df(x,y)$$

is an orthogonal transformation under the usual identification of $T_{(x,y)}H$ and $T_{f(x,y)}H$ with $\mathbb{R}^2$.

In particular, if $\ell : \mathbb{R}^2 \to \mathbb{R}^2$ is an orthogonal transformation, then $\ell|_H$ is an isometry of H because if $(x',y') = \ell(x,y)$, then $x'^2 + y'^2 = x^2 + y^2$, and $D\ell(x,y) = \ell$ for all (x,y). Thus the restriction to H of a rotation of $\mathbb{R}^2$ is a symmetry, and there are also symmetries derived from reflections like $(x,y) \mapsto (x,-y)$.

But it turns out there are other symmetries as well. One clue to finding them is the following consequence of the analysis above: if $f : H \to H$ is a symmetry and, for each $(x,y) \in H$, the determinant of $Df(x,y)$ is positive, then f is conformal. In Section 7.1 we saw that a map from an open subset of $\mathbb{R}^2$ to $\mathbb{R}^2$ is conformal if and only if its reinterpretation as a map from an open subset of $\mathbb{C}$ to $\mathbb{C}$ is holomorphic, so this suggests that we study the holomorphic diffeomorphisms between the unit disk in $\mathbb{C}$ and itself. It turns out that the diffeomorphisms we are interested in are a subclass of a set of diffeomorphisms of the Riemann sphere that is, in itself, quite interesting and important, and we will study these first.

Recall that the Riemann sphere S is one dimensional projective space over $\mathbb{C}$, i.e., the set of one dimensional linear subspaces of $\mathbb{C}^2$, and that if $(z,w) \in \mathbb{C}^2 \setminus \{(0,0)\}$, then $[z,w]$ denotes the subspace spanned by (z,w). Let $\ell : \mathbb{C}^2 \to \mathbb{C}^2$ be a nonsingular linear transformation whose matrix is $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. There is an associated **Möbius transformation** $\tau_\ell : S \to S$ given by

$$\tau_\ell([z,w]) := [\ell(z,w)] = [az+bw, cz+dw].$$

This is a well defined function, in the sense that the definition doesn't depend on the pair (z,w) chosen to represent $[z,w]$, because for any $\alpha \in \mathbb{C}^*$ we have

$$\tau_\ell([\alpha z, \alpha w]) = [\ell(\alpha z, \alpha w)] = [\alpha \ell(z,w)] = [\ell(z,w)] = \tau_\ell([z,w]).$$

If $\ell' : \mathbb{C}^2 \to \mathbb{C}^2$ is another nonsingular linear transformation, then

$$\tau_{\ell' \circ \ell}([z,w]) = [\ell'(\ell(z,w))] = \tau_{\ell'}([\ell(z,w)]) = \tau_{\ell'}(\tau_\ell([z,w]))$$

for any $[z,w] \in S$, so $\tau_{\ell'\circ\ell} = \tau_{\ell'} \circ \tau_\ell$. Since $\tau_{\mathrm{Id}_{\mathbb{C}^2}} = \mathrm{Id}_S$, setting $\ell' = \ell^{-1}$ reveals that $\tau_{\ell^{-1}} = \tau_\ell^{-1}$. Thus each Möbius transformation is invertible, with an inverse that is also a Möbius transformation. The set of Möbius transformations contains Id_S, compositions of any two of its elements, and the inverse of each of its elements, so it is a group.

In order to verify that τ_ℓ is holomorphic we need to look at it in relation to coordinate charts for the domain and range. Recall that S has the atlas consisting of the two coordinate charts

$$\varphi_0([z,w]) := w/z \quad \text{and} \quad \varphi_1([z,w]) := z/w$$

whose domains are

$$U_0 := \{\,[z,w] \in S : z \neq 0\,\} \quad \text{and} \quad U_1 := \{\,[z,w] \in S : w \neq 0\,\}.$$

Then $\tau_\ell(\varphi_1^{-1}(z)) = \tau_\ell([z,1]) = [az+b, cz+d]$, so

$$\varphi_1 \circ \tau_\ell \circ \varphi_1^{-1} : z \mapsto \frac{az+b}{cz+d}$$

at all points where this composition is defined, and similar formulas characterize the compositions $\varphi_0 \circ \tau_\ell \circ \varphi_0^{-1}$, $\varphi_0 \circ \tau_\ell \circ \varphi_1^{-1}$, and $\varphi_1 \circ \tau_\ell \circ \varphi_0^{-1}$. Because of these formulas, Möbius transformations are sometimes called **linear fractional transformations**. To prove that τ_ℓ is holomorphic we need to show that for any $[z,w]$ there are i and j such that $\varphi_i([z,w])$ is in the domain of $\varphi_j \circ \tau_\ell \circ \varphi_i^{-1}$. But there is at least one i such that $[z,w] \in U_i$ and at least one j such that $\tau_\ell([z,w]) \in U_j$, so this is obvious. Since its inverse is also a Möbius transformation, hence holomorphic, we say that τ_ℓ is **biholomorphic**.

The key to the geometry of Möbius transformations is that (say in the perspective afforded by the coordinate chart φ_1) a Möbius transformation maps each line to either a circle or a line, and it maps each circle to either a circle or a line. This is actually simpler than it sounds: a line is a circle that happens to contain the point ∞, and from the point of view of the group of Möbius transformations ∞ is not a special point. Let

$$C := \{\, [z,w] \in S : w \neq 0 \text{ and } |z/w| = 1 \,\}$$

be the unit circle centered at the origin with respect to the coordinate system given by φ_1. We define a **circle-or-line** to be any set of the form $c_\ell := \tau_\ell(C)$. The "proof" that Möbius transformations map circles-or-lines to circles-or-lines is now extremely simple: for any ℓ' we have

$$\tau_{\ell'}(c_\ell) = \tau_{\ell'}(\tau_\ell(C)) = \tau_{\ell'\circ\ell}(C) = c_{\ell'\circ\ell}.$$

But we need to show that our definition of a circle-or-line is satisfactory, by which we mean that $\varphi_1(c_\ell \cap U_1)$ is a circle or line in $\mathbb{C}$ in the usual sense, and every circle or line in $\mathbb{C}$ is $\varphi_1(c_\ell \cap U_1)$ for some ℓ.

Notationally, it is somewhat simpler to work with ℓ^{-1} because (in view of the fact that τ_ℓ is a bijection)

$$\varphi_1(c_{\ell^{-1}} \cap U_1) = \varphi_1\big(\{\, [z,w] \in U_1 : \tau_\ell([z,w]) \in C \,\}\big) = \Big\{\, z \in \mathbb{C} : \left|\frac{az+b}{cz+d}\right| = 1 \,\Big\}.$$

Suppose that $a \neq 0 \neq c$. (Everything goes through in the same way if $a = 0$ or $c = 0$, but the formulas are a bit messier; we'll leave it to you to work out the details. Of course $a = 0 = c$ is impossible because ℓ is nonsingular.) Then

$$\varphi_1(c_{\ell^{-1}} \cap U_1) = \{\, z \in \mathbb{C} : |z + b/a| = |c/a| \cdot |z + d/c| \,\}.$$

Let α, β, γ, δ, and ρ be the real numbers such that $-b/a = \alpha + i\beta$, $-d/c = \gamma + i\delta$, and $|c/a| = \rho$, and let $Q(x,y)$ be the quadratic polynomial

$$Q(x,y) := (x-\alpha)^2 + (y-\beta)^2 - \rho^2\big((x-\gamma)^2 + (y-\delta)^2\big).$$

Then $Q(x,y) = |z + b/a|^2 - |c/a|^2 \cdot |z + d/c|^2$, so

$$\varphi_1(c_{\ell^{-1}} \cap U_1) = \{\, x + iy \in \mathbb{C} : Q(x,y) = 0 \,\}.$$

The coefficients of x^2 and y^2 in Q are both $1 - \rho^2$, so $\varphi_1(c_{\ell^{-1}} \cap U_1)$ is a circle when $\rho \neq 1$. When $\rho = 1$ we have

$$\begin{aligned} Q(x,y) &= 2(\gamma-\alpha)x + \alpha^2 - \gamma^2 + 2(\delta-\beta)y + \beta^2 - \delta^2 \\ &= (\gamma-\alpha)(2x - \alpha - \gamma) + (\delta-\beta)(2y - \beta - \delta), \end{aligned}$$

so that $\varphi_1(c_{\ell^{-1}} \cap U_1)$ is a line. (Note that $\gamma = \alpha$ and $\beta = \delta$ would imply that $b/a = d/c$, which is impossible because $ad - bc \neq 0$.)

Is every line and circle a set of the form $\{\, z \in \mathbb{C} : |az + b|/|cz + d| = 1 \,\}$? It is easy to see that every circle in $\mathbb{C}$ has this form: if we want this set to be the circle centered at p with radius R we can simply set $a := 1$, $b := -p$, $c := 0$, and $d := R$. Since $\alpha + \gamma$, $\alpha - \gamma$, $\beta + \delta$, and $\beta - \delta$ can be any four real numbers, any line in $\mathbb{C}$ is the set where

$$(\gamma - \alpha)(2x - \alpha - \gamma) + (\delta - \beta)(2y - \beta - \delta)$$

vanishes for some α, β, γ, and δ. This line can be realized as $\varphi_1(c_{\ell^{-1}} \cap U_1)$ by setting $a := 1$, $b := -\alpha - i\beta$, $c := 1$, and $d := -\gamma - i\delta$. (Note that $ad - bc \neq 0$, so that ℓ is nonsingular, because $\alpha \neq \gamma$ or $\beta \neq \delta$.)

A subclass of the Möbius transformations map the unit disk onto itself, and it might seem like a small step to observe that the associated maps from H to itself are symmetries if we endow H with the Riemannian metric $\langle \cdot, \cdot \rangle^H$, as we shall see. But, according to Poincaré's own account of how the idea came to him, the process was much more roundabout, involving mysterious mental processes that led to a fully formed mathematical idea emerging in his conscious awareness at a specific moment:

> *For fifteen days I strove to prove that there could not be any functions like those I have since called Fuchsian functions. I was then very ignorant; every day I seated myself at my work table, stayed an hour or two, tried a great number of combinations and reached no results. One evening, contrary to my custom, I drank black coffee and could not sleep. Ideas rose in crowds; I felt them collide until pairs interlocked, so to speak, making a stable combination. By the next morning I had established the existence of a class of Fuchsian functions, those which come from the hypergeometric series; I had only to write out the results, which took but a few hours.*
>
> *Then I wanted to represent these functions by the quotient of two series; this idea was perfectly conscious and deliberate, the analogy with elliptic functions guided me. I asked myself what properties these series must have if they existed, and I succeeded without difficulty in forming the series I have called theta-Fuchsian.*
>
> *Just at this time I left Caen, where I was then living, to go on a geological excursion under the auspices of the school of mines.*

> *The changes of travel made me forget my mathematical work. Having reached Coutances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step the idea came to me, without anything in my former thoughts seeming to have paved the way for it, that the transformations I had used to define the Fuchsian functions were identical with those of non-Euclidean geometry. I did not verify this idea; I should not have had time, as, upon taking my seat in the omnibus, I went on with my conversation already commenced, but I felt a perfect certainty. On my return to Caen, for conscience sake I verified the result at my leisure.*

Let $D := \{\, z \in \mathbb{C} : |z| < 1 \,\}$ be the open unit disk in $\mathbb{C}$. The Möbius transformations we're interested in are those that map D onto itself in the frame of reference given by $\varphi_1 \circ \tau_\ell \circ \varphi_1^{-1}$, but it turns out that the logical structure of the analysis makes it easier to work with a condition that is, on its surface, slightly different. Eventually we'll show that it implies what we want.

Definition 9.16. *A Möbius transformation is* ***circular*** *if it maps* 0 *to a point in* D *and it maps the unit circle* $C := \{\, z \in \mathbb{C} : |z| = 1 \,\}$ *into itself.*

For $\theta \in \mathbb{R}$ let $m_\theta : z \mapsto e^{i\theta} z$ be the map that rotates the complex plane counterclockwise by θ radians. This is the Möbius transformation $z \mapsto (e^{i\theta} \cdot z + 0)/(0 \cdot z + 1)$, and obviously m_θ maps C to C and 0 to 0, so it is circular. There is a converse:

Lemma 9.17. *If* $\tau : z \mapsto \frac{az+b}{cz+d}$ *is a circular transformation mapping* 0 *to itself, then* $\tau = m_\theta$ *for some* θ*.*

Proof. Of course $b = 0$ because $\tau(0) = 0$. Since C is the set of z such that $z\overline{z} = 1$, for every $z \in C$ we have

$$\frac{cz+d}{az} = \frac{1}{\tau(z)} = \overline{\tau(z)} = \frac{\overline{az}}{\overline{cz}+\overline{d}} = \frac{\overline{a}\overline{z}z}{\overline{c}\overline{z}z+\overline{d}z} = \frac{\overline{a}}{\overline{c}+\overline{d}z},$$

so that

$$|a|^2 z = \overline{a}az = (cz+d)(\overline{c}+\overline{d}z) = c\overline{d}z^2 + (|c|^2+|d|^2)z + \overline{c}d.$$

Taking the difference between this equation with $z = 1$ and with $z = -1$ leads to $|a|^2 = |c|^2 + |d|^2$, so $c\overline{d}z^2 + \overline{c}d = 0$ for all $z \in C$. Setting $z = 1$ and

$z = i$ in the latter equation shows that $c\overline{d} = 0 = \overline{c}d$. The determinant of $\left(\begin{smallmatrix} a & 0 \\ c & d \end{smallmatrix}\right)$ is nonzero, so $d \neq 0$, and we conclude that $c = 0$. Now the equation $|a|^2 = |c|^2 + |d|^2$ simplifies to $|a| = |d|$, so $|a/d| = 1$ and consequently $a/d = e^{i\theta}$ for some θ. We have arrived at $\tau(z) = az/d = m_\theta(z)$. □

For any $\mu \in D$ the Möbius transformation

$$\sigma_\mu : z \mapsto \frac{z + \mu}{\overline{\mu}z + 1}$$

is circular because $\sigma_\mu(0) = \mu \in D$ and $\sigma_\mu(z)\overline{\sigma_\mu(z)} = 1$ whenever $z\overline{z} = 1$:

$$\frac{1}{\sigma_\mu(z)} = \frac{\overline{\mu}z + 1}{z + \mu} = \frac{(\overline{\mu}z + 1)\overline{z}}{(z + \mu)\overline{z}} = \frac{\overline{\mu} + \overline{z}}{1 + \mu\overline{z}} = \overline{\sigma_\mu(z)}.$$

In what follows we will often use the fact that $\sigma_\mu^{-1} = \sigma_{-\mu}$, which is established by direct computation:

$$\sigma_{-\mu}(\sigma_\mu(z)) = \frac{\frac{z+\mu}{\overline{\mu}z+1} - \mu}{-\overline{\mu}\frac{z+\mu}{\overline{\mu}z+1} + 1} = \frac{(z + \mu) - \mu(\overline{\mu}z + 1)}{-\overline{\mu}(z + \mu) + (\overline{\mu}z + 1)} = \frac{(1 - \mu\overline{\mu})z}{1 - \mu\overline{\mu}} = z.$$

In particular, $\sigma_{-\mu}(\mu) = \sigma_{-\mu}(\sigma_\mu(0)) = 0$. If τ is an arbitrary circular transformation, then $\sigma_{-\tau(0)} \circ \tau$ is a Möbius transformation because it is a composition of two Möbius transformations, it maps C into itself because it is a composition of two functions with this property, and it maps 0 to itself. Therefore $\sigma_{-\tau(0)} \circ \tau = m_\theta$ for some θ. Composing both sides of this equation with $\sigma_{\tau(0)}$ gives

$$\tau = \sigma_{\tau(0)} \circ m_\theta.$$

Proposition 9.18. *Each circular transformation maps C onto itself, D onto itself, and $S \setminus (C \cup D)$ onto itself. The circular transformations constitute a subgroup of the group of Möbius transformations.*

Proof. Above we saw that any circular transformation is $\sigma_\mu \circ m_\theta$ for some μ and θ. The inverse of $\sigma_\mu \circ m_\theta$ is $m_\theta^{-1} \circ \sigma_\mu^{-1} = m_{-\theta} \circ \sigma_{-\mu}$. This is a composition of Möbius transformations mapping the unit circle to itself, so it is a Möbius transformation that maps the unit circle to itself. In addition, it maps 0 to $-e^{-i\theta}\mu$, and $|-e^{-i\theta}\mu| = |\mu| < 1$, so it is circular. That is, *the inverse of a circular transformation is circular.* In particular, the inverse of a circular transformation maps C to C, so the circular transformation must map C *onto* itself.

Aiming at a contradiction, suppose that a circular τ maps some $z \in D$ to a point outside D. Since $\tau(0) \in D$, the intermediate value theorem, applied to the function $t \mapsto |\tau(tz)|$, implies that $|\tau(tz)| = 1$ for some t with $0 < t \leq 1$, which means that τ maps some point in the line segment $\{\, tz : 0 \leq t \leq 1 \,\}$ (which is contained in D because D is convex) to a point of C. But τ^{-1} maps C to itself, so this is impossible. We have shown that *a circular transformation maps D into itself.*

We can now see that a circular transformation maps $S \setminus (C \cup D)$ into itself because its inverse cannot map a point in $C \cup D$ to a point outside of $C \cup D$. Since Möbius transformations are bijective, if a Möbius transformation maps C into itself, D into itself, and $S \setminus (C \cup D)$ into itself, then it must map each of these sets onto itself.

In particular, the composition of two circular transformations maps the origin to a point in D. Of course it is a Möbius transformation that maps the circle into itself, so we now see (at long last!) that *the composition of two circular transformations is a circular transformation.* Since compositions and inverses of circular tranformations are circular, the circular transformations constitute a subgroup of the group of Möbius transformations. □

Of course we are interested in circular tranformations because they can be interpreted as transformations of H. It will be important to distinguish between a circular transformation and the induced map from H to itself, so we adopt the following notational convention: if τ is a circular transformation, then

$$\tilde{\tau} := \iota^{-1} \circ \tau \circ \iota|_H$$

is the associated map from H to itself. (Here $\iota : (x, y) \mapsto x + iy$ is the usual map from $\mathbb{R}^2$ to $\mathbb{C}$.) A map of this form will be called a **circular isometry**.

The first order of business is to show that:

Proposition 9.19. *A circular isometry $\tilde{\tau}$ is a symmetry of H.*

We need to show that each point of H is an isometry point of $\tilde{\tau}$. In principle we should be able to do this by computing the derivative of $\tilde{\tau}$ at an arbitrary point, but probably this would be a pretty messy calculation that wouldn't yield interesting insights. (To tell the truth, I haven't tried to find out.) Both because conceptual explanations are preferred to calculations, and as a simple matter of laziness, a mathematician confronted with this problem would reflexively look for ways to use basic properties of isometries (Lemmas 9.12 and 9.14) to minimize the burden of computation.

Proof. The only thing we will prove by explicit computation is that for any $\mu \in D$, 0 is an isometry point of $\tilde{\sigma}_\mu$. The formula for the derivative of a quotient gives

$$\sigma'_\mu(z) = \frac{(\overline{\mu}z + 1) - \overline{\mu}(z + \mu)}{(\overline{\mu}z + 1)^2} = \frac{1 - |\mu|^2}{(\overline{\mu}z + 1)^2},$$

so $\sigma'_\mu(0) = 1 - |\mu|^2$ and $D\tilde{\sigma}_\mu(0) = (1 - |\mu|^2)\mathrm{Id}_{\mathbb{R}^2}$. In view of our general characterization of isometry points of maps from H to itself, 0 is an isometry point of $\tilde{\sigma}_\mu$ if and only if

$$\frac{1 - |0|^2}{1 - |\sigma_\mu(0)|^2} D\tilde{\sigma}_\mu(0) = \frac{D\tilde{\sigma}_\mu(0)}{1 - |\mu|^2} = \mathrm{Id}_{\mathbb{R}^2}$$

is an orthogonal transformation, which is the case.

We wish to show that any $(x, y) \in H$ is an isometry point of $\tilde{\tau}$. Let $z := \iota(x, y) = x + iy$. As a composition of circular transformations, $\sigma_{-\tau(z)} \circ \tau \circ \sigma_z$ is a circular transformation, and $\sigma_{-\tau(z)}(\tau(\sigma_z(0))) = \sigma_{-\tau(z)}(\tau(z)) = 0$, so there is a θ such that

$$\sigma_{-\tau(z)} \circ \tau \circ \sigma_z = m_\theta.$$

Composing both sides of this equation with $\sigma_{\tau(z)}$ on the left and σ_{-z} on the right gives $\tau = \sigma_{\tau(z)} \circ m_\theta \circ \sigma_{-z}$, so

$$\tilde{\tau} = \iota^{-1} \circ \tau \circ \iota = (\iota^{-1} \circ \sigma_{\tau(z)} \circ \iota) \circ (\iota^{-1} \circ m_\theta \circ \iota) \circ (\iota^{-1} \circ \sigma_{-z} \circ \iota)$$

$$= \tilde{\sigma}_{\tau(z)} \circ \tilde{m}_\theta \circ \tilde{\sigma}_{-z}.$$

We can now show that (x, y) is an isometry point of $\tilde{\tau}$ by using Lemma 9.14 to combine the following facts:

(i) Since $(0, 0)$ is an isometry point of $\tilde{\sigma}_z$, Lemma 9.12 implies that (x, y) is an isometry point of $\tilde{\sigma}_z^{-1} = \tilde{\sigma}_{-z}$.

(ii) Any rotation $\tilde{m}_\theta$ is an isometry.

(iii) $\tilde{m}_\theta(\tilde{\sigma}_{-z}(x, y)) = (0, 0)$ is an isometry point of $\tilde{\sigma}_{\tau(z)}$.

□

After this extended digression we finally have the tools we need for the analysis of the geodesics of H. As is often the case in geometry, there will be an accumulation of "small" facts, after which the larger picture will be assembled by combining these details.

Lemma 9.20. *Suppose that p and q are distinct point of H and $(t,0) \in g_0$. Then there is a unique circular isometry $\tilde{\tau}$ such that $\tilde{\tau}(p) = (t,0)$ and $\tilde{\tau}(q)$ is an element of g_0 to the right of $(t,0)$.*

Proof. If $\mu := -\iota(p)$, then $\tilde{\sigma}_\mu(p) = (0,0)$, and there is a θ such that $\tilde{m}_\theta(\tilde{\sigma}_\mu(q))$ is in $\{\,(s,0) \in H : s > 0\,\}$. The circular isometry $\tilde{\sigma}_t : (s,0) \to (\frac{t+s}{ts+1}, 0)$ maps g_0 to itself while preserving its ordering, and it maps $(0,0)$ to $(t,0)$, so a satisfactory $\tilde{\tau}$ is given by setting

$$\tilde{\tau} := \tilde{\sigma}_t \circ \tilde{m}_\theta \circ \tilde{\sigma}_\mu.$$

If $\tilde{\tau}'$ also satisfies the required conditions, then $\tilde{\sigma}_t^{-1} \circ \tilde{\tau}' \circ \tilde{\tau}^{-1} \circ \tilde{\sigma}_t$ maps $(0,0)$ to itself, so it is $\tilde{m}_\rho$ for some ρ, and it takes $\tilde{\sigma}_t^{-1}(\tilde{\tau}(q))$, which is a point in g_0 to the right of the origin, to another point in g_0 to the right of the origin. Therefore $\tilde{m}_\rho = \mathrm{Id}_H$, so $\sigma_t^{-1} \circ \tilde{\tau}' = (\tilde{\tau}^{-1} \circ \tilde{\sigma}_t)^{-1} = \sigma_t^{-1} \circ \tilde{\tau}$. □

Lemma 9.21. *If a circular isometry $\tilde{\tau}$ maps two distinct elements of g_0 to points in g_0, then it maps g_0 onto g_0.*

Proof. Suppose that p and q are points in g_0 that are mapped by $\tilde{\tau}$ to points in g_0. Let s and t be the numbers such that $p = (s,0)$ and $\tilde{\tau}(p) = (t,0)$. Swapping p and q if necessary, we may suppose that q is to the right of p. If $\tilde{\tau}(q)$ is to the right of $\tilde{\tau}(p)$, then the uniqueness clause of the last result implies that $\tilde{\tau} = \tilde{\sigma}_t \circ \tilde{\sigma}_{-s}$, and if $\tilde{\tau}(q)$ is to the left of $\tilde{\tau}$, then it implies that $\tilde{m}_\pi \circ \tilde{\tau} = \tilde{\sigma}_{-t} \circ \tilde{\sigma}_{-s}$, so that $\tilde{\tau} = \tilde{m}_\pi \circ \tilde{\sigma}_{-t} \circ \tilde{\sigma}_{-s}$. □

The ultimate goal of the next few results is to show that any geodesic containing two points of g_0 is a subset of g_0. (This is Proposition 9.25 below.) Even though we already know that g_0 contains the image of any distance minimizing curve between two of its points, this is still a painstaking endeavor because the notion of distance minimization embedded in the definition of a geodesic is local.

Lemma 9.22. *If the image of a distance minimizing curve $\gamma : [a,b] \to H$ contains two points in g_0, then its image is contained in g_0.*

Proof. It suffices to show that if $a \le t_1 < t_2 < t_3 \le b$ and two of the three points $\gamma(t_1)$, $\gamma(t_2)$, and $\gamma(t_3)$ are contained in g_0, then so is the third. There is (by Lemma 9.20) a circular isometry $\tilde{\tau}$ such that $\tilde{\tau}(\gamma(t_1))$ and $\tilde{\tau}(\gamma(t_3))$ are contained in g_0. Since $\tilde{\tau}$ is an isometry, $\tilde{\tau} \circ \gamma|_{[t_1,t_3]}$ is distance minimizing, so Lemma 9.11 implies that $\tilde{\tau}(\gamma(t_2)) \in g_0$. Therefore $\tilde{\tau}$ maps all three points $\gamma(t_1)$, $\gamma(t_2)$, and $\gamma(t_3)$ to g_0. Since two of them are in g_0, the last result implies that $\tilde{\tau}$ maps g_0 onto itself, so it can't map any point outside of g_0 to g_0. Therefore all three points are in g_0, as desired. □

Corollary 9.23. *If g is a geodesic and $p \in g$, then there is a circular isometry that maps a neighborhood of p (in the relative topology of g) into g_0.*

Proof. Let $\gamma : [a, b] \to g$ be a distance minimizing curve whose image is a neighborhood of p in the relative topology of g, let q be another point in the image of γ, and let $\tilde{\tau}$ be a circular isometry such that $\tilde{\tau}(p), \tilde{\tau}(q) \in g_0$. Then $\tilde{\tau} \circ \gamma$ is distance minimizing, so Lemma 9.22 implies that its image is contained in g_0. □

Although connectedness is an intuitive geometric notion, the definition is topological, so using connectedness to prove something necessarily involves some fiddling around with certain carefully defined sets. It always feels a bit surprising when everything works out neatly and cleanly in the end, as in the next argument, but usually it does.

Lemma 9.24. *If a geodesic g contains a nonempty open subset of g_0, then $g \subset g_0$. Consequently g_0 is a complete geodesic.*

Proof. Let V_1 be the set of points $s \in g$ such that s has a neighborhood U with $g \cap U \subset g_0$. Let V_2 be the set of points $s \in g$ possessing a neighborhood U such $g \cap U \cap g_0$ is either empty or contains exactly one point. Clearly V_1 and V_2 and open, and $V_1 \cap V_2 = \emptyset$. Since V_1 is nonempty by assumption, if we can show that $V_1 \cup V_2 = g$, then (because a geodesic is connected) it will follow that $g = V_1 \subset g_0$.

Since g is a geodesic, a point $s \in g$ has a neighborhood U such that $g \cap U$ is contained in the image of a distance minimizing curve with nonzero derivative. If $g \cap U \cap g_0$ contains more than two points, then (by Lemma 9.22) $g \cap U \subset g_0$ and $s \in V_1$, and otherwise $s \in V_2$. □

Proposition 9.25. *If a geodesic g contains two points of g_0, then $g \subset g_0$.*

Proof. Suppose p and q are distinct points in $g \cap g_0$. Let $\tilde{\tau}$ be a circular isometry that maps a neighborhood of p in g to g_0. Then $\tilde{\tau}(g)$ is a geodesic containing a nonempty open subset that is contained in g_0, so the last result implies that $\tilde{\tau}(g) \subset g_0$. In particular, $\tilde{\tau}(q) \in g_0$. We now see that $\tilde{\tau}$ maps p and q to points in g_0, so Lemma 9.21 implies that $\tilde{\tau}$ restricts to a bijection from g_0 to itself. Since $\tilde{\tau}(g) \subset g_0$, it follows that $g \subset g_0$. □

After this patient accumulation of minor results, we are now ready to combine them in the proof of the following result, which establishes the most important properties of the geodesics in H.

Theorem 9.26. *If $\tilde{\tau}$ is a circular isometry, then $\tilde{\tau}(g_0)$ is a complete geodesic, and every complete geodesic is of this form. For any two distinct points p and q in H there is exactly one complete geodesic that contains them both. There is a distance minimizing curve between p and q, and this geodesic contains the image of any such distance minimizing curve.*

Proof. Of course any $\tilde{\tau}(g_0)$ is a geodesic. It must be complete because if it was a proper subset of a geodesic g, then g_0 would be a proper subset of $\tilde{\tau}^{-1}(g)$, but we know that g_0 is complete.

Let g be a complete geodesic. Then (by Corollary 9.23) there is a circular isometry $\tilde{\tau}$ that maps an open subset of g to g_0, and Lemma 9.24 implies that $\tilde{\tau}(g) \subset g_0$. Then $\tilde{\tau}^{-1}(g_0)$ is a complete geodesic that contains g, and since g is complete it follows that $g = \tilde{\tau}^{-1}(g_0)$.

For the given p and q Lemma 9.20 gives a circular isometry $\tilde{\tau}$ with $\tilde{\tau}(p), \tilde{\tau}(q) \in g_0$. If g is a complete geodesic containing p and q, then Proposition 9.25 implies that $\tilde{\tau}(g) \subset g_0$. Thus $g \subset \tilde{\tau}^{-1}(g_0)$, and in fact $g = \tilde{\tau}^{-1}(g_0)$ because g is complete.

Lemma 9.11 gives a distance minimizing curve $\gamma : [a, b] \to H$ with $\gamma(a) = \tilde{\tau}(p)$ and $\gamma(b) = \tilde{\tau}(q)$, so $\tilde{\tau}^{-1} \circ \gamma$ is a distance minimizing curve between p and q. Lemma 9.11 also implies that $\tilde{\tau}$ maps the image of any such curve to g_0, so the image of any such curve is contained in $\tilde{\tau}^{-1}(g_0)$. □

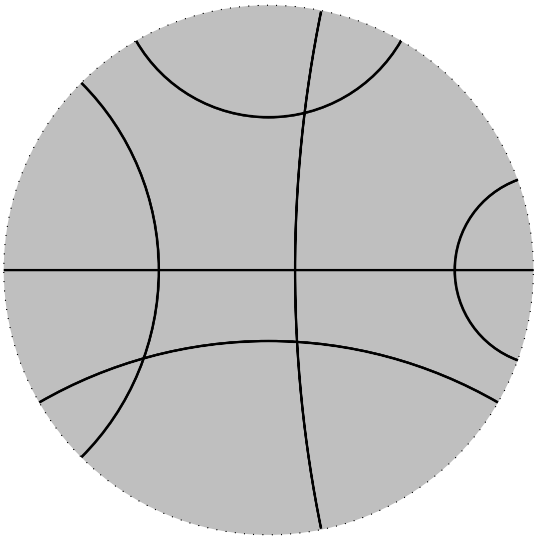


Figure 9.3

Since Möbius transformations map circles-or-lines to circles-or-lines, if $\tilde{\tau}$ is a circular isometry, then $\tilde{\tau}(g_0)$ must be the intersection of H with $\ell = \iota^{-1}(c)$ for some circle-or-line c. In addition, the x-axis is perpendicular to the unit circle at the two points where it intersects the unit circle. Since $\tilde{\tau}$ is conformal and maps the unit circle to itself, ℓ must be perpendicular to the unit circle at the two points where it intersects the unit circle. (See Figure 9.3.) In fact if $\ell = \iota^{-1}(c)$ for some circle-or-line c, and ℓ is perpendicular to the unit circle at both intersection points, then $\ell \cap H = \tilde{\tau}(g_0)$ for some circular isometry $\tilde{\tau}$, so $\ell \cap H$ is a geodesic. We won't bother to prove this, but you might enjoy giving it a try.

We are now finally in a position to compare the geometry of H with Euclidean geometry. The first five axioms of Euclid are:

(1) Any two distinct points are contained in a line.

(2) Any line segment can be extended indefinitely in a line.

(3) Given any two distinct points, there is exactly one circle centered at the first point that contains the second point.

(4) Any two right angles are congruent.

(5) Given a line ℓ_1 and a point not on the line, there is exactly one line ℓ_2 containing the point that does not intersect ℓ_1.

Each of these statements conjures up an unambiguous picture *if* you already have a clear visual understanding of Euclidean geometry, but from a modern point of view this axiom system (which is my rough transcription of various modern translations) is hopelessly ill posed. Any precise explication that would satisfy a modern mathematician would have to begin by declaring certain terms to be primitives that are not defined, after which definitions of the other terms would be provided using the machinery of formal logic and set theory. This is a very reasonable project, and a point of departure for an interesting line of research. However, Euclid didn't have a modern understanding of how to play "the set theory game," and an explanation of why the fifth axiom is not a logical consequence of the first four would be at least a bit unfair if it really depended on technical details of that sort. Since H closely resembles $\mathbb{R}^2$ in almost any intuitive sense of what geometry is about, it's more convincing to simply explain why H satisfies the first four axioms, but not the fifth, if we substitute 'complete geodesic' for 'line' throughout the list and interpret all other terms intuitively.

Theorem 9.26 tells us that H satisfies the first axiom because any two distinct points in H are contained in a complete geodesic. Of course it actually says something a bit stronger, insofar as there is exactly one complete geodesic that contains them.

The second axiom says that lines have infinite length. 'Length' is not a primitive concept here, so we have to deal with the question of how this idea should be expressed within the formal logical system. One way to do this, that is sometimes given as the formulation of this axiom, goes as follows: given two line segments AB and CD, the line ℓ that contains AB also contains a line segment BE that is congruent to CD and which is adjacent to AB in the sense that AB and BE intersect at the point B. We can repeat this maneuver, producing a line segment EF in ℓ that is congruent to CD and which is adjacent to BE, a line segment FG in ℓ that is congruent to CD and which is adjacent to EF, and so forth, so the length of the line containing AB is unbounded in the sense that it can contain any number of copies of CD lined up end to end. Here "congruent" means that there is an isometry that takes CD to BE.

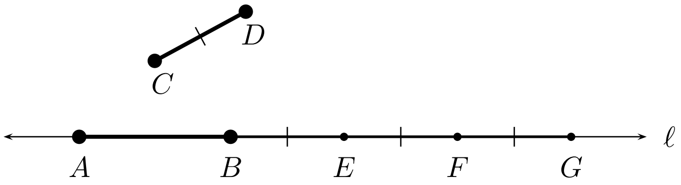


Figure 9.4

To see that H satisfies this condition, suppose that the line segment AB is contained in g_0 with B to the right of A. (Clearly we can transform the given situation by an isometry to bring this about.) Then Lemma 9.20 says that there is a circular isometry that takes C to B and takes D to a point in the portion of g_0 lying to the right of B.

In order to interpret the third axiom we need to say what a circle is. Suppose that two points p and q are given. One possibility is to define the circle with center p containing q to be the set of images of q under symmetries that leave p fixed. Another definition, that is valid even in a general metric space if there is a distance minimizing curve between p and q, say with length r, is that the circle c centered at p of radius r is the set of points that are the other endpoints of distance minimizing curves of length r that have p as one endpoint. The problem with these definitions is they are only definitions, so that they make the third axiom true automatically, simply because the circle centered at the first point and containing the second is whatever it is defined to be. For Euclid a circle was what you drew with a compass, and apparently the only sensible interpretation of the third axiom is that c is what we think a circle should be, namely a space homeomorphic or diffeomorphic to the unit circle in $\mathbb{R}^2$. This is the case for circles in H, automatically for a circle centered at the origin, and also for circles centered at any other point because there is a circular isometry taking that point to the origin.

The fourth axiom also requires us to interpret a piece of terminology, namely the word 'perpendicular.' The definition we've been using throughout this book is that two vectors are perpendicular if their inner product is zero, but the axiom system don't give us this kind of numerical information, nor do we know how to interpret such a definition in the hyperbolic context. Another possible definition is that if two geodesics intersect at a point, then they're perpendicular if there is a symmetry that takes that point to the origin while mapping the two geodesics to g_0 and the intersection of H with the y-axis. The definition that one often sees in connection with Euclid's axioms is that a right angle "divides a straight line in half. For example, the angle between the positive x-axis and the positive y-axis is congruent to the angle between the positive y-axis and the negative x-axis.

Actually, worrying about the details of what it means for two lines to be perpendicular is a bit besides the point because the real import of the fourth axiom is that the group of symmetries (or congruences, as they are usually called in geometry) is very large. An action of a group G on a set A is said to be **transitive** if for any two points a and a' in A there is a group element g such that $ga = a'$. (This usage of 'transitive' is unrelated to the notion of a transitive relation.) The isometries of $\mathbb{R}^2$ (that is, the Euclidean motions) act transitively on $\mathbb{R}^2$, and the circular isometries act transitively on H, but the fourth axiom says something much stronger. Suppose r_1 and r_2 are two perpendicular rays emanating from a point A, and s_1 and s_2 are two perpendicular rays emanating from a point B. What the fourth axiom says is that there is a symmetry taking A to B, r_1 to s_1, and r_2 to s_2. In the case of H Lemma 9.20 says that there is a circular isometry $\tilde{\tau}$ taking A to $(0,0)$ and r_1 to the right hand half of g_0 and another circular isometry $\tilde{\tau}_s$ taking B to $(0,0)$ and s_1 to the right hand half of g_0. If ρ is the reflection $(x,y) \mapsto (x,-y)$, then either $\tilde{\tau}_s^{-1} \circ \tilde{\tau}$ or $\tilde{\tau}_s^{-1} \circ \tilde{\rho} \circ \tilde{\tau}$ satisfies the condition demanded by the fourth axiom.

To show that H doesn't satisfy the fifth axiom we need only one example, and I think the nonintersecting geodesics in Figure 9.3 are certainly convincing enough. (If you feel the need for an analytic example you can try to prove that there is some $\varepsilon > 0$ such that $\tilde{m}_\theta(g_0) \cap \tilde{\sigma}_{i/2}(g_0) = \emptyset$ whenever $|\theta| < \varepsilon$.)

This completes the verification that H satisfies the first four axioms of Euclid, but not the fifth.

9.3 Curvature

There is another important concept that the Poincaré disk model can be used to illustrate, namely the notion of curvature, which was developed by Gauss for surfaces embedded in $\mathbb{R}^3$ and extended to higher dimensional manifolds by Riemann. If M is a two dimensional C^1 submanifold of $\mathbb{R}^3$, M "inherits" a Riemannian metric from $\mathbb{R}^3$ because each tangent space T_pM can be regarded as a subset of $T_p\mathbb{R}^3 = \mathbb{R}^3$ and endowed with the usual inner product. The definition of curvature given by Gauss refers to the embedding, but he was able to show that it really depends only on the surface's Riemannian metric, and he attached great significance to this result.

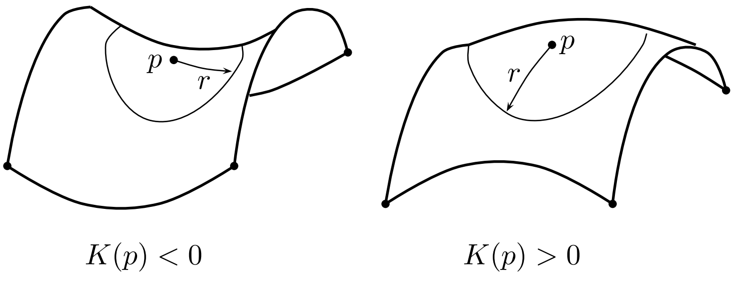


Figure 9.5

There are many ways of defining the curvature of a surface at a point; we'll describe only one, which is a matter of comparing the circumference of a circle centered at a point with its radius. In 1848 Joseph Bertrand (1822-1900) and Victor Puiseux (1820-1883) proved a formula concerning curvature, as it had been defined by Gauss, that we may take as a definition: if $C(r)$ is the circumference of the circle of radius r centered at a point p, then the **curvature** at p is

$$K(p) = \lim_{r \to 0} 3 \cdot \frac{2\pi r - C(r)}{\pi r^3}.$$

As a definition, this formula is superior to Gauss' definition because it is **intrinsic**, by which we mean that it refers only to things "inside" the manifold like the Riemannian metric, and does not refer to the artifacts of any embedding in another space. Thus we are relieved of any need to prove independence of the embedding (though it may still be interesting to prove that Gauss' original definition is equivalent to this one) and in addition this definition can be applied to two dimensional manifolds that have no embedding in $\mathbb{R}^3$. As it happens, there is no isometric embedding of H in $\mathbb{R}^3$, even locally—put another way, there is no C^1 surface $M \subset \mathbb{R}^3$ that is isometric to an open subset of H—and it seems sensible, though inherently speculative, to think that this accounts for the relatively late date of the discovery of non-Euclidean geometry.

The ratio of the circumference of a circle in $\mathbb{R}^2$ to its radius is always 2π, of course, so the curvature of the plane is zero. For the sphere S^2 the circle of radius r centered at $(1,0,0)$ is

$$\{\, (\cos r, \sin r \cos\theta, \sin r \sin\theta) : 0 \le \theta < 2\pi \,\},$$

which has circumference $C(r) = 2\pi \sin r$. Substituting this into the formula above, simplifying, then substituting the power series expansion

$$\sin r = r - \tfrac{1}{3!}r^3 + \tfrac{1}{5!}r^5 - \cdots$$

and taking limits, we find that the curvature of the sphere at $(1,0,0)$ is 1. It is visually obvious (and not hard to show formally) that for every pair of points in S^2 there is an isometry taking the first to the second—that is, the group of isometries acts transitively on S^2—so in fact the curvature of S^2 at each of its points is 1.

Now consider the circle centered at $(0,0) \in H$ that contains the point $(x,0)$. Under the identification of H with the unit disk, this is the circle of radius x, and as such its circumference is $2\pi x$. In comparison with the corresponding distances in the disk, distances in H near points in this circle are magnified by the factor $(1-x^2)^{-1}$, so its circumference as a circle in H is $2\pi x/(1-x^2)$. It radius as a circle in H is the length of the portion of g_0 lying between $(0,0)$ and $(x,0)$, which is[2]

$$r := \int_0^x \frac{1}{1-t^2}\,dt = \tfrac{1}{2}[\ln(1+t) + \ln(1-t)]\Big|_{t=0}^{x} = \tfrac{1}{2}[\ln(1+x) + \ln(1-x)].$$

[2] By the fundamental theorem of calculus, to show that the integral is evaluated correctly it suffices to show that if $f(t) = \frac{1}{2}[\ln(1+t) + \ln(1-t)]$, then $f'(t) = 1/(1-t^2)$. Differentiating both sides of the identity $t = \exp(\ln(t))$ using the chain rule gives $1 = \exp'(\ln(t))\ln'(t)$, and the exponential function is its own derivative, so $\ln'(t) = 1/t$. Now $f'(t)$ can be evaluated using this, the chain rule, and the other rules for differentiation.

Taking the exponential of both sides of this equation, then solving for x, leads to $x = (e^{2r} - 1)/(e^{2r} + 1)$. We substitute this expression for x into the formula for the circumference of the circle passing through $(x, 0)$, finding (after a little more algebra) that the circumference is

$$\begin{aligned} C(r) = \tfrac{\pi}{2}(e^{2r} - e^{-2r}) &= \tfrac{\pi}{2}\big((1 + 2r + \tfrac{1}{2!}(2r)^2 + \cdots) - (1 - 2r + \tfrac{1}{2!}(2r)^2 - \cdots)\big) \\ &= \pi\big(2r + \tfrac{1}{3!}(2r)^3 + \tfrac{1}{5!}(2r)^5 + \cdots\big). \end{aligned}$$

(All of a sudden it starts to makes sense that the function

$$t \mapsto \tfrac{1}{2}(e^t - e^{-t}) = t + \tfrac{1}{3!}t^3 + \tfrac{1}{5!}t^5 + \cdots$$

is called the *hyperbolic* sine function!) Substituting this quantity into the definition of curvature and taking the limit, we find that the curvature of H at $(0, 0)$ is -4. The group of symmetries of H acts transitively, so the curvature of H at each of its points is -4.

Riemann extended the definition of curvature to higher dimensional Riemannian manifolds. This is not the place to describe this work in any detail, but we can indicate some of the difficulties. It makes little sense to work with an embedding of the manifold in Euclidean space, so the definition should be intrinsic. We have seen one way to do this in the two dimensional case, and the formula above could be generalized to higher dimensional manifolds. However, roughly speaking, it is possible for the manifold to be curved to different extents in different directions, or along different two dimensional submanifolds, so a satisfactory notion of curvature will not be a single number, but instead some sort of vector.

In the two dimensional case there is an intuition that the geometry of the manifold is completely determined by the function taking each point to the curvature at that point. It's not so easy to formulate this idea precisely, but there is a result that points in this direction: if a surface has constant curvature in a neighborhood of a point, then there is a neighborhood of the point that is isometric to an open subset of a sphere of some radius, the plane, or a rescaling of H, according to whether the curvature is positive, zero, or negative. Riemann investigated when a point in an n-dimensional Riemannian manifold has a neighborhood that is isometric to an open subset of $\mathbb{R}^n$, finding a set of quantities that vanish throughout the neighborhood if and only if this is the case, and he took these quantities as the definition of the general notion of curvature. Much of the subsequent foundational work in differential geometry can be viewed as a search for abstractions that allow Riemannian curvature to be formulated in a manner that brings its structure and properties to the surface, so that they are visible in the notation.

9.4 Some Riemann Surfaces

A **Riemann surface** is a one dimensional holomorphic manifold over $\mathbb{C}$. (Here "one dimensional" means that there is one *complex* dimension and therefore two real dimensions. To add to the confusion, Riemann surfaces are sometimes called "complex curves.") That is, a Riemann surface is a Hausdorff space with an atlas $\{\, \varphi_i : U_i \to V_i \subset \mathbb{C} : i \in I \,\}$ of coordinate charts such that for all $i, j \in I$,

$$\varphi_j \circ \varphi_i^{-1}|_{\varphi_i(U_i \cap U_j)} \quad \text{and} \quad \varphi_i \circ \varphi_j^{-1}|_{\varphi_j(U_i \cap U_j)}$$

are inverse holomorphic diffeomorphisms. Riemann's Ph.D. thesis was, in part, about functions of a complex variable, and this led him to think about how the theory he had developed might fruitfully be extended. Many directions either suggest or inherently involve Riemann surfaces.

When the concept of a Riemann surface is first introduced it is often motivated as providing a response to the frustrations we encounter if we try to define things like a complex square root function or a complex logarithm function. As we labored to show in Chapter 3, any element of $\mathbb{C}^*$ has a unique representation of the form

$$re^{i\theta} = r(\cos\theta + i\sin\theta)$$

where $r > 0$ and $0 \le \theta < 2\pi$, so one could define the logarithm of $re^{i\theta}$ to be $\ln r + i\theta$, but of course this function is discontinuous along the positive real axis. This discontinuity is a somewhat arbitrary artifact of the range we chose for θ, and it seems to not represent any genuine discontinuity in the phenomenon that the logarithm function is trying to represent. We get a much better behaved picture if we let

$$M := \{\, (w, z) \in \mathbb{C} \times \mathbb{C}^* : z = \exp(w) \,\}$$

be the graph of the exponential function and define the logarithm function to be the projection $\pi_w : (w, z) \mapsto w$ from M to $\mathbb{C}$. Somehow M "feels" like the proper domain of the logarithm function, and it is well behaved in the following local sense: any (w_0, z_0) has a neighborhood $U \subset M$ such that the restrictions of $\pi_w : (w, z) \mapsto w$ and $\pi_z : (w, z) \mapsto z$ to U are holomorphic diffeomorphisms, in which case we have a local logarithm function $\pi_w \circ (\pi_z|_U)^{-1}$.

In general, whenever $U \subset \mathbb{C}$ is open and $f : U \to \mathbb{C}$ is holomorphic, the graph of f is a Riemann surface. An atlas satisfying the definition above is given by the single function $(z, f(z)) \mapsto z$ from $\mathrm{Gr}(f)$ to U. The space of graphs of holomorphic functions is already a very large and rather formless collection of objects, suggesting that the theory of *all* Riemann surfaces might be less interesting than the theories of certain types of Riemann surfaces.

With one exception, all the Riemann surfaces described in the remainder of the section are compact. The rigidity of holomorphic functions described in Section 7.6 suggests that compact Riemann surfaces might present an especially rich interaction between local analysis and global structure. In fact at this level we'll only be able to give the most superficial indication of the extent to which this is true, the importance of Riemann surfaces in contemporary mathematics, and the profound depth of the resulting theories.

The next Riemann surfaces we'll look at are embedded in higher dimensional projective spaces. Recall that if n is a positive integer, then $P^n(\mathbb{C})$ is the set of one dimensional linear subspaces of $\mathbb{C}^{n+1}$, the one dimensional subspace spanned by $z \in \mathbb{C}^{n+1} \setminus \{0\}$ is denoted by $[z]$, and $P^n(\mathbb{C})$ is given the structure of a complex manifold by the atlas of coordinate charts

$$\{\, \varphi_i : U_i \to \mathbb{C}^n : i = 0, \dots, n \,\}$$

where

$$U_i := \{\, [z] : z \in \mathbb{C}^{n+1}, z_i \neq 0 \,\}$$

and

$$\varphi_i([z]) := (\tfrac{z_0}{z_i}, \dots, \tfrac{z_{i-1}}{z_i}, \tfrac{z_{i+1}}{z_i}, \dots, \tfrac{z_n}{z_i}).$$

Of course the Riemann sphere $P^1(\mathbb{C})$ is itself a Riemann surface. Our discussion of Möbius transformations in Section 9.2 is the starting point of an important theme of this subject: for a compact Riemann surface the group of diffeomorphisms mapping that Riemann surface to itself contains important information about the Riemann surface, and is itself quite interesting.

The sets we'll be studying will be compact because (Theorem 3.38) they are closed subsets of $P^n(\mathbb{C})$ and this space is compact. To see this, for each $i = 0, \dots, n$ let

$$D_i := \{\, [z] \in P^n(\mathbb{C}) : |z_i| \geq |z_j| \text{ for all } j = 0, \dots, n \,\}.$$

(The condition used to define D_i is unaffected if we replace z with αz for some $\alpha \in \mathbb{C}^*$ because $|\alpha z_j| = |\alpha|\,|z_j|$ for all j, so this definition makes sense.) The map

$$[z] \mapsto \big(\tfrac{z_0}{z_i}, \dots, \tfrac{z_{i-1}}{z_i}, \tfrac{z_{i+1}}{z_i}, \dots, \tfrac{z_n}{z_i}\big)$$

from D_i to the n-fold cartesian product of the unit disk $\{ w \in \mathbb{C} : |w| \le 1 \}$ is continuous and has the continuous inverse

$$(w_0, \ldots, w_{i-1}, w_{i+1}, \ldots, w_n) \mapsto [w_0, \ldots, w_{i-1}, 1, w_{i+1}, \ldots, w_n],$$

so D_i is compact because (Theorem 3.42) it is homeomorphic to a cartesian product of compact sets. Each $[z]$ is contained in at least one D_i, so $P^n(\mathbb{C}) = D_0 \cup \ldots \cup D_n$, and of course a space is compact if it is covered by a finite collection of compact subsets.

We'll find a large collection of compact Riemann surfaces contained in $P^n(\mathbb{C})$ by applying the regular value theorem to polynomials in the system of variables

$$Z = (Z_0, \ldots, Z_n).$$

We think of a polynomial as a finite sum of monomials, where a **monomial** in these variable is an expression of the form $cZ_0^{e_0} \cdots Z_n^{e_n}$ in which c is a scalar and $e_0, \ldots, e_n$ are nonnegative integers. The **total degree** of this monomial is $e_0 + \cdots + e_n$.

A **homogeneous polynomial** of degree d is a polynomial

$$p(Z) = \sum_{j=1}^{k} c_j Z_0^{e_{j,0}} \cdots Z_n^{e_{j,n}} \in \mathbb{C}[Z_0, \ldots, Z_n]$$

whose monomials all have total degree d. A homogeneous polynomial doesn't define a function from $P^n(\mathbb{C})$ to $\mathbb{C}$, but in spite of this it does make sense to talk about whether such a p vanishes at a point in $P^n(\mathbb{C})$. To see what we mean by this consider that for any $z \in \mathbb{C}^{n+1}$ and any scalar $\alpha \in \mathbb{C}$ we have

$$p(\alpha z) = \sum_{j=1}^{k} c_j (\alpha z_0)^{e_{j,0}} \cdots (\alpha z_n)^{e_{j,n}} = \sum_{j=1}^{k} \alpha^{e_{j,0} + \cdots + e_{j,n}} c_j z_0^{e_{j,0}} \cdots z_n^{e_{j,n}}$$

$$= \alpha^d \sum_{j=1}^{k} c_j z_0^{e_{j,0}} \cdots z_n^{e_{j,n}} = \alpha^d p(z),$$

so for any $[z] \in P^n(\mathbb{C})$, p vanishes at one nonzero point in $[z]$ if and only if it vanishes at all the other points in this linear subspace. Let $V(p)$ be the set of points in $P^n(\mathbb{C})$ at which p vanishes in this sense.

We analyze $V(p)$ by studying its images under the coordinate charts $\varphi_0, \ldots, \varphi_n$. For each $i = 0, \ldots, n$ there is a polynomial

$$p^{-i}(Z) := \sum_{j=1}^{k} c_j Z_0^{e_{j,0}} \cdots Z_{i-1}^{e_{j,i-1}} Z_{i+1}^{e_{j,i+1}} \cdots Z_n^{e_{j,n}}$$

that we can think of as the result of substituting 1 for Z_i in p. If $z \in \mathbb{C}^{n+1}$ and $z_i \neq 0$, then

$$[z] \in V(p) \iff p(\tfrac{z_0}{z_i}, \dots, \tfrac{z_{i-1}}{z_i}, 1, \tfrac{z_{i+1}}{z_i}, \dots, \tfrac{z_n}{z_i}) = 0 \iff p^{-i}(\varphi_i([z])) = 0.$$

Therefore

$$V(p) = \bigcup_{i=0}^{n} (V(p) \cap U_i) = \bigcup_{i=0}^{n} (p^{-i} \circ \varphi_i)^{-1}(0).$$

It will often happen that $Dp^{-i}(w) \neq 0$ at all w such that $p^{-i}(w) = 0$, in which case the regular value theorem implies that $\{\, w \in \mathbb{C}^n : p^{-i}(w) = 0 \,\}$ is a codimension one submanifold of $\mathbb{C}^n$. We'll explain below that if this is true for every i, then $V(p)$ is a codimension one submanifold of $P^n(\mathbb{C})$.

More generally, suppose that $p_1, \dots, p_k$ are homogeneous polynomials. Let

$$\mathbf{p} = (p_1, \dots, p_k) : \mathbb{C}^{n+1} \to \mathbb{C}^k,$$

and set $V(\mathbf{p}) := \bigcap_{h=1}^{k} V(p_h)$. For each $i = 0, \dots, n$ let $\mathbf{p}^{-i} := (p_1^{-i}, \dots, p_k^{-i})$. Then

$$V(\mathbf{p}) = \bigcup_{i=0}^{n} (\mathbf{p}^{-i} \circ \varphi_i)^{-1}(0).$$

Proposition 9.27. *If, for each $i = 0, \dots, n$, 0 is a regular value of $\mathbf{p}^{-i}$, then $V(\mathbf{p})$ is a codimension k submanifold of $P^n(\mathbb{C})$.*

Proof. For each i, 0 is a regular value of $\mathbf{p}^{-i} \circ \varphi_i$ because φ_i is a diffeomorphism between U_i and $\mathbb{C}^n$, so the regular value theorem implies that

$$V(\mathbf{p}) \cap U_i = \{\, [z] \in U_i : \mathbf{p}^{-i}(\varphi_i([z])) = 0 \,\}$$

is a codimension k submanifold of U_i. In general, the property of being a submanifold is a local property, insofar as the definition asks for a neighborhood of each point satisfying a certain condition. This means that if M is a smooth (C^r, holomorphic, real analytic) manifold, $P \subset M$, and $\{U_i\}_{i \in I}$ is an open cover of M such that each $P \cap U_i$ is a smooth codimension k submanifold of U_i, then P is a smooth codimension k submanifold of M. □

In particular, if $k = n - 1$, then $V(\mathbf{p})$ is a one dimensional submanifold of $P^n(\mathbb{C})$. The technical lingo in this circumstance is that $V(\mathbf{p})$ is a **nonsingular projective curve**. (Here 'nonsingular' refers to the satisfaction of the hypotheses of the implicit function theorem.) A set defined by a finite system of algebraic equations is **irreducible** if it cannot be written as a union of two proper subsets, each of which is defined by a finite system of algebraic equations. In the 1850's Riemann conjectured that *every compact connected Riemann surface is holomorphically diffeomorphic to an irreducible nonsingular projective curve.* After about fifty years, with the development of enough tools, it became possible to prove this. This means that a compact Riemann surface that arises in some other way implicitly has an algebraic structure, but it may be very subtle and reflect deep properties.

A single homogeneous polynomial p in the variables X, Y, and Z defines a subset of $P^2(\mathbb{C})$. We'll now look in some detail at what happens in the simplest possible cases, which are those in which the degree of p is small. We are only interested in those p that define a Riemann surface by virtue of the appeal to the regular value theorem described in general above, which is the case if 0 is a regular value of p^{-i} for each $i = 0, 1, 2$. Suppose that $L : \mathbb{C}^3 \to \mathbb{C}^3$ is a nonsingular linear transformation. Then the map $[x, y, z] \mapsto [L(x, y, z)]$ is a holomorphic diffeomorphism from $P^2(\mathbb{C})$ to itself, and it induces a holomorphic diffeomorphism from $V(p \circ L)$ to $V(p)$. Our attitude is that we are interested in characterizing $V(p)$ up to holomorphic diffeomorphism, so for us $V(p)$ and $V(p \circ L)$ are the same.

Polynomials of degree one, namely linear polynomials, present no difficulties. Consider a linear equation

$$p(X, Y, Z) = aX + bY + cZ$$

with $(a, b, c) \neq (0, 0, 0)$. In this case $V(p)$ is holomorphically diffeomorphic to the Riemann sphere, and it is easy to give an explicit diffeomorphism. Supposing that $c \neq 0$ (of course everything is the same if $a \neq 0$ or $b \neq 0$) the map $[x, y, z] \mapsto [x, y]$ from $V(p)$ to $P^1(\mathbb{C})$ has the inverse

$$[x, y] \mapsto [x, y, -(ax + by)/c].$$

Next, consider a nonzero homogeneous polynomial of degree two, say

$$p(X, Y, Z) = aX^2 + bY^2 + cZ^2 + 2dXY + 2eXZ + 2fYZ.$$

We are only interested in those p such that 0 is a regular value of p^{-0}, p^{-1}, and p^{-2}, and our first task is to show that this implies that p cannot be a product of two linear polynomials. To see why, suppose otherwise, so that

$$p(X, Y, Z) = \ell_1(X, Y, Z)\ell_2(X, Y, Z).$$

Each of these linear functions vanishes on a two dimensional linear subspace of $\mathbb{C}^3$, and these subspaces either coincide or have a one dimensional intersection. Either way, there is a nonzero $(\overline{x}, \overline{y}, \overline{z}) \in \mathbb{C}^3$ with

$\ell_1(\overline{x},\overline{y},\overline{z}) = 0 = \ell_2(\overline{x},\overline{y},\overline{z})$. After permuting the components if need be, we can assume that $\overline{z} \neq 0$, and multiplying all components by $1/\overline{z}$ gives such a point with $\overline{z} = 1$. We have

$$p^{-2}(x,y) = \ell_1(x,y,1)\ell_2(x,y,1),$$

and of course $p^{-2}(\overline{x},\overline{y},1) = 0$. But the product rule gives

$$\frac{\partial p^{-2}}{\partial X}(\overline{x},\overline{y},1) = \frac{\partial \ell_1}{\partial X}(\overline{x},\overline{y},1)\ell_2(\overline{x},\overline{y},1) + \ell_1(\overline{x},\overline{y},1)\frac{\partial \ell_2}{\partial X}(\overline{x},\overline{y},1) = 0,$$

and

$$\frac{\partial p^{-2}}{\partial Y}(\overline{x},\overline{y},1) = \frac{\partial \ell_1}{\partial Y}(\overline{x},\overline{y},1)\ell_2(\overline{x},\overline{y},1) + \ell_1(\overline{x},\overline{y},1)\frac{\partial \ell_2}{\partial Y}(\overline{x},\overline{y},1) = 0,$$

which contradicts the assumption that 0 is a regular value of p^{-2}.

We're now going to show that if 0 is a regular value of p^{-0}, p^{-1}, and p^{-2}, then there is a nonsingular linear transformation $L : (X,Y,Z) \mapsto (\tilde{X},\tilde{Y},\tilde{Z})$ such that $p \circ L^{-1} = \tilde{X}^2 + \tilde{Y}^2 + \tilde{Z}^2$. We'll first show how things work out "typically," then discuss the exceptional cases.

We can write p as a matrix product:

$$p(X,Y,Z) = \begin{bmatrix} X & Y & Z \end{bmatrix} \begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}.$$

We would like to choose numbers α, β, γ, δ, ε, and ϕ such that

$$\begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} = \begin{bmatrix} \alpha & 0 & 0 \\ \delta & \beta & 0 \\ \varepsilon & \phi & \gamma \end{bmatrix} \begin{bmatrix} \alpha & \delta & \varepsilon \\ 0 & \beta & \phi \\ 0 & 0 & \gamma \end{bmatrix}.$$

Here are the equations that need to be solved, and how they give rise to a procedure for solving them that works if $a \neq 0$ and $ab \neq d^2$:

$$\begin{aligned} \alpha^2 &= a, & \alpha &:= \sqrt{a}, \\ \alpha\delta &= d, & \delta &:= d/\alpha, \\ \alpha\varepsilon &= e, & \varepsilon &:= e/\alpha, \\ \delta^2 + \beta^2 &= b, & \beta &:= \sqrt{b - \delta^2} = \sqrt{(ab - d^2)/a}, \\ \varepsilon\delta + \phi\beta &= f, & \phi &:= (f - \varepsilon\delta)/\beta, \\ \varepsilon^2 + \phi^2 + \gamma^2 &= c, & \gamma &:= \sqrt{c - \varepsilon^2 - \phi^2}. \end{aligned}$$

(In the definitions of α, β, and γ, either square root is acceptable.)

Assuming we've done this, let:

$$\tilde{X} := \alpha X + \delta Y + \varepsilon Z, \quad \tilde{Y} := \beta Y + \phi Z, \quad \tilde{Z} := \gamma Z.$$

Then

$$p(X,Y,Z) = \begin{bmatrix} X & Y & Z \end{bmatrix} \begin{bmatrix} \alpha & 0 & 0 \\ \delta & \beta & 0 \\ \varepsilon & \phi & \gamma \end{bmatrix} \begin{bmatrix} \alpha & \delta & \varepsilon \\ 0 & \beta & \phi \\ 0 & 0 & \gamma \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{X} & \tilde{Y} & \tilde{Z} \end{bmatrix} \begin{bmatrix} \tilde{X} \\ \tilde{Y} \\ \tilde{Z} \end{bmatrix} = \tilde{X}^2 + \tilde{Y}^2 + \tilde{Z}^2.$$

Since $\alpha \neq 0 \neq \beta$, if it's also the case that $\gamma \neq 0$, then the determinant $\alpha\beta\gamma$ of the matrix of the linear transformation $L : (X,Y,Z) \mapsto (\tilde{X},\tilde{Y},\tilde{Z})$ is nonzero, and

$$(p \circ L^{-1})(\tilde{X},\tilde{Y},\tilde{Z}) = \tilde{X}^2 + \tilde{Y}^2 + \tilde{Z}^2.$$

Reviewing what we did above, there are two ways that things could not work out. The first possibility is that $\gamma = 0$, so that $\tilde{Z} = 0$, but in this case p is a product of linear functions:

$$p(X,Y,Z) = \tilde{X}^2 + \tilde{Y}^2 = (\tilde{X} + i\tilde{Y})(\tilde{X} - i\tilde{Y}).$$

The other thing that could go wrong is that it may not be the case that $a \neq 0$ and $ab \neq d^2$. It is easy to show that as long as p is not identically zero, there is some linear change of coordinates that makes a, b, and c nonzero. Since we can permute X, Y, and Z, in order for there to be a problem it must be the case that all three of the equations

$$ab = d^2, \quad ac = e^2, \quad bc = f^2$$

hold. Let A, B, and C be square roots of a, b, and c respectively. Then AB is a square root of ab, so we may attain $AB = d$ by negating B if need be. Similarly, we can let C be the square root of c such that $AC = e$. Since f is a square root of bc, either $BC = f$ or $BC = -f$. If $BC = f$, then

$$\begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} = \begin{bmatrix} A \\ B \\ C \end{bmatrix} \begin{bmatrix} A & B & C \end{bmatrix},$$

but once again this implies that p is a product of linear factors:

$$p(X,Y,Z) = \begin{bmatrix} X & Y & Z \end{bmatrix} \left(\begin{bmatrix} A \\ B \\ C \end{bmatrix} \begin{bmatrix} A & B & C \end{bmatrix} \right) \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$
$$= \left(\begin{bmatrix} X & Y & Z \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} \right) \left(\begin{bmatrix} A & B & C \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \right) = (AX + BY + CZ)^2.$$

Therefore we may assume that $BC = -f$, and that $f \neq 0$. We have

$$\begin{bmatrix} a & d & e \\ d & b & f \\ e & f & c \end{bmatrix} = \begin{bmatrix} A \\ B \\ C \end{bmatrix} \begin{bmatrix} A & B & C \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2f \\ 0 & 2f & 0 \end{bmatrix}$$

and

$$p(X,Y,Z) = (AX + BY + CZ)^2 - 4fYZ.$$

Let κ be a square root of $-f$, and let

$$\tilde{X} := AX + BY + CZ, \quad \tilde{Y} := \kappa(Y + Z), \quad \tilde{Z} := -i\kappa(Y - Z).$$

Then

$$Y = (\tilde{Y} + i\tilde{Z})/2\kappa, \quad Z = (\tilde{Y} - i\tilde{Z})/2\kappa, \quad \text{and} \quad X = (\tilde{X} - B\tilde{Y} - C\tilde{Z})/A,$$

so the linear transformation $L : (X,Y,Z) \mapsto (\tilde{X}, \tilde{Y}, \tilde{Z})$ is invertible. In particular, $YZ = (\tilde{Y}^2 + \tilde{Z}^2)/(-4f)$, so

$$p(X,Y,Z) = p(L^{-1}(\tilde{X}, \tilde{Y}, \tilde{Z})) = \tilde{X}^2 + \tilde{Y}^2 + \tilde{Z}^2.$$

In $\mathbb{R}^2$ quadratic polynomials define ellipses (with circles as a special case) parabolas, and hyperbolas. In part because $\mathbb{C}$ is algebraicly complete, and in part because projective space is more symmetric, the classification of quadratic subsets of $P^2(\mathbb{C})$ is much simpler, and actually as simple as possible: there is (up to linear change of coordinates) only one quadratic algebraic curve in $P^2(\mathbb{C})$.

It turns out that nonsingular quadratic curves in $P^2(\mathbb{C})$ are also simple topologically, and as Riemann surfaces, because $V(X^2 + Y^2 + Z^2)$ is holomorphically diffeomorphic to the Riemann sphere. The diffeomorphism is the function $f : S \to V(X^2 + Y^2 + Z^2)$ given, on U_0 and U_1 respectively, by

$$[1, w] \mapsto [w, \tfrac{1}{2}(1 - w^2), \tfrac{i}{2}(1 + w^2)] \text{ and } [z, 1] \mapsto [z, \tfrac{1}{2}(z^2 - 1), \tfrac{i}{2}(z^2 + 1)].$$

Of course these formulas are holomorphic where they are defined. To see that they agree on $U_0 \cap U_1$ observe that when $w \neq 0 \neq z$ they amount to

$$[1, w] \mapsto [1, \tfrac{1}{2}(1/w - w), \tfrac{i}{2}(1/w + w)] \quad \text{and} \quad [1, 1/z] \mapsto [1, \tfrac{1}{2}(z - 1/z), \tfrac{i}{2}(z + 1/z)].$$

The inverse of f is the function $g : V(X^2 + Y^2 + Z^2) \to S$ given by

$$[a, b, c] \mapsto [b - ic, a] \quad \text{and} \quad [a, b, c] \mapsto [a, -b - ic],$$

and again these expressions define holomorphic functions, but of course we have to explain when each of these expressions is defined and why they agree when both are. They are both defined when $a \neq 0$, and when $a = 0$ we have $0 = b^2 + c^2 = (b + ic)(b - ic)$, so either $b + ic = 0$ or $b - ic = 0$, but not both, because $(a, b, c) = (0, 0, 0)$ does not define a point in $P^2(\mathbb{C})$. Therefore each of the formulas is defined on a set obtained by removing a single point (either $[0, 1, i]$ or $[0, 1, -i]$) from $V(X^2 + Y^2 + Z^2)$. For any $[a, b, c] \in V(X^2 + Y^2 + Z^2)$ we have

$$a^2 = -b^2 - c^2 = -(b + ic)(b - ic),$$

so if $a \neq 0$, then $b + ic \neq 0 \neq b - ic$, and

$$\frac{a}{b - ic} = \frac{-b - ic}{a}, \qquad (*)$$

whence $[b - ic, a] = [a, -b - ic]$. Therefore the two formulas agree whenever $a \neq 0$ and, by continuity, wherever they are both defined.

At first sight the verification that f and g are inverses looks like a tedious chore: each function has two formulas, so there are four cases to consider for $g \circ f$, and another four for $g \circ f$. But we can easily compute that:

$$f([1, 0]) = [0, 1, i], \quad g([0, 1, i]) = [1, 0],$$

$$f([0, 1]) = [0, 1, -i], \quad g([0, 1, -i]) = [0, 1].$$

For both functions both formulas are defined at all other points, so it suffices to give one verification for $g \circ f$ and one for $f \circ g$:

$$\begin{aligned} g(f([1, w])) &= g([w, \tfrac{1}{2}(1 - w^2), \tfrac{i}{2}(1 + w^2)]) \\ &= [\tfrac{1}{2}(1 - w^2) - i(\tfrac{i}{2}(1 + w^2)), w] = [1, w], \end{aligned}$$

and (applying $(*)$)

$$f(g([a, b, c])) = f([b - ic, a]) = [1, \tfrac{1}{2}(\tfrac{b-ic}{a} - \tfrac{a}{b-ic}), \tfrac{i}{2}(\tfrac{b-ic}{a} + \tfrac{a}{b-ic})]$$

$$= [1, \tfrac{1}{2}(\tfrac{b-ic}{a} - \tfrac{-b-ic}{a}), \tfrac{i}{2}(\tfrac{b-ic}{a} + \tfrac{-b-ic}{a})] = [1, \tfrac{b}{a}, \tfrac{c}{a}] = [a, b, c].$$

Having shown that every nonsingular quadratic curve in $P^2(\mathbb{C})$ is holomorphically diffeomorphic to the Riemann sphere, we now consider **elliptic curves**, which are Riemann surfaces $V(p)$ where p is a homogeneous polynomial of degree three. In contrast with degrees one and two, there are many elliptic curves, so we won't attempt the sort of analysis we saw above. Instead, we'll look at a different way to construct Riemann surfaces that happens to give all the elliptic curves.

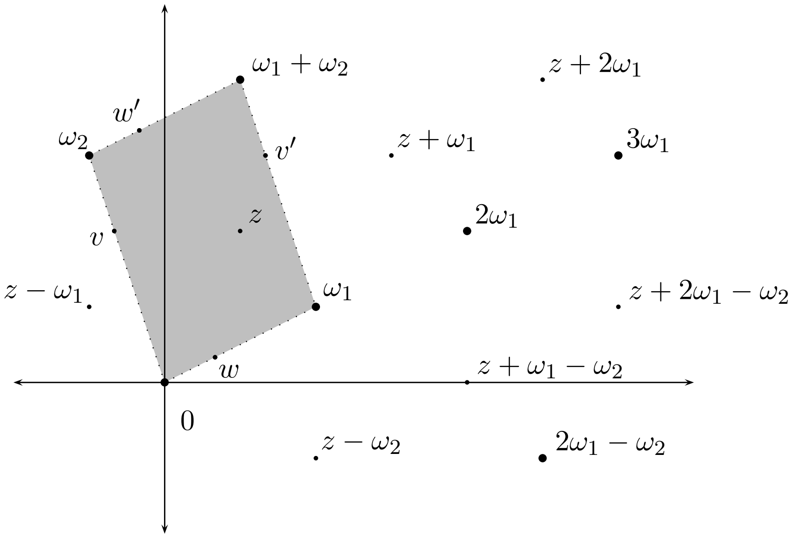


Figure 9.1

Suppose that ω_1 and ω_2 are two nonzero elements of $\mathbb{C}$ that are not on the same line through the origin, so that $\omega_2/\omega_1 \notin \mathbb{R}$. Then they are linear independent if we think of $\mathbb{C}$ as a vector space over $\mathbb{R}$, and their span $\{\, t_1\omega_1 + t_2\omega_2 : t_1, t_2 \in \mathbb{R} \,\}$ is all of $\mathbb{C}$. The **lattice** associated with ω_1 and ω_2 is

$$L(\omega_1, \omega_2) := \{\, n_1\omega_1 + n_2\omega_2 : n_1, n_2 \in \mathbb{Z} \,\}.$$

For any such lattice L there is a Riemann surface $C(L)$ obtained by identifying any two points in $\mathbb{C}$ whose difference is a lattice point, so that z is the same point as $z + \omega_1$, $z + \omega_2$, $z - 2\omega_1 + 5\omega_2$, etc. In group-theoretic terms, L is a subgroup of $\mathbb{C}$ (with addition as the group operation) which is necessarily normal because $\mathbb{C}$ is abelian, and $C(L)$ is the quotient group $\mathbb{C}/L$. Formally, the points of $C(L)$ are the cosets

$$z + L := \{\, z + \lambda : \lambda \in L \,\}.$$

We can also think of constructing $C(L)$ by starting with the parallelepiped region

$$\{\, t_1\omega_1 + t_2\omega_2 : 0 \le t_1, t_2 \le 1 \,\}$$

shown in Figure 9.1 and identifying points along the two edges, so that v is the same point as v' and w is the same point as w'. In particular, as a topological manifold, $C(L)$ is homeomorphic to a torus.

The holomorphic differentiable structure is given by an atlas with the following description. We say that an open set $V \subset \mathbb{C}$ is **L-small** if any coset $z + L$ intersects V at at most one point, so that $(z + L) \cap V = \{z\}$ for all $z \in V$. For such a V let $U_V := \{\, z + L : z \in V \,\}$, and define $\varphi_V : U_V \to V$ by requiring that $\varphi_V(z + L) = z$ for each $z \in V$. To show that this is a holomorphic atlas, suppose that V' is another L-small open set. Then φ_V and $\varphi_{V'}$ have holomorphic overlap because if $z + L = z' + L$ for some $z \in V$ and $z' \in V'$, then $\varphi_{V'} \circ \varphi_V^{-1}$ agrees with the map $w \mapsto w + (z' - z)$ on some neighborhood of z.

Sophisticated arguments show that any elliptic curve is homeomorphic to a torus, and that whenever a Riemann surface is homeomorphic to a torus, it is holomorphically diffeomorphic to some elliptic curve, and to $C(L)$ for some lattice L. This is a rather mystifying situation. For any lattice L, $C(L)$ is holomorphically diffeomorphic to $V(p)$ for some homogeneous polynomial p of degree three, but there is no simple or obvious way to derive p from L. Given a degree three homogeneous polynomial p such that 0 is a regular value of p^{-0}, p^{-1}, and p^{-2}, we know that $V(p)$ is holomorphically diffeomorphic to some $C(L)$, but it is not simple to pass from p to a suitable L.

In fact this is just the tip of a very large iceberg: the theory of elliptic curves is extensive, deep, and not completely understood. It was central to the proof of Fermat's last theorem. Briefly, if $p \neq 3$ is an odd prime and ℓ, m, and n are integers such that $\ell^p + m^p = n^p$, then the elliptic curve

$$V\big(Y^2 Z - X(X + \ell^p Z)(X - m^p Z)\big)$$

cannot have a certain property, but Andrew Wiles (with some help from Richard Taylor) showed that every relevant elliptic curve must have this property. One of the most famous open problems of contemporary mathematics, the conjecture of Birch and Swinnerton-Dyer, asserts that if C is an elliptic curve defined by a cubic polynomial with integer coefficients, then two attributes of C are the same. There are algorithms that compute these attributes for any given curve, and the conjecture is supported by a large body of computational evidence, but attempts to prove it have all run into dead ends.

Now consider a second lattice

$$L' = \{\, n_1\omega_1' + n_2\omega_2' : n_1, n_2 \in \mathbb{Z} \,\}.$$

When are $C(L)$ and $C(L')$ holomorphically diffeomorphic? The map

$$t_1\omega_1 + t_2\omega_2 + L \;\mapsto\; t_1\omega_1' + t_2\omega_2' + L'$$

is a real analytic diffeomorphism from $C(L)$ to $C(L')$, but it's usually not a holomorphic diffeomorphism because it's not conformal, in which case the Cauchy-Riemann equations don't hold. On the other hand, if there is a scalar $\alpha \in \mathbb{C}^*$ such that $L' = \alpha L$, then the map $z + L \mapsto \alpha z + L'$ *is* a holomorphic diffeomorphism; we say that L and L' are **homothetic** if this is the case. The proof is a bit beyond the tools we have at this point, but it turns out that this is the *only* way that $C(L)$ and $C(L')$ can be holomorphically diffeomorphic. That is, if $C(L)$ and $C(L')$ are holomorphically diffeomorphic, then L and L' are homothetic.

Homotheticity is clearly an equivalence relation, so the holomorphic diffeomorphism classes of elliptic curves are in one-to-one correspondence with the homotheticity classes of lattices. We're now going to investigate this space of equivalence classes. We always have $L(\omega_1, \omega_2) = L(\omega_2, \omega_1)$, and the imaginary part of a complex number is negative if and only if the imaginary part of its inverse is positive, so we can adopt the convention that whenever we represent a lattice as $L(\omega_1, \omega_2)$, the two generators are ordered so that ω_1/ω_2 is an element of the upper half plane

$$\mathcal{H} := \{\, x + iy : x \in \mathbb{R}, y > 0 \,\} \subset \mathbb{C}.$$

Since $L(\omega_1, \omega_2)$ and $L(\omega_1/\omega_2, 1)$ are homothetic, every lattice is homothetic to $L(\tau, 1)$ for some complex number $\tau \in \mathcal{H}$. For which $\tau, \tau' \in \mathcal{H}$ is it the case that $L(\tau, 1)$ and $L(\tau', 1)$ are homothetic?

Suppose that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is a nonsingular matrix whose entries are integers, and let $\tau' := (a\tau + b)/(c\tau + d)$. Then $L(\tau', 1)$ and $L(a\tau + b, c\tau + d) = (c\tau + d)L(\tau', 1)$ are homothetic, and $L(a\tau + b, c\tau + d) \subset L(\tau, 1)$ because $L(\tau, 1)$ contains $a\tau + b$ and $c\tau + d$. If the determinant $ad - bc$ of the matrix is 1, then $L(a\tau + b, c\tau + d) = L(\tau, 1)$ because $L(a\tau + b, c\tau + d)$ contains τ and 1:

$$d(a\tau + b) - b(c\tau + d) = \tau \quad\text{and}\quad -c(a\tau + b) + a(c\tau + d) = 1.$$

Let Γ be the set of 2×2 matrices with integer entries and determinant 1. The product of two matrices with integer entries has integer entries, so the multiplicative property of the determinant implies that Γ contains the product of any two of its elements. For any $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ the inverse $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)^{-1} = \left(\begin{smallmatrix} d & -b \\ -c & a \end{smallmatrix}\right)$ is an integer matrix whose determinant $da - (-b)(-c) = ad - bc$ is one, so it is in Γ. Therefore Γ is a group because it is a subset of the group of nonsingular 2×2 matrices that contains the products and inverses of its elements.

For $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ and $\tau \in \mathcal{H}$ let

$$g(\tau) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} (\tau) := \frac{a\tau + b}{b\tau + d}.$$

This is the restriction to $\mathcal{H}$ of a Möbius transformation. Since a, b, c, and d are all real, the Möbius transformation takes the "extended" real line $\mathbb{R} \cup \{\infty\}$ to itself, so it should come as no surprise that it maps $\mathcal{H}$ to itself. Nonetheless it will be useful to have the following rather clever calculation which gives a quantitative expression of this fact: the imaginary part of

$$\frac{a\tau + b}{c\tau + d} = \frac{(a\tau + b)(c\overline{\tau} + d)}{(c\tau + d)(c\overline{\tau} + d)} = \frac{ac\tau\overline{\tau} + bd + ad\tau + bc\overline{\tau}}{|c\tau + d|^2} \tag{$*$}$$

is $(ad - bc)/|c\tau + d|^2 = |c\tau + d|^{-2}$ times the imaginary part of τ, and is consequently necessarily positive.

So far we have seen that if $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$, $\tau \in \mathcal{H}$, and $\tau' = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)(\tau)$, then $\tau' \in \mathcal{H}$ and $L(\tau, 1)$ and $L(\tau', 1)$ are homothetic. It turns out that the converse also holds: if $\tau, \tau' \in \mathcal{H}$ and $L(\tau, 1)$ and $L(\tau', 1)$ are homothetic, then $\tau' = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)(\tau)$ for some $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$. To see this, suppose that $L(\tau', 1) = \alpha L(\tau, 1)$ for some α. Then there are integers a, b, c, d such that $\tau' = \alpha(a\tau + b)$ and $1 = \alpha(c\tau + d)$, and dividing the first equation by the second gives $\tau' = (a\tau + b)/(c\tau + d)$. This reasoning is equally valid with τ and τ' reversed, so we arrive at

$$\begin{pmatrix} \tau' \\ 1 \end{pmatrix} = \alpha \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \tau \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \tau \\ 1 \end{pmatrix} = \frac{1}{\alpha} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \begin{pmatrix} \tau' \\ 1 \end{pmatrix},$$

where a', b', c', d' are also integers, and these combine to give

$$\begin{pmatrix} \tau \\ 1 \end{pmatrix} = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \tau \\ 1 \end{pmatrix}.$$

If we think about this from the point of view of the identification of $\mathbb{C}$ with $\mathbb{R}^2$, so that $\tau = \tau_{\text{re}} + i\tau_{\text{im}}$ and $1 = 1 + i0$ are identified with $(\tau_{\text{re}}, \tau_{\text{im}})$

and $(1,0)$ respectively, then these two vectors are linearly independent, so the only way this equation can hold is if the matrix product is the identity matrix. Therefore the two matrices are inverses of each other, and their determinants are integers whose product is 1, so these determinants are either both 1 or both -1. The calculation $(*)$ shows that the imaginary part of τ' is $(ad-bc)/|c\tau+d|^2$ times the imaginary part of τ, so $ad-bc=1$ because τ and τ' are both elements of $\mathcal{H}$.

The map $\left(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right), \tau\right) \mapsto \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)(\tau)$ is actually a group action of Γ on $\mathcal{H}$; to see this we compute that

$$\begin{aligned} \begin{pmatrix} e & f \\ g & h \end{pmatrix} \left(\begin{pmatrix} a & b \\ c & d \end{pmatrix} (\tau) \right) &= \frac{e\frac{a\tau+b}{c\tau+d} + f}{g\frac{a\tau+b}{c\tau+d} + h} = \frac{e(a\tau+b)+f(c\tau+d)}{e(a\tau+b)+f(c\tau+d)} \\ &= \frac{(ea+fc)\tau + (eb+fd)}{(ga+hc)\tau+(gb+hd)} = \begin{pmatrix} ea+fc & eb+fd \\ ga+hc & gb+hd \end{pmatrix}(\tau) \\ &= \left(\begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right)(\tau). \end{aligned}$$

In general, whenever a group H acts on a set A, a set of the form $\{\, ha : h \in H \,\}$ is called the **orbit** of a. For example, any subgroup of a group acts on the group itself, and the orbits of this action are the cosets. The orbits of the action of Γ on $\mathcal{H}$ are the subsets of $\mathcal{H}$ of the form

$$O(\tau) := \left\{ \frac{a\tau+b}{c\tau+d} : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma \right\}.$$

We have given a bijection between the collection of all holomorphic diffeomorphism classes of elliptic curves and the set $M := \{\, O(\tau) : \tau \in \mathcal{H} \,\}$ of such orbits, and we now wish to study this orbit space.

For a lattice L the map $(\lambda, z) \mapsto \lambda + z$ is an action of L on $\mathbb{C}$, and the orbits of this action are the points of a new Riemann surface, namely the elliptic curve derived from L. The reasons for this seem quite general: if a group acts on a Riemann surface in a nice way (where "niceness" has not yet been defined precisely) then the space of orbits should be a new Riemann surface. These thoughts suggest trying to endow M with the structure of a Riemann surface.

Suppose that the group H acts on the set A, and let N be the set of group elements n such that $na = a$ for all $a \in A$. For such an n and all a we have $n^{-1}a = n^{-1}(na) = (n^{-1}n)a = ea = a$, so N contains the inverse of each of its elements. It obviously contains the product of any two of its elements, so it is a subgroup of H, and in fact it is a normal subgroup. To see this recall that for any group G and $g \in G$, C_g is the inner isomorphism $h \mapsto ghg^{-1}$, and observe that $C_h(n) \in N$ for any $n \in N$ and $h \in H$ because for all $a \in A$ we have

$$C_h(n)a = (h^{-1}nh)a = h^{-1}n(ha) = h^{-1}(ha) = (h^{-1}h)a = a.$$

It is natural to think that the action of H is *really* an action of the quotient group H/N.

If $(a\tau+b)/(c\tau+d) = \tau$ for all $\tau \in H$, then the polynomial $c\tau^2+(d-a)\tau-b$ vanishes identically, so that $b = 0 = c$ and $a = d$. Therefore $J := \{I, -I\}$ is the set of elements of Γ that leave every point of $\mathcal{H}$ fixed. We will say that an open set $V \subset \mathcal{H}$ is **Γ-small** if $g \in J$ (so that $g(\tau) = \tau$) whenever $g \in \Gamma$ and V contains both τ and $g(\tau)$. For such a V let

$$U_V := \{ O(\tau) : \tau \in V \} \subset M,$$

and define $\varphi_V : U_V \to V$ by specifying that $\varphi_V(O(\tau)) = \tau$ whenever $\tau \in V$. This definition is unambiguous: if $O(\tau) = O(\tau')$, then $\tau' = g(\tau)$ for some g, necessarily $g \in J$, and consequently $\tau' = \tau$.

We would like to show that these maps have holomorphic overlaps, so suppose that V and V' are Γ-small, and consider a point $\tilde{\tau}$ in the domain of $\varphi_{V'} \circ \varphi_V^{-1}$. Since $O(\tilde{\tau}) \in U_{V'}$ there is some $g \in \Gamma$ such that $g(\tilde{\tau}) \in V'$. The definitions of φ_V and $\varphi_{V'}$ state that $\varphi_V(O(\tau)) = \tau$ whenever $\tau \in V$ and $\varphi_{V'}(O(g(\tau))) = g(\tau)$ whenever $g(\tau) \in V'$, as will be the case (by continuity) for all τ in some neighborhood of $\tilde{\tau}$, in which case

$$\varphi_{V'}(\varphi_V^{-1}(\tau)) = \varphi_{V'}(O(\tau)) = \varphi_{V'}(O(g(\tau))) = g(\tau).$$

That is, $\varphi_{V'} \circ \varphi_V^{-1}$ agrees with the holomorphic map $\tau \mapsto g(\tau)$ in a neighborhood of $\tilde{\tau}$.

Since the maps φ_V do have holomorphic overlaps, they would constitute a holomorphic atlas for M if every point in $\mathcal{H}$ had a Γ-small neighborhood. *But this isn't true!* From this point on things will become a bit more complicated, which is not really to say that the material is truly "harder" or more "advanced," in a mathematical sense, than what we have done up to this point. But there is more detail, and some cumbersome computations, which will require a somewhat higher degree of concentration on your part.

The **stabilizer subgroup** of a point $\tau \in \mathcal{H}$ is

$$\Gamma_\tau := \{ g \in \Gamma : g(\tau) = \tau \}.$$

If $g, g' \in \Gamma_\tau$, then $g'g(\tau) = g'(g(\tau)) = g'(\tau) = \tau$ and $g^{-1}(\tau) = g^{-1}(g(\tau)) = \tau$, so Γ_τ contains products and inverses of its elements and therefore is, in fact, a subgroup of Γ. The next result gives a related generalization of the notion of a Γ-small open set.

Lemma 9.28. *Each $\tau \in \mathcal{H}$ has a neighborhood V such that for all $\tau' \in V$ and $g \in \Gamma$, if $g(\tau') \in V$, then $g \in \Gamma_\tau$.*

Proof. Otherwise there must exist sequences $\{\tau_n\}$ in $\mathcal{H}$ and $\{g_n\}$ in $\Gamma \setminus \Gamma_\tau$ such that $\tau_n \to \tau$ and $g_n(\tau_n) \to \tau$. Let $g_n := \left(\begin{smallmatrix} a_n & b_n \\ c_n & d_n \end{smallmatrix}\right)$. Since $\tau_n \to \tau$ and $g_n(\tau_n) = (a\tau_n + b_n)/(c_n\tau_n + d_n) \to \tau$, equation $(*)$ implies that $|c_n\tau_n + d_n|^{-2} \to 1$. Let $\tau_n = x_n + iy_n$ and $\tau = x + iy$. Then

$$|c_n\tau_n + d_n|^2 = (c_nx_n + d_n)^2 + c_n^2y_n^2 \to 1.$$

Since $y_n \to y > 0$, for large n there are only finitely many possibilities for c_n, and since $x_n \to x$, for any given value of c_n there are only finitely many possibilities for d_n when n is large. Passing to a subsequence, we may assume that there is a single pair (c, d) such that $(c_n, d_n) = (c, d)$ for all n. We now have $a_n\tau_n + b_n = \tau_n(c\tau_n + d) \to \tau(c\tau + d)$. Since a_n is an integer and the imaginary part of $a_n\tau_n$ converges to the imaginary part of $\tau(c\tau+d)$, there must be an integer a such that $a_n = a$ for all large n. Similarly, since $a\tau_n + b_n \to \tau(c\tau + d)$ there must be an integer b such that $b_n = b$ for all large n. Let $g := \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then $g_n = g$ for all large n, but $\tau_n \to \tau$ and $g(\tau_n) \to \tau$, so by continuity, $\tau = g(\tau)$, which contradicts our assumption that $g_n \notin \Gamma_\tau$. □

Note that J is always a subgroup of Γ_τ. A point $\tau \in \mathcal{H}$ is said to be **elliptic** if J is a proper subgroup of Γ_τ, so there is at least one $g \in \Gamma \setminus J$ such that $g(\tau) = \tau$. The last result implies that if a point isn't elliptic, then it has a Γ-small neighborhood. On the other hand, the definition of a Γ-small neighborhood immediately implies that an elliptic point cannot have such a neighborhood because any neighborhood contains the point itself. The remaining task in constructing the holomorphic differentiable structure on M is to figure out what the elliptic points are and what to do about them.

Suppose τ is elliptic, say because $g(\tau) = \tau$ for some $g \in \Gamma \setminus J$. If $h \in \Gamma$, then $h(\tau)$ is elliptic because $hgh^{-1}(h(\tau)) = hgh^{-1}h(\tau) = hg(\tau) = h(\tau)$. (Since J is a normal subgroup of Γ, $hgh^{-1} \notin J$.) That is, if τ is elliptic, then so is every element of $O(\tau)$. Every orbit has elements with certain properties that we describe next, and we will find the elliptic points by looking for elliptic points with these properties.

Fix a $\tau = x + iy \in \mathcal{H}$. For any integers c, d we have

$$|c\tau + d|^2 = (cx + d)^2 + c^2y^2.$$

There are only finitely many pairs of integers (c, d) with $|c\tau+d|^2 < 1$ because there are finitely many integers c with $c^2y^2 < 1$, and for each such c there are finitely many integers d such that $(cx + d)^2 < 1 - c^2y^2$. If $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$, $(*)$ implies that the imaginary part of $g(\tau)$ is $|c\tau + d|^{-2}$ times the imaginary part of τ. It follows that if α is the imaginary part of some element of $O(\tau)$, then the set of imaginary parts of elements of $O(\tau)$ contains only finitely many elements greater than α. In particular, *there is some element of $O(\tau)$ whose imaginary part is as large as the imaginary part of any other element of this orbit.*

Let $S := \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $T := \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. These are elements of Γ, of course, and we have $S(\tau) = -1/\tau = -\overline{\tau}/|\tau|^2$, so if the imaginary part of τ is as large as the imaginary part of $S(\tau)$, then $|\tau| \geq 1$. Direct computation shows that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & n+1 \\ 0 & 1 \end{smallmatrix}\right)$ for any integer n, so $T^{-1} := \left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$, and (by induction away from $n = 0$ in both directions) $T^n = \left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right)$. Since $T^n(\tau) = \tau + n$ has the same imaginary part as τ, some element of $O(\tau)$ whose imaginary part is maximal has a real part in the interval $(-\frac{1}{2}, \frac{1}{2}]$.

Now suppose that τ is elliptic, say because $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_\tau \setminus J$, and that τ has the other properties laid out above: y is as large as the imaginary part of any element of $O(\tau)$, $|\tau| \geq 1$, and $x \in (-\frac{1}{2}, \frac{1}{2}]$. Note that these conditions imply that $y \geq \sqrt{1 - x^2} \geq \sqrt{3}/2$.

The only possible values of c are -1, 0, and 1 because

$$1 = |c\tau + d|^2 = (cx + d)^2 + c^2y^2 \geq c^2y^2 \geq 3c^2/4,$$

If $c = 0$, then the first equality above implies that $d^2 = 1$, in which case $a = d$, because $1 = ad - bc = ad$, and $b = 0$, because $\tau = (a\tau + b)/(c\tau + d) = \tau + b/d$. That is, $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \pm\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, which contradicts the assumption that $g \notin J$, so $c = 0$ is impossible. In addition, we may assume that $c = 1$ because $g \in \Gamma_\tau \setminus J$ if and only if $-g \in \Gamma_\tau \setminus J$, and we can replace g with $-g$. The calculation above now gives $(x + d)^2 = 1 - y^2 \leq 1/4$, so $|x + d| \leq 1/2$. In view of the interval containing x, either (i) $d = 0$ or (ii) $x = 1/2$ and $d = -1$. We consider these two cases in turn.

Suppose $d = 0$. Then $b = bc = -(ad - bc) = -1$, so

$$\tau = (a\tau + b)/(c\tau + d) = (a\tau - 1)/\tau = a - 1/\tau.$$

In particular, the imaginary part of $a = \tau + 1/\tau$ is zero, so $|\tau|^2 = 1$ because the imaginary part of $1/\tau = \overline{\tau}/|\tau|^2$ is $-|\tau|^{-2}$ times the imaginary part of τ.

Therefore $1/\tau = \overline{\tau} = x - iy$ and $a = \tau + 1/\tau = 2x$, so a is either 0 or 1. If $a = 0$, so that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, then $\tau + 1/\tau = 0$, i.e., $\tau^2 = -1$, and $\tau = i$ is the square root of -1 in $\mathcal{H}$. If $a = 1$, so that $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, then $\tau + 1/\tau = 1$, i.e., $\tau^2 - \tau + 1 = 0$, and the root of this equation in $\mathcal{H}$ is

$$\rho := \frac{1 + i\sqrt{3}}{2}.$$

Finally, suppose that $c = 1$, $d = -1$, and $x = 1/2$. Then

$$\tau(c\tau + d) = (\tfrac{1}{2} + iy)(-\tfrac{1}{2} + iy) = -\tfrac{1}{4} - y^2,$$

and of course $\tau = (a\tau + b)/(c\tau + d)$, so

$$-\tfrac{1}{4} - y^2 = \tau(c\tau + d) = a\tau + b = (\tfrac{1}{2}a + b) + iay.$$

Equating imaginary parts yields $a = 0$, so $b = bc = -(ad - bc) = -1$. Therefore $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & -1 \end{smallmatrix}\right)$ and $-\frac{1}{4} - y^2 = -1$, so $y = \sqrt{3}/2$ and $\tau = \rho$.

We had to consider numerous cases and details, but the bottom line is pretty simple. Let $g_i := \pm\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $g_\rho := \pm\left(\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Either:

(a) $\tau = i$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \pm g_i$, or

(b) $\tau = \rho$ and either $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \pm g_\rho$ or $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \pm\left(\begin{smallmatrix} 0 & -1 \\ 1 & -1 \end{smallmatrix}\right)$.

Note that $\rho^2 = (1 + 2i\sqrt{3} - 3)/4 = (-1 + i\sqrt{3})/2 = \rho - 1$, and in particular $\rho = 1 - 1/\rho$. In view of this it is easy to see that i and ρ are, in fact, elliptic because $g_i(i) = -1/i = i$ and $g_\rho(\rho) = (\rho - 1)/\rho = \rho$. We conclude that the set of elliptic points is $O(i) \cup O(\rho)$.

To get a better view of what's going on here we observe that the stabilizer subgroups of i and ρ are

$$\Gamma_i = \left\{ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}$$

and

$$\Gamma_\rho = \left\{ \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}.$$

That is, $\Gamma_i = \{g_i, g_i^2, g_i^3, g_i^4\}$ and $\Gamma_\rho = \{g_\rho, g_\rho^2, g_\rho^3, g_\rho^4, g_\rho^5, g_\rho^6\}$. (Computing all the powers of g_i and g_ρ by hand is a big chore, and you don't need to do it if you don't want to.) We have $g_i^2 = -I = g_\rho^3$ and $g_i^4 = I = g_\rho^6$, and we can see that Γ_i/J is the unique (up to isomorphism) group with two elements while Γ_ρ/J is the unique group with three elements.

Well, all this is very nice, but what should we do about it? How are we supposed to construct coordinate charts for M on neighborhoods of $O(i)$ and $O(\rho)$?

Extending our terminology a bit, we will say that a neighborhood $\tau \in \mathcal{H}$ is **Γ_τ-small** if, for each $\tau' \in V$, the set of $g \in \Gamma$ such that $g(\tau') \in V$ is precisely Γ_τ. Lemma 9.28 gives an open neighborhood V' of τ such that for each $\tau' \in V'$, the set of $g \in \Gamma$ such that $g(\tau') \in V'$ is contained in Γ_τ. Let $V := \bigcap_{g \in \Gamma_\tau} g(V')$. Of course V contains τ, and we have shown that the stabilizer subgroup of every element of $\mathcal{H}$ is finite, so V is open. For any $g' \in \Gamma_\tau$ we have $g'(V) = \bigcap_{g \in \Gamma_\tau} g'g(V') = \bigcap_{g \in \Gamma_\tau} g(V') = V$, so V is Γ_τ-small. We now fix a Γ_i-small neighborhood V_i of i and a Γ_ρ-small neighborhood V_ρ of ρ. Let $U_i := \{\, O(\tau) : \tau \in V_i \,\}$ and $U_\rho := \{\, O(\tau) : \tau \in V_\rho \,\}$.

If τ is a point near i, but different from i, then $g_i(\tau) = -1/\tau$ is different from τ (because, after all, the roots of the equation $\tau = -1/\tau$ are i and $-i$) but $O(-1/\tau) = O(\tau)$. Then the restriction of the map $\tau \mapsto O(\tau)$ to the "punctured" neighborhood $V_i \setminus \{i\}$ is two-to-one. This should remind you of the map $z \mapsto z^2$.

In order to make this more than just a metaphor we use the function

$$\theta_i : \mathbb{C} \setminus \{-i\} \to \mathbb{C} \setminus \{1\} \quad \text{given by} \quad \theta_i(\tau) := \frac{\tau - i}{\tau + i}$$

to impose a new coordinate system on $\mathcal{H}$. This Möbius transformation is called the **Cayley transform**. It maps a point $\tau \in \mathbb{C}$ to a point in the unit disk $D = \{\, z \in \mathbb{C} : |z| < 1 \,\}$ if and only if $|\tau - i| < |\tau + i|$, and the set of points in $\mathbb{C}$ that are closer to i than to $-i$ is $\mathcal{H}$, so θ_i restricts to a bijection between $\mathcal{H}$ and D. Its inverse has the formula $z \mapsto -i(z + 1)(z - 1)$. That this is, in fact, the inverse can be verified directly by a couple rather cumbersome computations, but it is easier to observe that the product $\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}\begin{pmatrix} -i & -i \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 2i & 0 \\ 0 & 2i \end{pmatrix}$ of the matrices that define these Möbius transformations is a multiple of $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Of course it is important that $\theta_i(i) = 0$. Since θ_i maps the extended real line to the circle $C = \{\, z \in \mathbb{C} : |z| = 1 \,\}$, and g_i maps the extended real line to itself, the Möbius transformation $\theta_i \circ g_i \circ \theta_i^{-1}$ maps C into itself, and $\theta_i(g_i(\theta_i^{-1}(0))) = \theta_i(g_i(i)) = \theta_i(i) = 0$. Recall that in Section 9.2 we saw that a circular transformation mapping 0 to 0 is a rotation of $\mathbb{C}$ corresponding to multiplication by some element of the unit circle C. In fact setting $\tau =$

$\theta_i^{-1}(z)$ in equation $(**)$ below shows that $\theta_i(g_i(\theta_i^{-1}(z))) = -z$.

We now define $\varphi_i : U_i \to \mathbb{C}$ by setting

$$\varphi_i(O(\tau)) := \theta_i(\tau)^2.$$

In order for this to be a valid definition it must be the case that $\theta_i(\tau)^2 = \theta_i(\tau')^2$ whenever $\tau, \tau' \in V_i$ with $O(\tau) = O(\tau')$. If $O(\tau) = O(\tau')$, then $\tau' = g(\tau)$ for some $g \in \Gamma_i$, with the only potentially problematic case being $g = g_i$. But $\theta_i\big(g_i(\tau)\big)^2 = \theta_i(\tau)^2$ because

$$\theta_i\big(g_i(\tau)\big) = \theta_i(-1/\tau) = \frac{-\frac{1}{\tau} - i}{-\frac{1}{\tau} + i} = \frac{i(-1 - i\tau)}{i(-1 + i\tau)} = -\frac{\tau - i}{\tau + i} = -\theta_i(\tau). \qquad (**)$$

The map φ_i is compatible with our other charts in the following sense: if V is a Γ-small open set and $\tau \in V \cap V_i$, then on some neighborhood of $\varphi_V(O(\tau)) = \tau$ the change of coordinates $\varphi_i \circ \varphi_V^{-1}$ is holomorphic because it agrees with $\tau' \mapsto \theta_i(\tau')^2$, and on some neighborhood of $\varphi_i(O(\tau)) = \theta_i(\tau)^2$ the change of coordinates $\varphi_V \circ \varphi_i^{-1}$ is holomorphic because it agrees with $z \mapsto \theta_i^{-1}(\sqrt{z})$ for a suitable branch of the square root "function." (Recall our discussion of the logarithm function.)

Our treatment of the elliptic point ρ is similar, but before going into the details we should say a few things about the number ρ, which is actually pretty important. We have already seen that $\rho^2 = \rho - 1 = (-1 + i\sqrt{3})/2$. Going a step further, we have

$$\rho^3 = \rho \cdot \rho^2 = \tfrac{1}{2}(1 + i\sqrt{3}) \cdot \tfrac{1}{2}(-1 + i\sqrt{3}) = \tfrac{1}{4}(-1 + (i\sqrt{3})^2) = -1.$$

Therefore $\rho^6 = 1$, $1/\rho = \rho^5 = \rho^3\rho^2 = -\rho^2$, and similarly $\rho^4 = -\rho$. The multiplicative property of the norm implies that the norm of a root of unity is 1, and if z is a complex number with $|z| = 1$, then $z^{-1} = \overline{z}/|z|^2 = \overline{z}$. Therefore $1/\rho = \overline{\rho}$. All of these equations hold with $\overline{\rho}$ in place of ρ.

Let $\theta_\rho : \mathbb{C} \setminus \{\overline{\rho}\} \to \mathbb{C} \setminus \{1\}$ be the function

$$\theta_\rho(\tau) := \frac{\tau - \rho}{\tau - \overline{\rho}}.$$

This Möbius transformation restricts to a bijection between $\mathcal{H}$ (which is the set of points in $\mathbb{C}$ that are closer to ρ than to $\overline{\rho}$) and D. Its inverse is $z \mapsto (\overline{\rho}z - \rho)/(z - 1)$, as can be seen by computing that the product $\begin{pmatrix} 1 & -\rho \\ 1 & -\overline{\rho} \end{pmatrix}\begin{pmatrix} \overline{\rho} & -\rho \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} \overline{\rho} - \rho & 0 \\ 0 & \overline{\rho} - \rho \end{pmatrix}$ of the associated matrices is a multiple of the identity matrix. Clearly $\theta_\rho \circ g_\rho \circ \theta_\rho^{-1}$ takes D to itself, i.e., it is a circular transformation, and

$$\theta_\rho(g_\rho(\theta_\rho^{-1}(0))) = \theta_\rho(g_\rho(\rho)) = \theta_\rho(\rho) = 0.$$

Below it will become apparent that $\theta_\rho \circ g_\rho \circ \theta_\rho^{-1}$ is multiplication by $-\rho^2$.

Define $\varphi_\rho : U_\rho \to \mathbb{C}$ by setting

$$\varphi_\rho(O(\tau)) := \theta_\rho(\tau)^3.$$

This function is well defined because for $\tau, \tau' \in V_\rho$ we have $O(\tau) = O(\tau')$ if and only if $\tau' = \tau$, $\tau' = g_\rho(\tau)$, or $\tau' = g_\rho^2(\tau)$, and

$$\begin{aligned}\theta_\rho(g_\rho(\tau)) &= \frac{\frac{\tau-1}{\tau} - \rho}{\frac{\tau-1}{\tau} - \overline{\rho}} = \frac{(1-\rho)\tau - 1}{(1-\overline{\rho})\tau - 1} = \frac{-\rho^2\tau - 1}{-\overline{\rho}^2\tau - 1} \\ &= \frac{\rho^2}{\overline{\rho}^2} \cdot \frac{\tau + 1/\rho^2}{\tau + 1/\overline{\rho}^2} = \rho^{-2}\frac{\tau - \rho}{\tau - \overline{\rho}} = \rho^{-2}\theta_\rho(\tau),\end{aligned}$$

so that $\theta_\rho\big(g_\rho^2(\tau)\big)^3 = \theta_\rho\big(g_\rho(\tau)\big)^3 = \theta_\rho(\tau)^3$. It is compatible with our other coordinate charts: (a) if V is a Γ-small open set and $\tau \in V \cap V_\rho$, then on some neighborhood of $\varphi_V(O(\tau)) = \tau$ the change of coordinates $\varphi_\rho \circ \varphi_V^{-1}$ is holomorphic because it agrees with $\tau' \mapsto \theta_\rho(\tau')^3$, and on some neighborhood of $\varphi_\rho(O(\tau)) = \theta_\rho(\tau)^3$ the change of coordinates $\varphi_V \circ \varphi_\rho^{-1}$ is holomorphic because it agrees with $z \mapsto \theta_\rho^{-1}(\sqrt[3]{z})$ for a suitable branch of the cube root function; (b) by choosing small enough V_i and V_ρ we can insure that their intersection is empty, or we can argue that each point in their intersection is nonelliptic and consequently contained in some Γ-small V, so $\varphi_i \circ \varphi_\rho^{-1}$ agrees with $(\varphi_i \circ \varphi_V^{-1}) \circ (\varphi_V \circ \varphi_\rho^{-1})$ near this point, and similarly for $\varphi_\rho \circ \varphi_i^{-1}$.

This completes the demonstration that

$$\{\, \varphi_V : V \text{ is } \Gamma\text{-small} \,\} \cup \{\varphi_i, \varphi_\rho\}$$

is a holomorphic atlas for $M = \{\, O(\tau) : \tau \in \mathcal{H} \,\}$. Let's review and summarize. Each lattice determines an elliptic curve, and every elliptic curve is holomorphically diffeomorphic to one determined by some lattice. Two lattices are homothetic if and only if the associated elliptic curves are holomorphically diffeomorphic, so the homotheticity classes of lattices are in one-to-one correspondence with the holomorphic diffeomorphism classes of elliptic curves. The homotheticity classes of lattices are in natural bijection with the space of orbits of the action of Γ on $\mathcal{H}$, and this space of orbits can be endowed with the structure of a Riemann surface. In sum the construction above shows that the space of holomorphic diffeomorphism classes of elliptic curves is itself a Riemann surface!

9.5 The Fundamental Group

It's easy to see that the torus and the sphere are not homeomorphic, but how do you prove it?

One way to prove that two topological spaces X and Y are not homeomorphic is to find a point $x \in X$ such that the local nature of the space near x has properties that are different from the local nature of Y near any of its points. For example, every neighborhood of 0 in $\mathbb{R}$ becomes disconnected if we remove 0 itself from the neighborhood, but every point in $\mathbb{R}^2$ has a neighborhood that does not become disconnected if we remove the point itself. Therefore $\mathbb{R}$ and $\mathbb{R}^2$ cannot be homeomorphic. But any point in the torus and any point in the sphere have homeomorphic neighborhoods, so this certainly won't work.

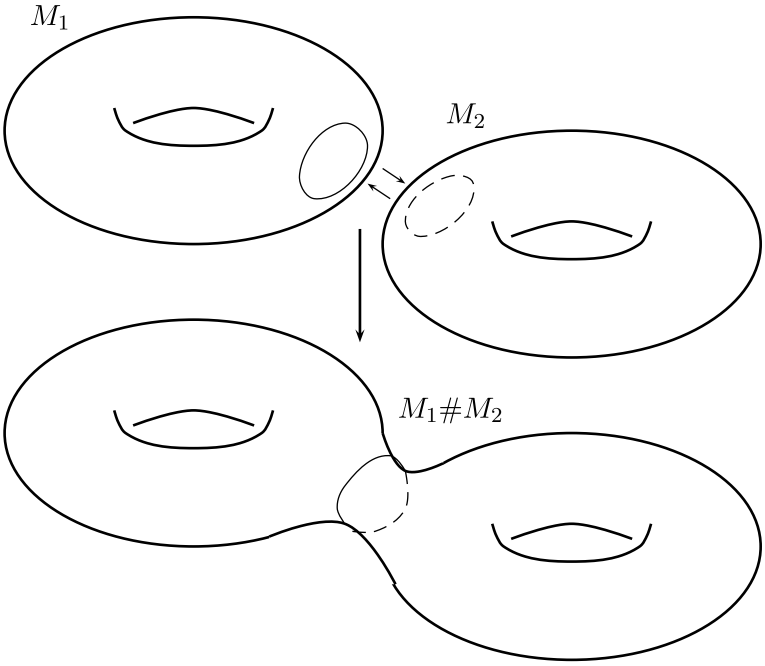


Figure 9.6

Riemann identified a topological property of compact Riemann surfaces called the **genus** that distinguishes some surfaces from others. This concept is hard to define but quite easy to describe. The genus of the sphere is zero, and the genus of the torus is one. Given two connected 2-manifolds M_1 and M_2, we form a new 2-manifold, called the **connected sum** and denoted by $M_1\#M_2$, by removing a small open disk from each and gluing the circles bounding these disks together. (It is clear, at least visually, that the homeomorphism type of the resulting object doesn't depend on where we remove the disks. A more subtle issue is that there are essentially two different ways to do the gluing, which we might call "clockwise" and "counterclockwise." Again, it turns out not to matter, but this is a consequence of the topological classification of compact surfaces described in the next section.) It is visually clear that the connected sum operation is associative and commutative. A surface of genus g is what you get by taking the connected sum of g copies of the torus.

If we cut a torus along a loop that wraps around it once, we obtain a tube. We can then cut along a line traversing the length of the tube, obtaining a square. (In general we can cut the surface of genus g along $2g$ circles or line segments connecting boundary points without disconnecting it.) Since the sphere has genus 0, the natural guess is that there is no way to remove a circle without disconnecting it, and it certainly *looks* like this is the case. This property of the sphere is known as the **Jordan curve theorem** after Camille Jordan (1838-1922) who stated it in a famous textbook in 1887, but Jordan's proof was completely wrong, and the first correct proof was given by Oswald Veblen (1880-1960) in 1905. While feasible in principle, using the Jordan curve theorem to prove that the sphere and the torus are not homeomorphic is looking rather difficult (to say the least) in practice. Instead, we're going to pursue a somewhat different approach to these issues pioneered by Poincaré.

The following notion is one of the most important concepts of topology. A **homotopy** is a continuous function

$$h : X \times [0,1] \to Y$$

where X and Y are topological spaces. Intuitively we think of deforming a function continuously over a unit interval of time, and the map

$$h(\cdot, t) : X \to Y$$

"at time t" is usually denoted by h_t. Two continuous functions $f, g : X \to Y$ are said to be **homotopic**, and we write $f \simeq g$, if there exists such an h with $h_0 = f$ and $h_1 = g$.

As usual the first order of business is to check that 'is homotopic to' is an equivalence relation. It is reflexive because for any f the "constant homotopy" $(x,t) \mapsto f(x)$ shows that $f \simeq f$, and it is symmetric because if h is a homotopy with $h_0 = f$ and $h_1 = g$, then $j : (x,t) \mapsto h(x, 1-t)$ has $j_0 = g$ and $j_1 = f$, so $g \simeq f$ whenever $f \simeq g$. To verify transitivity suppose that $e \simeq f$ and $f \simeq g$ by virtue of homotopies h and j with $e = h_0$, $h_1 = f = j_0$, and $j_1 = g$. The homotopy

$$(x,t) \mapsto \begin{cases} h(x,2t), & 0 \le t \le 1/2, \\ j(x,2t-1), & 1/2 \le t \le 1, \end{cases}$$

deforms e into f between time 0 and time 1/2, then deforms f into g between time 1/2 and time 1, showing that $e \simeq g$.

The **homotopy class** of a map f is its equivalence class, i.e., the set of all maps that are homotopic to it. Although the most important thing at the outset is to simply imagine a movie in which f deforms into g, you should also be aware that if f and g are *not* homotopic, then they differ in some sense that is *qualitative*. The space of homotopy classes of maps from X to Y can have quite interesting and useful properties and structure.

There is a simple but fundamental point that should be established right away, namely that the operation of passing from a function to its homotopy class commutes with composition. If $h : X \times [0,1] \to Y$ and $j : Y \times [0,1] \to Z$ are homotopies, then

$$(x,t) \mapsto j_t(h_t(x)) = j(h(x,t),t)$$

is a homotopy that shows that $j_0 \circ h_0 \simeq j_1 \circ h_1$. Any continuous function $g : Y \to Z$ induces a function $f \mapsto g \circ f$ taking continuous functions from X to Y to continuous functions from X to Z, so any homotopy class of functions from Y to Z induces a function taking homotopy classes of maps from X to Y to homotopy classes of maps from X to Z. Similarly, any homotopy class of maps from X to Y induces a function from the homotopy classes of maps from Y to Z to the homotopy classes of maps from X to Z.

A **closed path** or **loop** in a topological space Y is a continuous function

$$\ell : S^1 \to Y.$$

The space Y is **simply connected** if it is path connected and every loop in Y is homotopic to a constant path. That is, every loop in Y can be continuously deformed to a "path" that simply stays put at a certain point. Simple connectedness is a topological property that depends only on a space's homeomorphism type. To see this suppose that $j : Y \to \tilde{Y}$ is a homeomorphism. Then the maps

$$\ell \mapsto j \circ \ell \quad \text{and} \quad \tilde{\ell} \mapsto j^{-1} \circ \tilde{\ell}$$

induce inverse bijections between the set of loops in Y and the set of loops in $\tilde{Y}$. Since passage to homotopy classes commutes with composition, if ℓ

and ℓ' are homotopic, then so are $j \circ \ell$ and $j \circ \ell'$, so j induces a bijection between the set of homotopy classes of loops in Y and the set of homotopy classes of loops in $\tilde{Y}$. Consequently Y is simply connected if and only if $\tilde{Y}$ is. To show that the sphere and the torus are not homeomorphic it suffices to show that: a) the sphere is simply connected; b) the torus isn't.

We'll deal with the sphere first, applying a simple general principle: if Y is a topological space, $A \subset Y$ is simply connected, and every loop $\ell : S^1 \to Y$ is homotopic to a loop $\ell' : S^1 \to A$, then Y is simply connected, because there is a homotopy that deforms ℓ to ℓ' over the interval $[0, \frac{1}{2}]$, then deforms ℓ' to a constant loop over the interval $[\frac{1}{2}, 1]$. Recall that S^2 has a real analytic atlas consisting of the coordinate charts $\varphi_N : U_N \to \mathbb{R}^2$ and $\varphi_S : U_S \to \mathbb{R}^2$ where

$$U_N := S^2 \setminus \{(1,0,0)\} \quad \text{and} \quad U_S := S^2 \setminus \{(-1,0,0)\}.$$

Since φ_S is a homeomorphism, to show that S^2 is simply connected it suffices to show that any $\ell : S^1 \to S^2$ is homotopic to some $\tilde{\ell} : S^1 \to U_S$, and that $\mathbb{R}^2$ is simply connected.

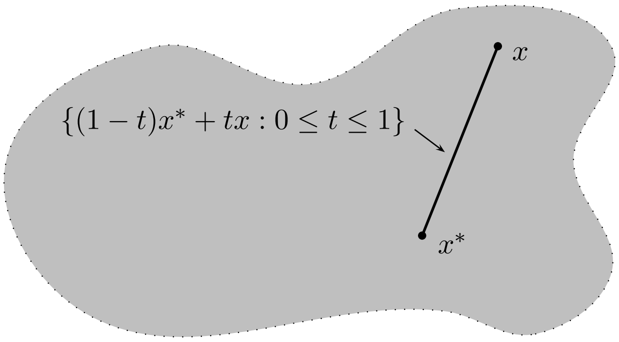


Figure 9.7

It is very easy to give a direct proof that $\mathbb{R}^2$ is simply connected, but there are some concepts and terminology here that everyone should know, so we'll drag it out a bit. A space Y is **contractible** if there is a homotopy $H : Y \times [0,1] \to Y$ between Id_Y and a constant function. Such a homotopy is called a **contraction**. A contractible space Y is simply connected because for any loop $\ell : S^1 \to Y$ there is the homotopy

$$h : (s,t) \mapsto H(\ell(s), t)$$

with $h_0 = \ell$ and h_1 a constant loop. For example, a set $S \subset \mathbb{R}^n$ is **star-shaped** (Figure 9.7) at a point $x^* \in S$ if, for each $x \in S$, the line segment

$\{\,(1-t)x + tx^* : 0 \le t \le 1\,\}$ is contained in S, in which case there is the contraction

$$H(x,t) := (1-t)x + tx^*.$$

Of course a set is convex if and only if it is star-shaped at each of its points, and $\mathbb{R}^n$ is convex, so it is simply connected.

Fix a loop $\ell : S^1 \to S^2$. To prove that S^2 is simply connected it suffices to show that ℓ is homotopic to a loop whose image doesn't contain $(-1,0,0)$, which seems quite obvious, but proving it formally involves some work. (There are methods for constructing "space filling curves" that can be used to show there are continuous surjections from S^1 to S^2, so it's not automatically the case that ℓ "misses" some point of the range.) It's actually rather common in topological reasoning that the interesting conceptual material can be expressed simply, but "simple" concrete constructions like our deformation of ℓ take a lot of space to explain, or involve a certain amount of complexity, even though it's obvious that such a construction is possible. (I'm afraid there will be other examples below.)

Of course the distance between $(-1,0,0)$ and $(1,0,0)$ is 2. Since S^1 is compact, ℓ is uniformly continuous (Proposition 9.3) so there is $\delta < 0$ such that the distance between $\ell(x)$ and $\ell(x')$ is less than 2 whenever $\|x'-x\| < \delta$. To get a concrete description of S^1 we treat it as the the image of the function

$$c : \theta \mapsto (\cos\theta, \sin\theta).$$

The domain of c is not compact, but since c is "periodic" it is not hard to show that it must be uniformly continuous because its restriction to a sufficiently large compact interval is uniformly continuous. Therefore there is $\gamma > 0$ such that $\|c(\theta') - c(\theta)\| < \delta$ whenever $|\theta' - \theta| < \gamma$. Choose an integer k with $2\pi/k < \gamma$, choose some $\theta_0 \in \mathbb{R}$ arbitrarily, and for $i = 1, \ldots, k$ let $\theta_i := \theta_0 + \frac{2\pi i}{k}$ and $A_i := c([\theta_{i-1}, \theta_i])$.

Consider a particular i, and suppose that $\ell(z) \neq (-1,0,0)$ for all $z \in A_i$. (Otherwise $(1,0,0) \notin \ell(A_i)$, and that case is handled symmetrically.) It is easy to construct a homotopy $h : A_i \times [0,1] \to \mathbb{R}^2$ that has $h_0 = \varphi_S \circ \ell|_{A_i}$ and $h_1(z) \neq (0,0)$ for all z in the interior of A_i, and that "holds endpoints fixed" in the sense that

$$h_t(c(\theta_{i-1})) = h_0(c(\theta_{i-1})) \quad \text{and} \quad h_t(c(\theta_i)) = h_0(c(\theta_i))$$

for all t. Let $j := \varphi_S^{-1} \circ h$. Then $j : A_i \times [0,1] \to S^2$ is a homotopy holding endpoints fixed with $j_0 = \ell|_{c(A_i)}$ and $j_1(z) \notin \{(-1,0,0),(1,0,0)\}$ for all z in the interior of A_i.

Since they hold endpoints fixed, we can combine these homotopies on the various arcs A_i to construct a homotopy between ℓ and a loop ℓ' that maps no points outside of $\{\, c(\theta_i) : 0 \le i \le k \,\})$ to $(-1, 0, 0)$ or $(1, 0, 0)$. But now observe that in the construction above, if we had known that ℓ mapped only finitely many points to $(-1, 0, 0)$ or $(1, 0, 0)$, then we could have chosen a θ_0 such that none of the points $c(\theta_i)$ are mapped to $(-1, 0, 0)$ or $(1, 0, 0)$, and in that case these points wouldn't be in the image of ℓ'. Therefore, by repeating this process, if necessary, we can deform ℓ to a loop whose image is contained in $U_N \cap U_S$. As we explained above, since $U_N \cap U_S \subset U_S$, this is enough to prove that S^2 is simply connected.

The rest of this section is devoted to showing that the torus is not simply connected. There are various ad hoc ways to do this, but our real agenda is to develop a powerful new concept, and this will involve several new definitions.

Recall that a path in a topological space X is a continuous function $p : [0, 1] \to X$. If the final endpoint $p(1)$ of a path p is the same as the initial endpoint $q(0)$ of a second path q, then we say they are **composable**. In this case their **composition** $p * q$ is the path given by following p at double the original speed, then following q at twice the pace:

$$p * q : s \mapsto \begin{cases} p(2s), & 0 \le s \le 1/2, \\ q(2s-1), & 1/2 \le s \le 1. \end{cases}$$

A **homotopy of paths holding the end points fixed** is a homotopy $h : [0, 1] \times [0, 1] \to X$ with

$$h(0, t) = h(0, 0) \quad \text{and} \quad h(1, t) = h(1, 0)$$

for all t. We will say that two paths p and q are **homotopic rel endpoints** if there is a homotopy holding endpoints fixed with $h_0 = p$ and $h_1 = q$. This is, in fact, an equivalence relation: the proof is a matter of noting that our earlier proof that 'is homotopic to' is an equivalence relation applies, with slight and obvious modifications, to homotopies that hold the end points fixed. The equivalence class of p is denoted by $[p]$.

We would like to define a composition operation on homotopy classes of composable paths by setting

$$[p] * [q] := [p * q]$$

when p and q are composable, but of course this doesn't make sense unless we can show that $[p * q]$ depends only on $[p]$ and $[q]$ and not on the particular representatives p and q. This is straightforward. Suppose that h and j are homotopies holding endpoints fixed that show, respectively, that p is homotopic rel endpoints to p' and q is homotopic rel endpoints to q'. Then h_t and j_t are composable for all t, and the homotopy

$$(s,t) \mapsto \begin{cases} h(2s,t), & 0 \le s \le 1/2, \\ j(2s-1,t), & 1/2 \le s \le 1, \end{cases}$$

shows that $p * q$ is homotopic rel endpoints to $p' * q'$.

As an operation on paths, composition is not very well behaved. Among other things, it's not even associative. But composition of homotopy classes of paths has more appealing properties. We'll illustrate this by developing several useful facts.

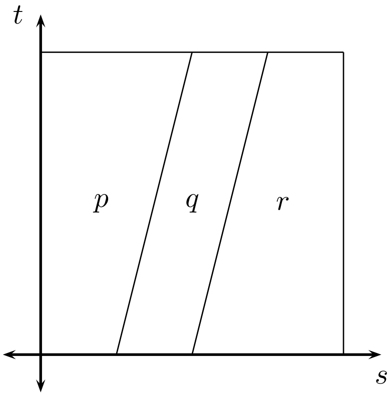


Figure 9.7

Suppose p, q, and r are paths with $p(1) = q(0)$ and $q(1) = r(0)$. Then the definition of composition of homotopy classes implies that $([p]*[q])*[r] = [(p*q)*r]$ and $[p]*([q]*[r]) = [p*(q*r)]$, and if we can show that $(p*q)*r$ and $p*(q*r)$ are homotopic rel endpoints, so that $[(p*q)*r] = [p*(q*r)]$, then it will follow that composition of homotopy classes is associative. The function

$$(s,t) \mapsto \begin{cases} p\big(4s/(t+1)\big), & 0 \le s \le (t+1)/4, \\ q\big(4s-t-1\big), & (t+1)/4 \le s \le (t+2)/4, \\ r\big((4s-t-2)/(2-t)\big), & (t+2)/4 \le s \le 1, \end{cases}$$

is an explicit homotopy with endpoints fixed that demonstrates this, thereby fulfilling the author's obligation to be rigorous and all that, but I'd like to suggest that you not study it. Instead, think about why the assertion should be true, and consider Figure 9.7. If you think that the basic idea is clear and that, with a little work, you could come up with a homotopy that would do the job, feel free to spare yourself the tedious details of my construction.

For any $x \in X$ let ω_x be the constant path that maps every element of $[0,1]$ to x. Then $\omega_{p(0)} * p$ and $p * \omega_{p(1)}$ are homotopic rel endpoints to p, so that

$$[\omega_{p(0)}] * [p] = [p] = [p] * [\omega_{p(1)}].$$

The homotopies that prove this are

$$(s,t) \mapsto \begin{cases} x_{p(0)}, & 0 \le s \le (1-t)/2, \\ p\big((2s+t-1)/(t+1)\big), & (1-t)/2 \le s \le 1, \end{cases}$$

and

$$(s,t) \mapsto \begin{cases} p\big(s/(t+1)\big), & 0 \le s \le (1+t)/2, \\ p(1), & (1+t)/2 \le s \le 1, \end{cases}$$

but again a visual understanding is primary, with the algebra being a rendering of that.

For any path p let $p^- : s \mapsto p(1-s)$ be the path that traverses the route taken by p in the opposite direction. Then $p * p^-$ is homotopic rel endpoints to $\omega_{p(0)}$, while $p^- * p$ homotopic rel endpoints to $\omega_{p(1)}$, so that

$$[p] * [p^-] = [\omega_{p(0)}] \quad \text{and} \quad [p^-] * [p] = [\omega_{p(1)}].$$

Once again, for the sake of completeness, we give an algebraic homotopy establishing the first equation (since $(p^-)^- = p$, the second equation follows from the first) expecting you to pass over it lightly if the idea seems clear:

$$(s,t) \mapsto \begin{cases} p\big(2s(1-t)\big), & 0 \le s \le 1/2, \\ p^-\big(1-2(1-s)(1-t)\big), & 1/2 \le s \le 1. \end{cases}$$

Composition of homotopy classes becomes even better behaved if we restrict to those paths that begin and end at some particular point, because any two such paths are composable. A **pointed space** is a pair (X, x_0) in which X is a topological space and x_0 is an element of X called the **base point**. In this context we'll think of a **loop based at** x_0 as a path γ in X whose endpoints are both x_0: $\gamma(0) = x_0 = \gamma(1)$. Reviewing our results above, we see that composition of homotopy classes of loops based at x_0 is associative, that $[\omega_{x_0}]$ is a two sided identity for this operation, and that for any loop based at x_0, say γ, $[\gamma^-]$ is a two sided inverse of $[\gamma]$.

We have defined a group, called the **fundamental group** of (X, x_0) and denoted by $\pi_1(X, x_0)$, consisting of homotopy classes of loops based at x_0 with composition as the group operation.

To what extent does $\pi_1(X, x_0)$ depend on the choice of x_0? The definition refers only to objects that lie entirely in the path component of x_0. (Recall that this is the set of all points $x_1 \in X$ such that there is a continuous $p : [0, 1] \to X$ with $p(0) = x_0$ and $p(1) = x_1$.) For this reason there is not much sense in applying this concept to pointed spaces that are not path connected.

So suppose that X is path connected, x_1 is another point in X, and p is a path in X with $p(0) = x_0$ and $p(1) = x_1$. There is a function

$$\iota_p : \pi_1(X, x_0) \to \pi_1(X, x_1) \quad \text{given by} \quad \iota_p([\gamma]) := [p * \gamma * p^-].$$

If γ and η are any two loops with base point x_0, then the difference between $p * (\gamma * \eta) * p^-$ and $(p * \gamma * p^-) * (p * \eta * p^-)$ (aside from details of timing) is that the latter path has an extra trip from x_1 to x_0 and back in the middle, and this can be deformed to a constant loop. More formally, the various results above allow the calculation

$$\begin{aligned}[(p * \gamma * p^-) * (p * \eta * p^-)] &= [p * \gamma * (p^- * p) * \eta * p^-] \\ &= [p * \gamma * \omega_{x_1} * \eta * p^-] = [p * (\gamma * \eta) * p^-].\end{aligned}$$

This allows us to compute that ι_p is a homomorphism:

$$\begin{aligned}\iota_p([\gamma]) * \iota_p([\eta]) &= [p * \gamma * p^-] * [p * \eta * p^-] = [(p * \gamma * p^-) * (p * \eta * p^-)] \\ &= [p * (\gamma * \eta) * p] = \iota_p([\gamma * \eta]) = \iota_p([\gamma] * [\eta]).\end{aligned}$$

(Here the first and fourth equality are from the definition of ι_p, and the second and last are from the definition of composition of homotopy classes.) In fact ι_p is an isomorphism:

$$\begin{aligned}\iota_{p^-}(\iota_p([\gamma])) &= \iota_{p^-}([p * \gamma * p^-]) = [p^- * (p * \gamma * p^-) * p] \\ &= [(p^- * p) * \gamma * (p^- * p)] = [\omega_{x_0} * \gamma * \omega_{x_0}] = [\gamma].\end{aligned}$$

In sum:

Proposition 9.29. *If X is path connected and $p : [0, 1] \to X$ is continuous with $p(0) = x_0$ and $p(1) = x_1$, then*

$$\iota_p : \pi_1(X, x_0) \to \pi_1(X, x_1) \quad \textit{and} \quad \iota_{p^-} : \pi_1(X, x_1) \to \pi_1(X, x_0)$$

are inverse isomorphisms.

One interesting possibility is that $x_1 = x_0$. Then p is a loop based at x_0, and

$$\iota_p([\gamma]) = [p] * [\gamma] * [p^-] = [p] * [\gamma] * [p]^{-1} = C_{[p]}([\gamma]).$$

Since the isomorphism type of $\pi_1(X, x_0)$ doesn't depend on x_0 when X is path connected (and this is the only interesting case) we typically talk about (and think about) "the fundamental group of X" without mentioning the base point.

It's a good idea to pause briefly and contemplate the significance of the fundamental group. We have a general method for passing from a pointed space to a group. As we'll see below, if two pointed spaces are homeomorphic (in the appropriate pointed sense) then their fundamental groups are isomorphic. If we want to prove that two pointed spaces are *not* homeomorphic, it suffices to show that their fundamental groups are not isomorphic.

As a general matter, in order to prove that two mathematical objects are different one has to define some attribute that is potentially different, then "compute" the attributes of the two objects and show that they're actually different. The fundamental group seems promising insofar as it passes from a pointed space to a quite different sort of object, but this wouldn't amount to much if there weren't powerful and systematic ways to compute it. To get a better sense of this we now study its most basic properties.

First of all, the relationship between the fundamental group and simple connectedness is what one would naively expect, so if we can show that the fundamental group of the torus is not trivial, it will follow that the torus is not simply connected.

Proposition 9.30. *Let X be a path connected space. Then X is simply connected if and only if, for any $x_0 \in X$, $\pi_1(X, x_0)$ is the trivial group with one element.*

Proof. First suppose that the fundamental group is trivial, and let a loop $\ell : S^1 \to X$ be given. Since $\pi_1(X, x_0)$ is trivial and isomorphic to $\pi_1(X, \ell(1,0))$, the latter group is trivial. (We take $(1,0)$ to be the base point of S^1.) Therefore ℓ is homotopic (by a homotopy that holds endpoints fixed, not that that matters) to $\omega_{\ell(1,0)}$. Thus X is simply connected.

Now suppose that X is simply connected. Let γ be a loop based at x_0. Since X is simply connected there is a homotopy $h : [0,1] \times [0,1] \to X$ with $h_0 = \gamma$, $h_t(0) = h_t(1)$ for all t, and h_1 a constant function. Let p be the path $p(t) := h_0(t) = h_1(t)$, and let j be the homotopy

$$j(s,t) := \begin{cases} p\big(2s\big), & 0 \le s \le \frac{1}{2}t, \\ h\big((s - \frac{1}{2}t)/(1-t), t\big), & \frac{1}{2}t \le s \le 1 - \frac{1}{2}t, \\ p\big(2(1-s)\big), & 1 - \frac{1}{2}t \le s \le 1. \end{cases}$$

(Even though the function $(s,t) \mapsto (s - \frac{1}{2}t)/(1-t)$ is discontinuous at $(\frac{1}{2}, 1)$, j is continuous because h_1 is a constant function. But, as usual, it is better to convince yourself that the result is correct without studying this construction.) This is a homotopy holding endpoints fixed between $h_0 = \gamma$ and $p * p^-$. Thus $[\gamma] = [p * p^-] = [\omega_{x_0}]$, which shows that $\pi_1(X, x_0)$ is trivial. □

A **pointed map** from (X, x_0) to another pointed space (Y, y_0) is a continuous function $f : X \to Y$ with $f(x_0) = y_0$. It will come as no surprise that there is a category of pointed spaces and pointed maps, as I am sure you can verify for yourself. If f is a pointed map from (X, x_0) to (Y, y_0) and γ is a loop in X based at x_0, then $f \circ \gamma$ is a loop in Y based at y_0, and if $h : [0,1] \times [0,1] \to X$ is a homotopy holding endpoints fixed that shows that γ and η are homotopic rel endpoints, then $f \circ h$ is a homotopy holding endpoints fixed that shows that $f \circ \gamma$ and $f \circ \eta$ are homotopic rel endpoints. This means that there is a well defined map

$$\pi_1(f) : [\gamma] \mapsto [f \circ \gamma].$$

The identity $f \circ (\gamma * \eta) = (f \circ \gamma) * (f \circ \eta)$ is a straightforward consequence of the definitions, so

$$\pi_1(f)([\gamma * \eta]) = [f \circ (\gamma * \eta)] = [f \circ \gamma] * [f \circ \eta] = \pi_1(f)([\gamma]) * \pi_1(f)([\eta]),$$

and consequently $\pi_1(f) : \pi_1(X, x_0) \to \pi_1(Y, y_0)$ is a homomorphism.

We now come to perhaps the most important property of the fundamental group:

Theorem 9.31. *π_1 is a functor from the category of pointed spaces and pointed maps to the category of groups and homomorphisms.*

Proof. For any pointed space (X, x_0) we have $\pi_1(\mathrm{Id}_{(X,x_0)}) = \mathrm{Id}_{\pi_1(X,x_0)}$ because if γ is a loop based at x_0, then

$$\pi_1(\mathrm{Id}_{(X,x_0)})([\gamma]) = [\mathrm{Id}_X \circ \gamma] = [\gamma] = \mathrm{Id}_{\pi_1(X,x_0)}([\gamma]).$$

If $f : (X, x_0) \to (Y, y_0)$ and $g : (Y, y_0) \to (Z, z_0)$ are pointed maps, then $\pi_1(g \circ f) = \pi_1(g) \circ \pi_1(f)$ because

$$\pi_1(g \circ f)([\gamma]) = [g \circ f \circ \gamma] = \pi_1(g)([f \circ \gamma]) = \pi_1(g)(\pi_1(f)([\gamma])).$$

□

This result can be used in a variety of ways to extract information about the fundamental group of one space if you know something about another space's fundamental group, and the particular example considered here is, in this respect, fairly typical. We think of the torus as the cartesian product $S^1 \times S^1$ of two circles. Let $i : S^1 \to S^1 \times S^1$ and $r : S^1 \times S^1 \to S^1$ be the maps

$$i(p) := (p, (1, 0)) \quad \text{and} \quad r(p, q) := p.$$

Then $r \circ i = \mathrm{Id}_{S^1}$, so

$$\pi_1(r) \circ \pi_1(i) = \pi_1(r \circ i) = \pi_1(\mathrm{Id}_{S^1}) = \mathrm{Id}_{\pi_1(S^1)}.$$

(Here we are letting the fact that the choice of base point doesn't matter creep into the notation. I expect that you can see how to make everything kosher.) The point is that if the fundamental group of the torus was trivial, the image of $\pi_1(r)$ would also be trivial, and this is impossible if the fundamental group of S^1 is not trivial. That is, if we can show that the fundamental group of S^1 is not trivial, then it will follow that the fundamental group of the torus is not trivial.

We'll need a couple more concepts. Let X and $\tilde{X}$ be topological spaces, and let $c : \tilde{X} \to X$ be a continuous function. An open set $U \subset X$ is **c-small** if $c^{-1}(U)$ is a disjoint union of copies of U:

$$c^{-1}(U) = \bigcup_{\alpha \in A} \tilde{U}_\alpha$$

where A is an index set, the various U_α are open and pairwise disjoint, and each $c|_{\tilde{U}_\alpha}$ is a homeomorphism between $\tilde{U}_\alpha$ and U. The map c is a **covering space** for X if the c-small open sets cover X, so each $x \in X$ has a c-small open neighborhood. An obvious and pertinent example is the map $\theta \mapsto (\cos\theta, \sin\theta)$ from $\mathbb{R}$ to S^1, and it is easy to give many others, e.g., the "double covering" $z \mapsto z^2$ that maps $\mathbb{C}^*$ onto itself.

Even though it plays no role in what follows, it is still worth mentioning that if X is a smooth (C^r, real analytic, or holomorphic) manifold, then there is an induced differential structure that makes $\tilde{X}$ a smooth manifold.

Let $\{\varphi_\beta : U_\beta \to V_\beta\}_{\beta \in B}$ be a smooth atlas for X. The collection $\{\varphi_\beta|_{U_\beta \cap W_\gamma}\}$ is a smooth atlas for X whenever $\{W_\gamma\}_{\gamma \in C}$ is an open cover of X, so X has a smooth atlas whose domains are all c-small, and consequently we may assume that each U_β is c-small. Whenever $\tilde{U}_{\beta,\alpha}$ is one of the copies of U_β in $c^{-1}(U_\beta)$ let

$$\tilde{\varphi}_{\beta,\alpha} := \varphi_\beta \circ c|_{\tilde{U}_{\beta,\alpha}} : \tilde{U}_{\beta,\alpha} \to V_\beta.$$

Then the collection of all such maps is a smooth atlas for $\tilde{X}$ because $\tilde{\varphi}_{\beta',\alpha'} \circ \tilde{\varphi}_{\beta,\alpha}^{-1}$ always agrees with $\varphi_{\beta'} \circ \varphi_\beta^{-1}$ on its domain of definition. This idea is particularly important in the theory of Riemann surfaces.

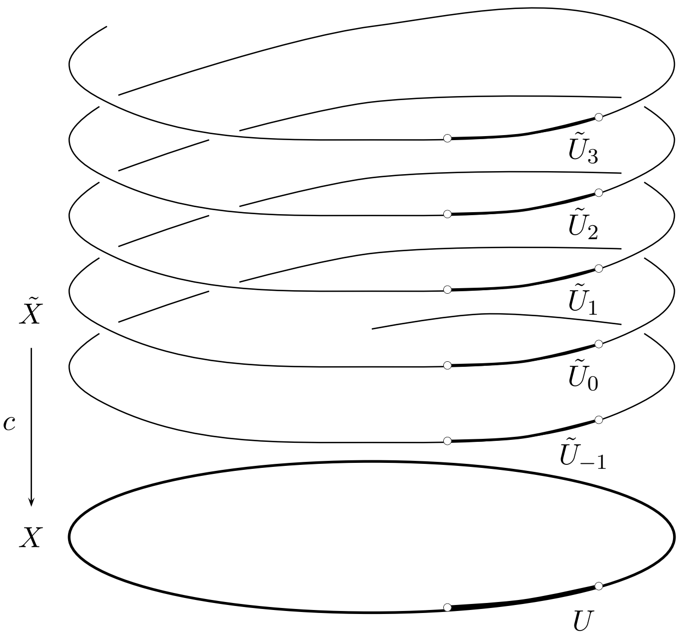


Figure 9.8

And its significance is quite a bit more general: for a wide variety of structures on X "lifting" can be used to induce a structure on $\tilde{X}$. The example of particular interest to us is as follows. A **lift** of a map $f : Y \to X$ is a continuous function $\tilde{f} : Y \to \tilde{X}$ such that $c \circ \tilde{f} = f$. Described intuitively, the basic idea of the next result is as follows. Suppose that $p : [0,1] \to X$ is a path, and that $\tilde{x}_0 \in c^{-1}(0)$ is given. We would like to show that there is a unique lift $\tilde{p} : [0,1] \to \tilde{X}$ with $\tilde{p}(0) = \tilde{x}_0$. Supposing that we have already shown that $p|_{[0,s]}$ has a unique lift $\tilde{p} : [0,s] \to \tilde{X}$, let U be a c-small set containing $p(s)$, and let $\tilde{U}_\alpha \subset c^{-1}(U)$ be the copy of U containing $\tilde{p}(s)$. Then for some $\varepsilon > 0$ we can use $(c|_{\tilde{U}_\alpha})^{-1} \circ p$ to extend $\tilde{p}$ to $[0, s+\varepsilon]$, and this extension is unique because continuity prevents an extension from jumping from $\tilde{U}_\alpha$ to some other $\tilde{U}_{\alpha'}$.

Theorem 9.32 (Homotopy Lifting Property)**.** *If $c : \tilde{X} \to X$ is a covering space, $h : Y \times [0,1] \to X$ is a homotopy, and $\tilde{h}_0 : Y \to \tilde{X}$ is a lift of h_0, then there is a unique lift $\tilde{h} : Y \times [0,1] \to \tilde{X}$ that extends $\tilde{h}_0$.*

In a sense we will prove the homotopy lifting property three times, preparing the general argument by passing through two special cases that, together, amount to a more formal and precise rendering of the intuition described above.

Lemma 9.33. *If $c : \tilde{X} \to X$ is a covering space, $p : [s_0, s_1] \to U$ is a path where $U \subset X$ is c-small, and $\tilde{x}_0 \in p^{-1}(s_0)$, then there is a unique lift $\tilde{p} : [s_0, s_1] \to \tilde{X}$ with $\tilde{p}(s_0) = \tilde{x}_0$.*

Proof. Suppose that $c^{-1}(U) = \{\tilde{U}_\alpha\}_{\alpha \in A}$, as per the definition of a covering space, and that $\tilde{x}_0 \in \tilde{U}_\beta$. Then $(c|_{\tilde{U}_\beta})^{-1} \circ p$ is a lift of $p|_{[s_0,s_1]}$ that maps s_0 to $\tilde{x}_\beta$. If $\tilde{p}$ is any lift of $p|_{[s_0,s_1]}$, then each of the sets $\tilde{p}^{-1}(\tilde{U}_\alpha)$ is an open subset of $[s_0, s_1]$, but $[s_0, s_1]$ is connected, so only one of them can be nonempty. Therefore $(c|_{\tilde{U}_\beta})^{-1} \circ p$ is the only lift that maps s_0 to $\tilde{x}_\beta$. □

To avoid repetition we give a separate a technical fact that will be applied in each of the following two proofs.

Lemma 9.34. *Under the hypotheses of Theorem 9.32, for any $y \in Y$ there is an integer K, c-small open sets $U_1, \ldots, U_K \subset X$, an open set $V \subset Y$ containing y, and numbers $0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = 1$, such that*

$$h(V \times [t_{k-1}, t_k]) \subset U_k \quad \textit{for each } k = 1, \ldots, K.$$

Proof. For each $t \in [0,1]$ let U_t be a c-small set containing $h(y,t)$. Since h is continuous, $Z_t := h^{-1}(U_t)$ is open. The definition of the product topology implies that there an open set $V_t \subset Y$ and an open (in the relative topology of $[0,1]$) interval I_t such that $(y,t) \in V_t \times I_t \subset Z_t$. Since $\{y\} \times [0,1]$ is compact, the open cover $\{\, V_t \times I_t : t \in [0,1] \,\}$ has a finite subcover $V_1 \times I_1, \ldots, V_K \times I_K$. Let $V := V_1 \cap \ldots \cap V_K$. Then $V \times I_1, \ldots, V \times I_K$ is an open cover of $\{y\} \times [0,1]$, and for each $k = 1, \ldots, K$ there is a c-small set U_k with $V \times I_k \subset h^{-1}(U_k)$.

By reindexing we may achieve $0 \in I_1$ and make it the case that I_2 contains the least upper bound of I_1, I_3 contains the least upper bound of

I_2, and so forth, terminating the process (and throwing out any redundant elements of the cover) as soon as $1 \in I_K$. Choose $t_1 \in I_1 \cap I_2$, choose $t_2 \in I_2 \cap I_3$ with $t_2 > t_1$, choose $t_3 \in I_3 \cap I_4$ with $t_3 > t_2$, and so forth until t_{K-1} has been chosen. □

Lemma 9.35. *If $c : \tilde{X} \to X$ is a covering space, $p : [0,1] \to X$ is a path, and $\tilde{x}_0 \in p^{-1}(0)$, then there is a unique lift $\tilde{p} : [0,1] \to \tilde{X}$ with $\tilde{p}(0) = \tilde{x}_0$.*

Proof. Thinking of $[0,1]$ as a cartesian product $\{y\} \times [0,1]$, let $U_1, \ldots, U_K$, and $t_0, \ldots, t_K$ be as above. Applying Lemma 9.33 repeatedly shows that there is a unique function $\tilde{p} : [0,1] \to \tilde{X}$ such that $\tilde{p}(0) = \tilde{x}_0$ and each $\tilde{p}|_{[t_{k-1},t_k]}$ is a lift of $p|_{[t_{k-1},t_k]}$. (More concretely, once $\tilde{p}|_{[t_0,t_1]}$ has been determined we know what $\tilde{p}(t_1)$ is, there is a unique $\tilde{p}|_{[t_1,t_2]}$ consistent with this datum that determines $\tilde{p}(t_2)$, and so forth.) Since the sets $[t_{k-1}, t_k]$ constitute a finite closed cover of $[0,1]$, Proposition 3.26 implies that $\tilde{p}$ is continuous, hence a lift. □

In the situation described in the hypotheses of Theorem 9.32, the last result can be applied to each point in Y, so there is a unique function $\tilde{h} : Y \times [0,1] \to \tilde{X}$ extending $\tilde{h}_0$ such that for each $y \in Y$, $\tilde{h}|_{\{y\}\times[0,1]}$ is a lift of $h|_{\{y\}\times[0,1]}$. Any lift has to agree with $\tilde{h}$ on each $\{y\} \times [0,1]$, so there can't be another lift. But we still need to show that $\tilde{h}$ is a lift, and the one remaining piece of that is showing that $\tilde{h}$ is continuous. Since continuity is a local property, it suffices to show that each $y \in Y$ has an open neighborhood W such that $\tilde{h}|_{W\times[0,1]}$ is continuous, and we will do this by showing that for some such W, $h|_{W\times[0,1]}$ has a lift: the restriction of $\tilde{h}$ to $W \times [0,1]$ must agree with this lift, and consequently must be continuous.

Proof of Theorem 9.32. Fixing a point $y \in Y$, let the c-small sets $U_1, \ldots, U_K$, the open set V, and the numbers $t_0, \ldots, t_K$ be as in Lemma 9.34. Let $W_0 := V$, and define $\tilde{j}_0 : W_0 \times [0, t_0] \to \tilde{X}$ by setting

$$\tilde{j}_0(y', t') := \tilde{h}_0(y').$$

(Of course $t_0 = 0$, so this seems a bit silly, but it conforms with the general pattern below.) Of course $\tilde{j}_0$ is a lift of $h|_{W_0\times[0,t_0]}$.

Proceeding inductively, suppose that for some $k = 1, \ldots, K$ we have already found an open set $W_{k-1} \subset Y$ containing y and a lift $\tilde{j}_{k-1}$ of $h|_{W_{k-1}\times[0,t_{k-1}]}$ such that $\tilde{j}_{k-1}(y', 0) = \tilde{h}_0(y')$ for all $y' \in W_{k-1}$. Let $c^{-1}(U_k) = \{\tilde{U}_{k,\alpha}\}_{\alpha\in A}$, as per the definition of a covering space, and let $\tilde{U}_{k,\beta}$ be the copy of U_k that contains $\tilde{j}_{k-1}(y,t_{k-1})$. It may not be the case that $\tilde{j}_{k-1}(y',t_{k-1}) \in \tilde{U}_{k,\beta}$ for all $y' \in W_{k-1}$, but we can set

$$W_k := \{\, y' \in W_{k-1} : \tilde{j}_{k-1}(y',t_{k-1}) \in \tilde{U}_{k,\beta} \,\}.$$

Then W_k contains y, and it's open because $\tilde{j}_{k-1}$ is continuous. Define $\tilde{j}_k : W_k \times [0,t_k] \to \tilde{X}$ by setting

$$\tilde{j}_k(y',t') := \begin{cases} \tilde{j}_{k-1}(y',t'), & 0 \le t' \le t_{k-1}, \\ (c|_{\tilde{U}_{k,\beta}})^{-1}(h(y',t')), & t_{k-1} \le t' \le t_k. \end{cases}$$

Then $\tilde{j}_k$ is well defined, and it is continuous on $W_k \times [0,t_{k-1}]$ and on $W_k \times [t_{k-1},t_k]$. These sets constitute a finite closed cover of $W_k \times [0,t_k]$, so Proposition 3.26 implies that $\tilde{j}_k$ is continuous and thus a lift of $h|_{W_k \times [0,t_k]}$. By induction this construction is possible for all k. Assuming this has been done, let $W := W_K$ and $\tilde{j} := \tilde{j}_K$. The uniqueness clause of the last result implies that $\tilde{j} = \tilde{h}|_{W \times [0,1]}$, so $\tilde{h}$ is continuous on $W \times [0,1]$. Since y was an arbitrary element of Y, this completes the proof. □

Here is a useful and surprisingly simple consequence of the homotopy lifting property.

Lemma 9.36. *If $c : \tilde{X} \to X$ is a covering space, $h : [0,1] \times [0,1] \to X$ is a homotopy that holds endpoints fixed, and $\tilde{h}$ is a lift of h, then $\tilde{h}$ holds endpoints fixed.*

Proof. The homotopy lifting property implies that there are unique lifts of $h(0,\cdot)$ and $h(1,\cdot)$ mapping 0 to $\tilde{h}(0,0)$ and $\tilde{h}(1,0)$ respectively. Of course $\tilde{h}(0,\cdot)$ and $\tilde{h}(1,\cdot)$ are such lifts, but so are the constant functions $t \mapsto \tilde{h}(0,0)$ and $t \mapsto \tilde{h}(1,0)$. □

We will now apply what we have learned to the covering space $c : \mathbb{R} \to S^1$ where $c(\theta) := (\cos\theta, \sin\theta)$. Let $\gamma : [0,1] \to S^1$ is a loop based at $(1,0)$. The homotopy lifting property gives a unique lift $\tilde{\gamma} : [0,1] \to \mathbb{R}$ with $\tilde{\gamma}(0) = 0$, and since $\gamma(1) = (1,0)$ we have $c(\tilde{\gamma}(1)) = (1,0)$ and $\tilde{\gamma}(1) = 2\pi n_\gamma$ for some integer n_γ. We claim that n_γ depends only on the associated element $[\gamma]$ of $\pi_1(S^1,(1,0))$. To see this observe that if $h : [0,1] \times [0,1] \to S^1$ is a homotopy of loops based at $(1,0)$ that holds endpoints fixed with $h_0 = \gamma$, and $\tilde{h} : [0,1] \times [0,1] \to \mathbb{R}$ is a lift, then $\tilde{h}$ holds endpoints fixed, so

$$2\pi n_{h_0} = \tilde{h}_0(1) = \tilde{h}_1(1) = 2\pi n_{h_1}.$$

Thus there is a well defined function $[\gamma] \mapsto n_\gamma$ from $\pi_1(S^1,(1,0))$ to $\mathbb{Z}$. For each integer N the function $\gamma_N : [0,1] \to S^1$ taking t to $c(2\pi N t)$ is a loop in S^1 based at $(1,0)$ with $n_{\gamma_N} = N$, so this function is surjective. *We have now produced a function with domain $\pi_1(S^1,(1,0))$ whose image contains more than one element of the range, thereby completing the proof that $\pi_1(S^1,(1,0))$ is nontrivial, and that consequently the torus is not simply connected and not homeomorphic to S^2.*

Of course it would make little sense to come this far without also showing that $[\gamma] \mapsto n_\gamma$ is an isomorphism between $\pi_1(S^1,(1,0))$ and $\mathbb{Z}$ when $\mathbb{Z}$ is regarded as an additive group. If γ and η are two loops in S^1 based at $(1,0)$, and $\tilde\gamma$ and $\tilde\eta$ are the lifts with $\tilde\gamma(0) = 0$ and $\tilde\eta(0) = 0$, then

$$\tilde\kappa : s \mapsto \begin{cases} \tilde\gamma(2s), & 0 \le s \le \frac{1}{2}, \\ \tilde\gamma(1) + \tilde\eta(2s-1), & \frac{1}{2} \le s \le 1, \end{cases}$$

is the lift of $\gamma * \eta$ mapping 0 to 0, and $\tilde\kappa(1) = \tilde\gamma(1) + \tilde\eta(1)$. Therefore

$$n_{\gamma*\eta} = n_\gamma + n_\eta,$$

so $[\gamma] \mapsto n_\gamma$ is a homomorphism. To establish that it's injective we need to show that its kernel is trivial. Let $\tilde\gamma$ be the lift of γ with $\tilde\gamma(0) = 0$, and suppose that $n_{[\gamma]} = 0$, so that $\tilde\gamma(1) = 0$. Then

$$h : (s,t) \mapsto c((1-t)\tilde\gamma(s))$$

is a homotopy holding endpoints fixed between $h_0 = \gamma$ and $h_1 = \omega_{(1,0)}$, so $[\gamma] = [\omega_{(1,0)}]$ is the identity element of $\pi_1(S^1,(1,0))$.

To further illustrate the power of the fundamental group we'll prove the two dimensional case of Brouwer's fixed point theorem. Let

$$D^2 = \{\, (x,y) \in \mathbb{R}^2 : x^2 + y^2 \le 1 \,\}$$

be the unit disk. Aiming at a contradiction, suppose that $f : D^2 \to D^2$ is a continuous function without a fixed point. Let $r : D^2 \to S^1$ be the function taking each p to the point[3] where the ray emanating from $f(p)$ and passing

[3] It is visually obvious that the ray in question intersects S^1 exactly once, but a formal proof allows us to review several concepts. To see that there is at least one such point apply the intermediate value theorem to the restriction of the function $t \mapsto \|p + t(p - f(p))\|$ to $[0,\infty)$. Since a quadratic equation has at most two real roots, the line ℓ containing p and $f(p)$ intersects S^1 at most twice. Since D^2 and ℓ are both convex, their intersection is convex. The continuity of the norm implies that the boundary points (in the relative topology of ℓ) of $D \cap \ell$ are in S^1, and $f(p) \in D^2 \cap \ell$, so it can't be the case that both points in $S^1 \cap \ell$ are past p on the ray emanating from $f(p)$ and passing through p.

through p intersects S^1. Clearly r is continuous[4]. If $p \in S^1$, then $r(p) = p$, so $r \circ i = \mathrm{Id}_{S^1}$ where $i : S^1 \to D^2$ is the inclusion. Consequently

$$\pi_1(r) \circ \pi_1(i) = \mathrm{Id}_{\pi_1(S^1,(1,0))},$$

but D^2 is convex, hence contractible and simple connected, with trivial fundamental group (for any base point) so this is impossible.

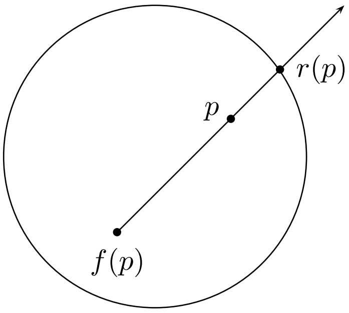


Figure 9.9

9.6 Classification of Compact Manifolds

There are a few more things to say about functors like the fundamental group and how they contribute to the study of manifolds. One can define functors π_n for $n = 2, 3, \ldots$ that associate with each pointed space (X, x_0) a group of homotopy classes of pointed maps from $(S^n, (1, 0, \ldots, 0))$ to (X, x_0). There are many other functors from the category of (unpointed) topological spaces to the categories of abelian groups and commutative rings. Of course these are useful as ways of distinguishing manifolds, but, as often happens with good ideas, the study of these functors has evolved into a subject in and of itself, with a major impact on abstract algebra and an increasing number of applications in areas other than topology.

[4] In this case the formal proof is just an annoyance. Suppose $p_n \to p$. For each n there is some $t_n \geq 0$ such that $r(p_n) = p_n + t_n(p_n - f(p_n))$, and the sequence $\{t_n\}$ is bounded because the sequence $\{\|f(p_n) - p_n\|\}$ is bounded away from 0, so (Theorem 3.44) it has a convergent subsequence. If t is the limit of a convergent subsequence, then $p + t(p - f(p)) \in S_1$, so, by the argument in the previous footnote, there is at at most one such limit point. If t is the unique limit point, then $t_n \to t$ (if t_n was outside some neighborhood of t for infinitely many n there would be a subsequence with a different limit) and $r(p_n) \to p + t(p - f(p)) = r(p)$.

The question we posed and answered in the last section is one element of the analysis leading to the classification of compact connected surfaces up to homeomorphism. Although Riemann's concept of genus clearly points toward a classification of oriented two dimensional manifolds, the first paper explicitly considering this project was an 1870 paper of Möbius. He considered only surfaces that were embedded in $\mathbb{R}^3$ (which implies that the surface is orientable) finding that any such surface is either the sphere or a direct sum of finitely many toruses. This means that Riemann's notion of genus classifies surfaces embedded in $\mathbb{R}^3$: two such surfaces are homeomorphic if and only if they have the same genus. In fact this result doesn't depend on embeddability, as we'll explain below.

The theorem classifying all compact surfaces, including those that are not orientable, was correctly stated by Walther van Dyke (1856-1934) in 1888, but he did not give a complete proof. Any connected compact surface is either the sphere, a connected sum of finitely many copies of the torus, a connected sum of the Klein bottle and finitely many copies of the torus, or the connected sum of $P^2(\mathbb{R})$ and finitely many copies of the torus. The facts that drive this result are: (a) the connected sum of two copies of $P^2(\mathbb{R})$ is the Klein bottle; (b) the connected sum of three copies of $P^2(\mathbb{R})$ is the same as the connected sum of one copy of the torus and one copy of $P^2(\mathbb{R})$. (The proofs of these facts are in some sense elementary, but nonetheless they involve sophisticated geometric imagination.) We can think of creating an oriented surface of genus g as a matter of starting with a sphere and attaching g "handles" by taking connected sums with torii. It is natural to think of also taking a connected sum with finitely many copies of $P^2(\mathbb{R})$, but the third copy of $P^2(\mathbb{R})$ fails to produce anything you can't get by attaching handles.

Once one has a complete and nonredundant list of surfaces, in order to establish the classification theorem you must prove two things: (a) any two elements of the list are not homeomorphic; (b) any connected compact surface is homeomorphic to some element of the list. The fundamental group is an adequate tool for this job if one can show that: (i) any two elements of the list have different fundamental groups; (ii) any connected compact surface has the same fundamental group as some element of the list; (iii) if two connected compact surfaces have the same fundamental group, then they are homeomorphic. All these things are true, but it would be hasty to jump to the conclusion that this was the method employed in the first rigorous proof of the classification theorem, which was given by Max Dehn (1878-1952) and Poul Heegaard (1871-1948) in 1907, under the hypothesis that the surface can be triangulated. (To tell the truth, I don't know how the Dehn-Heegaard argument works.)

A **simplex** is a point, line segment, triangle, tetrahedron, or higher dimensional analogue. A (finite) **simplicial complex** is a finite collection of simplices in some Euclidean space with the following properties:

(a) each face (including $\emptyset$) of a member of the collection is also a member of the collection;

(b) the intersection of any two simplices in the collection is a (possibly empty) face of each of them.

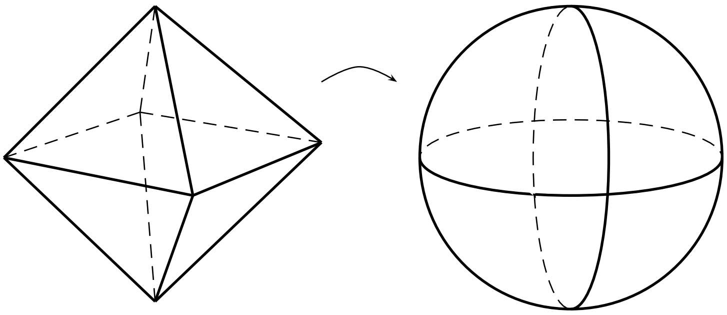


Figure 9.10

The **space** of a simplicial complex is the union of all its members. A topological space is **triangulable** if it is homeomorphic to the space of a simplicial complex. For example, the boundary of a tetrahedron, the boundary of a octahedron (Figure 9.10) and the boundary of an icosahedron are each the space of a simplicial complex that is homeomorphic to S^2. The topology of a simplicial complex is entirely determined once one specifies the list of simplices, their dimensions, and their containment relations, so (up to homeomorphism) a simplicial complex is a finite combinatorial object, which presents many advantages, and for this reason simplicial complexes are quite important in topology. Among other things, tools like the fundamental group are most easily defined and computed when the space can be triangulated. In 1925 Tibor Rado (1895-1965) proved that any compact surface is triangulable, so that the Dehn-Heegaard argument classifies all compact surfaces. Around 1940 Stewart Cairns (1904-1982) and J. H. C. Whitehead (1904-1960) proved that all compact C^1 manifolds can be triangulated, but there are topological manifolds with no triangulation.

With the topology of compact surfaces completely understood, the natural next step is to consider compact connected three dimensional manifolds.

After some early missteps Poincaré formulated what would seem to be the most basic question in the direction of a classification result: is a simply connected compact three manifold necessarily homeomorphic to S^3? Although Poincaré didn't express a strong view one way or the other, this became known as the **Poincaré Conjecture**, and over the course of the 20th century its stature increased until it became one of the most famous unsolved problems of mathematics. In particular, it was one of the Clay Mathematics Institute's seven Millenium Prize Problems.

We're going to use categories to provide a partial explanation of the difficulty. Let $\mathcal{T}$ be the category of pointed spaces and pointed maps. There is a closely related category $\hat{\mathcal{T}}$ whose objects are pointed spaces and whose morphisms are not pointed maps, but are instead homotopy classes of pointed maps. If $f : (X, x_0) \to (Y, y_0)$ is a pointed map, let $\{f\}$ be the equivalence class of f under the relation 'is homotopic to by a homotopy keeping base points fixed.' There is a binary operation on homotopy classes that we'll describe as **composition**: if $f : (X, x_0) \to (Y, y_0)$ and $g : (Y, y_0) \to (Z, z_0)$ are pointed maps, then

$$\{g\} \circ \{f\} := \{g \circ f\}.$$

As usual we have to show that this definition doesn't depend on the choices of reresentatives, i.e., $\{g \circ f\}$ is the same as $\{g' \circ f'\}$ for any $f' \in \{f\}$ and $g' \in \{g\}$. The argument is just as before except that now the homotopies $h : X \times [0, 1] \to Y$ and $j : Y \times [0, 1] \to Z$ hold the base points fixed: $h_t(x_0) = y_0$ and $j_t(y_0) = z_0$ for all t. Then $j_t(h_t(x_0)) = z_0$ for all t, so $(x, t) \mapsto j_t(h_t(x))$ is a homotopy holding the base point fixed between $j_0 \circ h_0$ and $j_1 \circ h_1$.

Composition of homotopy classes of pointed maps is associative by virtue of the associativity of composition of pointed maps:

$$(\{h\} \circ \{g\}) \circ \{f\} = \{h \circ g\} \circ \{f\} = \{(h \circ g) \circ f\} = \{h \circ (g \circ f)\}$$

$$= \{h\} \circ \{g \circ f\} = \{h\} \circ (\{g\} \circ \{f\}).$$

If $f : (X, x_0) \to (Y, y_0)$ is a pointed map, then

$$\{\mathrm{Id}_{(Y,y_0)}\} \circ \{f\} = \{\mathrm{Id}_{(Y,y_0)} \circ f\} = \{f\} = \{f \circ \mathrm{Id}_{(X,x_0)}\} = \{f\} \circ \{\mathrm{Id}_{(X,x_0)}\},$$

so the homotopy equivalence classes of the identity functions are two sided identities for composition. As promised, we have shown that $\hat{\mathcal{T}}$ has all the properties required of a category.

We now arrive at one of the most important concepts in topology. Two pointed spaces (X, x_0) and (X', x'_0) are **homotopy equivalent** if they are isomorphic in $\hat{\mathcal{T}}$. That is, there are pointed maps

$$\epsilon : (X, x_0) \to (X', x'_0) \quad \text{and} \quad \epsilon' : (X', x'_0) \to (X, x_0)$$

such that $\{\epsilon'\} \circ \{\epsilon\} = \{\mathrm{Id}_{(X,x_0)}\}$ and $\{\epsilon\} \circ \{\epsilon'\} = \{\mathrm{Id}_{(X',x'_0)}\}$, so that $\epsilon' \circ \epsilon$ and $\epsilon \circ \epsilon'$ are homotopic to $\mathrm{Id}_{(X,x_0)}$ and $\mathrm{Id}_{(X',x'_0)}$ by homotopies that keep the base points fixed. A topological concept, say a functor on topological spaces, can be used to prove that two spaces are not homeomorphic if it takes different values on the two spaces. One reason homotopy equivalence is important is that many topological concepts are "crude" measures of such differences in the sense that they take the same value on any two spaces that are homotopy equivalent. The following discussion illustrates this concretely.

There is a rather trivial functor F from $\mathcal{T}$ to $\hat{\mathcal{T}}$ that takes each pointed space to itself and each pointed map to its homotopy class. The conditions defining a functor, namely that

$$F(g \circ f) = \{g \circ f\} = \{g\} \circ \{f\} = F(g) \circ F(f)$$

and $F(\mathrm{Id}_{(X,x_0)}) = \{\mathrm{Id}_{(X,x_0)}\}$, are automatic.

There is also a covariant functor $\hat{\pi}_1$ from $\hat{\mathcal{T}}$ to the category of groups and homomophism given by setting $\hat{\pi}_1(X, x_0) := \pi_1(X, x_0)$ whenever (X, x_0) is a pointed space and setting $\hat{\pi}_1(\{f\}) := \pi_1(f)$ whenever $f : (X, x_0) \to (Y, y_0)$ is a pointed map. The definition of $\hat{\pi}_1(\{f\})$ is independent of the choice of representative f because if $f' \in \{f\}$, then $\pi_1(f') = \pi_1(f)$: concretely, if γ is a loop in X based at x_0, then $f' \circ \gamma$ is homotopic to $f \circ \gamma$, by a homotopy holding the base point fixed, so that $[f' \circ \gamma] = [f \circ \gamma]$. The functorial properties of $\hat{\pi}_1$ are automatic consequences of the definitions and the functorial properties of π_1:

$$\hat{\pi}_1(\{g\} \circ \{f\}) = \hat{\pi}_1(\{g \circ f\}) = \pi_1(g \circ f) = \pi_1(g) \circ \pi_1(f) = \hat{\pi}_1(\{g\}) \circ \hat{\pi}_1(\{f\})$$

and

$$\hat{\pi}_1(\{\mathrm{Id}_{(X,x_0)}\}) = \pi_1(\mathrm{Id}_{(X,x_0)}) = \mathrm{Id}_{\pi_1(X,x_0)} = \mathrm{Id}_{\hat{\pi}_1(X,x_0)}.$$

These definitions give $\hat{\pi}_1(F(X, x_0)) = \pi_1(X, x_0)$ for each pointed space (X, x_0) and $\hat{\pi}_1(F(f)) = \hat{\pi}_1(\{f\}) = \pi_1(f)$ for each pointed map $f : (X, x_0) \to (Y, y_0)$, so

$$\pi_1 = \hat{\pi}_1 \circ F.$$

(Recall the discussion of composition of functors in Section 6.6.) Mathematicians are inclined to think of functors like F as "forgetting" about the differences between homotopic pointed maps. The existence of a functor $\hat{\pi}_1$ satisfying this equation demonstrates that π_1 doesn't depend on the forgotten information.

Let M be a compact connected 3-manifold, and let x_0 be an arbitrary point of M. Some fairly advanced tools are involved, but it's not terribly difficult to show that if $\pi_1(M, x_0)$ is trivial, then (M, x_0) is homotopy equivalent to $(S^3, (1, 0, 0, 0))$. Therefore the Poincaré Conjecture is equivalent to the following assertion: if (M, x_0) and $(S^3, (1, 0, 0, 0))$ are homotopy equivalent, then M and S^3 are homeomorphic. For this reason a 3-manifold with base point that is homotopy equivalent to $(S^3, (1, 0, 0, 0))$ is called a **homotopy 3-sphere**, and a homotopy 3-sphere that's not homeomorphic to S^3 is called a **fake 3-sphere**. The Poincaré conjecture boils down to the assertion that there are no fake 3-spheres. Given the success that we had using the fundamental group to prove that S^2 and the torus are not homeomorphic, it might seem natural to look for another functor that necessarily has a different value for $(S^3, (1, 0, 0, 0))$ and for (M, x_0). Unfortunately there are systematic reasons for being pessimistic about this approach.

Any covariant functor K maps isomorphic objects in the domain category to isomorphic objects in the range category: if $f : X \to Y$ and $g : Y \to X$ are inverse isomorphisms, then $K(f)$ and $K(g)$ are inverse isomorphisms because

$$K(g) \circ K(f) = K(g \circ f) = K(\mathrm{Id}_X) = \mathrm{Id}_{K(X)}$$

and

$$K(f) \circ K(g) = K(f \circ g) = K(\mathrm{Id}_Y) = \mathrm{Id}_{K(Y)}.$$

We succeeded in showing that the torus and S^2 are not homeomorphic because we were able to show that π_1 maps them to nonisomorphic groups. But this argument actually shows that the sphere and the torus are not even homotopy equivalent because $\hat{\pi}_1$ would map them to isomorphic groups if they were. If a functor from $\mathcal{T}$ to the category of groups is "really" a functor on $\hat{\mathcal{T}}$, in the sense that it factors into a composition of another functor and F, then the natural strategy for using the the functor to prove that two spaces are not homeomorphic will not work if the two spaces are homotopy equivalent. Unfortunately, pretty much all the functors studied in algebraic topology are like this, or have similar factorizations in the corresponding categories of unpointed spaces. This is systematically related to the fact that they are designed to detect discrete or qualitative differences between functions, whereas two maps that are homotopic differ only in a quantitative sense.

That S^2 is the only compact simply connected surface follows from the classification: we have a nonredundant list of all compact two manifolds, and (using tools for computing π_1 that we haven't covered) one can show that S^2 is the only simply connected element of the list. If one had a similar classification of compact three manifolds, and adequate tools for computing their fundamental groups, we could resolve the Poincaré conjecture, but this is at best a distant prospect. Some new idea is needed.

For any n it makes sense to ask whether there can be a compact n-dimensional manifold that is homotopy equivalent to S^n but not homeomorphic to it, and the **Generalized Poincaré Conjecture** is the assertion that this cannot happen. (It does makes sense to ask whether an n-dimensional simply connected compact manifold is necessarily homeomorphic to S^n, but in general the answer is no.) In 1961 Stephen Smale (b. 1930) shocked everyone by proving the Generalized Poincaré Conjecture when the dimension is at least five. It may seem surprising that the situation becomes simpler in higher dimensions, but an admittedly inaccurate analogy may help. A loop of string in $\mathbb{R}^3$ can be knotted in various ways (and in fact the study of knots is an important and thriving area of topology) but there are no nontrivial knots in $\mathbb{R}^n$ for $n \geq 4$ because there are no "obstructions" to deforming a knot into the so-called unknot.

In 1982 Michael Freedman (b. 1951) proved the four dimensional version of the conjecture. Again, it may seem surprising that his methods don't work in three dimensions, but one of the main findings of low dimensional topology is that dimensions three and four are *very* different. In honor of these results Smale was awarded the Fields Medal in 1966, and Freedman received it in 1986.

In late 2002 and 2003 Grigori Perelman (b. 1966) posted 3 papers on the internet that sketched a proof of the original Poincaré Conjecture, completing a specific research program developed by Richard Hamilton (b. 1943) that was in turn based on a geometric approach to the topology of three manifolds originated by William Thurston (b. 1946). The papers were very long, with numerous new ideas. Over the next three years six mathematicians, working in three teams of two, set about filling in the details and verifying the validity of the argument. In the middle of 2006 each team posted a paper reporting its findings. These papers were 200, 327, and 493 pages long respectively. Each of the three teams attained a complete proof that followed Perelman's original outline, and although many of Perelman's arguments were expressed in extremely terse fashion, leaving numerous details to be filled in, all three teams agreed that all gaps in his argument were minor. In August 2006 he was awarded the Fields Medal, which he refused as an expression of alienation from the mathematical community.

Like Andrew Wiles' work on Fermat's Last Theorem, Perelman's proof has a degree of depth, scope, and raw length that had never been seen in a single piece of research prior to the late 20$^{\text{th}}$ century. Ultimately this is a reflection of the power and precision of the abstractions that became possible as a result of the set theory revolution. Mathematicians are now like sure footed mountain goats who can cover great distances while travelling along narrow ledges and craggy ridges at the highest levels of the subject.

Wiles and Perelman both worked in secret, and in isolation, for over half a decade, minimizing contact with other mathematicians and people outside their immediate families. Certainly one should have no confidence in one's ability to imagine what they felt at the ends of their journeys, and there is no reason to doubt that achieving the ultimate goal was in many ways quite exhilarating, but I am inclined to suspect that for each there was a moment when he walked out into what computer programmers call the Big Blue Room, took a deep breath, looked at the sky, and sadly began to come to terms with the fact that he could never return to the paradise in which he had spent the preceeding several years.

Chapter 10

More and More Math

Now vee may perhaps to begin.

—The "punch line" of Philip Roth's novel *Portnoy's Complaint*

This book has tried to explain little bits about a great many things while giving some sort of organized picture of the very beginnings of contemporary mathematics. The idea has been to stimulate your imagination while doing very little to satisfy your curiosity. If I've succeeded, you're now eager to continue, but you have to realize that you don't know very much about anything, so many subjects are simply inaccessible, while others are sufficiently self-contained in a logical sense that you could study them "in principle," but in practice they presume more mathematical maturity than you currently possess.

Whether you're aiming at graduate school or just want to take a couple more courses that sound fun, it's a good idea to maintain a balance between studying and recreational reading. To an unfortunate extent current instructional practices are based on an assumption that studying is the only way. Sadly, this is a largely self fulfilling prophecy because it leads to an enormous proliferation of textbooks, while the books intended for recreational reading are few in number and hard to learn about. It is especially difficult for those at the beginning level to know which are suitable, and of high quality. You can, of course, try to read textbooks for fun, but textbooks tend to be verbose, comprehensive (because the student wouldn't be taking the course if she didn't "need" to know all that stuff) and weighed down with the task of supplying the student with a substantial amount of work over a period of several months.

It's also a good idea to aim for a good balance between reading and working problems. This book has deemphasized problem solving in part because it's a book about concepts, not puzzles, and in part in reaction to my experience with students who *only* want to learn to do the problems that will be the basis of their grade, perhaps because they never had instructors who expected anything else. In my own study I don't do very many problems, but that's partly because I'm already pretty good at it, and I also do a fair amount of problem solving in the course of my writing and research. But a large part of the pleasure of mathematics comes from figuring things out for yourself, and you can't develop a truly firm grasp of a subject without doing some of this. If you like solving problems, by all means indulge yourself. If it's not really your cup of tea, try to solve a few more than might otherwise be your inclination.

If you're serious about this stuff, it's also a good idea to work on writing. Historically it has been difficult to give writing assignments in mathematics because the subject isn't that well suited to it, but even more fundamentally because the technology was totally inadequate. TEX and LATEX have changed all that: using online tutorials you can fairly quickly get to the point where you are producing typesetting of English prose that is as clear and beautiful as what you see in the best books. After that there is still a great deal to learn about formatting, and about typesetting mathematics, but most of it can be picked up as you go along by looking things up in references (many are available online) as the need arises.

Writing clear, aesthetically pleasing mathematics requires all the skills that constitute good prose style, and much more. An excellent way to practice is to rewrite a proof of moderate length that seems a bit foggy or confusing. Begin by typing up the proof verbatim using LATEX. Then start changing anything and everything about it you don't like, trying to highlight the key ideas, guide the reader's mind along an easy path, and generally make everything as simple and elegant as you can. Continue until you can't think of a single way to make things even slightly better. Unless the author of the proof you started with is *very* good, you'll find that the process goes on for quite a while, with improvements in one aspect revealing opportunities to improve other things, or necessitating minor adjustments here and there, and that the final result is quite different from what you started with. You'll also find that this type of exercise does at least as much as problem solving to solidify your understanding of the material.

10.1 Some Other Books

There are, of course, a huge number of math books in the world, many of which are quite good in at least some sense. One good way to get recommendations is to ask your instructors, since they'll know about your current level and your interests, and they'll have their own favorites. In order to be included in the list below, a book had to be about truly interesting mathematics, it had to be accessible to readers with very little background, and it had to *not* be a textbook.

In rough order of increasing difficulty:

- M. Aigner and G.M. Ziegler, *Proofs from the Book*, third edition, Springer-Verlag, 1992.
    - Paul Erdös (1913-1996) was one of the great mathematicians of the 20th century, and an extremely beloved figure in the mathematical community of his era. Partly as a result of persecution during the McCarthy era, he became unemployed, and took up a life of constant travel around the world, staying with friends wherever he went, always doing mathematics, to the end of his life. A mathematician's **Erdös number** is the number of steps from her to Erdös in the coauthor graph. For example, if she wrote a paper with Jones, and Jones and Smith have a paper together, and Smith once wrote a paper with Erdös, then her Erdös number is three unless there's a shorter path. One of Erdös' concepts is *God's Book of Proofs* where the Supreme Fascist (his term for the deity) records perfect and wonderful arguments. *Proofs from the Book* is a collection of elementary proofs, from the fields of mathematics he was most interested in, that are, in the opinion of Erdös and the authors, "bookworthy."
- C.C. Adams, *The Knot Book: an Elementary Introduction to the Mathematical Theory of Knots*, American Mathematical Society, 2001.
    - Two ways of arranging a loop of string in space are equivalent if you can move the string from one position to the other without cutting it, and a **knot** is an equivalence class of this relation. That is, a knot is pretty much what you've always thought it was. Knot theory is a thriving field of research, and *The Knot Book* gives a relaxed and informal (but rigorous) description of some parts of it that can be explained with pictures and a bit of

elementary algebra. A remarkable wealth of material meets these conditions, including some advanced concepts of low dimensional topology and the work for which Vaughan Jones (b. 1952) was awarded the Fields Medal in 1990. The subject is full of easily stated open problems, some of which just might be cracked by a clever amateur.

- J.D. Sally and P.J. Sally, *Roots to Research: a Vertical Development of Mathematical Problems*, American Mathematical Society, 2007.

  - The standard approach to mathematics education is "horizontal," laying down one layer of bricks at a time. This seems logical, and is necessary to at least some extent, but an exclusively horizontal approach leads to tunnel vision while frustrating students who want to obtain some glimmer of more advanced topics and contemporary research. Like this book, *Roots to Research* is vertical, aiming to take students quickly to a high level, but instead of emphasizing generalities, the Sally's study five particular problems. In each case they begin with aspects that can be understood using only high school algebra, then slowly build up the theory, eventually applying more advanced concepts, and finally arriving at recent results. Their expectation is that most students below the level of graduate studies will not make it all the way, but can still benefit from trying to go a bit past their comfort levels. Reading their book will make it easier to read this one because you will see the concepts developed here applied again and again. Reading this book will give you the tools you need to go a bit further with each of the topics they study. *Going back and forth, according to your mood, will make both books more enjoyable.*

- V. Klee and S. Wagon, *Old and New Unsolved Problems in Plane Geometry and Number Theory*, Mathematical Association of America Dolciani Mathematical Expositions No. 11, 1991.

  - Beginning with any positive integer n_0, form the sequence $n_1, n_2, \ldots$ according to the rule

    $$n_{k+1} := \begin{cases} n_k/2, & \text{if } n_k \text{ is even,} \\ 3n_k + 1, & \text{if } n_k \text{ is odd.} \end{cases}$$

    Is it the case that for any n_0 the sequence eventually settles into the cycle $\ldots, 1, 4, 2, 1, 4, 2, 1, \ldots$? Legend has it that this problem passed from one university mathematics department to the next after it was discovered, in each case preventing anyone from getting anything done for about a month, after which everyone gave up and went back to whatever they had been doing. Klee and Wagon talk about some of the biggies like the Riemann hypothesis, but there are also lots of simple, seductive, totally baffling problems like this that are good to think about even if you're very unlikely to solve one.

- J.W. Milnor, *Topology from the Differentiable Viewpoint*, University Press of Virginia, 1965.
    - Milnor's slender monograph is an ideal next step into differential topology for those who made it through this book. Technicalities are minimized by focusing on the simplest and most ideal cases, allowing a rapid approach to ideas that were then at the forefront of research, and are now fundamental in the study of the topology of manifolds. Milnor is universally acknowledged to be one of the greatest expositors ever, and this book is as simple, lucid, and perfect in all its details, as anything he ever wrote.
- G.H. Hardy and E.M. Wright, *An Introduction to the Theory of Numbers*, fifth edition, Oxford University Press, 1979.
    - Over its 2500 year history number theory has accumulated many interesting results and larger theories that are beautiful and surprising, but require little more than high school algebra and some logical thought. Hardy and Wright were at the top of the field during their lives, and this book is an undisputed masterpiece. The material it covers permeates the rest of mathematics and is fundamental to all aspects of number theory.
- J.H. Silverman and J. Tate, *Rational Points on Elliptic Curves*, Springer-Verlag, 1992.
    - The subject matter of this book is (or at least is thought to be) more advanced than what the other books on this list treat, but the authors do a marvelous job of keeping everything concrete. The subject itself is extremely beautiful, a meeting place of number theory, algebra, and analysis, and highly relevant to contemporary research.

Appendix A

Problems

A problem below is described as an "Exercise" if its primary purpose is to give some practice with computations, or perhaps to illustrate some concept. The primary purpose of an "Extension" is to introduce an important concept or result. In some cases the distinction is not clear cut. In many cases the solution of a problem depends on the earlier problems for the same chapter. (If the solution depends on an earlier problem for a different chapter, that will be noted.)

Problems for Chapter 1

Extension 1.1: The **order** of a group G is the cardinality of G. A group is **finite** if its order is finite, in which case the order of G is denoted by $o(G)$. Show that if G is a finite group and $g \in G$, then there is a natural number n such that $g^n = e_G$. The least such natural number is called the **order** or **period** of g.

Extension 1.2: A group is **cyclic** if there is an element γ such that

$$G = \{\, \gamma^n : n \text{ is an integer} \,\}.$$

An element $\gamma \in G$ with this property is called a **generator** of G.

(a) Prove that for every integer n there is a cyclic group C_n with $o(C_n) = n$.

(b) Prove that if d is a natural number that divides n, then C_n has a subgroup H with $o(H) = d$.

(c) When is a cyclic group simple?

Extension 1.3: Prove that if G and H are groups, then $G \times H$ is a group if we define the group operation by the formula

$$(g, h)(g', h') := (gg', hh').$$

Exercise 1.4: A finite group can be described by its multiplication table. The multiplication tables for the unique (up to isomorphism) groups with two and three elements are:

	e	a
e	e	a
a	a	e

and

	e	a	b
e	e	a	b
a	a	b	e
b	b	e	a

Find the two possible multiplication tables for groups with four elements. If you *really* enjoy this sort of thing, show that there are three possible multiplication tables for groups with six elements, so that C_6, $C_2 \times C_3$, and S_3 are (again, up to isomorphism) the only groups with six elements.

Exercise 1.5: Let C be the cube with vertices $(\pm 1, \pm 1, \pm 1)$.

(a) Describe the group G of symmetries of C.

(b) Identify four vertices of C that are the vertices of a regular tetrahedron T. Describe the group H of symmetries of T.

(c) Determine which elements of G induce symmetries of T, and show that they constitute a subgroup of G. Is this subgroup normal?

Extension 1.6: Prove that for any group G the map $(g, a) \mapsto ga$ is an action of G on itself. For any set X let $\mathrm{Sym}(X)$ be the set of bijections $f : X \to X$.

(a) Prove that if the product gf of two elements of $\mathrm{Sym}(X)$ is defined to be the composition $g \circ f$, then $\mathrm{Sym}(X)$ is a group.

(b) Let G be a group, and define $\varphi : G \to \mathrm{Sym}(G)$ by letting $\varphi(g) : G \to G$ be the function $a \mapsto ga$. Prove that φ is a homomorphism.

(c) Prove that φ is injective, hence an isomorphism between G and $\varphi(G)$.

We have proved **Cayley's theorem**, which is the assertion that every group G is isomorphic is a subgroup of $\mathrm{Sym}(G)$. In particular, if G is finite, then it is isomorphic to a subgroup of $S_{o(G)}$.

Exercise 1.7: Use induction to prove that

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

for every $n = 1, 2, \ldots$.

Exercise 1.8: Prove that if X is an uncountable infinite set and C is a countable subset, then $X \setminus C$ is uncountable. (Hint: proof by contradiction.)

Exercise 1.9: Prove that the set $\mathbf{Q}$ of rational numbers is countable.

Extension 1.10: For any set X the **power set** of X is the set of subsets of X. Often the power set of X is denoted by 2^X, but sometimes you will see $P(X)$ or $\mathcal{P}(X)$. Prove that for any function $f : X \to 2^X$ the set $\{\, x \in X : x \notin f(x) \,\}$ is not in the image of f. (Note the similarity to Russell's Paradox.) Roughly, this implies that for any set X, there is another set, namely 2^X, whose cardinality is greater. But there is some work to do (defining "greater" for cardinalities, and proving that it has certain properties) before this becomes precise.

Problems for Chapter 2

Extension 2.1: Prove that a finite integral domain is a field.

Extension 2.2: A **division ring** is a possibly noncommutative ring with unit whose nonzero elements have unique multiplicative inverses. That is, it satisfies (R1)-(R7) and the two sided version of (F8) (for each $r \in R \setminus \{0\}$ there is a unique s such that $rs = sr = 1$) but multiplication need not be commutative. The ring of **quaternions** is

$$\mathbb{H} := \{ a + bi + cj + dk : a, b, c, d \in \mathbb{R} \}$$

with addition and multiplication defined as follows: if $\alpha = a + bi + cj + dk$ and $\alpha' = a' + b'i + c'j + d'k$, then

$$\alpha + \alpha' := (a + a') + (b + b')i + (c + c')j + (d + d')k$$

and

$$\alpha\alpha' := (aa' - bb' - cc' - dd') + (ab' + a'b + cd' - c'd)i + (ac' + a'c - bd' + b'd)j$$
$$+ (ad' + a'd + bc' - b'c)k.$$

(a) The quaternions were discovered by William Rowan Hamilton in a flash of inspiration as he was walking past Brougham Bridge in Dublin, and he celebrated the moment by carving the formula

$$i^2 = j^2 = k^2 = ijk = -1$$

in a stone of the bridge. Prove that this formula implies all the multiplicative relations between i, j, and k that are embedded in the formula for multiplication. (Here associativity of multiplication, 1 being the multiplicative identity, and $(-1)^2 = 1$ are taken as given.)

(b) Prove that $\mathbb{H}$ satisfies (R1)-(R7). (It is easier to prove (R5) if you have already established (R6).)

(c) The **conjugate** of $\alpha = a + bi + cj + dk$ is $\alpha^* := a - bi - cj - dk$, and the **norm** of α is

$$\|\alpha\| := \sqrt{\alpha\alpha^*} = \sqrt{\alpha^*\alpha} = \sqrt{a^2 + b^2 + c^2 + d^2}.$$

Prove that if $\alpha \neq 0$, then $\alpha^*/\|\alpha\|^2$ is a two sided multiplicative inverse of α, and the unique multiplicative inverse from either side.

(d) If you feel like a challenge you can try to prove that $\|\alpha\alpha'\| = \|\alpha\| \, \|\alpha'\|$ for all $\alpha, \alpha' \in \mathbb{H}$. (Writing out the entire calculation would be extremely tedious; if you think about what would happen *if* you used the distributive and associative laws to expand $(\alpha\alpha')(\alpha\alpha')^*$ you should be able to see that certain terms cancel. To see this more concretely write out the calculation when $\alpha = a + bi$ and $\alpha' = a' + c'j$.)

Extension 2.3: Let R be a commutative ring. Prove that if I and J are ideals of R, then so are:

(a) $I \cap J$;

(b) $I + J := \{ i + j : i \in I \text{ and } j \in J \}$;

(c) $IJ := \{ i_1, j_1 + \cdots + i_n j_n : n = 1, 2, \ldots, i_1, \ldots, i_n \in I, \text{ and } j_1, \ldots, j_n \in J \}$;

(d) $(I : J) := \{ r \in R : rJ \subset I \}$. (This is called the **ideal quotient** of I and J.)

Extension 2.4: Suppose that R is a commutative ring with unit, and I is an ideal that is a proper subset of R.

(a) Prove that the quotient module R/I is a commutative ring with unit if we define multiplication by the formula $(a + I)(b + I) := ab + I$.

(b) We say that I is a **prime ideal** if, whenever $a, b, \in R$ and $ab \in I$, either $a \in I$ or $b \in I$.

(i) Prove that the prime ideals of $\mathbb{Z}$ are the principal ideals (p) where p is a prime number.

(ii) Prove that if I is a prime ideal if and only if R/I is an integral domain.

(c) Prove that if J is an ideal of R that contains I, then J/I is an ideal of R/I. Prove that if A is an ideal of R/I, then $J := \{ a \in R : a + I \in A \}$ is an ideal of R that contains I, and $A = J/I$.

(d) We say that I is a **maximal ideal** of R if it is not a proper subset of another ideal of R that is, in turn, a proper subset of R.

(i) Prove that a maximal ideal is prime. (Hint: If I is maximal and $a \in R \setminus I$, then the smallest ideal containing a and I is all of R, so $1 = ra + i$ for some $r \in R$ and $i \in I$. Since $1 \notin I$ because I is a proper subset of R, it is enough to prove that if $a, b \in R \setminus I$ and $ab \in I$, then $1 \in I$.)

(ii) Use Theorem 2.19 to prove that I is maximal if and only if R/I is a field.

Extension 2.5: Let R be an integral domain. A set $S \subset R$ is **multiplicative** if $1 \in S$ and $st \in S$ whenever $s, t \in S$. For such a set we define an equivalence relation on $R \times S$ by specifying that (a, s) and (b, t) are equivalent if $at = bs$.

(a) Prove that this is, in fact, an equivalence relation. We denote the equivalence class of (a, s) by a/s, and the set of equivalence classes is $S^{-1}R$.

(b) Addition and multiplication of equivalence classes are defined by the formulas $a/s + b/t = (at + bs)/st$ and $a/s \cdot b/t = ab/st$. Prove that these definitions are independent of the choices of representatives, and that they turn $S^{-1}R$ into a ring.

(c) Give two or three "interesting" multiplicative subsets $S \subset \mathbb{Z}$, and for each describe the derived ring of fractions $S^{-1}\mathbb{Z}$.

(d) Show that the function $\varphi : R \to S^{-1}R$ given by the formula $\varphi(r) := r/1$ is an injective homomorphism.

(e) If I is an ideal of R, then $R \setminus I$ is multiplicative if and only if I is prime. (This is simply what the definition of a prime ideal says.) When P is a prime ideal we write R_P instead of $(R \setminus P)^{-1}R$. Observing that (0) is prime, show that $R_{(0)}$ is a field, and conclude that any integral domain is isomorphic to a subring of a field.

(f) If $\varphi : R^k \to R^\ell$ is an R-module homomorphism, then the matrix of φ is also the matrix of an $S^{-1}R$-module homomorphism $\tilde{\varphi} : (S^{-1}R)^k \to (S^{-1}R)^\ell$.

(i) Prove that $\tilde{\varphi}$ is injective (surjective) if and only if φ is injective (surjective).

(ii) Prove that if $k \neq \ell$, then R^k and R^ℓ are not isomorphic. (This requires a result from Chapter 4, so you should probably return to this later if you don't see how things work.)

Exercise 2.6: Let R be a principal ideal domain. Prove that any two nonzero elements of R, say a and b, have a **least common multiple**. That is, there is a nonzero $c \in R$ such that $a|c$ and $b|c$, and $c|d$ for all d such that $a|d$ and $b|d$.

Extension 2.7: Let G be a finite group. Prove that if H is a subgroup of G, then $o(G/H) = o(G)/o(H)$. (Whenever X is a finite set that has a partition into subsets that all have the same number of elements, the number of "cells" in the partition times the number of elements in each cell is the number of elements of X.) Observing that the order $o(g)$ of any element $g \in G$ is the number of elements of the cyclic subgroup $\{g, g^2, \ldots, g^{o(g)} = e_G\}$ generated by g, conclude that $g^{o(G)} = e_G$ for every $g \in G$. Applying this to the multiplicative group $\mathbb{Z}_p^*$, where p is a prime, conclude that for every integer a it is the case that $a^{p-1} \equiv 1 \bmod p$ and $a^p \equiv a \bmod p$. This result was discovered by Fermat, and is known as **Fermat's little theorem**.

Exercise 2.8: Let R be a commutative ring, let M be an R-module, let N and P be submodules of M, and define

$$(N : P) := \{\, r \in R : rP \subset N \,\}.$$

Prove that $(N : P)$ is an ideal of R. The **annihilator** of M is $(\{0\}, M)$; this is the set of r such that $rm = 0$ for all $m \in M$.

Exercise 2.9: Prove that if M is an R-module, where R is a commutative ring, and N and P are submodules of M, then $(N + P)/P$ is isomorphic to $N/(N \cap P)$.

Problems for Chapter 3

Extension 3.1: Let τ be the set of $U \subset \mathbb{R}$ such that for every $a \in U$ there is an $\varepsilon > 0$ such that $[a, a + \varepsilon) \subset U$. Prove that τ is a topology. The space $(\mathbb{R}, \tau)$ is called the **Sorgenfrey line**.

Extension 3.2: Let (X, τ) be a topological space, and let $f : X \to Y$ be a function.

(a) Prove that $\sigma := \{\, V \subset Y : f^{-1}(V) \in \tau \,\}$ is (the collection of open sets of) a topology for Y. This is called the **quotient topology** induced by f.

(b) If σ' and σ'' are topologies for Y with $\sigma' \subset \sigma''$, so that every σ'-open set is σ''-open, then we say that σ' is **coarser** than σ'', and that σ'' is **finer** than σ'. Prove that if f is continuous when Y has the topology σ', then σ' is coarser than the quotient topology.

(c) Recall that f is an open map (closed map) if it is continuous and $f(U)$ is open whenever $U \subset X$ is open ($f(C)$ is closed whenever $C \subset X$ is closed). Prove that if f is surjective and either an open map or a closed map, then Y has the quotient topology.

Exercise 3.3: Let $f : [0, \infty) \to [0, \infty)$ be a function with $f(0) = 0$, $f(t) > f(s)$ whenever $t > s$, and $f(s, t) \leq f(s) + f(t)$ for all s and t. Let (X, d) be a metric space, and let $\tilde{d} := f \circ d : X \times X \to [0, \infty)$.

(a) Prove that $\tilde{d}$ is a metric.

(b) Prove that the topology induced by $\tilde{d}$ is coarser than the topology induced by d.

(c) Prove that if f is continuous at 0, then d and $\tilde{d}$ induce the same topology.

Extension 3.4: Prove that if X and Y are topological spaces, then the product topology on $X \times Y$ is the coarsest topology such that the projections $\pi_X : (x, y) \mapsto x$ and $\pi_Y : (x, y) \mapsto y$ are continuous. Prove that the projections are open maps, and give an example showing that they are not necessarily closed maps.

Exercise 3.5: Let X and Y be topological spaces, and let $X \times Y$ have the product topology. Suppose that $A \subset X$ and $B \subset Y$.

(a) Prove that the interior of $A \times B$ is the cartesian product of the interior of A and the interior of B, and the closure of $A \times B$ is the cartesian product of the closure of A and the closure of B.

(b) Describe the boundary $A \times B$ in terms of the closures and boundaries of A and B.

(c) Prove that the topology $A \times B$ inherits as a subspace of $X \times Y$ coincides with the product topology derived from the subspace topologies of A and B.

Extension 3.6: If R is a commutative ring, an R-module M is **Noetherian** if each of its submodules is finitely generated. (Since the ideals of R are precisely the submodules, R is a Noetherian ring if and only if it is a Noetherian R-module.) Prove that the following are equivalent:

(a) M is Noetherian;

(b) every increasing sequence of submodules $M_1 \subset M_2 \subset \ldots$ "stabilizes": there is some K such that $M_k = M_K$ for all $k \geq K$;

(c) any collection of submodules of M has an element that is maximal in the sense of not being a subset of some other element of the collection.

Extension 3.7: A collection of subsets of a set has the **finite intersection property** if any finite subcollection has a nonempty intersection. Prove that a topological space is compact if and only if, whenever a collection of closed sets has the finite intersection property, the intersection of all elements of the collection is nonempty.

Extension 3.8: Let X and Y be topological spaces. A bijection $f : X \to Y$ is a **homeomorphism** if it is continuous and its inverse is also continuous. Prove that if X is compact, Y is a Hausdorff space, and $f : X \to Y$ is a continuous bijection, then f is a homeomorphism. (Hint: prove that if C is a closed subset of X, then $f(C)$ is closed.)

Extension 3.9: In addition to "Hausdorffness," there are several "separation" conditions that are similar but logically distinct. A topological space X is a T_1 **space** (terrible terminology, but it seems that the world is stuck with it) if, for each $x \in X$, the singleton $\{x\}$ is a closed set. It is **regular** if, for any $x \in X$ and any open set U containing x, there is a closed neighborhood of x that is contained in U. A topological space X is **normal** if, for any disjoint closed sets C and D, there are disjoint open sets U and V with $C \subset U$ and $D \subset V$.

(a) Prove that the space X_n discussed at the beginning of Section 3.4 is not T_1 or regular when $n \geq 2$, but it is always normal.

(b) Prove that any metric space is T_1, regular, and normal.

(c) Prove that if X is T_1 and regular, then it is Hausdorff.

(d) Prove that if X is Hausdorff, $K \subset X$ is compact, and $x \in X \setminus K$, then there are disjoint open sets U and V with $K \subset U$ and $x \in V$.

(e) Prove that if X is Hausdorff and compact, then it is normal.

(f) Prove that if X is regular, $K \subset X$ is compact, and U is an open superset of K, then there is an open V that contains K and whose closure is contained in U.

(g) Prove that if X is regular and compact, then it is normal.

(h) Prove that if X and Y are T_1 (or regular, or Hausdorff) then so is $X \times Y$. (The product of two normal spaces is not necessarily normal. The standard example is the **Sorgenfrey plane**, which is the product of two copies of the Sorgenfrey line; we won't discuss it further here, but you might like to look at the *Wikipedia* entry.)

Exercise 3.10: Prove that if A is a connected subset of a topological space X, then the closure of A is connected.

Exercise 3.11: Let X and Y be topological spaces, and let $X \times Y$ have the product topology.

(a) Prove that for each $x \in X$ the map $y \mapsto (x, y)$ is a homeomorphism between Y and $\{ (x, y) : y \in Y \}$ (with its relative topology as a subspace of $X \times Y$).

(b) Prove that if X and Y are connected, then so is $X \times Y$.

Problems for Chapter 4

Exercise 4.1: Show that if the characteristic of the field k is 0, then $(1, 1, 0, 0)$, $(2, 1, -1, 0)$, and $(0, 0, 0, 5)$ are linearly independent elements of k^4. For which primes p is it the case that these vectors are *not* linearly independent when the characteristic of k is p.

Exercise 4.2: Suppose that W is a linear subspace of the vector space V. Prove that $X \longleftrightarrow X/W$ is a bijection between the linear subspaces of V that contain W and the linear subspaces of V/W.

Extension 4.3: Let V and W be vector spaces, and let $L(V, W)$ be the set of linear transformations from V to W.

(a) Prove that $L(V, W)$ is a vector space if the vector operations are defined naturally: $(\ell_1 + \ell_2)(v) := \ell_1(v) + \ell_2(v)$ and $(\alpha\ell)(v) := \alpha\ell(v)$.

(b) Prove that if U and X are vector spaces and $k : U \to V$ and $m : W \to X$ are linear transformations, then the map $\ell \mapsto m \circ \ell \circ k$ is a linear transformation from $L(V, W)$ to $L(U, X)$.

Extension 4.4: If V is a vector space over the field k, then the **dual space** of V is

$$V^* := L(V, k).$$

Elements of V^* are called **linear functionals**. If W is a second vector space and $\ell : V \to W$ is linear, define $\ell^* : W^* \to V^*$ by letting $\ell(w^*)$ be the linear functional $v \mapsto w^*(\ell(v))$.

(a) Prove that $(\mathrm{Id}_V)^* = \mathrm{Id}_{V^*}$.

(b) Prove that if X is a third vector space and $m \in L(W, X)$, then $(m \circ \ell)^* = \ell^* \circ m^*$.

(c) Suppose that $v_1, \ldots, v_n$ is a basis of V, and for $i = 1, \ldots, n$ define $v_i^* \in V^*$ by the formula

$$v_i^*(\alpha_1 v_1 + \cdots + \alpha_n v_n) := \alpha_i.$$

Then $v_1^*, \ldots, v_n^*$ is the **dual basis** associated with $v_1, \ldots, v_n$. Prove that it is, in fact, a basis of V^*.

Since a vector space and its dual have the same dimension, the distinction between the two might seem insignificant, but in applications it is obvious. In economics, for example, V might be the set of vectors consisting of quantities of iron, rubber, and coal (with negative quantities understood as debts) in which case an element of V^* is naturally interpreted as a triple consisting of a price of iron, a price of rubber, and a price of coal.

Extension 4.5: Let V be a vector space over $\mathbb{R}$. An **inner product** on V is a function

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

such that for all $v, v', w \in W$ and $\alpha \in \mathbb{R}$:

(a) $\langle w, v \rangle = \langle v, w \rangle$,

(b) $\langle v + v', w \rangle = \langle v, w \rangle + \langle v', w \rangle$,

(c) $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$, and

(d) $\langle v, v \rangle \geq 0$, with equality if and only if $v = 0$.

Recall that any inner product induces a norm given by the formula $\|v\| := \langle v, v \rangle^{1/2}$. If $\|v\| = 1$, then we say that v is a **unit vector**, and when $\langle v, w \rangle = 0$ we say that v and w are **orthogonal**. A collection $\{ v_i : i \in I \} \subset V$ is an **orthonormal set** if its elements are unit vectors and distinct elements are orthogonal, so that $\langle v_i, v_j \rangle = \delta_{ij}$ for all $i, j \in I$. (Here δ_{ij} is the Kronecker delta.) An **orthonormal basis** is an orthonormal set that is also a basis. If W is a linear subspace, then the **orthogonal complement** of W is

$$W^\perp := \{ v \in V : \langle v, w \rangle = 0 \text{ for all } w \in W \}.$$

(a) Prove that if $v_1, \ldots, v_n$ is an orthonormal basis of V and $v_1, \ldots, v_k$ is a basis of W, then $v_{k+1}, \ldots, v_n$ is a basis of $W^\perp$.

(b) Let $w_1, \ldots, w_k$ be linearly independent. Define $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ inductively by the formulas

$$u_i := w_i - \langle w_i, v_1 \rangle v_1 - \cdots - \langle w_i, v_{i-1} \rangle v_{i-1} \quad \text{and} \quad v_i := \frac{u_i}{\|u_i\|}.$$

(This procedure is called the **Gram-Schmidt orthonormalization process**.) Prove that $\{v_1, \ldots, v_k\}$ is an orthonormal set with the same span as $w_1, \ldots, w_k$.

(c) Prove that if V is n-dimensional and W is a k-dimensional subspace, then there is an orthonormal basis $v_1, \dots, v_n$ such that $v_1, \dots, v_k$ an orthonormal basis of W. Conclude that $\dim W + \dim W^{\perp} = \dim V$.

(d) Prove that if $v_1, \dots, v_k$ is an orthonormal set and $w \in V$, then

$$\sum_{i=1}^{k} \langle w, v_i \rangle^2 \leq \|w\|^2.$$

This is called **Bessel's inequality**. When does it hold with equality?

Extension 4.6: When F and K are fields, with K an extension of F, we say that K is a **finite extension** of F if K is a finite dimensional vector space over F. In this case the **degree** of K over F, denoted by $[K : F]$, is the dimension of K. Suppose that K is a finite extension of F, and that L is an intermediate field, so that $F \subset L \subset K$.

(a) Prove that K is a finite extension of L. (Hint: the L-span of any subset of K contains the F-span.)

(b) Prove that $[K : F] = [K : L][L : F]$. (Hint: show that if $x_1, \dots, x_m$ is a basis of L, as a vector space over F, and if $y_1, \dots, y_n$ is a basis of K, as a vector space over L, then $\{ x_i y_j : i = 1, \dots, m, j = 1, \dots, n \}$ is a basis of K as a vector space over F.)

(c) A **minimal polynomial** for $x \in K$ is a polynomial $f \in F[X]$ such that $f(x) = 0$ and $g(x) \neq 0$ for all polynomials $g \in F[X]$ of lower degree. Prove that for each $x \in K$ there is a unique minimal polynomial that is monic.

(d) Prove that the degree of a minimal polynomial of x divides $[K : F]$. (Hint: if n is the degree of minimal polynomials of x, then $1, x, \dots, x^{n-1}$ spans a field.)

Problems for Chapter 5

Exercise 5.1: Compute the determinants of each of the following matrices twice, first by applying the formula defining the determinant, then using row and column operations.

$$\begin{pmatrix} 1 & 0 & 2 \\ 3 & 4 & 1 \\ 0 & 3 & 1 \end{pmatrix} \qquad \begin{pmatrix} 2 & 3 & 7 \\ 1 & 0 & 5 \\ 6 & 3 & 1 \end{pmatrix}$$

Extension 5.2: Let $\ell : V \to V$ be a linear transformation, where V is a finite dimensional vector space over a field k. We say that ℓ is **nilpotent** if $\ell^e = 0$ for some integer e. Prove that if e is the least such integer, then the minimal polynomial of ℓ is X^e.

Extension 5.3: Let R be a ring, and let M be an R-module. We say that m is a **torsion element** if $\mathrm{Ann}(m) \neq (0)$. The **torsion submodule** of M, denoted by M_T, is the set of all torsion elements of M. We say that M is a **torsion R-module** if $M_T = M$, and we say that M is **torsion free** if $M_T = (0)$. Prove the M_T is actually a submodule of M.

Extension 5.4: Suppose that R is a PID and M is an R-module.

(a) Suppose that g and h generate M and a and b are ring elements with $a \neq 0$ whose greatest common divisor is 1, so that $as + bt = 1$ for some s and t. (When we say that "the greatest common divisor of $a, \dots, a_r$ is 1," what we really mean is that any greatest common divisor is a unit, so that the ideal generated by $a, \dots, a_r$ is all of R. There will be similar expressions below, which should be interpreted in the same spirit.) Prove that M is generated by $ag + bh$ and $-tg + sh$.

(b) Prove that if $g_1, \dots, g_r$ is a system of generators for M, $a_1, \dots, a_r$ are elements of R whose greatest common divisor is 1, $a_1 \neq 0$, and $g := a_1 g_1 + \cdots + a_r g_r$, then there are $g_2', \dots, g_r'$ such that $g, g_2', \dots, g_r'$ is a system of generators for M. (Hint: Let N be the submodule generated by $g_2, \dots, g_r$, let d be the greatest common divisor of $a_2, \dots, a_r$, and let $h := (a_2/d)g_2 + \cdots + (a_r/d)g_r$.)

The underlying idea of the proof is similar to Gaussian elimination, but this becomes a bit obscure when we simplify things by combining induction with the 2×2 case.

Extension 5.5: An R-module M is **free** if it is an internal direct sum of copies of R. That is, there is a system of generators $\{ g_i : i \in I \}$, which may be infinite, such that each $m \in M$ has a unique representation of the form

$$m = a_1 g_{i_1} + \cdots + a_r g_{i_r}.$$

(a) Prove that if R is a PID and M is a finitely generated torsion free R-module, then M is free.

(b) Prove that if R is a PID and M is a finitely generated free R-module, then any submodule N is finitely generated and free.

It is actually the case that if R is a PID, then any submodule of a free R-module is free, even when R is not finitely generated, but the proof requires a sophisticated application of the axiom of choice that is used to produce a minimal set of generators.

Exercise 5.6: Prove that a PID is Noetherian.

Extension 5.7: (Structure Theorem for PID's I: Existence) Let R be a PID, let M be a finitely generated R-module, and let r be the minimal number of elements of any system of generators for M. We will say that a system of generators $g_1, \dots, g_r$ is **taut** if $\mathrm{Ann}(g_1)$ not a proper subset of $\mathrm{Ann}(g_1')$ for some other system of generators $g_1', \dots, g_r'$ with r elements.

(a) Prove there there is a taut system of generators.

(b) Suppose $g_1, \dots, g_r$ is a taut system of generators, let $N := Rg_2 + \cdots + Rg_r$, and assume that $N = Rg_2 \oplus \cdots \oplus Rg_r$. Prove that $a_1 = 0$ whenever $a_1 g_1 + a_2 g_2 + \cdots + a_r g_r = 0$. (Hint: apply Problem 5.4.)

(c) Prove that if $g_1, \dots, g_r$ is a taut system of generators, then $M = Rg_1 \oplus \cdots \oplus Rg_r$.

(d) Prove that if $g_1, \dots, g_r$ is a taut system of generators and $\mathrm{Ann}(g_2) \supset \cdots \supset \mathrm{Ann}(g_r)$, then $\mathrm{Ann}(g_1) \supset \mathrm{Ann}(g_2)$. (Hint: apply Problem 5.4.)

Prove that there is a system of generators $g_1, \dots, g_r$ with $M = Rg_1 \oplus \cdots \oplus Rg_r$ and $\mathrm{Ann}(g_1) \supset \cdots \supset \mathrm{Ann}(g_r)$.

Extension 5.8: (Structure Theorem for PID's II: Uniqueness) Let R be a PID, let M be a finitely generated R-module, and let r be the minimal number of elements of any system of generators for M. Suppose that $g_1, \ldots, g_r$ is a system of generators with $M = Rg_1 \oplus \cdots \oplus Rg_r$ and $\mathrm{Ann}(g_1) \supset \cdots \supset \mathrm{Ann}(g_r)$, and let $g'_1, \ldots, g'_s$ be another system of generators with $M = Rg'_1 \oplus \cdots \oplus Rg'_s$ and $\mathrm{Ann}(g'_1) \supset \cdots \supset \mathrm{Ann}(g'_s)$. We wish to show that $s = r$ and $\mathrm{Ann}(g'_i) = \mathrm{Ann}(g_i)$ for all i, so we may assume we are dealing with a counterexample for which r is minimal.

(a) Let $\mathrm{Ann}(g_1) = (d_1)$, and show that $d_1 \in \mathrm{Ann}(g'_1)$ by considering d_1M.

(b) Prove that in a PID R any proper (in the sense of not being all of R) ideal I is contained in a proper prime ideal.

(c) Prove that if P is a proper prime ideal that contains $\mathrm{Ann}(g_1)$, then M/PM is isomorphic to the direct sum of r copies of R/P, and also isomorphic to the direct sum of s copies of R/P. Applying Problem 2.5.f.ii, conclude that $s = r$.

(d) Let $\mathrm{Ann}(g'_1) = (d'_1)$, and show that $\mathrm{Ann}(g'_i) = \mathrm{Ann}(g_i)$ for all i by considering d'_1M.

Extension 5.9: Let G be a finitely generated abelian group.

(a) Prove that G is isomorphic to a unique group of the form

$$\mathbb{Z}^n \oplus \mathbb{Z}_{d_1} \oplus \cdots \oplus \mathbb{Z}_{d_k}$$

where $d_1, \ldots, d_k \geq 2$ with $d_i | d_j$ whenever $i < j$.

(b) Prove that if m and n are relatively prime, then $\mathbb{Z}_{mn}$ is isomorphic to $\mathbb{Z}_m \oplus \mathbb{Z}_n$.

(c) Prove that G is isomorphic to a unique (up to reordering of the pairs (p_i, e_i) group of the form

$$\mathbb{Z}^n \oplus \mathbb{Z}_{p_1^{e_1}} \oplus \cdots \oplus \mathbb{Z}_{p_m^{e_m}}$$

where $p_1, \ldots, p_m$ are (not necessarily distinct) primes.

This is called the **structure theorem for finitely generated abelian groups**.

Extension 5.10: (Proof of Theorem 5.22) Let V be a finite dimensional vector space, and let $\ell \in \mathrm{End}(V)$ be an endomorphism whose minimal polynomial has the prime factorization $p = p_1^{e_1} \cdots p_k^{e_i}$. For each i let U_i be the kernel of $p_i^{e_i}(\ell)$, and let Z_i be the image of $q_i(\ell)$ where $q_i := \prod_{j \neq i} p_j^{e_j}$.

(a) Prove that the ideal of $k[X]$ generated by $q_1, \ldots, q_k$ is all of $k[X]$, so there are polynomials $h_1, \ldots, h_k$ such that $h_1q_1 + \cdots h_kq_k = 1$.

(b) Prove that $Z_1 + \cdots + Z_k = V$ because $q_1(\ell)h_1(\ell) + \cdots + q_k(\ell)h_k(\ell) = \mathrm{Id}_V$.

(c) Prove that each U_i is an invariant subspace of ℓ because $p_i^{e_i}(\ell)\ell = \ell p_i^{e_i}(\ell)$.

(d) Prove that $Z_i \subset U_i$ for all i.

(e) Prove that if $u_1 + \cdots + u_k = 0$ with $u_i \in U_i$ for all i, then for each i we have $q_j(\ell)u_i = 0$ for all $j \neq i$ because $p_i^{e_i} | q_j$, and consequently $0 = q_i(\ell)(u_1 + \cdots + u_k) = q_i(\ell)u_i$, so $u_i = 0$ because $h_1(\ell)q_1(\ell) + \cdots + h_k(\ell)q_k(\ell) = \mathrm{Id}_V$. Conclude that $V = U_1 \oplus \cdots \oplus U_k$, and that $Z_i = U_i$ for all i.

Problems for Chapter 6

Extension 6.1: Recall that $M_n(k)$ is the ring of $n \times n$ matrices with entries in k, and let $f : M_n(k) \to M_n(k)$ be the function $f(A) = A^2$.

(a) Prove that $Df(A)B = AB + BA$.

(b) Generalizing this result, find the derivative of the function $A \mapsto p(A)$ for an arbitrary polynomial $a_0 + a_1X + \cdots + a_mX^m \in k[X]$.

Exercise 6.2: Suppose that $x, u, v \in \mathbb{R}^n$, let $p : \mathbb{R} \to \mathbb{R}^n$ be the function $p(s) := x + su$, and let $q : \mathbb{R} \to \mathbb{R}^n$ be the function $q(t) := tv$.

(a) Express the function $D(s,t) := \|p(s) - q(t)\|^2$ in terms of the inner product.

(b) Compute $\frac{\partial D}{\partial s}(s,t)$ and $\frac{\partial D}{\partial t}(s,t)$.

(c) When is there a unique pair (s^*, t^*) such that $\frac{\partial D}{\partial s}(s^*,t^*) = 0 = \frac{\partial D}{\partial t}(s^*,t^*)$? Relate you answer to the Cauchy-Schwartz inequality.

Extension 6.3: Suppose that $P \subset \mathbb{R}^m$ and $X \subset \mathbb{R}^n$ are open, and that $U : P \times X \to \mathbb{R}$ is a C^1 function. For $(p, x) \in P \times X$ let $D_pU(p,x)$ and $D_xU(p,x)$ be the derivatives of the functions $U(\cdot, x) : P \to \mathbb{R}$ and $U(p, \cdot) : X \to \mathbb{R}$ at p and x respectively. Suppose that $C : P \to X$ is a C^1 function, and let $M : P \to \mathbb{R}$ be the function $M(p) := U(p, C(p))$. Use the chain rule to prove that if $D_xU(p, C(p)) = 0$, then

$$DM(p) = D_pU(p, M(p)).$$

This result is known in economics as the **envelope theorem**. The idea is that U is a quantity that you want to maximize (e.g., utility or profits) while p is a parameter (e.g., prices) that you don't control, and x is something you get to choose. If, for each p, $C(p)$ is the optimal choice, then $D_xU(p, C(p)) = 0$. When we then ask how the optimized value of the problem $M(p)$ changes as p changes we don't need to know how changes in p affect $C(p)$. For example, the change in profits resulting from a change in the price you charge your customers is, to a fair approximation, simply the price change times the amount you are currently selling, and the change resulting from a change in some input price is the input price change times the amount of that input you are consuming. The specific applications of the envelope theorem to the theory of production are known as **Hotelling's lemma**.

Exercise 6.4: For a ring R let $G(R)$ be the underlying commutative group given by addition, and for a ring homomorphism $\varphi : R \to S$ let $G(\varphi)$ be φ understood as a function from $G(R)$ to $G(S)$. Prove that G is a covariant functor from the category of rings and homomorphisms to the category of groups and homomorphisms.

Extension 6.5: If G is a group and $g, h \in G$, the **commutator** of g and h is

$$[g, h] := ghg^{-1}h^{-1}.$$

The **commutator subgroup** of G is the smallest subgroup of G containing all the commutators. It is denoted by $[G, G]$.

(a) Show that $[g, h]^{-1} = [h, g]$, so $[G, G]$ is the collection of all finite products of commutators.

(b) Prove that the commutator subgroup is normal. (Hint: first show that $f[g, h]f^{-1} = [f, [g, h]] \cdot [g, h]$.)

(c) The **abelianization** of G is $G^{\mathrm{ab}} := G/[G,G]$. Prove that G^{ab} is abelian.

(d) Define a sequence of subgroups $G^{(0)}, G^{(1)}, G^{(2)}, \ldots$ inductively by letting $G^{(0)} := G$ and

$$G^{(n)} := [G^{(n-1)}, G^{(n-1)}]$$

for all $n \geq 1$. The group G is **solvable** if $G^{(n)} = \{e_G\}$ for some n. Prove that if $\varphi : G \to H$ is a homomorphism, then $\varphi(G^{(n)}) \subset H^{(n)}$ for all n, so there are homomorphisms

$$\varphi^{(n)} : G^{(n-1)}/G^{(n)} \to H^{(n-1)}/H^{(n)}$$

given by $\varphi^{(n)}(G^{(n)}g) := H^{(n)}\varphi(g)$.

(e) Prove that for each $n = 1, 2, \ldots$ there is covariant functor from the category of groups and homomorphisms to the category of abelian groups and homomorphisms that takes each G to $G^{(n-1)}/G^{(n)}$ and each φ to $\varphi^{(n)}$.

Extension 6.6: Use the univariate version of Taylor's theorem to prove that if $f : (a,b) \to \mathbb{R}$ is C^2 and $f(t^*) \geq f(t)$ for all $t \in (a,b)$, then $f''(t^*) \leq 0$.

Extension 6.7: An $n \times n$ matrix A with entries in $\mathbb{R}$ is said to be **negative semidefinite** if $v^T A v \leq 0$ for all $v \in \mathbb{R}^n$. Suppose that $U \subset \mathbb{R}^n$ is open, $f : U \to \mathbb{R}$ is C^2, and $f(x^*) \geq f(x)$ for all $x \in U$. Prove that the matrix

$$\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x^*) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x^*) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_2^2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x^*) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x^*) & \frac{\partial^2 f}{\partial x_n \partial x_2}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(x^*) \end{pmatrix}$$

is negative semidefinite by applying the last problem to the function $t \mapsto f(x^* + tv)$.

Extension 6.8: Suppose that $f : [a,b] \to \mathbb{R}$ and $g : [a,b] \to \mathbb{R}$ are continuous and differentiable at each $t \in (a,b)$. Apply Rolle's theorem to the function

$$h(t) := \big(f(b) - f(a)\big)\big(g(x) - g(a)\big) - \big(g(b) - g(a)\big)\big(f(x) - f(a)\big)$$

to obtain a $t \in (a,b)$ such that

$$\big(f(b) - f(a)\big)g'(t) - \big(g(b) - g(a)\big)f'(t) = 0.$$

If $g(b) - g(a)$ and $g'(t)$ are both nonzero, then

$$\frac{f'(t)}{g'(t)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

This is known as **Cauchy's mean value theorem**.

Extension 6.9: Suppose that $f : (a,b) \to \mathbb{R}$ and $g : (a,b) \to \mathbb{R}$ are differentiable with $g'(t) \neq 0$ for all t, and $\lim_{t \to a} f(t) = 0 = \lim_{t \to a} g(t)$. Use Cauchy's mean value theorem to prove **L'Hopital's rule**:

$$\lim_{t \to a} \frac{f(t)}{g(t)} = \lim_{t \to a} \frac{f'(t)}{g'(t)}$$

whenever the right hand side limit exists.

Problems for Chapter 7

Exercise 7.1: Prove that compositions of conformal maps are conformal. Use this to prove the Cauchy-Riemann equations for the polar representation $z = re^{i\theta}$:

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta} \quad \text{and} \quad \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}, \qquad \text{which imply that} \qquad \frac{\partial f}{\partial r} = \frac{1}{ir}\frac{\partial f}{\partial \theta}.$$

Extension 7.2: The **Laplacian** is the differential operator $\Delta := \frac{\partial^2}{\partial x_1^2} + \cdots + \frac{\partial^2}{\partial x_n^2}$. What we mean by this is that if $U \subset \mathbb{R}^n$ is open and $g : U \to \mathbb{R}$ is C^2, then $\Delta g : U \to \mathbb{R}$ is the function

$$\Delta g(x) := \frac{\partial^2 g}{\partial x_1^2}(x) + \cdots + \frac{\partial^2 g}{\partial x_n^2}(x).$$

If $\Delta g(x) = 0$ for all $x \in U$, then g is said to be **harmonic**. Use the Cauchy-Riemann equations to prove that if $U \subset \mathbb{C}$ is open, $f = u + iv : U \to \mathbb{C}$ is C^1 in the complex sense, and u and v are C^2 in the real sense, then u and v are harmonic.

Extension 7.3: (Schwarz lemma) Let $D := \{\, z \in \mathbb{C} : |z| < 1 \,\}$, and let $f : D \to D$ be an analytic function with $f(0) = 0$.

(a) Define $g : D \to \mathbb{C}$ by setting $g(0) = f'(0)$ and $g(z) = f(z)/z$ if $z \neq 0$. Prove that g is analytic.

(b) For each $r < 1$ apply the maximum modulus principle to the restriction of g to the set of z with $|z| \leq r$, arriving at the conclusion that $|f(z)| \leq |z|$ for all z, and that if $|f(z)| = |z|$ for some z, then g is constant, so $f(z) = \alpha z$ for some α with $|\alpha| = 1$.

(c) Prove that if f is a bijection and $f^{-1} : D \to D$ is analytic, then there is an α with $|\alpha| = 1$ such that $f(z) = \alpha z$ for all z.

Extension 7.4: An **Hermitian inner product** on $\mathbb{C}^n$ is a function $\langle \cdot, \cdot \rangle : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ such that for all $u, v, w \in \mathbb{C}^n$ and all $\alpha \in \mathbb{C}$:

(1) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$;

(2) $\langle \alpha u, v \rangle = \alpha \langle u, v \rangle$;

(3) $\langle v, u \rangle = \overline{\langle u, v \rangle}$;

(4) $\langle u, u \rangle \geq 0$ with equality if and only if $u = 0$.

Prove that, in addition, $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$ and $\langle u, \alpha v \rangle = \overline{\alpha}\langle u, v \rangle$. The **standard Hermitian inner product** is $\langle u, v \rangle := u_1\overline{v}_1 + \cdots + u_n\overline{v}_n$; verify that it satisfies (1)-(4). Discuss the extension of all the concepts and results of Problem 4.5 to this setting. Prove that the standard Hermitian inner product is *not* differentiable in the complex sense. For $A \in M_n(\mathbb{C})$ let $\overline{A}$ denote the matrix whose entries are the complex conjugates of the entries of A. We say that A is **unitary** if $A\overline{A}^T = I$. Prove that A is unitary if and only if it preserves the standard Hermitian inner product in the sense that $\langle Au, Av \rangle = \langle u, v \rangle$ for all $u, v \in \mathbb{C}^n$.

Problems for Chapter 8

Exercise 8.1: Give a formal proof that if $f : M \to N$ is a C^r function, where M and N are C^r manifolds, and $P \subset M$ is a C^r submanifold, then $f|_P : P \to N$ is C^r.

Extension 8.2: A **Lie group** is a C^∞ (or complex analytic, if the underlying field is $\mathbb{C}$) manifold G that is also a group with group operations $(g, h) \mapsto gh$ and $g \mapsto g^{-1}$ that are C^∞ (complex analytic) functions.

(a) Prove that if $H \subset G$ is both a subgroup and a C^∞ (complex analytic) submanifold, then H is also a Lie group.

(b) Use the regular value theorem to prove that the circle $C = \{ z \in \mathbb{C} : |z| = 1 \}$ and the set $\{ \alpha \in \mathbb{H} : |\alpha| = 1 \}$ of unit quaternions are Lie groups with multiplication as the group operation.

(c) Prove that **general linear groups** $Gl_n(\mathbb{R}) := \{ A \in M_n(\mathbb{R}) : |A| \neq 0 \}$ and $Gl_n(\mathbb{C}) := \{ A \in M_n(\mathbb{C}) : |A| \neq 0 \}$ are Lie groups with matrix multiplication as the group operation. (Hint: Cramer's rule.)

(d) Use the regular value theorem to prove that the **special linear groups** $SL_n(\mathbb{R}) := \{ A \in M_n(\mathbb{R}) : |A| = 1 \}$ and $SL_n(\mathbb{C}) := \{ A \in M_n(\mathbb{C}) : |A| = 1 \}$ are Lie groups.

(e) A matrix $A \in M_n(\mathbb{R})$ is **symmetric** if $A^T = A$. Let $Sym_n(\mathbb{R})$ be the set of symmetric matrices in $M_n(\mathbb{R})$. Observe that $Sym_n(\mathbb{R})$ is a $\frac{1}{2}(n+1)n$-dimensional linear subspace of $M_n(\mathbb{R})$, and in this sense a C^∞ manifold. Prove that every $A \in M_n(\mathbb{R})$ is a regular point of the function $A \mapsto A + A^T$ from $M_n(\mathbb{R})$ to $Sym_n(\mathbb{R})$.

(f) Use the regular value theorem to prove that if J is a nonsingular symmetric matrix, then $\{ A \in M_n(\mathbb{R}) : A^T J A = J \}$ is a Lie group with matrix multiplication as the group operation. When $J = I$ is the identity matrix this yields the **orthogonal group** $O(n)$, when

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$$

it gives the **symplectic group** $Sp(2n)$, and when

$$J = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

the result is the **Lorentz group**.

(g) The **unitary group** $U(n)$ is the set of $n \times n$ unitary matrices. Use the fact that it is the set of matrices that preserve the Hermitian inner product to prove that it is a Lie group (over $\mathbb{R}$, but *not* over $\mathbb{C}$).

Problems for Chapter 9

Extension 9.1: The integral of a continuous function $g = x + iy : [a, b] \to \mathbb{C}$ is defined to be

$$\int_a^b g(t)\,dt := \int_a^b x(t)\,dt + i \int_a^b y(t)\,dt.$$

If $f : U \to \mathbb{C}$ is continuous, where $U \subset \mathbb{C}$ is open, and $\gamma : [a,b] \to U$ is C^1, let

$$\int_\gamma f := \int_a^b f(\gamma(t))\gamma'(t)\,dt.$$

This is called a **contour integral**.

(a) Prove that the definition of $\int_\gamma f$ does not depend on the parameterization of γ in the sense that if $\lambda : [c,d] \to [a,b]$ is a C^1 function with $\lambda(c) = a$ and $\lambda(d) = b$, then $\int_{\gamma\circ\lambda} f = \int_\gamma f$.

A 1-**form** on an open set $V \subset \mathbb{C}$ is an expression of the form $f(z)dz$ where $f : V \to \mathbb{C}$ is continuous. A 1-**form** ω on a Riemann surface C is defined to be a specification of a 1-form $f_\varphi(z)dz$ on V for each analytic coordinate chart $\varphi : U \to V$, where these 1-forms are required to satisfy the following consistency condition:

$$f_{\varphi_2}(\varphi_2(p)) = \left(\varphi_1 \circ \varphi_2^{-1}\right)'(\varphi_2(p)) \cdot f_{\varphi_1}(\varphi_1(p))$$

whenever $\varphi_1 : U_1 \to V_1$ and $\varphi_2 : U_2 \to V_2$ are analytic coordinate charts and $p \in U_1 \cap U_2$. If ω is a 1-form on C and $\gamma : [a,b] \to C$ is a path, let

$$\int_\gamma \omega := \int_{\varphi_1\circ\gamma|_{[t_0,t_1]}} f_{\varphi_1} + \cdots + \int_{\varphi_k\circ\gamma|_{[t_{k-1},t_k]}} f_{\varphi_k}$$

where $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ and each $\varphi_i : U_i \to V_i$ is an analytic coordinate chart with $\gamma([t_{i-1}, t_i]) \subset U_i$.

(b) Prove that the definition of $\int_\gamma \omega$ is independent of the choices of $t_0, \ldots, t_k$ and $\varphi_1, \ldots, \varphi_k$.

Exercise 9.2: Suppose you are at the origin in hyperbolic space. We will say that objects at a point p in hyperbolic space have **apparent distance** d if, in the limit as $\varepsilon \to 0$, the amount of your visual field that a small object at p occupies is ε/d times the amount of your visual field that it would occupy if it was at distance ε from the origin. Describe the relationship between apparent distance and actual distance. Consider an object moving away from the origin at constant speed. Describe its “apparent speed” as a function of time.

Extension 9.3: There is a way to connect two Riemann surfaces C_1 and C_2 to each other, along the lines suggested by Figure 9.6. Let $D := \{ z \in \mathbb{C} : |z| < 2 \}$ be the open disk of radius 2 centered at the origin in $\mathbb{C}$, let $B := \{ z \in \mathbb{C} : |z| \leq 1/2 \}$ be the closed disk of radius 1/2, and let $A := D \backslash B$. (Such a set is called an **annulus**.) Let $\iota : A \to A$ be the map $\iota(z) := 1/z$; of course ι is an analytic diffeomorphism. Suppose that $\varphi_1 : U_1 \to D$ and $\varphi_2 : U_2 \to D$ are analytic coordinate charts for $U_1 \subset C_1$ and $U_2 \subset C_2$. The idea is to create a new Riemann surface $C_1 \# C_2$ by using ι to “glue” $C_1 \backslash \varphi_1^{-1}(B)$ and $C_2 \backslash \varphi_2^{-1}(B)$ to each other. Specifically, for each $z \in A$ we identify $\varphi_1^{-1}(z)$ and $\varphi_2^{-1}(\iota(z))$. Give a formal description of this construction by specifying a system of coordinate charts, verifying that they have analytic overlap, and showing that the space they describe is Hausdorff.

Exercise 9.4: Suppose that $a_0 + a_1X + \cdots + a_mX$ and $b_0 + b_1X + \cdots + b_mX$ are polynomials in $\mathbb{C}[X]$ of the same degree that don't have any common roots in $\mathbb{C}$. Prove that the function

$$[z, w] \mapsto [a_0w^m + a_1zw^{m-1} + \cdots + a_nz^m, b_0w^m + b_1zw^{m-1} + \cdots + b_nz^m]$$

from the Riemann sphere to itself is complex analytic.

Exercise 9.5: Use the homotopy lifting property to prove that if $c : \tilde{X} \to X$ is a covering space with $\tilde{X}$ connected, then (for any base points) $\pi_1(c) : \pi_1(\tilde{X}) \to \pi_1(X)$ is injective.

Extension 9.6: A topological space is **locally simply connected** if each point has a simply connected neighborhood. Let X be a topological space that is connected and locally simply connected, and fix a base point $x_0 \in X$. Let $\tilde{X}$ be the set of equivalence classes $[p]$ of paths $p : [0, 1] \to X$ with $p(0) = x_0$, where two paths p, q are equivalent if $p(1) = q(1)$ and the element of $\pi_1(X, x_0)$ corresponding to $p * q^-$ is the identity, so that $p * q^-$ is homotopic, by a homotopy holding endpoints fixed, to the constant path at x_0. Let $c : \tilde{X} \to X$ be the function $[p] \mapsto p(1)$, and endow $\tilde{X}$ with the topology in which the open sets are the sets $\tilde{U} \subset \tilde{X}$ such that $c(\tilde{U})$ is open in X.

(a) Prove that this collection of sets actually is a topology for $\tilde{X}$.

(b) Fix $x \in X$, a simply connected neighborhood U, and $\tilde{x} = [p] \in c^{-1}(x)$. Let $\tilde{U}$ be the set of $[p * q]$ where $q : [0, 1] \to U$ is a path with $q(0) = p(1)$. Prove that $c|_{\tilde{U}}$ is a bijection and consequently a homeomorphism.

(c) Prove that if $\tilde{x}' = [p']$ is a second element of $c^{-1}(x)$ and $\tilde{U}'$ is the set of $[p*q]$ where $q : [0, 1] \to U$ is a path with $q(0) = p'(1)$, then $\tilde{U} \cap \tilde{U}' = \emptyset$. Conclude that c is a covering space.

(d) Use the homotopy lifting property to prove that $\tilde{X}$ is simply connected.

(e) Use the homotopy lifting property to prove that if $\hat{c} : \hat{X} \to X$ is a another covering space, then there is a covering space $d : \tilde{X} \to \hat{X}$ such that $\hat{c} \circ d = c$.

In recognition of property (e), c is called the **universal covering space** of X. Now let H be a subgroup of $\pi_1(X, x_0)$, let $\tilde{X}_H$ be the set of equivalence classes $[p]$ of paths $p : [0, 1] \to X$ with $p(0) = x_0$, where two paths p, q are equivalent if $p(1) = q(1)$ and the element of $\pi_1(X, x_0)$ corresponding to $p * q^-$ is an element of H, and let $c_H : \tilde{X}_H \to X$ be the function $[p] \mapsto p(1)$. Repeat all the steps above, with suitable modifications. Describe $\pi_1(\tilde{X}_H, \tilde{x}_H)$, where $\tilde{x}_H$ may be any element of $c_H^{-1}(x_0)$, and describe

$$\pi_1(c_H) : \pi_1(\tilde{X}_H, \tilde{x}_H) \to \pi_1(X, x_0).$$

Index